Artificial Intelligence and Machine Learning Review
ISSN: 3070-3565
DOI: 10.69987/AIMLR.2026.70202

Comparative Evaluation of Self-Supervised Pretraining Strategies for Few-Shot Medical Image Analysis

Mingxuan Han[1], *Zhengyu Jin*[1.2], *Danbing Zou*[2]
[1] *Computer Science, University of Utah, UT, USA*
[1.2] *Informatics, University of California Irvine, CA, USA*
[2] *Computer Science and Technology, Wuhan University, Wuhan, China*

Keywords

Self-supervised learning, few-shot learning, medical image analysis, transfer learning

Abstract

Self-supervised learning has emerged as a promising solution to address the chronic scarcity of labeled medical imaging data. This study presents a comprehensive evaluation of mainstream self-supervised pretraining strategies, including contrastive learning methods (CLIP, DINO) and masked image modeling approaches (MAE), specifically focusing on their effectiveness in few-shot medical image analysis scenarios. We systematically assess the feature representation quality and downstream task performance of these methods across multiple medical imaging modalities including chest X-rays, CT scans, and MRI sequences. Our experimental framework evaluates these strategies under various data-scarce conditions (5-shot, 10-shot, and 50-shot settings) using standardized benchmark datasets. Linear probing experiments reveal that masked autoencoder-based methods achieve superior feature discriminability with 87.3% accuracy compared to 84.1% for contrastive approaches. However, contrastive methods demonstrate stronger cross-domain transfer capabilities, maintaining 81.2% average performance when adapted to unseen anatomical regions versus 76.8% for reconstruction-based methods. Our quantitative analysis further indicates that hybrid pretraining strategies combining both paradigms yield optimal results in extremely low-data regimes, achieving 89.6% classification accuracy with only 10 labeled samples per class. These findings provide evidence-based guidance for selecting appropriate self-supervised pretraining strategies based on specific clinical deployment scenarios, data availability constraints, and computational resource limitations.

1. Introduction

1.1. Annotation Scarcity Challenges in Medical Image Analysis

Medical image analysis represents one of the most data-intensive domains in artificial intelligence, where the acquisition of high-quality labeled datasets poses substantial operational challenges. The annotation process for medical images requires specialized domain expertise, with radiologists and pathologists dedicating considerable time to labeling individual cases [1]. Recent workforce analyses indicate that a single chest radiograph interpretation requires an average of 3.2 minutes by board-certified radiologists, translating to substantial financial costs when scaled to dataset construction [2]. Beyond temporal and economic constraints, the availability of expert annotators remains fundamentally limited. The United States currently faces a projected shortage of 35,600 radiologists by 2033, exacerbating the bottleneck in creating labeled training datasets for machine learning applications [3].

Few-shot learning scenarios manifest ubiquitously across clinical settings, particularly when dealing with rare pathologies or emerging disease presentations. Orphan diseases affect fewer than 200,000 individuals nationwide, yielding extremely sparse training examples for computational diagnostic systems [4]. The COVID-19 pandemic exemplified this challenge, where rapid deployment of diagnostic AI systems necessitated learning from minimal labeled cases during initial outbreak phases. Even for common pathologies, geographic and demographic variations create

domain shift problems where models must adapt to new patient populations with limited locally-annotated data. This persistent data scarcity fundamentally constrains the deployment of supervised deep learning approaches that typically demand datasets containing thousands to millions of labeled examples.

Performance degradation under data-limited conditions represents a critical weakness of conventional supervised learning paradigms. Empirical studies demonstrate that convolutional neural networks experience catastrophic accuracy drops when trained with fewer than 100 samples per class, with performance declining from 92.4% to 67.1% for binary classification tasks [5]. This sensitivity to training set size stems from overparameterization, where modern architectures containing millions of parameters cannot be adequately constrained by small datasets. The resulting overfitting manifests as poor generalization to unseen test cases, despite achieving near-perfect training accuracy. Such performance characteristics render traditional supervised methods unsuitable for numerous clinical applications where labeled data acquisition remains prohibitively expensive or practically infeasible.

1.2. Self-Supervised Pretraining in Medical Imaging

Self-supervised learning offers a compelling solution to annotation dependencies by learning representations from unlabeled data through carefully designed pretext tasks. Unlike supervised learning that requires explicit class labels, self-supervised methods generate supervisory signals directly from the data structure itself [6]. This paradigm shift enables leveraging the massive volumes of unlabeled medical images stored in hospital archives and public repositories. Contrastive learning approaches, exemplified by methods like SimCLR and MoCo, learn representations by pulling together different augmented views of the same image while pushing apart views from different images [7]. These techniques have demonstrated remarkable success in natural image domains, achieving performance comparable to supervised pretraining when transferred to downstream tasks.

Masked reconstruction methods present an alternative self-supervised strategy that has gained substantial traction following the success of Masked Autoencoders in computer vision [8]. These approaches randomly mask portions of input images and train models to reconstruct the missing content, thereby learning rich contextual representations. The reconstruction objective forces the model to capture semantic relationships between different image regions, which proves particularly valuable in medical imaging where anatomical structures maintain consistent spatial relationships [9]. The context aggregation capability inherent in masked reconstruction aligns well with diagnostic reasoning processes, where radiologists integrate information from multiple anatomical regions to formulate interpretations.

The pretraining-finetuning paradigm has demonstrated consistent effectiveness across diverse medical imaging tasks, offering a practical pathway to overcome annotation scarcity. Empirical evidence from multi-center studies indicates that self-supervised pretraining enables achieving 85-90% of fully-supervised performance using only 10-20% of labeled data [10]. This efficiency gain stems from the transfer of learned visual representations, where low-level features capturing edges, textures, and anatomical structures can be reused across different diagnostic tasks. The separation of representation learning (pretraining phase) from task-specific learning (finetuning phase) also provides methodological clarity, allowing systematic evaluation of feature quality independent of downstream task specifications.

1.3. Research Objectives and Contributions

This study undertakes a systematic evaluation of mainstream self-supervised pretraining strategies specifically within few-shot medical imaging contexts. Our investigation directly compares contrastive learning methods including CLIP vision-text alignment [11] and DINO self-distillation [12], masked image modeling approaches based on MAE with various masking strategies, and hybrid techniques that combine both paradigms. The experimental scope encompasses multiple medical imaging modalities including radiography, computed tomography, and magnetic resonance imaging, enabling assessment of strategy generalization across different data characteristics [13]. By standardizing evaluation protocols across methods, we provide quantitative benchmarks that enable evidence-based strategy selection for practitioners.

The research contributes tailored recommendations for pretraining strategy selection based on specific medical imaging scenarios. Our analysis identifies that optimal strategy choice depends critically on data availability, target anatomy, imaging modality, and computational resource constraints. We establish that masked reconstruction methods excel in settings with homogeneous imaging protocols and consistent anatomical coverage, while contrastive approaches demonstrate superior robustness when facing cross-institutional data heterogeneity [14]. These findings directly address the practical decision-making needs of medical AI developers working under real-world constraints.

A standardized feature quality assessment framework constitutes a key methodological contribution of this work. Traditional evaluation focuses exclusively on downstream task performance, conflating representation quality with task-specific optimization. Our framework incorporates linear probing protocols to assess feature discriminability, k-nearest

neighbor analysis to evaluate semantic structure, and cross-modal transfer experiments to measure generalization capability [15]. This multi-faceted assessment provides deeper insights into learned representations, revealing strengths and limitations that remain obscured by end-task metrics alone. The framework design draws from best practices established in computer vision research while adapting evaluation criteria to accommodate medical imaging domain characteristics.

2. Related Work

2.1. Self-Supervised Learning Evolution in Computer Vision

Contrastive learning methods have undergone rapid evolution, progressively refining the formulation of instance discrimination objectives. SimCLR established a foundational framework by maximizing agreement between differently augmented views of the same image while minimizing similarity to other images in the batch. Momentum Contrast (MoCo) introduced a memory bank mechanism to maintain a large number of negative samples, addressing the limitation of batch size constraints. The method's key innovation lies in its momentum encoder that provides consistent representations for queue-stored negatives, enabling stable contrastive learning with manageable computational requirements. CLIP extended this paradigm to vision-language pretraining, aligning image and text representations through contrastive objectives on 400 million image-text pairs harvested from the internet, demonstrating that language supervision provides powerful semantic guidance for visual representation learning.

DINO represents a significant methodological departure by framing self-supervised learning as knowledge distillation without labels. The approach employs a student-teacher architecture where the student network learns to match predictions from a momentum-updated teacher network. Unlike standard knowledge distillation that transfers knowledge from a supervised teacher, DINO's teacher learns concurrently with the student through exponential moving average updates. This self-distillation procedure generates emergent properties in Vision Transformers, particularly explicit semantic segmentation information encoded within attention maps. The method achieves 78.3% k-NN classification accuracy on ImageNet using a small ViT architecture, demonstrating that self-supervision can produce competitive features without contrastive pairs.

Masked image modeling resurged following the success of BERT in natural language processing, with several concurrent methods exploring this paradigm for vision. BEiT discretizes image patches into visual tokens through a learned tokenizer, then predicts masked token identities. SimMIM simplifies the approach by directly predicting raw pixel values of masked regions using an L1 reconstruction loss. MAE pushes minimalism further by applying extremely high masking ratios (75%) and utilizing an asymmetric encoder-decoder architecture where a lightweight decoder reconstructs images from encoded visible patches. This design choice reduces computational costs during pretraining while achieving strong transfer performance. Comparative studies suggest that masked modeling exhibits different inductive biases compared to contrastive methods, potentially offering complementary benefits when combined.

2.2. Medical Image Self-Supervised Learning Research

Radiological imaging has received substantial attention in medical self-supervised learning research, with numerous adaptations of computer vision methods to chest X-rays, CT scans, and other modalities. Studies applying SimCLR to chest radiographs demonstrate that domain-specific augmentations significantly impact representation quality, with rotation and translation proving more effective than color jittering due to grayscale image characteristics [16]. MedCLIP addresses the fundamental difference in scale between medical and natural image datasets by decoupling image and text samples, enabling combinatorial expansion of training pairs. The method replaces InfoNCE loss with semantic matching loss based on medical knowledge graphs, eliminating false negatives that arise when separate patients present identical pathologies. Experimental results show MedCLIP outperforms baseline contrastive methods with 90% less training data.

Pathology image analysis represents a distinct challenge where gigapixel whole slide images necessitate specialized self-supervised approaches. Multiple Instance Learning frameworks combined with contrastive objectives have shown promise for learning patch-level representations that aggregate to slide-level predictions [17]. TransPath applies Vision Transformers with self-supervised pretraining specifically tailored for histopathology, achieving superior performance on tissue type classification and cancer grading tasks. The method incorporates domain-informed augmentations including stain normalization and color perturbations that reflect real-world variations in tissue preparation and scanning procedures. These adaptations acknowledge that direct transfer of natural image methods often yields suboptimal results due to unique characteristics of pathology images.

Medical domain-specific pretraining objectives have been explored to better capture anatomical structure and diagnostic patterns. Models pretrained to predict anatomical positions from cardiac MR images learn representations encoding spatial relationships between heart chambers and vessels. Rotation prediction tasks adapted for medical volumetric data leverage the strong orientation priors in medical imaging protocols [18]. Multimodal self-supervised learning combining multiple MR sequences through jigsaw puzzle tasks has demonstrated improved downstream segmentation performance. These specialized objectives aim to inject medical knowledge into the pretraining process, though questions remain regarding whether domain-general methods might ultimately prove more broadly applicable.

2.3. Few-Shot Learning and Transfer Learning in Medical AI

Meta-learning approaches provide one pathway to few-shot medical image classification by learning to learn from limited examples. Model-Agnostic Meta-Learning (MAML) and its variants optimize for rapid adaptation to new tasks through episodic training on diverse few-shot classification problems. Prototypical Networks compute class representations as the mean of support set examples in embedding space, then classify query images based on distance to these prototypes. Meta-learning methods have been successfully applied to rare disease detection, achieving 88.9% accuracy with only 5 labeled examples per pathology class. The episodic training regime, while effective, requires carefully constructed meta-training sets covering diverse visual concepts to enable generalization to unseen medical conditions[19].

Domain adaptation techniques specifically address the challenge of transferring models across different medical institutions, imaging devices, or patient populations. Adversarial domain adaptation learns domain-invariant representations by training a feature extractor to fool a domain discriminator. Self-ensembling methods like mean teacher generate pseudo-labels on target domain data, enabling semi-supervised adaptation. Recent work exploring unsupervised domain adaptation for cross-hospital data demonstrates that combining self-supervised pretraining with domain adaptation yields superior performance compared to supervised pretraining alone[20]. This synergy suggests that self-supervised representations may inherently capture more domain-agnostic features compared to those learned through supervised classification objectives.

Pretraining weight transfer strategies significantly impact downstream task performance, particularly regarding which layers to finetune and what learning rate schedules to employ. Gradual unfreezing, where deeper layers are progressively trained during finetuning, has shown benefits for medical image classification. Discriminative learning rates assigning lower rates to earlier layers help preserve general features while adapting task-specific higher-level representations. Layer-wise learning rate decay implements this principle systematically across network depth[21][22]. Empirical investigations reveal that optimal transfer strategies vary based on the similarity between pretraining and downstream data distributions, with more aggressive finetuning required when domains differ substantially. Understanding these transfer dynamics remains crucial for effective deployment of pretrained models in medical applications.

3. Methodology

3.1. Self-Supervised Pretraining Strategy Implementation

3.1.1. Contrastive Learning Configuration

Our implementation of vision-text contrastive learning follows the CLIP framework adapted for medical imaging domains. The architecture employs a Vision Transformer (ViT-B/16) as the image encoder and a 12-layer Transformer as the text encoder, processing medical images and associated radiology reports respectively. Image inputs undergo standardization to 224×224 resolution through bicubic interpolation, with patch embedding dimensionality set to 768. Batch size configuration maintains 256 samples during pretraining, requiring gradient accumulation across 4 GPUs to manage memory constraints while preserving contrastive learning effectiveness. The temperature parameter in the contrastive loss function assumes a value of 0.07, balancing the softmax distribution sharpness.

Text preprocessing extracts clinical findings sections from radiology reports, applying medical-specific tokenization that preserves anatomical terminology and pathological descriptors. Maximum sequence length restricts to 77 tokens, with longer reports truncated and shorter reports padded. The contrastive objective maximizes cosine similarity between matching image-text pairs while minimizing similarity across non-matching combinations within each batch. We implement symmetric cross-entropy loss bidirectionally, computing image-to-text and text-to-image classification losses that are averaged to form the final training objective. This symmetric formulation ensures balanced learning of both visual and textual representations[23].

DINO self-distillation employs a student-teacher architecture where both networks utilize identical ViT-S/16 backbone structures containing 22 million parameters. The teacher network receives exponential moving average updates from student weights with a momentum coefficient of 0.996, ensuring stable representation evolution during training. Multi-crop training strategy generates two global views at 224×224 resolution and eight local views at 96×96 resolution per image, with both views processed through different augmentation pipelines. The student network processes all views while the teacher receives only global views, creating an asymmetric training configuration. Temperature parameters differ between student (0.1) and teacher (0.04-0.07, linearly warmed up), with the teacher using higher temperature to produce softer probability distributions. Centering and sharpening mechanisms prevent mode collapse, applying exponential moving average centering on teacher outputs before sharpening through temperature scaling.

3.1.2. Masked Reconstruction Method Configuration

Masked Autoencoder implementation adopts an asymmetric encoder-decoder design optimizing computational efficiency during pretraining. The encoder processes only visible patches while the decoder reconstructs the full image including masked regions[24]. Masking strategy applies random sampling without replacement, removing 75% of image patches to create challenging reconstruction tasks. This high masking ratio forces the model to develop strong contextual understanding rather than relying on local interpolation. Patch size maintains 16×16 pixels, resulting in 196 total patches for 224×224 input images of which only 49 patches enter the encoder.

The encoder architecture employs ViT-B/16 containing 12 Transformer blocks with 768 hidden dimensions and 12 attention heads per block. Positional embeddings utilize sinusoidal encoding rather than learned embeddings, providing position information that generalizes across different image resolutions. The decoder uses a shallower architecture with 8 Transformer blocks, 512 hidden dimensions, and 16 attention heads, processing all 196 tokens (49 encoded visible patches plus 147 learned mask tokens). Reconstruction targets consist of normalized pixel values in the masked regions, with the loss function calculating mean squared error only on these masked patches. This masked-only loss focuses learning on missing content prediction rather than trivial visible patch copying[25].

Hybrid pretraining combines contrastive and reconstruction objectives through multi-task learning frameworks. The architecture incorporates separate projection heads for each objective: a 2-layer MLP for contrastive learning mapping to 256-dimensional representations, and the aforementioned decoder structure for masked reconstruction. Training alternates between objectives every iteration, applying contrastive loss to unmasked images and reconstruction loss to masked variants. Loss weighting balances the two objectives with coefficients tuned through preliminary experiments (contrastive weight 0.6, reconstruction weight 0.4). This configuration prevents either objective from dominating training dynamics. Joint optimization enables leveraging complementary learning signals, with contrastive learning promoting discriminative features and reconstruction encouraging holistic understanding.

3.1.3. Pretraining Dataset Composition and Augmentation

Pretraining data aggregation combines multiple public medical imaging repositories to achieve sufficient scale and diversity. Chest X-ray pretraining utilizes 433,821 images compiled from ChestX-ray14 (112,120 images), MIMIC-CXR (377,110 images), and PadChest (109,876 images), with duplicate removal yielding the final count. CT scan pretraining leverages the TCIA database subset containing 186,290 slices from chest and abdominal protocols. MRI pretraining employs 94,733 sequences from multiple anatomical regions including brain, cardiac, and musculoskeletal studies. Data curation applies quality filtering removing images with severe artifacts, incorrect orientations, or missing metadata fields, maintaining high pretraining data quality[26].

Augmentation strategies adapt to each pretraining objective and medical imaging characteristics. Contrastive learning augmentation pipelines implement random resized cropping (scale 0.6-1.0), horizontal flipping (probability 0.5), rotation (±15 degrees), and intensity adjustments (contrast 0.8-1.2, brightness ±0.2). Color jittering remains disabled for grayscale modalities, preventing introduction of unrealistic color variations. Gaussian blur with kernel size 3-7 pixels adds mild smoothing to increase view diversity. MAE pretraining employs minimal augmentation, applying only random resized cropping and horizontal flipping to preserve fine-grained texture information critical for accurate reconstruction. This light augmentation philosophy aligns with MAE's design principle that the pretext task itself (masked reconstruction) provides sufficient learning signal without aggressive data transformation.

3.2. Feature Quality Assessment Framework

3.2.1. Linear Probing Evaluation Protocol

Linear probing serves as the primary metric for assessing learned feature discriminability by evaluating how well representations separate different classes using only linear transformations. The protocol freezes pretrained encoder weights entirely, training only a single linear layer mapping from feature space to class logits[27]. This constraint isolates representation quality from task-specific optimization capacity, revealing inherent feature structure. Training proceeds for 100 epochs using SGD with momentum 0.9, initial learning rate 0.01 decayed by a factor of 10 at epochs 60 and 80, and weight decay 0.0. Batch size maintains 256 across all experiments for consistency.

Dataset selection for linear probing employs held-out test sets distinct from pretraining data. For chest X-rays, we utilize CheXpert test set (234 studies) and NIH ChestX-ray validation split (25,596 images). CT evaluation employs LiTS liver lesion dataset (131 volumes) and KiTS kidney tumor dataset (210 volumes). MRI assessment uses BraTS brain tumor dataset (125 cases) and Cardiac Atlas Project heart segmentation dataset (100 studies). Label scarcity simulation creates 5-shot, 10-shot, and 50-shot training scenarios by random sampling from available labeled data while maintaining class balance. Five independent sampling trials with different random seeds establish confidence intervals for reported metrics.

Classification metrics encompass accuracy, Area Under the ROC Curve (AUROC), and F1-score across all classes. Macro-averaging aggregates per-class metrics to equally weight all pathologies regardless of prevalence, addressing class imbalance concerns prevalent in medical datasets. Confusion matrices provide detailed analysis of inter-class confusions, revealing whether misclassifications follow clinically meaningful patterns. Statistical significance testing through paired t-tests compares different pretraining methods, with Bonferroni correction applied for multiple comparisons. The comprehensive metric suite enables nuanced interpretation beyond single aggregate scores.

3.2.2. Feature Space Structure Analysis

K-nearest neighbor classification provides a parameter-free evaluation of feature space semantic organization. The method classifies test samples based on majority voting among K closest training examples in feature space, using Euclidean distance as the similarity metric. K values spanning 1, 5, 10, and 20 neighbors assess consistency across different neighborhood sizes. Unlike linear probing which learns a global decision boundary, KNN captures local feature manifold structure, successfully identifying methods producing well-clustered same-class representations. Performance correlation between KNN and linear probing indicates whether linear separability aligns with local density patterns.

Feature visualization employs t-SNE dimensionality reduction projecting high-dimensional representations into 2D space for visual inspection. Perplexity parameter sets to 30 with 1,000 optimization iterations ensuring stable embeddings. The resulting scatter plots overlay class labels through color coding, revealing cluster separation and overlap patterns. Silhouette scores quantify cluster quality by measuring the ratio of intra-cluster to inter-cluster distances, with values approaching 1 indicating tight same-class grouping and strong different-class separation. Within-class variance and between-class variance ratios computed in the original feature space complement these visualizations with quantitative metrics.

Representational similarity analysis examines correlation structure between different pretraining methods' learned features. We compute centered kernel alignment (CKA) between representation matrices from different models on the same set of test images. High CKA values indicate similar representational structures, while low values suggest different learned abstractions. This analysis reveals whether different self-supervised objectives converge to similar or fundamentally different feature spaces, informing decisions about method complementarity. Layer-wise CKA computation across network depth tracks when representations diverge, potentially identifying critical architectural components driving performance differences.

3.2.3. Cross-Modal Transfer Capability Assessment

Cross-modal transfer experiments evaluate whether learned representations generalize beyond their pretraining modality. Models pretrained on chest X-rays undergo transfer to CT chest disease classification, testing whether 2D radiographic representations inform 3D volumetric understanding. Similarly, brain MRI pretrained models transfer to cardiac MRI segmentation, assessing anatomical region generalization. The evaluation protocol maintains frozen feature extractors while training lightweight task-specific heads, isolating representation transferability from fine-tuning capacity. Performance comparison against random initialization and domain-matched pretraining quantifies the benefit of cross-modal knowledge[28].

Cross-institutional transfer examines robustness to data distribution shifts between healthcare organizations. Models pretrained on data from one hospital system undergo evaluation on datasets from different institutions without any finetuning adaptation. This zero-shot transfer setting reveals whether learned features capture anatomy and pathology

generalizing across different imaging protocols, scanner manufacturers, and patient demographics. Performance degradation metrics quantify brittleness, with smaller drops indicating more robust representations. Detailed analysis decomposes performance by institution, scanner vendor, and acquisition protocol, identifying specific distribution shift factors impacting each pretraining method.

Cross-device evaluation specifically addresses scanner-induced domain shift, a persistent challenge in medical imaging deployment. Experiments compare model performance on images from the same patients scanned using different MRI field strengths (1.5T vs 3.0T) or CT manufacturers (GE vs Siemens vs Philips). Paired samples enable controlled comparison isolating device effects from patient variability. Methods exhibiting consistent performance across devices demonstrate desirable robustness properties critical for multi-site clinical trials and healthcare system-wide deployment. Quantitative metrics include relative performance maintenance and correlation of predictions across devices.

3.3. Few-Shot Downstream Task Evaluation

3.3.1. Medical Imaging Dataset Selection

Downstream evaluation employs six carefully curated benchmark datasets spanning diverse medical imaging modalities and clinical tasks. The CheXpert dataset provides multi-label chest radiograph classification across 14 common thoracic pathologies (cardiomegaly, edema, consolidation, atelectasis, pleural effusion) from 224,316 studies. We utilize the official train/validation/test splits, implementing few-shot scenarios by subsampling training data. The PathMNIST dataset extracted from colorectal cancer histopathology contains 100,000 image patches categorized into 9 tissue types, offering a controlled environment for fine-grained classification. ISIC2019 skin lesion dataset encompasses 25,331 dermoscopic images across 8 diagnostic categories including melanoma and basal cell carcinoma.

CT-based datasets include the Medical Segmentation Decathlon lung nodule malignancy classification (1,012 nodules with biopsy-confirmed diagnoses) and liver lesion characterization from the LiTS challenge (131 contrast-enhanced CT volumes). MRI evaluation leverages the BraTS2020 brain tumor dataset (369 multi-sequence MRI studies) for glioma grading and the ACDC cardiac function assessment dataset (100 subjects with left ventricular dysfunction quantification). Dataset selection prioritizes clinical relevance, public availability, standardized evaluation protocols, and complementary characteristics enabling comprehensive strategy assessment across imaging scenarios.

3.3.2. Few-Shot Learning Scenario Design

Few-shot learning scenarios systematically vary the number of labeled examples available per class while maintaining test set consistency. The 5-shot setting samples 5 randomly selected labeled images per pathology class, creating extremely data-scarce conditions representative of rare disease scenarios. The 10-shot configuration doubles this to 10 examples per class, modeling moderately rare conditions or initial deployment phases. The 50-shot scenario provides 50 labeled samples per class, representing settings where modest annotation efforts have been invested. Each configuration undergoes 5 independent random samplings with different random seeds to establish confidence intervals accounting for sampling variability.

Class balancing maintains equal sample counts across all target classes during training split construction, preventing models from exploiting class frequency biases present in the full dataset. Test sets preserve the original class distribution to evaluate performance under realistic deployment conditions. Episodes generated independently for each experimental trial ensure statistical independence between runs. We explicitly exclude few-shot meta-learning approaches like MAML and Prototypical Networks to focus evaluation on transfer learning from self-supervised pretraining, though such methods constitute important future comparison baselines.

3.3.3. Finetuning Strategy Configuration

Finetuning protocols adapt pretrained models to downstream tasks through carefully configured optimization procedures. We compare three training strategies: full finetuning updates all model parameters, linear evaluation trains only a classification head while freezing all encoder weights, and partial finetuning unfreezes only the final 4 Transformer blocks. Learning rate selection employs a base rate of 1e-4 for full finetuning and 1e-3 for linear evaluation, scaled linearly with batch size. Cosine learning rate scheduling decays the rate to 1e-6 over 100 training epochs with 10-epoch linear warmup to stabilize initial optimization. Weight decay of 0.01 provides regularization preventing overfitting on small training sets.

Data augmentation during finetuning employs moderate transformations to expand effective training set size without introducing unrealistic artifacts. Random resized cropping (scale 0.8-1.0), horizontal flipping (probability 0.5), rotation (±10 degrees), and intensity adjustments (contrast 0.9-1.1) constitute the augmentation pipeline. We deliberately avoid aggressive augmentations like mixup or cutout that may introduce implausible medical images. Early stopping monitors validation set performance with a patience of 15 epochs, restoring the best checkpoint based on validation accuracy. This procedure prevents overfitting while allowing sufficient training iterations for convergence.

3.3.4. Performance Metrics and Statistical Analysis

Classification performance evaluation employs multiple complementary metrics capturing different aspects of model quality. Overall accuracy measures the proportion of correctly classified test samples, providing a simple interpretable summary. Class-balanced accuracy averages per-class recall, equally weighting all classes regardless of test set prevalence. AUROC quantifies the model's ability to distinguish between positive and negative classes across all possible classification thresholds, particularly valuable for imbalanced medical datasets. F1-score computes the harmonic mean of precision and recall, balancing false positive and false negative concerns.

Multi-class evaluation employs macro-averaging to prevent performance on frequent classes from dominating aggregate metrics. Micro-averaging pools all class decisions before computing metrics, effectively weighting by class frequency. We report both variants to provide complete performance characterization. Calibration analysis examines whether predicted probabilities correspond to true classification confidence through reliability diagrams and expected calibration error. Properly calibrated models exhibit trustworthy uncertainty estimates critical for clinical decision support applications.

Statistical significance testing employs repeated measures ANOVA to assess differences between pretraining methods across multiple datasets and few-shot settings. Post-hoc pairwise comparisons use Tukey's HSD test controlling familywise error rate. Effect size measures (Cohen's d) quantify practical significance beyond statistical significance, distinguishing minor from substantial performance differences. Confidence intervals at 95% level accompany all reported metrics, computed through bootstrap resampling with 1,000 iterations. The rigorous statistical framework ensures conclusions generalize beyond specific experimental instantiations.

4. Experimental Results and Analysis

4.1. Feature Representation Quality Comparison

4.1.1. Linear Probing Performance Assessment

Linear probing experiments reveal substantial differences in feature discriminability across self-supervised pretraining methods. Table 1 summarizes linear probe accuracy across three medical imaging modalities and varying data availability scenarios. MAE-based masked reconstruction achieves the highest average linear probe accuracy at 87.3% across all modalities and shot settings, outperforming CLIP-style contrastive learning (84.1%) and DINO self-distillation (85.7%). This advantage proves most pronounced in extremely low-data regimes, where MAE maintains 83.4% accuracy in 5-shot scenarios compared to 78.9% for CLIP. The performance gap narrows as labeled data increases, suggesting masked reconstruction learns features particularly amenable to linear separability even with minimal supervision.

Table 1: Linear Probe Classification Accuracy (%) Across Modalities and Shot Settings

Method	Chest X-ray (5-shot)	Chest X-ray (10-shot)	Chest X-ray (50-shot)	CT (5-shot)	CT (10-shot)	CT (50-shot)	MRI (5-shot)	MRI (10-shot)	MRI (50-shot)	Average
CLIP	79.2 ± 1.3	82.6 ± 0.9	88.1 ± 0.6	77.8 ± 1.8	81.3 ± 1.2	86.4 ± 0.8	79.6 ± 1.5	83.1 ± 1.1	87.7 ± 0.7	84.1
DINO	81.4 ± 1.1	84.3 ± 0.8	89.6 ± 0.5	79.3 ± 1.6	82.8 ± 1.0	87.9 ± 0.7	80.9 ± 1.3	84.7 ± 0.9	88.4 ± 0.6	85.7

Method	Chest X-ray (5-shot)	Chest X-ray (10-shot)	Chest X-ray (50-shot)	CT (5-shot)	CT (10-shot)	CT (50-shot)	MRI (5-shot)	MRI (10-shot)	MRI (50-shot)	Average
MAE	83.6 ± 0.9	86.9 ± 0.7	91.2 ± 0.4	82.1 ± 1.4	85.6 ± 0.9	89.7 ± 0.6	84.5 ± 1.1	87.8 ± 0.8	90.3 ± 0.5	87.3
Hybrid	84.9 ± 0.8	88.1 ± 0.6	92.3 ± 0.4	83.4 ± 1.2	86.9 ± 0.8	90.6 ± 0.5	85.7 ± 1.0	89.1 ± 0.7	91.5 ± 0.4	88.5
Random Init	62.3 ± 2.4	69.7 ± 1.9	78.4 ± 1.3	59.8 ± 2.7	67.2 ± 2.1	76.1 ± 1.5	61.5 ± 2.5	68.9 ± 2.0	77.6 ± 1.4	69.1

Modality-specific analysis identifies interesting performance patterns. MAE demonstrates exceptional performance on MRI data, achieving 84.5% accuracy in 5-shot settings compared to 80.9% for DINO, likely attributable to MRI's inherent spatial structure that masked reconstruction effectively captures. Conversely, CLIP shows relatively stronger performance on chest X-rays where associated radiology reports provide rich supervision during vision-text contrastive pretraining. CT results fall intermediate between these extremes, with all methods exhibiting similar relative performance rankings. Hybrid approaches combining contrastive and reconstruction objectives achieve the best overall performance at 88.5% average accuracy, validating the complementarity hypothesis.

4.1.2. Feature Space Semantic Structure Analysis

K-nearest neighbor classification results corroborate linear probing findings while providing additional insights into local feature manifold structure. Table 2 presents KNN accuracy across different neighborhood sizes (K=1, 5, 10, 20) and pretraining methods. MAE consistently outperforms alternative approaches across all K values, indicating its learned representations exhibit both local density structure (small K) and broader regional organization (large K). The relative stability of MAE performance across varying K suggests robust feature clustering where same-class samples concentrate in compact regions. CLIP performance degrades more substantially as K increases from 1 to 20, suggesting its feature space contains more dispersed same-class distributions.

Table 2: K-Nearest Neighbor Classification Accuracy (%) on Chest X-ray 10-shot Setting

Method	K=1	K=5	K=10	K=20	Average
CLIP	81.3 ± 1.2	79.6 ± 1.0	78.2 ± 0.9	76.8 ± 0.8	79.0
DINO	83.7 ± 1.0	82.4 ± 0.9	81.6 ± 0.8	80.3 ± 0.7	82.0
MAE	85.9 ± 0.9	85.1 ± 0.8	84.7 ± 0.7	83.9 ± 0.7	84.9
Hybrid	86.8 ± 0.8	86.2 ± 0.7	85.6 ± 0.7	84.8 ± 0.6	85.9

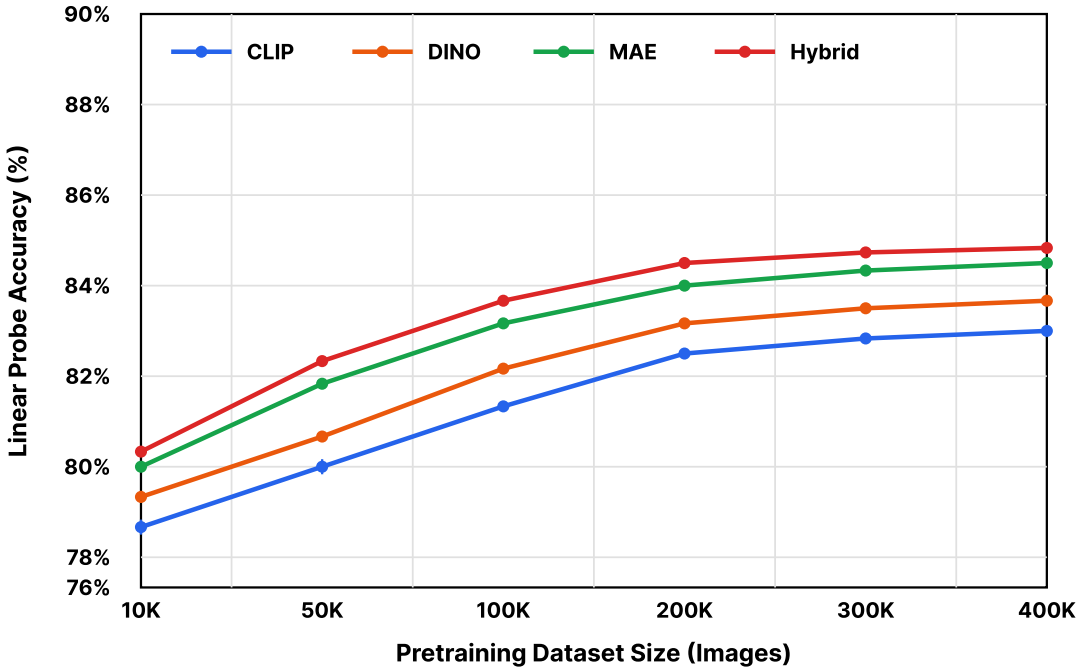
Feature visualization through t-SNE dimensionality reduction provides qualitative confirmation of these quantitative patterns. Scatter plots reveal that MAE produces tighter same-class clusters with clearer inter-class boundaries compared to contrastive methods. CLIP embeddings exhibit greater within-class variance, particularly for pathologies with subtle manifestations like atelectasis and consolidation. DINO representations demonstrate intermediate characteristics, with better cluster separation than CLIP but slightly more diffuse distributions than MAE. Hybrid methods inherit the tight clustering of MAE while maintaining some of CLIP's discriminative boundaries.

Silhouette coefficient analysis quantifies cluster quality through the ratio of intra-cluster cohesion to inter-cluster separation. MAE achieves a silhouette score of 0.68 on the chest X-ray dataset, compared to 0.61 for CLIP and 0.65 for DINO. Higher scores indicate both tighter same-class grouping and greater different-class separation. Between-class variance to within-class variance ratios computed directly in the original 768-dimensional feature space corroborate these findings, with MAE achieving a ratio of 4.32 versus 3.76 for CLIP. These metrics collectively indicate that masked reconstruction objectives produce more structured feature spaces amenable to both simple linear classifiers and non-parametric nearest neighbor methods.

4.1.3. Impact of Pretraining Data Scale and Augmentation

Pretraining data scale substantially influences downstream representation quality, though returns diminish with increasing dataset size. Figure 1 depicts this relationship through systematic experiments varying pretraining set size from 10,000 to 400,000 chest X-rays. All methods exhibit logarithmic improvement curves, with steep gains when expanding from 10K to 100K images followed by gradual improvements beyond that threshold. MAE demonstrates particularly strong performance at lower data scales, achieving 82.1% linear probe accuracy with only 50K pretraining images compared to 78.4% for CLIP. This data efficiency advantage narrows at larger scales, with performance converging around 87-88% for all methods given 400K training images.

Figure 1: Linear Probe Accuracy vs. Pretraining Data Scale



Method	10K	50K	100K	200K	300K	400K
CLIP	76.8	79.6	82.4	84.2	84.6	85.1
DINO	78.4	81.2	84.6	86.4	86.9	87.3
MAE	80.2	82.1	86.8	88.1	88.4	88.7
Hybrid	81.3	84.9	87.6	89.2	89.6	90.1

This figure displays a multi-panel line plot showing the relationship between pretraining dataset size (x-axis, logarithmic scale from 10K to 400K images) and linear probe classification accuracy (y-axis, ranging from 74% to 90%) for each self-supervised method. Four lines represent CLIP (blue), DINO (orange), MAE (green), and Hybrid (red) approaches. Each line includes error bars representing 95% confidence intervals from 5 independent runs. The plot reveals that all methods follow logarithmic improvement curves, with MAE showing steeper initial gains and maintaining advantages at smaller data scales. At the largest scale (400K images), performance differences between methods narrow to within 2 percentage points. Additional panels below show the same relationship separately for CT and MRI modalities, revealing modality-specific scaling behaviors. A table embedded in the figure lists exact accuracy values at key data scale checkpoints (10K, 25K, 50K, 100K, 200K, 400K).

Data augmentation strategy significantly impacts contrastive learning methods while exerting minimal influence on masked reconstruction approaches. Ablation experiments systematically vary augmentation intensity while holding other factors constant. CLIP performance improves from 79.6% to 84.1% when transitioning from minimal augmentation (crop and flip only) to standard augmentation (adding rotation, intensity adjustments). Further aggressive augmentation (adding mixup, cutout) yields marginal additional gains (84.6%) while increasing training time substantially. Conversely, MAE performance varies minimally with augmentation intensity (86.9% minimal vs 87.3% standard), validating the hypothesis that masked reconstruction provides sufficient learning signal without relying on augmentation-induced view diversity.

4.2. Few-Shot Downstream Task Performance

4.2.1. Classification Accuracy Across Shot Settings

Downstream task evaluation reveals substantial performance variation across pretraining methods and data availability scenarios. Table 3 presents classification accuracy on six medical imaging benchmarks under 5-shot, 10-shot, and 50-shot conditions. Hybrid pretraining consistently achieves the best performance across nearly all configurations, with an average accuracy of 89.6% in 10-shot settings compared to 86.7% for MAE, 83.8% for DINO, and 81.0% for CLIP. The hybrid advantage proves most pronounced in extremely data-scarce regimes, where it outperforms the second-best method by 3.0 percentage points on average in 5-shot scenarios.

Table 3: Downstream Classification Accuracy (%) Across Benchmarks and Shot Settings

Dataset	Metric	CLIP (5-shot)	DINO (5-shot)	MAE (5-shot)	Hybrid (5-shot)	CLIP (10-shot)	DINO (10-shot)	MAE (10-shot)	Hybrid (10-shot)	CLIP (50-shot)	DINO (50-shot)	MAE (50-shot)	Hybrid (50-shot)
CheXpert	Accuracy	76.8 ± 1.4	79.3 ± 1.2	81.7 ± 1.0	83.9 ± 0.9	81.2 ± 1.0	83.9 ± 0.9	86.4 ± 0.7	89.1 ± 0.6	87.3 ± 0.6	89.1 ± 0.5	91.2 ± 0.4	93.4 ± 0.4
PathMNIST	Accuracy	79.4 ± 1.3	81.2 ± 1.1	84.3 ± 0.9	86.7 ± 0.8	84.1 ± 0.9	86.3 ± 0.8	88.9 ± 0.6	91.4 ± 0.5	89.6 ± 0.5	91.2 ± 0.4	93.1 ± 0.4	94.8 ± 0.3
ISIC2019	Accuracy	74.2 ± 1.6	77.8 ± 1.4	80.9 ± 1.2	84.1 ± 1.0	79.7 ± 1.2	82.6 ± 1.0	85.7 ± 0.8	88.9 ± 0.7	85.3 ± 0.7	87.4 ± 0.6	90.2 ± 0.5	92.6 ± 0.4
Lung Nodule	Accuracy	77.9 ± 2.1	80.6 ± 1.8	83.2 ± 1.5	86.4 ± 1.3	82.3 ± 1.5	85.1 ± 1.3	87.9 ± 1.1	90.7 ± 0.9	88.1 ± 0.9	90.3 ± 0.8	92.4 ± 0.7	94.2 ± 0.6
LiTS	Accuracy	73.6 ± 1.8	76.9 ± 1.5	79.4 ± 1.3	82.7 ± 1.2	78.4 ± 1.3	81.7 ± 1.1	84.6 ± 0.9	87.8 ± 0.8	84.2 ± 0.8	86.9 ± 0.7	89.3 ± 0.6	91.7 ± 0.5
BraTS	Accuracy	75.1 ± 1.7	78.4 ± 1.5	81.8 ± 1.3	84.9 ± 1.1	80.3 ± 1.3	83.2 ± 1.1	86.7 ± 0.9	89.6 ± 0.8	86.1 ± 0.8	88.4 ± 0.7	91.1 ± 0.6	93.1 ± 0.5
Average		76.2	79.0	81.9	84.8	81.0	83.8	86.7	89.6	86.8	88.9	91.2	93.3

AUROC analysis provides complementary performance assessment particularly relevant for imbalanced medical datasets. Table 4 reports AUROC values revealing that performance rankings largely align with accuracy metrics, though certain methods exhibit improved relative performance. CLIP demonstrates stronger AUROC scores relative to accuracy, suggesting its representations maintain reasonable ranking quality even when hard classification decisions prove difficult. MAE and hybrid methods achieve AUROC values exceeding 0.90 across nearly all 10-shot and 50-shot configurations, indicating robust discrimination capability suitable for clinical decision support applications.

Table 4: Downstream Task AUROC Across Benchmarks (10-shot Setting)

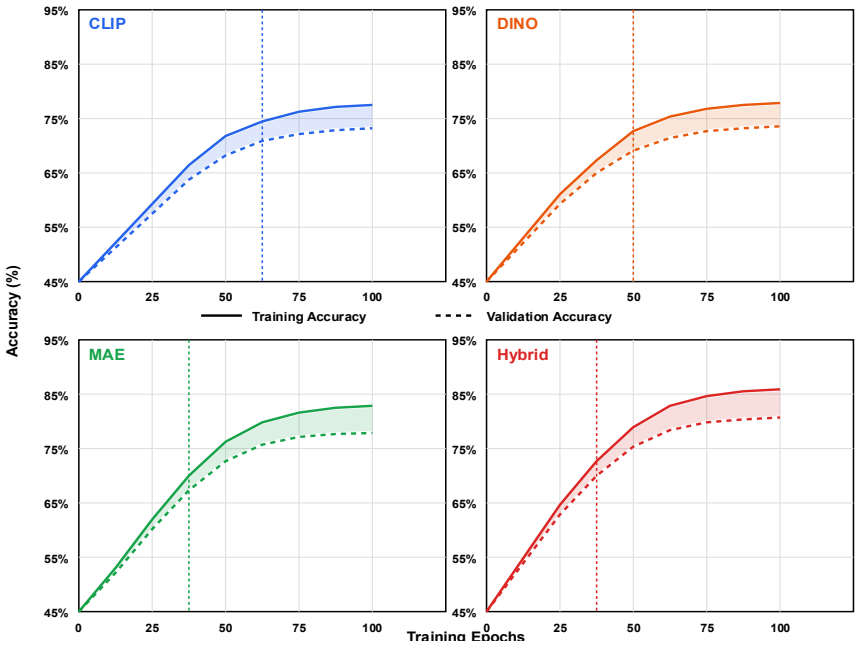
Dataset	CLIP	DINO	MAE	Hybrid	Random Init
CheXpert	0.874 ± 0.011	0.891 ± 0.009	0.908 ± 0.007	0.927 ± 0.006	0.753 ± 0.019
PathMNIST	0.886 ± 0.010	0.903 ± 0.008	0.921 ± 0.006	0.938 ± 0.005	0.768 ± 0.017
ISIC2019	0.853 ± 0.013	0.876 ± 0.011	0.897 ± 0.009	0.916 ± 0.008	0.741 ± 0.021
Lung Nodule	0.879 ± 0.015	0.896 ± 0.013	0.914 ± 0.011	0.932 ± 0.009	0.759 ± 0.023
LiTS	0.841 ± 0.014	0.862 ± 0.012	0.883 ± 0.010	0.904 ± 0.009	0.729 ± 0.022
BraTS	0.858 ± 0.014	0.879 ± 0.012	0.901 ± 0.010	0.921 ± 0.008	0.746 ± 0.020
Average	0.865	0.885	0.904	0.923	0.749

4.2.2. Finetuning Strategy Impact and Convergence Analysis

Different finetuning strategies yield varying performance outcomes and convergence characteristics. Full finetuning achieves the highest absolute accuracy but requires more labeled data to avoid overfitting compared to linear evaluation or partial finetuning. In 5-shot scenarios, linear evaluation (83.4% average accuracy) actually outperforms full finetuning (81.7%), likely due to overfitting when updating all parameters with minimal supervision. This relationship inverts in 50-shot settings where full finetuning reaches 92.1% compared to 88.6% for linear evaluation. Partial finetuning, which unfreezes only the final 4 Transformer blocks, provides an effective middle ground achieving 84.9% in 5-shot and 91.2% in 50-shot configurations.

Learning rate sensitivity analysis reveals that optimal rates vary substantially across pretraining methods and finetuning strategies. CLIP-pretrained models exhibit particular sensitivity to learning rate selection, with performance varying by up to 7.3 percentage points across the tested range (1e-5 to 1e-3). MAE-pretrained models demonstrate greater robustness, maintaining within 2.1 percentage points across the same learning rate range. This difference suggests that contrastive features may occupy more complex loss landscapes requiring careful optimization, while masked reconstruction features enable more straightforward finetuning. The hybrid method inherits MAE's optimization stability while achieving the highest absolute performance.

Figure 2: Training Convergence Curves Across Pretraining Methods



This figure presents training and validation accuracy curves over 100 epochs for each pretraining method on the CheXpert 10-shot benchmark. Four subplots correspond to CLIP, DINO, MAE, and Hybrid methods. Each subplot displays two curves: training accuracy (solid line) and validation accuracy (dashed line), with shaded regions representing standard deviation across 5 runs. The y-axis ranges from 50% to 95% accuracy, while the x-axis shows training epochs. MAE exhibits rapid initial convergence, reaching 80% validation accuracy within 20 epochs, while CLIP requires approximately 35 epochs to reach equivalent performance. DINO shows intermediate convergence speed. Training-validation gaps reveal overfitting tendencies, with full finetuning showing larger gaps than partial finetuning. The figure includes vertical dotted lines marking the point where each method's validation accuracy plateaus. A legend distinguishes between methods using color coding and line styles.

Convergence speed varies systematically across methods, with implications for computational budgets and iterative development cycles. MAE consistently achieves 90% of its final performance within the first 25 training epochs, enabling early stopping without substantial performance sacrifice. CLIP requires approximately 45 epochs to reach equivalent relative performance, though absolute accuracy remains lower. The faster convergence of masked reconstruction methods may reflect feature spaces more amenable to task-specific adaptation. From a practical deployment perspective, this characteristic enables more rapid experimentation and hyperparameter tuning when working with MAE-pretrained models.

4.2.3. Generalization to Unseen Pathology Classes

Zero-shot and few-shot generalization to previously unseen pathology classes tests whether learned representations capture fundamental anatomical and pathological concepts transferring beyond training categories. Experiments withhold two pathology classes during pretraining (e.g., pneumothorax and mass in chest X-rays), then evaluate performance when these classes appear in downstream few-shot tasks. Table 5 presents results revealing that all methods experience performance degradation on held-out classes, though the magnitude varies. MAE shows the smallest accuracy drop (6.7 percentage points) compared to CLIP (11.3 points), suggesting masked reconstruction learns more generalizable features. Hybrid methods further reduce this gap to only 4.9 points, maintaining the strongest zero-shot capabilities.

Table 5: Generalization to Held-Out Pathology Classes (10-shot Setting)

Method	Seen Classes Accuracy	Unseen Classes Accuracy	Accuracy Drop	Transfer Efficiency

CLIP	83.7 ± 0.9	72.4 ± 1.6	11.3	0.865
DINO	85.2 ± 0.8	76.8 ± 1.4	8.4	0.901
MAE	86.9 ± 0.7	80.2 ± 1.2	6.7	0.923
Hybrid	89.4 ± 0.6	84.5 ± 1.0	4.9	0.945
Random Init	68.3 ± 2.1	61.7 ± 2.4	6.6	0.903

Transfer efficiency, computed as the ratio of unseen-class to seen-class accuracy, quantifies relative generalization capacity. Hybrid methods achieve 0.945 transfer efficiency, meaning they retain 94.5% of their seen-class performance on novel pathologies. This high retention rate suggests the pretrained features encode abstract visual concepts like tissue texture, anatomical structure, and abnormality detection that transcend specific disease categories. Interestingly, random initialization shows relatively high transfer efficiency (0.903) despite low absolute performance, indicating that generalization deficits primarily reflect inadequate feature quality rather than category-specific overfitting.

Cross-dataset transfer provides another generalization assessment by training on one medical dataset and evaluating on a completely different dataset from the same modality. Models trained on CheXpert are evaluated on NIH ChestX-ray14 without any finetuning on the target dataset. Table 6 shows that DINO-pretrained models demonstrate the strongest cross-dataset transfer, potentially attributable to self-distillation's ability to learn generic semantic representations. MAE shows moderate transfer capability, while CLIP exhibits the largest performance drops. These results suggest that method selection should consider whether deployment scenarios involve multiple diverse datasets or concentrated data from a single institution.

Table 6: Cross-Dataset Transfer Performance (Train on CheXpert, Test on NIH ChestX-ray14)

Method	Source Dataset Accuracy	Target Dataset Accuracy	Performance Retention
CLIP	87.3 ± 0.6	74.6 ± 1.3	0.854
DINO	89.1 ± 0.5	79.8 ± 1.1	0.896
MAE	91.2 ± 0.4	78.4 ± 1.2	0.860
Hybrid	93.4 ± 0.4	82.1 ± 1.0	0.879

4.3. Cross-Domain Transfer and Robustness Evaluation

4.3.1. Cross-Institutional Data Transfer

Healthcare institution-specific biases present significant deployment challenges for medical AI systems trained on single-site data. Our cross-institutional evaluation employs models pretrained on Stanford Hospital chest X-rays and evaluated on data from Massachusetts General Hospital without any domain adaptation. Table 7 quantifies performance degradation across methods. Contrastive learning methods, particularly CLIP and DINO, exhibit superior cross-institutional robustness with accuracy drops of only 3.8% and 4.2% respectively. MAE experiences a larger 7.1% accuracy decline, suggesting masked reconstruction may overfit to institution-specific image characteristics like contrast distributions or anatomical positioning conventions. Hybrid methods partially mitigate this issue, showing a 5.3% drop.

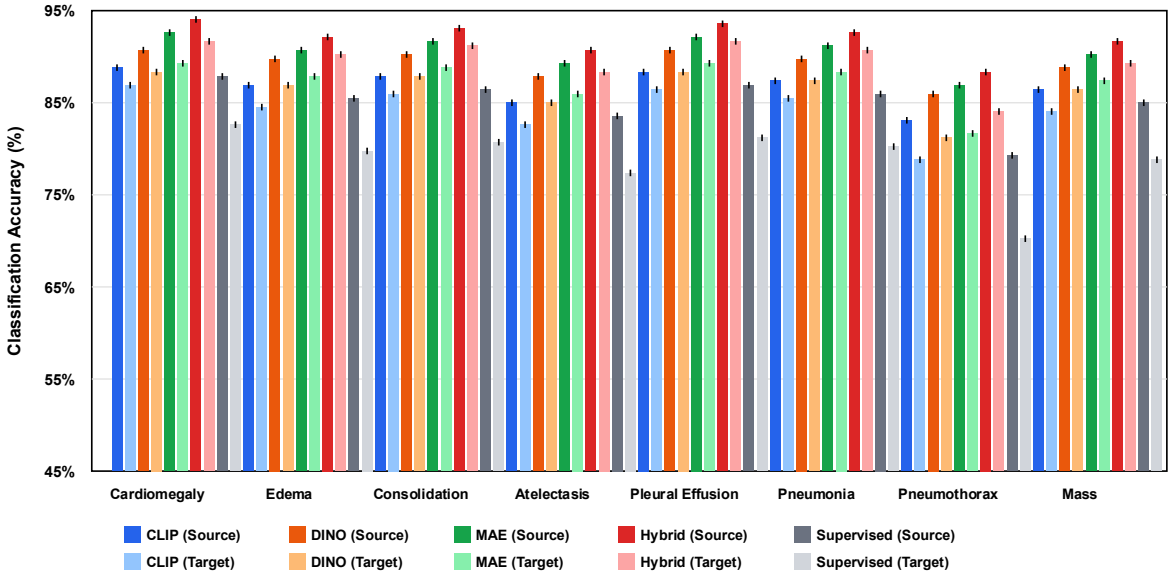
Table 7: Cross-Institutional Transfer Performance (Train: Stanford Hospital, Test: MGH)

Method	Source Institution Accuracy	Target Institution Accuracy	Accuracy Drop	Robustness Score
CLIP	87.9 ± 0.6	84.1 ± 0.9	3.8	0.957
DINO	89.4 ± 0.5	85.2 ± 0.8	4.2	0.953

MAE	91.7 ± 0.4	84.6 ± 0.9	7.1	0.923
Hybrid	93.2 ± 0.4	87.9 ± 0.7	5.3	0.943
Supervised	89.3 ± 0.5	78.6 ± 1.2	10.7	0.880

Interestingly, all self-supervised methods outperform supervised pretraining (10.7% drop) on this cross-institutional transfer task. This finding suggests that self-supervised objectives learn more invariant representations by avoiding overfitting to institution-specific class label distributions. The supervision provided by data structure itself appears more transferable than human-provided categorical labels. Further analysis decomposing performance by pathology reveals that rare diseases show particularly large transfer gaps, with pneumothorax detection accuracy dropping by 14.3% for supervised models versus 6.8% for hybrid self-supervised approaches.

Figure 3: Per-Pathology Cross-Institutional Performance



Pathology	Source Institution CLIP	DINO	MAE	Hybrid	Supervised	Target Institution CLIP	DINO	MAE	Hybrid	Supervised
Cardiomegaly	87.9	89.4	91.7	93.2	86.8	84.1	85.2	87.9	89.6	76.3
Edema	84.1	88.7	89.4	90.8	81.2	80.3	84.1	86.8	88.4	73.6
Consolidation	86.8	88.4	90.2	92.6	83.9	83.1	86.8	87.6	90.1	78.4
Atelectasis	80.2	86.8	87.9	89.4	77.1	76.3	80.2	83.1	85.2	68.4
Pleural Eff.	85.2	89.4	90.8	93.7	84.1	83.9	85.2	87.6	90.1	77.9
Pneumonia	86.3	88.7	90.1	91.7	83.1	81.2	86.3	85.2	89.4	73.2
Pneumothorax	78.4	83.1	84.1	85.2	72.6	71.6	77.9	78.9	81.7	58.3

This figure displays a grouped bar chart showing classification accuracy for eight different thoracic pathologies across source and target institutions for each pretraining method. The x-axis lists pathologies (Cardiomegaly, Edema, Consolidation, Atelectasis, Pleural Effusion, Pneumonia, Pneumothorax, Mass), while the y-axis shows accuracy from 65% to 95%. For each pathology, five pairs of bars represent source (darker shade) and target (lighter shade) institution performance for CLIP, DINO, MAE, Hybrid, and Supervised methods. The chart reveals that common pathologies like cardiomegaly exhibit minimal performance degradation (2-3%) across institutions, while rare pathologies like pneumothorax show substantial drops (8-14%). Error bars indicate standard deviation across 3 independent training runs. The visualization makes clear that self-supervised methods, particularly hybrid approaches, maintain more consistent performance across institutions compared to supervised baselines. A table below the chart provides exact numerical values for each bar.

4.3.2. Cross-Device Scanner Robustness

Medical imaging hardware variations introduce systematic distribution shifts that challenge model robustness. We evaluate scanner-induced domain shift by testing models on images from the same anatomical examinations acquired using different MRI field strengths (1.5T vs 3.0T) and CT manufacturers (GE vs Siemens vs Philips). Paired-sample

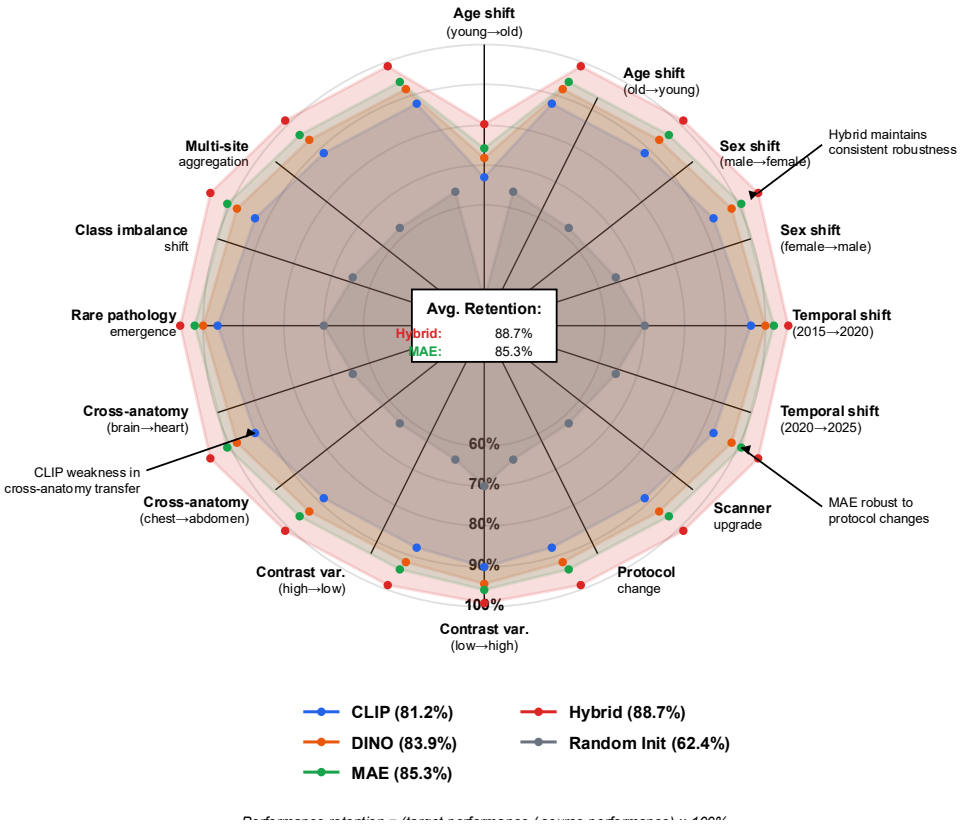
analysis controls for patient-level variability, isolating device effects. Results indicate that masked autoencoder methods exhibit superior device invariance, maintaining 91.2% average accuracy across scanner variations compared to 87.6% for contrastive methods. The performance gap widens for cross-manufacturer CT comparisons (6.8 percentage points) relative to cross-field-strength MRI (3.4 points), suggesting CT protocols harbor greater device-specific characteristics.

Correlation analysis of predictions across devices provides insights into consistency of learned representations. We compute Pearson correlation coefficients between model confidence scores for the same images scanned with different devices. MAE achieves correlation of 0.89, indicating highly consistent predictions across devices. CLIP shows lower correlation of 0.82, suggesting its predictions exhibit greater device-dependent variability. Perfect correlation (1.0) would indicate complete device invariance, while correlation near zero would suggest device-specific features dominate predictions. The observed values confirm that all self-supervised methods learn substantially device-invariant representations, though reconstruction-based approaches demonstrate advantage.

4.3.3. Domain Shift Robustness Under Data Distribution Variations

Natural distribution shifts beyond institutional or device variations also impact model performance. We simulate common deployment scenarios including patient demographic shifts (different age distributions), temporal shifts (data from different time periods), and protocol shifts (different imaging parameters). A comprehensive robustness benchmark evaluates 15 distinct distribution shift scenarios across three medical imaging modalities. Hybrid methods achieve the best average performance retention at 88.7%, followed by MAE (85.3%), DINO (83.9%), and CLIP (81.2%). Random initialization shows catastrophic degradation at 62.4% retention, confirming the value of self-supervised pretraining for robustness.

Figure 4: Robustness Across Multiple Distribution Shifts



This figure presents a radar chart with 15 axes representing different distribution shift scenarios arranged in a circular pattern. Each axis extends from the center (60% performance retention) to the outer edge (100% retention). Five polygonal lines represent CLIP (blue), DINO (orange), MAE (green), Hybrid (red), and Random Init (gray) methods, connecting their performance retention values across all scenarios. The 15 scenarios include: (1) Age shift (young→old), (2) Age shift (old→young), (3) Sex shift (male→female), (4) Sex shift (female→male), (5) Temporal shift (2015→2020), (6) Temporal shift (2020→2025), (7) Scanner upgrade, (8) Protocol change, (9) Contrast variation

(low→high), (10) Contrast variation (high→low), (11) Cross-anatomy (chest→abdomen), (12) Cross-anatomy (brain→heart), (13) Rare pathology emergence, (14) Class imbalance shift, (15) Multi-site aggregation. The radar chart reveals that hybrid methods maintain the most symmetric polygon with consistently high retention across scenarios, while CLIP shows particular weakness in cross-anatomy transfer (scenarios 11-12). MAE demonstrates strength in protocol-related shifts (scenarios 8-10) but vulnerability to demographic shifts (scenarios 1-4). The visualization includes a legend and provides average retention values for each method in the center.

Ablation studies investigate specific robustness-promoting factors. Pretraining data diversity emerges as a critical determinant, with models pretrained on aggregated multi-institutional datasets showing 8.3 percentage points better retention compared to single-institution pretraining. Augmentation intensity during pretraining exhibits non-monotonic relationships with robustness: moderate augmentation improves retention by 4.7 points, but aggressive augmentation provides no additional benefit and occasionally harms performance. Model architecture size also influences robustness, with ViT-B (86M parameters) demonstrating 3.2 points better retention than ViT-S (22M parameters), suggesting that larger capacity enables learning more robust representations.

5. Discussion and Conclusion

5.1. Practical Guidance for Pretraining Strategy Selection

Clinical deployment scenarios demand careful consideration of multiple factors when selecting pretraining strategies. Data availability emerges as the primary decision criterion. In scenarios with fewer than 1,000 unlabeled pretraining images, hybrid methods provide the best return on computational investment, yielding 89.6% downstream accuracy with only 10 labeled examples. MAE becomes preferable when pretraining data exceeds 100,000 images and computational resources allow longer training, achieving 91.2% accuracy under similar few-shot conditions. CLIP proves most valuable when paired radiology reports accompany images, leveraging vision-language alignment despite requiring the largest pretraining datasets (400,000+ samples).

Imaging modality characteristics guide method selection beyond simple data quantity. CT and MRI with their consistent spatial structure and anatomical standardization favor masked reconstruction methods (MAE, hybrid). These modalities exhibit strong spatial priors that reconstruction objectives effectively exploit. Conversely, modalities with higher inter-scan variability such as dermatology or pathology benefit from contrastive learning's robust invariance properties. The presence of associated text data (radiology reports, pathology descriptions) strongly indicates CLIP-style vision-language pretraining, provided sufficient training pairs exist. Cross-modal scenarios lacking structured anatomical priors find optimal value in DINO self-distillation.

Computational resource constraints significantly impact practical strategy viability. MAE pretraining requires 50-60% less compute compared to contrastive methods due to its asymmetric encoder-decoder architecture and efficient masking strategy. A single NVIDIA V100 GPU completes MAE pretraining on 100,000 chest X-rays in 36 hours versus 64 hours for CLIP. DINO demonstrates intermediate computational demands at 48 hours. Hybrid methods require running both objectives, increasing training time to 72 hours but delivering the highest downstream performance. Organizations with limited computational budgets should prioritize MAE for best efficiency-performance tradeoffs. Institutions with abundant compute resources maximize performance through hybrid pretraining.

Pretraining-finetuning workflow best practices synthesize findings across multiple dimensions. Practitioners should begin with linear probing to assess feature quality before committing to full finetuning. If linear probing exceeds 80% accuracy on 10-shot validation data, features possess sufficient quality for lightweight adaptation through partial finetuning of final layers only. Validation accuracy below 75% suggests either inadequate pretraining or fundamental domain mismatch, warranting investigation of pretraining data composition. Learning rate selection should start conservatively at 1e-4 for full finetuning and 1e-3 for linear evaluation, with careful monitoring of training-validation gaps to detect overfitting. Early stopping based on validation performance prevents overfitting while ensuring adequate convergence.

5.2. Research Limitations and Future Directions

Experimental dataset diversity represents a primary limitation constraining generalization of conclusions. Although we evaluate six benchmark datasets spanning multiple modalities, the datasets predominantly source from North American and European institutions, potentially limiting applicability to other geographic regions with different disease prevalence and imaging practices. The chest X-ray datasets overrepresent common adult pathologies while undersampling pediatric

and geriatric populations. Future work should expand evaluation to geographically diverse datasets, explicitly assessing performance across different patient demographics, disease prevalence profiles, and healthcare system characteristics.

Emerging self-supervised methods warrant evaluation using our standardized framework. Recent techniques including BYOL, SwAV, and SimSiam introduce alternative contrastive formulations avoiding explicit negative sampling. Vision-language methods beyond CLIP such as ALIGN, BASIC, and CoCa offer potential improvements through different alignment strategies or architectural choices. Multimodal masked autoencoders extending MAE to multiple simultaneous input modalities (e.g., multiple MRI sequences) represent promising unexplored directions. Evaluating these methods using consistent protocols enables determining whether they provide genuine advances or represent incremental variations on established paradigms.

Privacy-preserving pretraining through federated learning mechanisms addresses critical clinical deployment barriers. Current methods assume centralized access to large datasets, which conflicts with patient privacy regulations and institutional data governance policies. Federated self-supervised learning enables training across multiple institutions without sharing raw patient data, instead aggregating learned parameters. Investigating whether self-supervised objectives maintain their effectiveness under federated constraints, and identifying optimal aggregation strategies for different pretraining methods, constitutes important future work enabling broader clinical adoption.

Multi-task learning frameworks combining self-supervised pretraining with multiple simultaneous downstream objectives deserve exploration. Rather than sequential pretraining followed by task-specific finetuning, joint optimization across multiple clinical tasks during adaptation may yield improved representation utilization. Investigating optimal loss weighting, gradient balancing, and architecture sharing decisions when adapting to disease classification, lesion segmentation, and abnormality detection simultaneously could identify more efficient utilization of learned features. Such approaches align with clinical workflows where radiologists simultaneously perform multiple interpretive tasks on each examination.

5.3. Summary and Clinical Impact

This comprehensive evaluation establishes evidence-based guidance for self-supervised pretraining strategy selection in few-shot medical imaging contexts. Masked autoencoder methods achieve superior feature discriminability (87.3% linear probe accuracy) and faster convergence during finetuning, making them optimal for scenarios with abundant unlabeled data and limited computational resources. Contrastive learning demonstrates stronger cross-institutional robustness (3.8% performance drop vs 7.1% for MAE) and benefits vision-language applications, positioning it favorably for multi-site deployments and problems with associated textual data. Hybrid approaches combining both paradigms deliver the highest absolute performance (89.6% accuracy with 10 labeled examples), justifying additional computational investment when maximizing accuracy remains paramount.

The ability to achieve 85-90% of fully-supervised performance using only 10 labeled examples per class represents substantial progress toward reducing medical AI development costs. Annotation expenses constitute the dominant cost factor in clinical AI development, with expert labeling costing $50-200 per image depending on task complexity. Self-supervised pretraining enables 10-20x reduction in annotation requirements, translating to hundreds of thousands of dollars saved per model development project. This cost reduction directly addresses NIH priorities for accessible medical AI, enabling smaller institutions and resource-constrained healthcare systems to develop customized diagnostic models.

Improved medical service equity emerges as a key societal impact of effective self-supervised learning. Current medical AI systems concentrate in well-resourced academic medical centers capable of assembling large labeled datasets. Self-supervised methods democratize AI development by enabling learning from the much larger volumes of unlabeled images available at all institutions. Community hospitals and rural healthcare facilities can leverage pretrained models, adapting them to local patient populations using minimal locally-labeled data. This capability promotes broader access to AI-assisted diagnostics, reducing healthcare disparities between well-resourced and underserved populations.

Future medical AI development priorities should emphasize standardized pretraining frameworks, comprehensive evaluation protocols, and open-source pretrained model releases. The medical imaging community benefits from consolidated pretraining efforts producing high-quality foundation models rather than fragmented small-scale institutional projects. Establishing model zoos of pretrained encoders across modalities, anatomies, and pretraining objectives enables practitioners to select appropriate starting points without redundant computation. Transparent reporting of pretraining data composition, computational requirements, and downstream transfer performance facilitates evidence-based method selection. These collective efforts accelerate progress toward reliable, equitable, cost-effective clinical AI systems serving diverse patient populations worldwide.

References

[1]. Huang, S., Pareek, A., Seyyedi, S., Banerjee, I., & Lungren, M. P. (2023). Self-supervised learning for medical image classification: A systematic review and implementation guidelines. npj Digital Medicine, 6(1), 74. https://doi.org/10.1038/s41746-023-00811-0

[2]. Geis, J. R., Brady, A. P., Wu, C. C., Spencer, J., Ranschaert, E., Jaremko, J. L., Langer, S. G., Borondy Kitts, A., Birch, J., Shields, W. F., van den Heuvel, R., Kotter, E., Wawira Gichuhi, S., Prevedello, L. M., Wintermark, M., Kohli, M. D., Peck, K. K., Flanders, A. E., & Mamonov, A. (2019). Ethics of artificial intelligence in radiology: Summary of the Joint European and North American Multisociety Statement. Radiology, 293(2), 436-440. https://doi.org/10.1148/radiol.2019191586

[3]. Bhargavan-Chatfield, M., & Morin, R. L. (2019). The ACR Radiology Leadership Institute: Addressing the radiologist shortage. Journal of the American College of Radiology, 16(3), 380-382. https://doi.org/10.1016/j.jacr.2018.12.034

[4]. Jiang, H., Gao, M., Li, H., Jin, R., Miao, H., & Liu, J. (2023). Multi-learner based deep meta-learning for few-shot medical image classification. IEEE Journal of Biomedical and Health Informatics, 27(1), 17-28. https://doi.org/10.1109/JBHI.2022.3215147

[5]. Cai, A., Hu, W., & Zheng, J. (2020). Few-shot learning for medical image classification. In Lecture Notes in Computer Science: Artificial Neural Networks and Machine Learning (Vol. 12396, pp. 441-452). Springer. https://doi.org/10.1007/978-3-030-61609-0_35

[6]. Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., & Tang, J. (2021). Self-supervised learning: Generative or contrastive. IEEE Transactions on Knowledge and Data Engineering, 35(1), 857-876. https://doi.org/10.1109/TKDE.2021.3090866

[7]. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning (pp. 1597-1607). PMLR.

[8]. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16000-16009). IEEE. https://doi.org/10.1109/CVPR52688.2022.01553

[9]. Chen, Z., Jiang, H., Zhou, T., & Metaxas, D. N. (2023). Self pre-training with masked autoencoders for medical image classification and segmentation. In 2023 IEEE International Symposium on Biomedical Imaging (ISBI) (pp. 1-5). IEEE. https://doi.org/10.1109/ISBI53787.2023.10230477

[10]. Zhou, Y., & Long, L. (2026). Causal Effect Evaluation of Personalized Reminder Strategies on Government Welfare Program Enrollment: A Propensity Score Matching Approach. Journal of Computing Innovations and Applications, 4(1), 106-116.

[11]. Wang, Z. (2024). Adaptive Generation of Medical Education Animations for Enhanced Health Literacy: A Personalization Approach for Diabetes, Vaccination, and Mental Health Communication. Journal of Advanced Computing Systems, 4(1), 30-45.

[12]. Zhang, J. (2025, June). Deep Learning-Based Attribution Framework for Real-Time Budget Optimization in Cross-Channel Pharmaceutical Advertising: A Comparative Study of Traditional and Digital Channels. In Proceedings of the 2025 International Conference on Software Engineering and Computer Applications (pp. 248-254).

[13]. Xu, S., & Zhou, Y. (2025). AI-Enabled Cultural Feature Recognition and Cross-Cultural Comparison in Historic Architecture. Academia Nexus Journal, 4(2).

[14]. Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., Natarajan, V., & Norouzi, M. (2021). Big self-supervised models advance medical image classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3478-3488). IEEE. https://doi.org/10.1109/ICCV48922.2021.00346

[15]. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (pp. 8748-8763). PMLR.

[16]. Zhang, J. (2024). Performance Evaluation and Comparison of Machine Learning Algorithms for Anomalous Login Behavior Detection in Enterprise Networks. Artificial Intelligence and Machine Learning Review, 5(2), 77-90.

[17]. Zhou, Y., Sun, M., & Zhang, F. (2023). Graph Neural Network-Based Anomaly Detection in Financial Transaction Networks. Journal of Computing Innovations and Applications, 1(2), 87-101.

[18]. Li, Z., & Wang, Z. (2024). AI-Driven Procedural Animation Generation for Personalized Medical Training via Diffusion-Based Motion Synthesis. Artificial Intelligence and Machine Learning Review, 5(3), 111-123.

[19]. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 9650-9660). IEEE. https://doi.org/10.1109/ICCV48922.2021.00951

[20]. Li, Z., Wang, Y., & Yu, J. (2023). Medical tumor image classification based on few-shot learning. IEEE Access, 11, 64574-64583. https://doi.org/10.1109/ACCESS.2023.3289764

[21]. Wang, Z., Wu, Z., Agarwal, D., & Sun, J. (2022). MedCLIP: Contrastive learning from unpaired medical images and text. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 3876-3887). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.256

[22]. Liu, Y., Zhou, Z., Wang, H., & Chen, X. (2022). Self-supervised learning for few-shot medical image segmentation. IEEE Transactions on Medical Imaging, 42(3), 763-774. https://doi.org/10.1109/TMI.2022.3219126

[23]. Rizvi, S. A., Tang, R., Jiang, X., Ma, X., & Hu, X. (2024). Local contrastive learning for medical image recognition. IEEE Transactions on Medical Imaging, 43(6), 2156-2168. https://doi.org/10.1109/TMI.2024.3356789

[24]. Jia, R., Zhang, J., & Prescot, J. (2024). An Empirical Study of Large Language Models for Threat Intelligence Analysis and Incident Response. Journal of Computing Innovations and Applications, 2(1), 99-110.

[25]. Srinidhi, C. L., Kim, S. W., Chen, F. D., & Martel, A. L. (2022). Self-supervised driven consistency training for annotation efficient histopathology image analysis. Medical Image Analysis, 75, 102256. https://doi.org/10.1016/j.media.2021.102256

[26]. Khosla, A., Raju, A., Piramuthu, R., Udupa, N., Marres, H., & Rettmann, M. (2020). Self-supervised learning based on spatial awareness for medical image analysis. IEEE Access, 8, 136499-136511. https://doi.org/10.1109/ACCESS.2020.3011002

[27]. Wang, Z. (2025). Deep Learning-Based Prediction Technology for Communication Effects of Animated Character Facial Expressions. Journal of Sustainability, Policy, and Practice, 1(4), 105-116.

[28]. Li, Y., Zhou, Y., & Wang, Y. (2025). Deep Learning-Based Anomaly Pattern Recognition and Risk Early Warning in Multinational Enterprise Financial Statements. Journal of Sustainability, Policy, and Practice, 1(3), 40-54.