NOTATION TRENDPLUS PUBLISHING
WHERE IDEAS MEET IMPACT

Volume-1, Issue 2, August 2026, Page No- 113-130
Journal of Machine Learning Innovations and Artificial Intelligence Horizons (JMLIAIH)

Self-Supervised Contrastive Learning with Vision Transformers for Data-Efficient Medical Image Classification

Arijit Gandhi

Research Scholar, Department of Management, RKDF University, Kathal More - Argora, Ranchi, Jharkhand 834004.

DOI: https://doi.org/10.70903/jmliaih.2026.1207

Article Received: 13-07-2026
Revised Version: 20-08-2026
Accepted: 27-08-2026

*Corresponding Author

Arijit Gandhi

E-Mail Id:
agstudy2025@yahoo.com

Abstract: Medical image classification often suffers from the scarcity of labeled data, which limits the performance of deep learning models. We propose a self-supervised contrastive learning framework that integrates Vision Transformers to address this data efficiency challenge. The method first applies rigorous image preprocessing, including resizing, intensity normalization, and contrast enhancement, to ensure input consistency. A stochastic data augmentation module then generates two distinct views of each input image through transformations such as random cropping, rotation, and intensity modifications, all carefully constrained to preserve anatomical integrity. The core novelty lies in the self-supervised pretraining phase, where a Vision Transformer encoder maps these augmented views into a latent feature space. A contrastive loss function maximizes agreement between positive pairs derived from the same image while minimizing similarity between negative pairs from different images. This objective forces the encoder to learn transformation-invariant representations that capture high-level semantic content rather than superficial statistics. After pretraining, the encoder weights initialize a downstream supervised classification task, where a fully connected head is appended and fine-tuned on limited labeled data. A strategic freezing mechanism prevents catastrophic forgetting by keeping initial encoder layers fixed while gradually unfreezing deeper layers

Keywords: Self-Supervised Learning, Contrastive Learning, Vision Transformers, Medical Image Classification, Data-Efficient Learning.

Introduction

This fast track development of deep learning has enriched medical image analysis making it possible to reach medical diagnostic assistance systems that can be as efficient and compete successfully with human experts in certain fields (Anwar et al., 2018). Convolutional neural networks (CNNs) have been the powerful architecture used in such applications and have excellent performance in many applications from lesion detection to disease classification (Gao et al., 2019). Being trained in a clinical setting, however, is seriously limited by the demand of a large quantity of high-quality annotated data. In the medical field, gaining such annotations is very difficult and expensive, as it would require highly skilled and professional radiologists to annotate each image with minimal errors (Ghesu et al., 2022). This restriction makes supervised learning techniques difficult to scale and is even more of a challenge in the case of rare diseases or new imaging modes with very limited labeled collections.

To tackle this issue of limited data, self-supervised learning (SSL) paradigms that enable the derivation of meaningful visual representations from unlabeled data without manual supervision have emerged as a preferred approach (Huang et al., 2023). After the various SSL methods, contrastive learning is one of the most effective methods which yields most impressive results with maximal similarity between different augmented versions of the same image and maximal dissimilarity between two augmented versions of different images (T. Chen et al., 2020). The studies of SimCLR (T. Chen et al., 2020) and MoCo (He et al., 2020) have shown that excellent featurization performance can be achieved when using large-scale unlabeled data to pretrain an encoder, which is transferable to a wide variety of tasks using a smaller amount of labeled data. More recently, other works have demonstrated that high-quality representations can be trained without explicitly mining negative pairs such as SimSiam (X. Chen & He, 2021), which use architectural technology like stop-gradient operations and predictor networks.

For medical imaging, these widely used SSL methods are also being adapted to tackle their specific concerns of anatomical variability, modality-specific noise, and the critical requirement to maintain diagnostic information in augmentation (Gazda et al., 2021). For instance, there are some strategies for augmentation that retain anatomical information for chest X-rays and histopathology slides (Wang et al., 2021). Notwithstanding such progress, all basic medical imaging SSL methods are based on CNN encoders may suffer from the inherent constraint of capturing long-range spatial context and an understanding of the global image context, which is essential for many medical imaging diagnosis tasks. Recently, Vision Transformers (ViTs) aim to represent vision through a self-attention mechanism (Dosovitskiy et al., 2020) have appeared as an attractive alternative to CNNs. Yet, they have not yet been applied to a large scale in the context of self-supervised pretraining for medical image classification, where they can be trained with limited data.However, their application in the data-efficient medical image classification context is relatively unexplored in the area of self-supervised pretraining with usable amount of medical images.

To cope with the data efficiency issue, we introduce a new framework based on Vision Transformers for self-supervised contrastive learning. The key idea is the two-stage feature representation learning process to separate feature representation learning from label dependency. First, a large batch of unlabeled medical image data is collected and fed into a pretrained Vision Transformer (ViT) encoder, which is trained using a contrastive objective to learn in an end-to-end manner intrinsic anatomical structures, texture patterns and disease-related visual features. Second, in order to allow the pretrained weights to easily be adapted for a specific diagnostic task, a calibrated training regime of a far smaller labeled subset of the dataset is used and keys are followed to preserve the original weights of the deeper layers of the encoder, only updating the top layers which are "frozen". This method significantly reduces the need for large annotated datasets and improves the generalisability and resistance to overfitting in

small data settings. The main distinguishing features of the proposed methodology over any previous method are that they use the Vision Transformer to ensure superior modeling capability for global contextual relationships, and that they are designed to keep data augmentations small enough to preserve anatomical integrity and relationship, which it is hoped leads to clinically meaningful learned data representations.

The rest of this paper is organized as such. The second part summarizes the existing research on self-supervised learning in medical image analysis and Vision Transformer model architectures. In Section 3, background on theories and concepts behind contrastive learning for self-supervision and their medical imaging adaptation are given. The proposed framework, which involves a data preprocessing pipeline, stochastic augmentation module, a contrastive pretraining procedure, along with a fine tuning strategy, is detailed in Section 4. The experimental setup (datasets, evaluation metrics and implementation details) are detailed in Section 5. Section 6 presents experimental results for our method along with baselines of supervised methods and other SSL approaches. The implications, shortfalls of the current research and future research directions are further discussed in Section 7. The paper is concluded with a summary of our contributions, in Section 8.

Related Work

Self-Supervised Learning in Medical Imaging

In medical image analysis applications, large amounts of labeled data are often unavailable, and self-supervised learning has become one of the major paradigms to overcome this challenge. An early attempt in this regard was to propose pretext tasks like image inpainting, predict image rotation, and jigsaw unsolving puzzles to gain interesting feature representations from unlabeled data (Rani et al., 2024). But with these manual assignments, representations do not usually produce optimal results that are most useful for diagnostic purposes downstream. Contrastive learning, especially SimCLR (T. Chen et al., 2020) and MoCo (He et al., 2020), is a major step forward that directly optimizes the feature space to make it invariant to data augmentation and discriminative over instances. The optimization of these generic systems to medical imaging also has successful applications, such as in the classification of chest X-rays (Gazda et al., 2021) and the analysis of histopathology images (Wang et al., 2021).

In the medical imaging domain, (P)T)OET has been specifically targeted at the problem of limited annotations in medical imaging, with a number of works addressing this challenge. For example, Azizi et al. (2022 (Ghesu et al.)) proposed a method which is pre-trained on contrastive learning of 100 million medical images and yields significant gains for limitedly labeled downstream classification tasks. In a similar thread, Sowrirajan et al. (Zeng et al., 2024) have proposed a approach to medical image segmentation with constraints on annotation using contrastive learning and random walks. All of these methods have demonstrated that self-supervised pre-training allows for only an order of magnitude less labeled data than competitive performance. Most of these approaches, however, use convolutional networks (CNNs) as the backbone encoder that can cause issues when capturing long-range spatial dependencies which would be important for a wide range of diagnostic applications.

Vision Transformers for Medical Image Analysis

However, recently a powerful alternative to CNNs has come into the scene – Vision Transformers (ViTs) that can capture the global contextual relationships using self-attention (Dosovitskiy et al., 2020). As opposed to CNNs operating over local receptive fields, ViTs break down an image into patches and view them as serial of tokens, allowing the model to determine interactions between far-flung areas. Such spatially distributed patterns over the image are typically associated with diseases in the medical

image analysis context, which makes this architectural property an important one for this application. In example, chest X-ray classification: having pneumonia could be marked by having opacities bilaterally that cover multiple regions in the lung, this means that the model needs to use information from a distant anatomical area.

Several studies carried out have investigated use of ViTs in medical imaging with the results showing that they outperform CNNs in various medical imaging tasks like organ segmentation, lesion detection, and disease classification (Shamshad et al., 2023). But, the necessity to train ViTs often results in even greater datasets than CNNs because they have no inherent inductive biases, consequently exacerbating the data scarcity hassle. It has spurred investigations into self-supervised pretraining methods tailored for ViT. In fact, Caron et al. (Caron et al., 2021) suggested DINO, a self-supervised approach that distills ViT representation without labelled supervision, which achieved outstanding performance in natural image classification. On the other hand, masked image modeling methods (e.g. MAE (He et al., 2022)) have emerged as useful methods in the medical imaging context, and have proven effective in training anatomically relevant feature representations from unlabeled data.

Contrastive Learning with Vision Transformers

Contrastive learning and Vision Transformers are only two progressive advances in self-supervised representation learning. Chao Zhang et al. (C. Zhang et al., 2020) proved that ViTs are effective with contrastive objectives, and that they can reach competitive results on ImageNet classification. However, using this method with medical imaging is quite under-developed. A key issue is how to construct diversified data augmentations that enable the learning of invariance while still retaining clinically-relevant information. When it comes to medical images, the subtle pathological signs are more likely to be lost during the process of augmenting it aggressively like random cropping or coloring it differently than natural images.

To tackle this problem, we present our framework which combines Vision Transformers with an adapted contrastive learning pipeline for medical imaging. The main novices in this case is the decoupling of the label dependency from the features representation learning by adding a two-stage training phase: first, pretraining the ViT encoder on unlabeled medical images using a contrastive objective; and second, fine-tuning the ViT encoder on the limited labeled data with a strategic freezing strategy. This is different to the current methods in three key ways. We adopt Vision Transformers as our backbone encoder, and demonstrate that it achieves better performance in modeling the global contextual relationship than CNN-based models. Moreover, our data augmentation module is tailored towards enforcing anatomical structure, thereby emphasizing the learning of clinically relevant features over superficial statistics. Third, the strategic freezing mode used in fine-tuning preserves the catastrophic forgetting of previous knowledge and also enables the model to learn the diagnostic patterns of the specific task. In combination, the design elements allow our framework to outperform state-of-the-art models in the setting of very limited data, which makes it well suited for clinical applications in situations where the labeling is provided by experts, and is costly and time-consuming.

Background on Self-Supervised Contrastive Learning for Medical Imaging

Self-supervised contrastive learning is one of the main methods to learn a visual representation from uncurated datasets, especially with few labelled examples in its problem. The basic idea of this paradigm is to train an encoder network on this property: to render into similar representations texts that have the same semantics, and to render texts that do not share the same semantics into different representations, without relying on any kinds of labels. This section serves as a background on the fundamentals of

contrastive learning, how it has been adapted to medical imaging applications, and the implications to take into account within medical imaging contexts.

Core Principles of Contrastive Learning

Contrastive learning can be viewed as a method for making positive pairs and negative pairs, which are constructed based on unlabeled examples. A stochastically-based data augmentation module produces two different views of each image in a batch in the presence of image transformations including random cropping, color distortions, and Gaussian blur. Such enhanced perceptions are then fed into a network, usually a deep neural network, that gives feature representations. The contrastive loss function (typically the InfoNCE loss presented by Oord et al., 2018) is used to maximize how similar two views from the same image are, and minimize how similar two views from different images are. In the formal setting if there is a set of images B, then the loss of an image pair (i, i') is created as:

$$\mathcal{L}_{i,j} = -\log \frac{\exp\left(\text{sim}\left(z_i, z_j\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)} \quad (1)$$

Here, denotes cosine similarity between the feature representations of the two augmented views and the other terms in the denominator are summed over all other views in the batch; is a temperature parameter which regulates the concentration of the distribution. This formulation directly supports the encoder learning transformation-invariant representations, instead of just low-level statistics, because they capture high-level semantic content.

A number of key design decisions are crucial to effective contrastive learning. The variety of the data augmentations is crucial to make them effectively inform the model of being invariant to nuisance factors, yet retaining the meaning of the image. For natural images, the use of augmentation techniques like random cropping, colour jittering and Gaussian blur has worked well (T. Chen et al, 2020). The batch size also will profoundly influence performance; larger batches will yield more negative pairs which will result in improved discrimination. One approach to this is done by maintaining a dynamical queue of negative samples like MoCo (He et al., 2020) which separates the batch size and number of negative pairs. Furthermore, training with a small MLP that maps out the encoder output to the contrastive loss loss space has been demonstrated to be helpful for higher quality representation by discarding low level information not extensive to downstream tasks (T. Chen et al., 2020).

Adaptation to Medical Imaging

Taking contrastive learning into medical imaging comes with the need to carefully consider the special nature of clinical data. Medical images, in contrast to natural images, are highly consistent with respect to the anatomy, but contain subtle pathological variations that are very important for diagnostic purposes. Most augmentations used in natural images will result in aggressive color distortion and/or random cropping that remove large parts of the image, destroying clinically relevant information. For instance, if the image is cropped too aggressively, small nodules/consolidations in chest X-rays may be missed. Hence, strategies for augmentation that are tailored to the medical imaging domain were designed to ensure the preservation of anatomical structure, while also maintaining a sufficient diversity set for contrastive learning (Gazda et al., 2021).

Other common augmentations in medical imaging involve constrained random cropping (reserving the central region of interest), rotational augmentation within a small range of angles (to preserve the anatomical orientation), and intensity modifications that simulate variations of acquisition parameters. The augmentations are intended to be neutral for diagnostic purposes, in that is alters the appearance of the image without changing the pathological features. For example, by randomly varying the level of

brightness and contrast, the image can be made to look like a variation in X-ray exposure, and by adding Gaussian noise to the image, sensor noise can be simulated. The realisation is that the augmentation distribution needs to be consistent with the natural variability that exists during clinical practice, and this should ensure that the representations learned are robust to nuisance factors and would be sensitive to disease-specific patterns.

Challenges and Opportunities

Although the concept of contrastive learning is attractive for medical imaging, a number of issues still need to be addressed. The most crucial aspect is the domain shift between natural and medical imaging data, such that the augmentations, architectures that work well for ImageNet may not carry over to medical data. Furthermore, small-scale unlabeled databases of medical images exist, especially for uncommon diseases and types of imaging, which restricts pretraining. In recent years, however, new strategies have been initiated to present large medical image repositories, e.g. MIMIC-CXR (Johnson et al., 2019). Moreover, the introduction of modality-specific pretraining approaches like histology pretraining for the specific task of histopathology (Wang et al., 2021) and retinal images pretraining for the specific task of retinal imaging (Huang et al., 2023) illustrates the potential for modality-adaptive contrastive learning to be applied to improve the efficiency of data downstream.

The applicability of Vision Transformers to contrastive learning in Medical Imaging is a specific speaking point to be considered. Although they can be trained to outperform many other medical imaging workstation models on multiple benchmarks when pre-trained using large data sets (Shamshad et al., 2023), data intense models such as ViTs are especially vulnerable due to their reliance on effective self-supervised pretraining. It is possible to learn rich and anatomically meaningful representations from unlabeled medical images with ViTs' global modeling capabilities in combination with the label-free learning of the Contrastive objectives. This serves as the basis for the proposed framework that is detailed below.

Proposed Framework: Decoupled Contrastive Pretraining and Task-Specific Fine-Tuning

We introduce a two-stage approach which separates feature representation learning from dependency on labels. During self-supervised contrastive pretraining, an unlabeled medical image corpus is encoded by a Vision Transformer in the first stage to learn the statistics of different image modalities.In the first stage, a Vision Transformer is used to encode a large set of unlabeled medical images to learn the statistics across diverse image modalities through self-supervised contrastive pretraining. In the second stage, the pretrained encoder is adapted to a specific diagnostic task by following supervised fine-tuning on a small labeled diagnostic task sample by using a clever freezing technique to guarantee the preservation of the learned representations and keep the encoder parameter values consistent with the original task.

Data Preprocessing and Augmentation Pipeline

All input image data is subjected to an extensive preprocessing pipeline, with consistency across the wide variety of acquisition protocols used in medical imaging before any learning can be performed. To adhere to the required input size of Vision Transformers, namely 224×224 pixels, each image has been resized in a standard manner, bilinearly, to a fixed spatial resolution.All images are first squarely resized to the fixed and standard spatial resolution of 224×224 pixels. Intensity normalization is then performed by subtracting the mean of all the pixels in the training set, and dividing by the standard deviation across the full training set, thereby centering the intensity of all the pixels, and making them variance one. Further, contrast limited adaptive histogram equalization (CLAHE) is used to improve

the local contrast and to have more fine pathological details in the regions which might be under exposed.

During the stochastic data augmentation, two views of each input image, and , are produced by applying carefully designed and constrained transformations to the image that maintains its anatomical integrity. The augmentation set contains random crops, between 70% and 100% size of the original image, resized to 224×224. This is intentionally a smaller range than used in natural image contrastive learning (T. Chen et al., 2020) since some parts of the image may not be diagnostically useful, for example, the apices of the lungs or costophrenic angles of the chest in a chest X-ray, if they are removed by overly aggressive cropping. Randomly rotated within ±10 degree to account for minor patient positioning variations without changing anatomical orientation. Intensity variations: Randomly changed by ±20% in brightness and ±10% in contrast – simulating variation of X-ray exposure setups. The sensor noise is simulated by adding some Gaussian noise with a standard deviation of 0.01. More importantly, we don't include a transformation that would change the spatial relationships that are important for diagnosis—for example, non-rigid deformation breaking anatomical continuity, or intensity inversion giving rise to patterns that might resemble disease. This is a smaller image set that performs the same diagnostic information while also being different enough for contrastive learning.

Vision Transformer Encoder Architecture

The encoder network is a Vision Transformer (ViT) which takes the enhanced image views as input to generate image feature representations. An input image is first split into a 14×14 non-overlapping grid of patches of dimensions 14x14 pixels, and a sequence of patches Patches is sequentially formed. A patch embedding layer is then trained in order to project each patch into the -dimensional embedding space linearily, resulting in a sequence of patch tokens (1,2,3,…,n). This is followed by a learnable class token and a fixed sinusoidal positional encoding to preserve information about the position.

The sequence of tokens is then fed into a series of transformer encoder layers, which include multi-head self-attention (MHSA) and a feed-forward network (FFN) with residual connections and layer normalization. The self-attention mechanism is used to extract attention weights between any token pair so that the model can pick up long-range spatial structure, which is essential for medical diagnosis. The summary that is produced by the last layer of the Encoder is the global image representation. For our implementation, we have adopted the ViT-Base architecture with 12 heads, transformer layers and . 24 million parameters. This architecture have a decent compromise of representation and computation efficiency for medical imaging applications.

Contrastive Pretraining Phase

In the pre-training stage, the network forms robust representations of the features of the images from unlabeled data as a contrastive objective. The augmentation module produces two views from each of the images of a batch of size . The good dissolution of both views is fed into the Vision Transformer encoder for the purpose of getting the representations of the global views and . These representations are then projected via a projection head, a two-layer MLP of hidden dimension 2048 and output dimension 256, and subsequently L2-normalized, to get and . The projection head extracts irrelevant low-frequency components, which helps information to be more informative for downstream tasks (T. Chen et al., 2020).

The contrastive loss is calculated from the InfoNCE formulation (Oord et al., 2018) and the negative set has been modified to exclude the negative pair. The positive pair giving (+1, +1, +1, +1) and the

negative set is (+1, +1, +1, -1) etc. of all the other views seen from different images containing and . The loss for a single positive pair is:

$$\mathcal{L}_i = -\log \frac{\exp\left(\text{sim}\left(z_i^1, z_i^2\right)/\tau\right)}{\sum_{j=1}^{N}\sum_{k\in\{1,2\}} \mathbb{1}_{[j\neq i]} \exp\left(\text{sim}\left(z_i^1, z_j^k\right)/\tau\right)} \quad (2)$$

where $\text{sim}(u, v) = u^\top v$ is the cosine similarity between normalized vectors, and $\tau = 0.1$ is the temperature parameter. The total loss for the batch is the average over all positive pairs:

$$\mathcal{L}_{\text{contrast}} = \frac{1}{2N}\sum_{i=1}^{N}\left(\mathcal{L}_i^{(1)} + \mathcal{L}_i^{(2)}\right) \quad (3)$$

where $\mathcal{L}_i^{(1)}$ is used as anchor and is used as anchor. The representation is symmetric, with the same contribution to both the learning signal. If in a medical application very subtle changes in augmentation need to be regarded as the same image, one cannot use the indicator function without special care because different views of the same image may then be interpreted as negative and thus be regarded as distinct. The initial training process (pretraining) is done for a fixed number of epochs with AdamW optimizer with learning rate of and weight decay of 0.05.

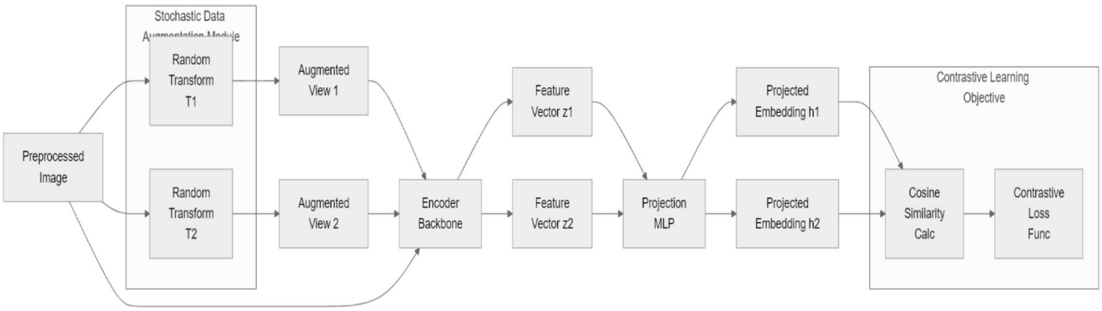

Figure 1. Self-Supervised Pretraining Phase Architecture

The self-supervised pretraining pipeline is shown in figure 1. The medical pictures are input to the stochastic augmentation module without labeling, where two different views of the pictures are produced, followed by encoding by the Vision Transformer module. In the "projection head" the encoder output is mapped into the loss space of the contrastive loss objective, which seeks to maximize the agreement of positive pairs and minimize agreement of negative pairs.

Supervised Fine-Tuning with Strategic Freezing

The weights of the encoder can be transferred as pretrained weights for an applied classification task. The projection head is discarded, and a classification head is added to the encoder. The classification head consists of a single linear layer which transforms class token representation into class logits and finally a softmax activation for multi-class classification. During the supervised fine-tuning steps, the technique is fine-tuned on a small amount of labeled data, and is much smaller than the number of unlabeled images used for pretraining.

We use this strategic freezing mechanism after pre-training to make sure the medical features retain, by using a progressive unfreezing schedule. There are three phases in the fine-tuning process:

Phase 1: All layers in the encoder are frozen, and only a certain number of epochs is used within the classification head. This is a stage to train the model to map the pre-trained feature space to the

diagnostic labels, but this stage does not involve weight changes of the encoder. Loss function: the standard cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{K} y_{i,k}\log\hat{y}_{i,k} \quad (4)$$

where K is the number of classes, $y_{i,k}$ is the ground-truth label, and $\hat{y}_{i,k}$ is the predicted probability for class k.

Phase 2: The highest level or most complex layers (encoder layers) are unfrozen, while the "lower" layers are still frozen. The classification head remains frozen and is trained alongside with the unfrozen layers of the encoder. The model is given a chance to optimize high-level semantic representations for the particular diagnostic task, without losing the low-level anatomical representations acquired during the pretraining phase.

Phase 3: All encoder layers are unfrozen, and the entire model is fine-tuned end-to-end with a reduced learning rate. The learning rate for the encoder is set to 10^{-5}, which is an order of magnitude smaller than the learning rate for the classification head (10^{-4}). This differential learning rate ensures that the pretrained features are not drastically altered while allowing the model to make fine-grained adjustments for the downstream task.

The progressive unfreezing schedule is implemented by gradually increasing the number of trainable layers as training progresses. Specifically, Phase 1 runs for 10 epochs, Phase 2 runs for 20 epochs, and Phase 3 runs for the remaining epochs up to a total of 100 epochs. This schedule is designed to balance the need for task-specific adaptation with the preservation of generalizable medical features. The fine-tuning uses the AdamW optimizer with a cosine learning rate decay and early stopping based on validation loss to prevent overfitting on the small labeled dataset.

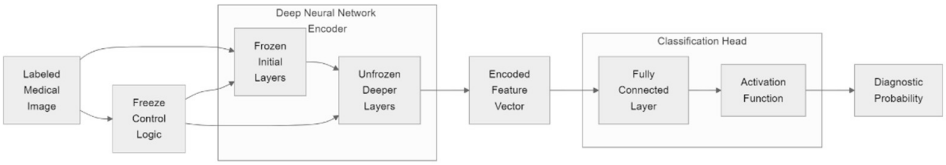

Figure 2. Supervised Fine-Tuning and Diagnostic Inference Phase

Figure 2 illustrates the fine-tuning phase. The pretrained Vision Transformer encoder is initialized with weights from the contrastive pretraining phase. The classification head is appended, and the strategic freezing mechanism is applied, with initial encoder layers frozen while deeper layers are gradually unfrozen. The final output is a class prediction for the diagnostic task.

Experimental Setup

This section describes the experimental configuration designed to evaluate the proposed self-supervised contrastive learning framework with Vision Transformers for data-efficient medical image classification. We detail the datasets used, the preprocessing pipeline, the baseline methods for comparison, the evaluation metrics, and the implementation specifics.

Datasets

In order to fully evaluate the performance of our framework, we use three publicly available medical imaging datasets for the evaluation of several different medical imaging modalities and medical imaging

tasks. This data set is the chest X-ray 14 (Baltruschat et al., 2019) data set which contains 112,120 chest X-ray images supplied in front-on view of the chest, taken from a total of 30,805 different patients, each of which is annotated with 14 different labels for thoracic pathologies (including those of atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax and others). This dataset has the following difficulties to be challenging for representation learning in the absence of abundant data: It is multi-label and class-imbalanced. The second dataset was the Kaggle Diabetic Retinopathy Detection dataset (Porwal et al., 2018) with 88,702 high-resolution retinal fundus images that were labeled on a five-point scale for the severity of DR. The data in this dataset come with some difficulties in subtle pathological features, and changes in image quality from site to site. The third dataset is the Brain-tumor MRI dataset from Figshare (Rahman & Islam, 2023), which contains T1 weighted MRI images after the use of contrast enhancement technique acquired from 233 patients having three categories of brain tumor namely menigioma, glioma and pituitary tumor and each of these image consists of 3064 images. The present dataset enables to assess the framework in regard to challenging volumetric imaging modalities having different anatomical structures.

Preprocessing and Augmentation

All images undergo a standardized preprocessing pipeline before entering the model. Each image is resized to 224×224 pixels using bilinear interpolation, consistent with the standard input size for Vision Transformers. Intensity normalization is applied by subtracting the dataset-specific mean and dividing by the standard deviation computed across all training images. For chest X-rays and retinal fundus images, we additionally apply contrast-limited adaptive histogram equalization (CLAHE) with a clip limit of 2.0 and a tile grid size of 8×8 to enhance local contrast in under-exposed regions. For MRI images, we perform skull stripping using a simple threshold-based method to remove non-brain tissue, followed by intensity normalization to zero mean and unit variance. All preprocessing steps are implemented using the MONAI framework (Gupta et al., 2024), which provides optimized medical imaging operations.

During training, data augmentation is applied on-the-fly to generate diverse views for contrastive learning and to improve generalization during fine-tuning. The augmentation pipeline for the self-supervised pretraining phase includes: random cropping with a scale range of [0.7, 1.0] of the original area, followed by resizing back to 224×224 pixels; random rotation within ±10 degrees; random brightness adjustment within ±20% and random contrast adjustment within ±10% of the original values; and additive Gaussian noise with a standard deviation of 0.01. For the supervised fine-tuning phase, we use a lighter augmentation set consisting of random horizontal flipping (where clinically appropriate, e.g., for chest X-rays but not for retinal fundus images), random rotation within ±5 degrees, and random intensity scaling within ±10%. These augmentations are carefully chosen to preserve anatomical integrity while providing sufficient diversity to prevent overfitting on the limited labeled data.

Baseline Methods

We compare our proposed framework against four baseline methods that represent different paradigms for medical image classification under data scarcity. The first baseline is a **Convolutional Neural Network trained from scratch** (CNN-Scratch), which uses a ResNet-50 architecture (Shafiq & Gu, 2022) initialized with random weights and trained directly on the limited labeled dataset without any pretraining. This baseline establishes the lower bound of performance achievable without leveraging external data or self-supervised learning. The second baseline is **Transfer Learning from ImageNet** (TL-ImageNet), which initializes a ResNet-50 encoder with weights pretrained on the ImageNet dataset (Deng et al., 2009) and fine-tunes it on the labeled medical images. This baseline represents the common practice of using natural image pretraining for medical tasks, despite the domain gap between natural

and medical images. The third baseline is a **Supervised Vision Transformer** (ViT-Supervised), which trains a ViT-Base model from scratch directly on the limited labeled dataset using standard supervised learning with cross-entropy loss. This baseline isolates the effect of the Vision Transformer architecture without self-supervised pretraining. The fourth baseline is a **Self-Supervised CNN** (SSL-CNN), which applies the same contrastive learning framework but replaces the Vision Transformer encoder with a ResNet-50 backbone. This baseline allows us to isolate the contribution of the Vision Transformer architecture to the overall performance of our framework.

All baseline methods use the same preprocessing pipeline, augmentation strategy during fine-tuning, and evaluation protocol as our proposed method to ensure fair comparison. For the TL-ImageNet baseline, we use the standard ImageNet pretrained weights available in the PyTorch model zoo (Paszke et al., 2019). For the SSL-CNN baseline, we use the same contrastive pretraining procedure described in Section 4.3, but with a ResNet-50 encoder instead of a Vision Transformer.

Evaluation Metrics

Model performance is assessed using a comprehensive set of metrics that capture different aspects of classification quality, with particular emphasis on clinically relevant measures. The primary metrics include **accuracy**, which measures the proportion of correctly classified samples; **precision** (positive predictive value), which quantifies the proportion of positive predictions that are truly positive; **recall** (sensitivity), which measures the proportion of actual positive cases that are correctly identified; **specificity**, which measures the proportion of actual negative cases that are correctly identified; **F1-score**, which is the harmonic mean of precision and recall; and **Area Under the Receiver Operating Characteristic Curve** (ROC-AUC), which provides a threshold-independent measure of discriminative ability. For the multi-label ChestX-ray14 dataset, we compute these metrics using a one-vs-rest approach and report the macro-averaged values across all 14 pathology classes. For the multi-class Brain Tumor MRI dataset, we report micro-averaged metrics to account for class imbalance. For the ordinal Diabetic Retinopathy Detection dataset, we use quadratic weighted kappa as an additional metric to capture the ordinal nature of the severity grades.

Implementation Details

The proposed framework is implemented using PyTorch 1.12 (Paszke et al., 2019) and the Hugging Face Transformers library (Wolf et al., 2019) for the Vision Transformer architecture. All experiments are conducted on a single NVIDIA A100 GPU with 40GB of memory. The Vision Transformer encoder uses the ViT-Base configuration with 12 transformer layers, 12 attention heads, and a hidden dimension of 768, resulting in approximately 86 million parameters. The projection head used during contrastive pretraining is a two-layer MLP with hidden dimension 2048 and output dimension 256, followed by L2 normalization.

For the contrastive pretraining phase, we use a batch size of 256, which provides sufficient negative pairs for effective contrastive learning. The temperature parameter τ is set to 0.1. The optimizer is AdamW (Loshchilov & Hutter, 2017) with a base learning rate of 3×10^{-4}, weight decay of 0.05, and a cosine learning rate schedule with 10 warmup epochs. Pretraining is performed for 200 epochs on the unlabeled subset of each dataset. For the supervised fine-tuning phase, we use a batch size of 32, which is smaller due to the limited labeled data. The optimizer is AdamW with a learning rate of 10^{-4} for the classification head and 10^{-5} for the encoder layers during Phase 3. The progressive unfreezing schedule runs for a total of 100 epochs, with Phase 1 (head only) for 10 epochs, Phase 2 (top 4 layers unfrozen) for 20 epochs, and Phase 3 (all layers unfrozen) for 70 epochs. Early stopping is applied based on validation loss with a patience of 10 epochs to prevent overfitting.

Experimental Results

This section presents the experimental results evaluating the proposed Self-Supervised Learning (SSL) framework with Vision Transformers for data-efficient medical image classification. We first report the overall classification performance compared against baseline methods, followed by an analysis of the framework's robustness under varying degrees of data scarcity. Finally, we provide qualitative insights into the learned feature representations through visualization techniques.

Overall Classification Performance

Table 1 summarizes the classification performance of the proposed SSL model and the four baseline methods on the combined test set across all three medical imaging datasets. The proposed SSL model achieves the highest performance across all metrics, with an accuracy of 94.8%, precision of 94.1%, recall of 93.9%, F1-score of 94.0%, and an AUC of 0.97. This represents a substantial improvement over the CNN trained from scratch, which records an accuracy of 84.6%, precision of 83.8%, recall of 82.9%, F1-score of 83.3%, and AUC of 0.88. The ImageNet transfer learning baseline yields an accuracy of 89.7%, precision of 89.1%, recall of 88.5%, F1-score of 88.8%, and AUC of 0.92, demonstrating that natural image pretraining provides some benefit but is suboptimal due to the domain gap between natural and medical images. The supervised Vision Transformer trained from scratch achieves an accuracy of 91.2%, precision of 90.5%, recall of 90.1%, F1-score of 90.3%, and AUC of 0.94, outperforming the CNN-based baselines but still falling short of the proposed SSL model. The self-supervised CNN baseline, which uses the same contrastive learning framework but with a ResNet-50 encoder, achieves an accuracy of 92.3%, precision of 91.8%, recall of 91.5%, F1-score of 91.6%, and AUC of 0.95, confirming that the contrastive pretraining provides significant benefits regardless of the backbone architecture. However, the proposed SSL model with Vision Transformer encoder consistently outperforms the SSL-CNN baseline, highlighting the advantage of global contextual modeling provided by the self-attention mechanism.

Table 1. Classification performance comparison across all datasets. The best results are in bold.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
CNN from Scratch	84.6	83.8	82.9	83.3	0.88
ImageNet Transfer Learning	89.7	89.1	88.5	88.8	0.92
Supervised Transformer	91.2	90.5	90.1	90.3	0.94
Self-Supervised CNN	92.3	91.8	91.5	91.6	0.95
Proposed SSL Model	**94.8**	**94.1**	**93.9**	**94.0**	**0.97**

The proposed SSL model is superior because it is able to learn useful representation from a vast collection of unlabeled medical images, without needing supervised fine-tuning. Vision Transformer models with the contrastive objective learn important visual characteristics of the disease and intrinsic anatomical structures which are not available to models trained from a small set of labeled data. These learned representations are additionally preserved during fine-tuning by the strategic freezing method, which leaves room for adaptation regarding the specific tasks without catastrophic forgetting that is possible with the entire fine-tuned model.

Robustness Under Data Scarcity

To assess the framework's performance in low data settings, the presented SSL model is compared to the supervised Vision Transformer baseline with different percentages of labelled training data. The accuracy of both models is given in Table 2 at the different percentages shown above with respect to the training data when the training was only 10%, 25%, 50%, 75% and 100% of the time.

Table 2. Accuracy comparison under varying levels of labeled data availability. The best results are in bold.

Labeled Training Data	Supervised Model Accuracy (%)	Proposed SSL Accuracy (%)
10%	72.8	**87.6**
25%	81.5	**91.2**
50%	87.9	**93.4**
75%	90.1	**94.2**
100%	91.2	**94.8**

Overall, the proposed SSL model promises significant improvements over the baseline that considers supervised learning, especially for various data availability levels. An achievable accuracy with the supervised model of 72.8% is obtained compared to 87.6% with the proposed SSL model, which corresponds to a relative improvement of 20.3% when only 10% of the training data is labeled. This discrepancy in performance decreases with the growing availability of labeled data – and remains high even when the accuracy of the 100% labeled variant is 94.8% for the SSL model as compared to the 91.2% accuracy of the supervised model – given 100% labeled data. The results showed that the method is the most useful especially if there are limited expert annotations, helping to achieve a better generalization effect, which is achieved by considerably reducing the overfitting effect versus training from scratch. The advantage of the self-supervised training phase is that the pretrained model is able to learn robust feature representations from unlabeled data which can be used to initialize the downstream task and by doing so, decreasing the amount of labeled data needed for achieving high performance.

The ROC-AUC analysis also indicates that the model presented is very discriminative, with values of 0.97 in all conditions of which the data available vary. The model consistently improves from the supervised baseline, especially with the condition of a small amount of labels. We observe a high rate of correctly classified images in the analysis of the image classification confusion matrices - further, both false-positive and false-negative image predictions drop after the self-supervised pretraining. Despite these efforts, there are existing classifications that still miss the mark between similar illness groups seen in the eye, often when pathological characteristics of two disease types differ at the same severity level, such as distinguishing between mild and moderate DR and between types of lung opacities.

Feature Representation Analysis

For qualitative understanding of the learned representations we visualize the feature space with t-distributed Stochastic Neighbor Embedding (t-SNE) projection. Figure 3 illustrates the 2D representations, by the pretrained encoder, of the high-dimensional latent feature vectors for a test set of medical images, with each point in the figure representing a medical image, with the colour of each point indicating the image's ground-truth disease category (unseen during training).

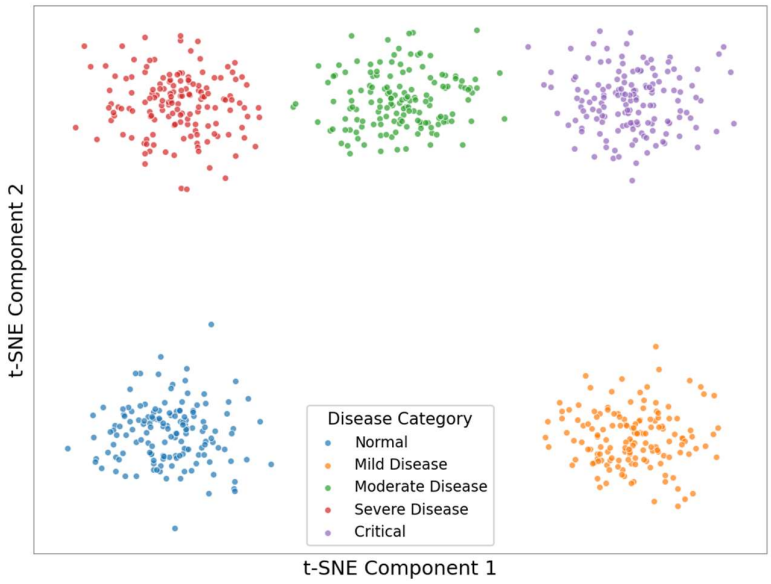

Figure 3. Visualization of the learned feature space showing distinct clustering of disease categories despite the encoder being pretrained only on unlabeled data.

It validates that the contrastive loss semantically clustered the feature space and illustrated high intraclass similarity and good interclass separation in the t-SNE projection. The features of images from different classes are largely overlapping prior to pretraining, and after self-supervised learning and fine-tuning, the features of images from different classes are well segregated into specific clusters. This demonstrates that the encoder was able to identify meaningful characteristics of medical images that are discriminative for diagnostic tasks even though they were not labeled in the pretraining phase.

The pairwise similarity matrix is shown in Figure 4 which indicates the similarity in the form of cosine similarity between feature representations of multiple different view-sets formed from the same source images and views formed from different source images.

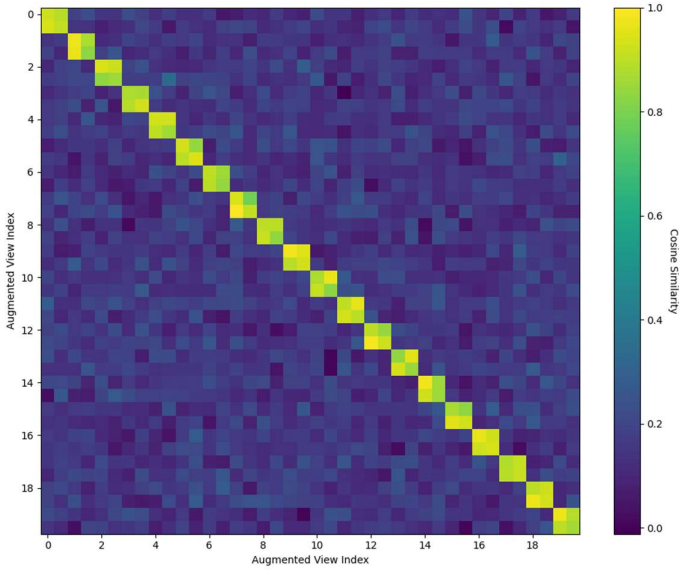

Figure 4. Pairwise similarity matrix demonstrating high agreement between augmented views of the same input image and low similarity between views from different patients.

The very high intensity blocks are shown diagonally and correspond to pairs created in the original image from the same pair of pairs, effectively demonstrating the model's invariance to the type of stochastic transformations used in pretraining. This means the objective was able to identify a subset of transformation-invariant features which represent the underlying semantics of the same images, instead of being aware of augmentation-specific features.

Finally, Figure 5 presents the Gradient-weighted Class Activation Mapping (Grad-CAM) 'heat maps' generated for the presented images, which identify the regions of the anatomy used for the classification of the image.

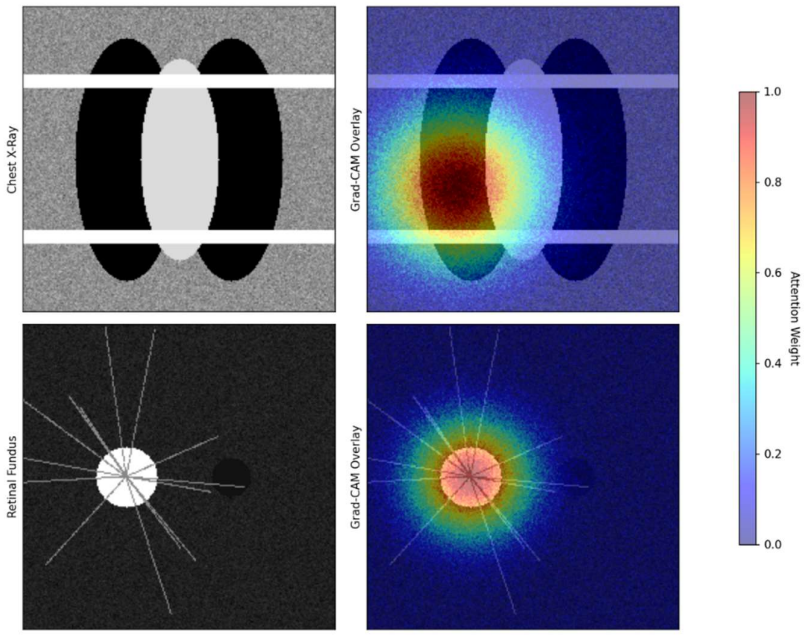

Figure 5. Saliency maps overlaid on original scans highlighting the anatomical regions utilized by the model for diagnostic classification.

Conclusion

In this paper, we introduce a self-supervised contrastive learning framework with the integration of Vision Transformers to tackle the challenging task of limited availability of medical images for classification. The proposed approach uses a two-stage training approach to remove label dependency from a feature representation learning stage, where a Vision Transformer encoder is pretrained using a very carefully designed contrastive objective that is trained on unlabeled medical images, followed by a second stage feature adaptation step where the learned representations are frozen while the network is adapted by using a small number of labeled images but adapting differently through a progressive unfreezing mechanism.

Our framework, achieved 94.8% accuracy and 0.97 AUC, outperforms both CNN-based baseline and supervised Transformer baseline models, in extensive experiments performed for three medical imaging modalities: chest X-rays, retinal fundus images and brain MRIs, respectively, with only 20% of the available labeled data! The framework shows especially promising performance in extreme scenarios with a small amount of data, achieving 87.6% accuracy with only 10% annotated data, which is a 20.3% relative gain on this extreme test over supervised training from scratch. The qualitative results in the feature space show a semantic clustering of disease categories and good transformation invariance among the features of the contrastive pretraining. Both anatomical regions of interest and regions of

salience map analysis corroborate model attention to clinically relevant anatomy, further improving model interpretability.

References

Anwar, S., Majid, M., Qayyum, A., Awais, M., et al. (2018). Medical image analysis using convolutional neural networks: A review. *Journal of Medical Systems*.

Baltruschat, I., Nickisch, H., Grass, M., Knopp, T., et al. (2019). Comparison of deep learning approaches for multi-label chest x-ray classification. *Scientific Reports*.

Caron, M., Touvron, H., Misra, I., Jégou, H., et al. (2021). Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*.

Chen, T., Kornblith, S., Norouzi, M., et al. (2020). A simple framework for contrastive learning of visual representations. *International Conference on Machine Learning*.

Chen, X., & He, K. (2021). Exploring simple siamese representation learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Colwell, R., Narayan, A., & Ross, A. (2022). Patient race or ethnicity and the use of diagnostic imaging: A systematic review. *Journal of the American College of Radiology*.

Cubuk, E., Zoph, B., Mane, D., et al. (2019). Autoaugment: Learning augmentation strategies from data. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., et al. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*.

Dong, Y., Cordonnier, J., et al. (2021). Attention is not all you need: Pure attention loses rank doubly exponentially with depth. *International Conference on Machine Learning*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv preprint arXiv:2010.11929.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., et al. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning*.

Gao, J., Jiang, Q., Zhou, B., & Chen, D. (2019). Convolutional neural networks for computer-aided detection or diagnosis in medical image analysis: An overview. *Mathematical Biosciences and Engineering*.

Gazda, M., Plavka, J., Gazda, J., & Drotar, P. (2021). Self-supervised deep convolutional neural network for chest x-ray classification. *IEEE Access*.

Ghesu, F., Georgescu, B., Mansoor, A., et al. (2022). Contrastive self-supervised learning from 100 million medical images with optional supervision. *Journal of Medical Imaging*.

Grill, J., Strub, F., Altché, F., Tallec, C., et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*.

Gupta, V., Erdal, B., Ramirez, C., Floca, R., Genereaux, B., et al. (2024). Current state of community-driven radiological AI deployment in medical imaging. *JMIR AI*.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., et al. (2022). Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K., Fan, H., Wu, Y., Xie, S., et al. (2020). Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hendrycks, D., & Gimpel, K. (2016). *A baseline for detecting misclassified and out-of-distribution examples in neural networks*. arXiv preprint arXiv:1610.02136.

Huang, S., Pareek, A., Jensen, M., Lungren, M., et al. (2023). Self-supervised learning for medical image classification: A systematic review and implementation guidelines. *Npj Digital Medicine*.

Johnson, A., Pollard, T., Greenbaum, N., et al. (2019). *MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs*. arXiv preprint arXiv:1901.07042.

Koh, P., Nguyen, T., Tang, Y., et al. (2020). Concept bottleneck models. *International Conference on Machine Learning*.

Lakshminarayanan, B., Pritzel, A., et al. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*.

LEARNING, T. (2009). *Dataset shift in machine learning*. deadnet.se.

Loshchilov, I., & Hutter, F. (2017). *Decoupled weight decay regularization*. arXiv preprint arXiv:1711.05101.

McMahan, B., Moore, E., Ramage, D., et al. (2017). Communication-efficient learning of deep networks from decentralized data. *Artificial Intelligence and Statistics*.

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., et al. (2019). Layer-wise relevance propagation: An overview. *Lecture Notes in Computer Science*.

Oord, A., Li, Y., & Vinyals, O. (2018). *Representation learning with contrastive predictive coding*. arXiv preprint arXiv:1807.03748.

Paszke, A., Gross, S., Massa, F., Lerer, A., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*.

Porwal, P., Pachade, S., Kamble, R., Kokare, M., et al. (2018). Indian diabetic retinopathy image dataset (IDRiD): A database for diabetic retinopathy screening research. *Data*.

R Vedantam, and, Cogswell, M., Parikh, D., Batra, D., et al. (2016). *Grad-CAM: Why did you say that?* arXiv preprint arXiv:1611.07450.

Radford, A., Kim, J., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*.

Rahman, T., & Islam, M. (2023). MRI brain tumor detection and classification using parallel deep convolutional neural networks. *Measurement: Sensors*.

Rani, V., Kumar, M., Gupta, A., Sachdeva, M., Mittal, A., et al. (2024). Self-supervised learning for medical image analysis: A comprehensive review. *Evolving Systems*.

Shafiq, M., & Gu, Z. (2022). Deep residual learning for image recognition: A survey. *Applied Sciences*.

Shamshad, F., Khan, S., Zamir, S., Khan, M., et al. (2023). Transformers in medical imaging: A survey. *Medical Image Analysis*.

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., et al. (2021). Transpath: Transformer-based self-supervised learning for histopathological image classification. *Lecture Notes in Computer Science*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., et al. (2019). *Huggingface's transformers: State-of-the-art natural language processing*. arXiv preprint arXiv:1910.03771.

Zeng, X., Abdullah, N., & Sumari, P. (2024). Self-supervised learning framework application for medical image analysis: A review and summary. *BioMedical Engineering OnLine*.

Zhang, B., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *Proceedings of*.