Self-Supervised Learning for Medical Image Classification: A Systematic Review and Implementation Guidelines

Shih-Cheng Huang[1,2], Anuj Pareek[1,2]*, Malte Jensen[1], Matthew P. Lungren[1,2,3], Serena Yeung[1,2,4,5,6‡], Akshay S. Chaudhari[1,2,3,7‡]*

[1] Department of Biomedical Data Science, Stanford University, Stanford, CA, USA
[2] Center for Artificial Intelligence in Medicine & Imaging, Stanford University, Stanford, CA, USA
[3] Department of Radiology, Stanford University, Stanford, CA, USA
[4] Department of Computer Science, Stanford University, Stanford, CA, USA
[5] Department of Electrical Engineering, Stanford University, Stanford, CA, USA
[6] Clinical Excellence Research Center, Stanford University School of Medicine, Stanford, CA, USA
[7]Stanford Cardiovascular Institute, Stanford University, CA, USA
{*,‡} Equal Contribution

Abstract

Advancements in deep learning and computer vision provide promising solutions for medical image analysis, potentially improving healthcare and patient outcomes. However, the prevailing paradigm of training deep learning models requires large quantities of labeled training data, which is both time-consuming and cost-prohibitive to curate for medical images. Self-supervised learning (SSL) has the potential to make significant contributions to the development of robust medical imaging models through its ability to learn useful insights from copious medical datasets without labels. In this review, we provide consistent descriptions of different self-supervised learning strategies and compose a systematic review of papers published between 2012 and 2022 on PubMed, Scopus, and ArXiv that applied self-supervised learning to medical imaging classification. We screened a total of 412 relevant studies and included 79 papers for data extraction and analysis. With this comprehensive effort, we synthesize the collective knowledge of prior work and provide implementation guidelines for future researchers interested in applying self-supervised learning to their development of medical imaging classification models.

Main

The utilization of medical imaging technologies has become an essential part of modern medicine, enabling diagnostic decisions and treatment planning. The importance of medical imaging is exemplified by the consistent rate of growth in medical imaging utilization in modern healthcare[1,2]. However, as the number of medical imaging relative to the available radiologists continues to become more disproportionate, the workload for radiologists continues to increase. Studies have shown that an average radiologist now needs to interpret one image every 3-4 seconds to keep up with clinical workloads[3–5]. With such a huge cognitive burden placed on radiologists, delays in diagnosis and diagnostic errors are unavoidable[6,7]. Thus, there is an urgent need to integrate automated systems into the medical imaging workflow, which will improve both efficiency and accuracy of diagnosis.

In recent years, deep learning models have demonstrated diagnostic accuracy comparable to that of human experts in narrow clinical tasks for several medical domains and imaging modalities, including chest and extremity X-rays[8–10], computed tomography (CT)[11], magnetic resonance imaging (MRI)[12], whole slide images (WSI)[13,14], and dermatology images[15]. While deep learning provides promising solutions for improving medical image interpretation, the current success has been largely dominated by supervised learning frameworks, which typically require large-scale labeled datasets to achieve high performance. However, annotating medical imaging datasets requires domain expertise, making large-scale annotations cost-prohibitive and time-consuming, which fundamentally limits building effective medical imaging models across varying clinical use cases.

Besides facing challenges with training data, most medical imaging models underperform in generalizing to external institutions or in attempting to repurpose for other tasks[16]. The inability to generalize can be largely due to the process of supervised learning, which encourages the model to mainly learn features heavily correlated with specific labels rather than general features representative of the whole data distribution. This creates specialist models that can perform well only on the tasks they were trained to do[17]. In a healthcare system where myriad opportunities and possibilities for automation exist, it is practically impossible to curate labeled datasets for all tasks, modalities, and outcomes for training supervised models. Therefore, it is important to develop strategies for training medical AI models that can be fine-tuned for many downstream tasks, while remaining pragmatic regarding the challenges in curating large-scale labeled datasets.

Self-supervised learning (SSL), the process of training models to produce meaningful representations using unlabeled data, is a promising solution to challenges caused by difficulties in curating large-scale annotations. Unlike supervised learning, self-supervised learning can create generalist models that can be finetuned for many downstream tasks without large-scale labeled datasets. Self-supervised learning was first popularized in the field of natural language processing (NLP), when researchers leveraged copious amounts of unlabeled text scraped from the internet to improve the performance of their models. These pretrained large language models[18,19], are capable of achieving state-of-the-art results for a wide range of NLP tasks, and have shown the ability to perform well on new tasks with only a fraction of the labeled data that traditional supervised learning techniques require. Motivated by the initial success of SSL in NLP, there is great interest in translating similar techniques of SSL to computer vision tasks. Such work in computer vision has already demonstrated performance for natural images that is superior to that achieved by supervised models, especially in label-scarce scenarios[20].

Reducing the number of manual annotations required to train medical imaging models will significantly reduce both the cost and time required for model development, making automated systems more accessible to different specialties and hospitals, thereby reducing workload for radiologists and potentially improving patient care. While there is already a growing trend in recent medical imaging AI literature to leverage self-supervised learning (Figure 1), as well as a few narrative reviews[21,22], the most suitable strategies and best practices for medical images have not been sufficiently investigated. The purpose of this work is to present a comprehensive review of deep learning models that leverage self-supervised learning for medical image classification, define and consolidate relevant terminology, and summarize the results from state-of-the-art models in relevant current literature. We focus on medical image classification tasks because many clinical

tasks are based on classification, and thus our research may be directly applicable to deep learning models for clinical workflows. This review intends to help inform future modeling frameworks and serve as a reference for researchers interested in the application of self-supervised learning in medical imaging.

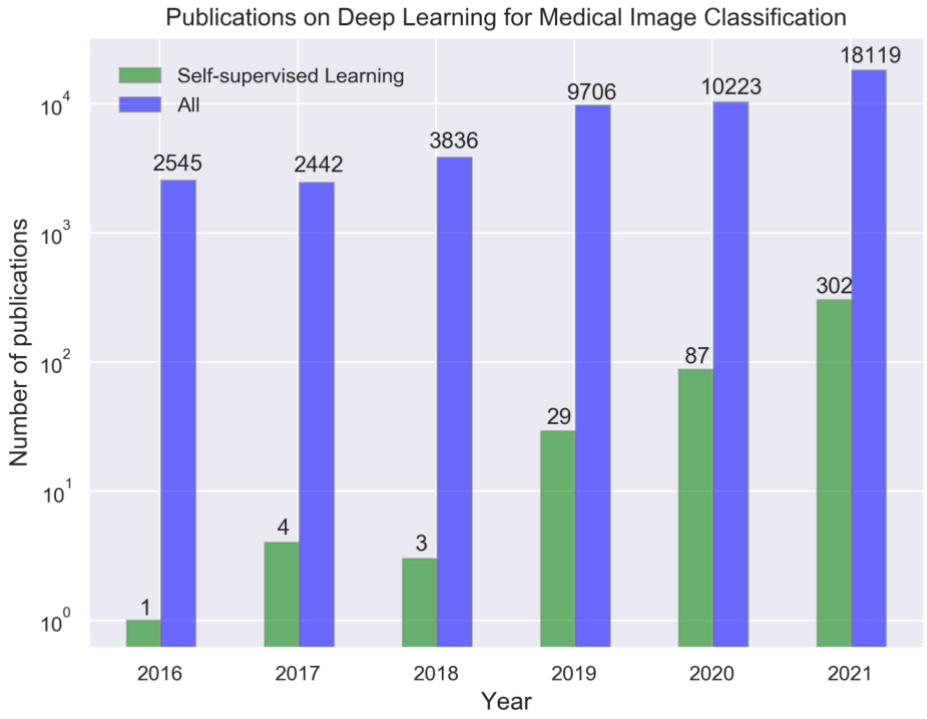


Figure 1. Timeline showing number of publications on deep learning for medical image classification per year, found by using the same search criteria on PubMed, Scopus and ArXiv. The figure shows that self-supervised learning is a rapidly growing subset of deep learning for medical imaging literature.

Terminology and strategies in self-supervised learning

Here we provide definitions for different categorizations of self-supervision strategies, namely innate relationship, generative, contrastive, and self-prediction (Figure 2)[23]. The relative ordering of these self-supervision strategies is based on the chronological order in which they were popularized. It is worth noting that some definitions can be overlapping since clear distinctions between each method can not always be made.

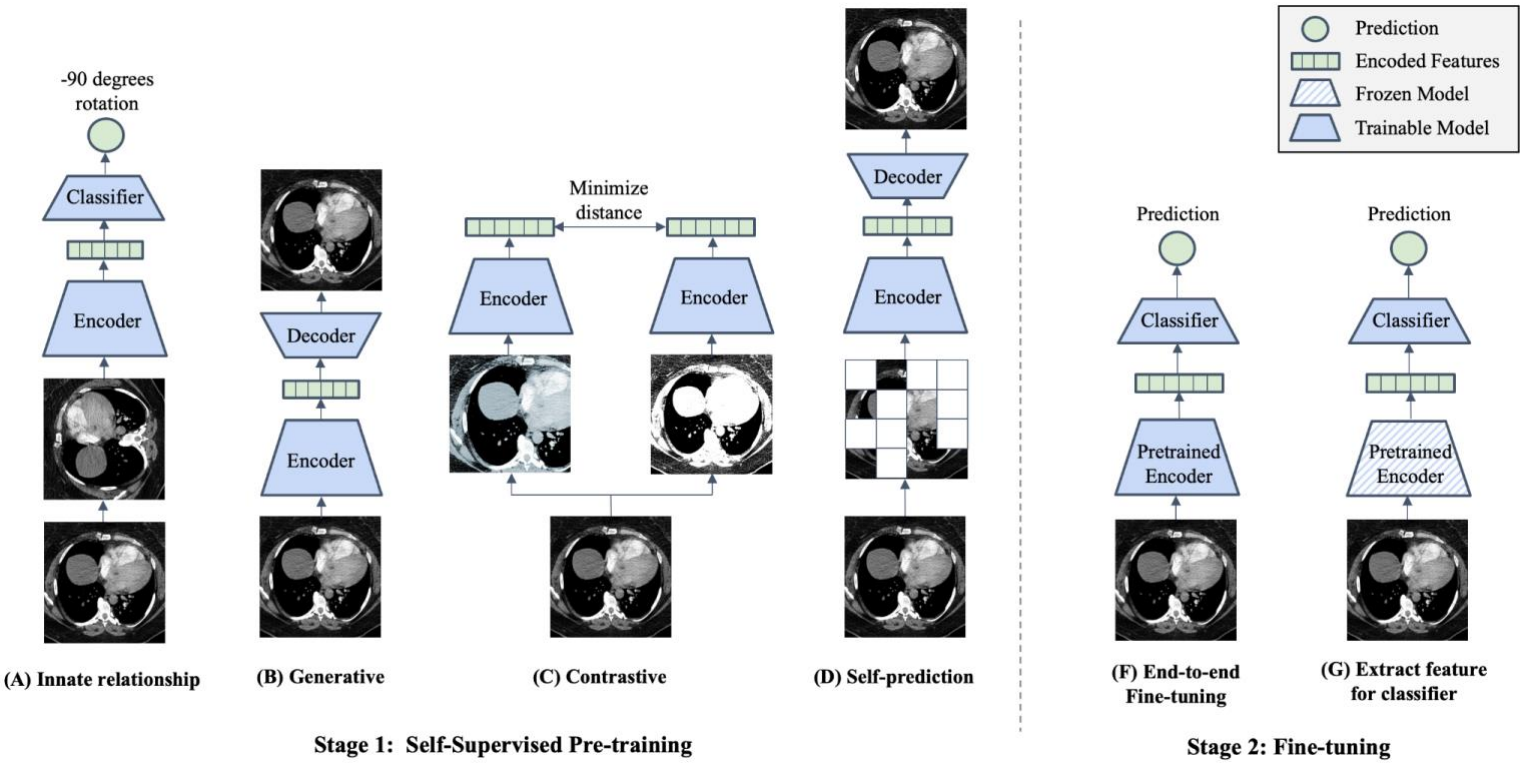


*Figure 2: Illustration of different self-supervised learning and fine-tuning strategies. During **Stage 1** a model is pre-trained using one or more of the following self-supervised learning strategies: (a) Innate relationship SSL pretrains a model on a hand-crafted task by leveraging the internal structure of the data, (b) Generative SSL learns the distribution of training data, enabling reconstruction of the original input (c) Contrastive SSL forms positive pairs between different augmentations of the same image and minimizes representational distances of positive samples closer together (d) Self-prediction augments or masks out random portions of an image, and reconstructs the original image based on the unaltered parts of the original image. During **Stage 2**, the pre-trained model can be fine-tuned using one of the following strategies: (f) end-to-end fine-tuning of the pretrained model and classifier, or (g) train a classifier which uses extracted features from the SSL pretrained model.*

Innate relationship

Innate relationship SSL is the process of pretraining a model on a hand-crafted task which can leverage the internal structure of the data, without acquiring additional labels. In the most general sense, innate relationship models perform classification or regression based on the hand-crafted task instead of optimizing based on the model's ability to reconstruct (Generative and Self-prediction) or represent the original image (Contrastive). Specifically, these methods are optimized using classification or regression loss derived from the given task. Pretraining the model on such a hand-crafted task makes the model learn visual features as a starting point. However, innate relationship SSL can lead to visual features that are effective only for the hand-crafted task but have limited benefits for the downstream task. Examples of innate relationship for visual inputs include predicting image rotation angle[24], solving jigsaw puzzles of an image[25], or determining the relative positions of image patches[26].

Generative

Generative models, popularized through the advent of traditional autoencoders[27], variational autoencoders[28] and generative adversarial networks (GANs)[29], are able to learn the distribution of training data, and thereby reconstruct the original input or create new synthetic data instances. By using readily available data as the target, generative models can be trained to automatically learn useful latent representations without the need for explicit labels, and thus constitute a form of self-supervision. Early work that leverages generative models for self-supervised learning rely on autoencoders, where an encoder converts inputs into latent representations and a decoder reconstructs the representation back to the original image[30]. Subsequently, these models are optimized based on how closely the reconstructed images resemble the original image. More recent work has explored utilizing GANs for generative self-supervised learning, with improvement in performance over prior work[31,32].

Contrastive

Contrastive self-supervised methods are based on the assumption that variations caused by transforming an image do not alter the image's semantic meaning. Therefore different augmentations of the same image constitute a so-called positive pair, while the other images and their augmentations are defined to be negative pairs in relation to the current instance. Subsequently a model is optimized to minimize the latent space distances between the positive pairs and push apart negative samples. Separating representations for positive and negative pairs can be based on arbitrary distance metrics incorporated into the contrastive loss function. One pioneering contrastive-based method is SimCLR[20], which outperformed supervised models on ImageNet benchmark using 100 times fewer labels. However, SimCLR requires a very large batch size to perform well, which can be computationally prohibitive for most researchers. To reduce the large batch size required by SimCLR to ensure enough informative negative samples, Momentum Contrast (MoCo) introduced a momentum encoded queue to keep negative samples[33]. More recently, a subtype of contrastive self-supervised learning called instance discrimination, which includes methods such as DINO[34], BYOL[35] and SimSiam[36], further eliminates the need for negative samples. Instead of contrastive augmented pairs from the same image, several studies have explored contrasting clustering assignments of augmented versions of the same image[37–39].

Self-prediction

Self-prediction SSL is the process of masking or augmenting portions of the input and using the unaltered portions to reconstruct the original input. The idea of self-prediction self-supervised learning originated from the field of Natural Language Processing (NLP), where state-of-the-art models were pre-trained using the Masked Language Modeling approach by predicting missing words in a sentence[18,19]. Motivated by the success in NLP, early work in the field of computer vision made similar attempts by masking out or augmenting random patches of an image and training Convolutional Neural Networks (CNNs) to reconstruct the missing regions as a pre-training strategy[40] but only with moderate success. Recently, the introduction of Vision Transformers (ViT) allowed computer vision models to also have the same transformer-based architecture. Studies such as BERT Pre-Training of Image Transformers (BEiT) and Masked Auto-encoders (MAE), which combine ViT with self-prediction pre-training objective, achieve state-of-the-art results when fine-tuned across several natural image benchmarks[41,42]. Similar to generative SSL, self-prediction models are optimized using the reconstruction loss. The key difference between self-prediction and generative self-supervised learning methods is that self-prediction applies masking or

augmentations only to portions of the input image, and uses the remaining, unaltered portions to inform reconstruction. On the other hand, generative based self-supervised learning either applies augmentations on the whole image or does not apply any augmentations.

Strategies for fine-tuning

There are two main strategies for fine-tuning models that have been pre-trained using SSL (Figure 2). If we consider any imaging model to be composed of an encoder part and a classifier part, then these two strategies are 1) end-to-end fine-tuning vs. 2) extract features from the encoder first and subsequently train an additional classifier. In end-to-end fine-tuning all the weights of the encoder and classifier are unfrozen and can be adjusted through optimization using supervised learning in the fine-tuning phase. In the feature-extraction strategy, the weights of the encoder are kept frozen to extract features as inputs to the downstream classifier. While many previous work uses linear classifiers with trainable weights (also known as linear probing), any type of classifier or architecture can be used, including SVMs or even non-trainable classifiers such as k-nearest neighbor[43]. It is worth emphasizing that SSL is task agnostic, and the same SSL pretrained model can be fine-tuned for different types of downstream tasks, including classification, segmentation, and object detection.

Results

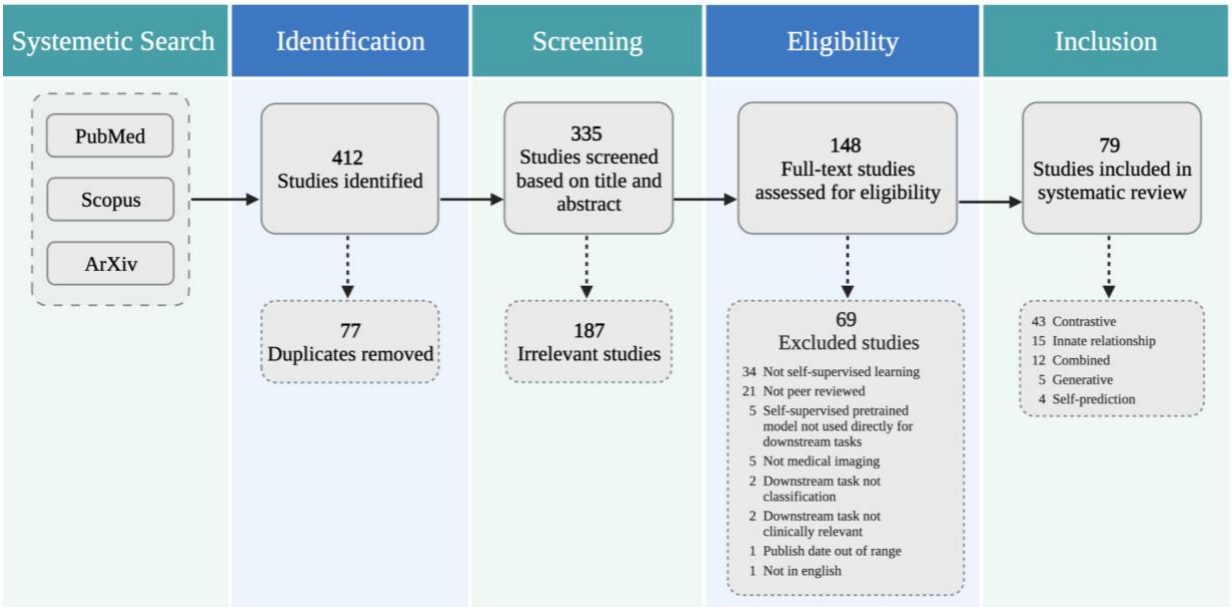
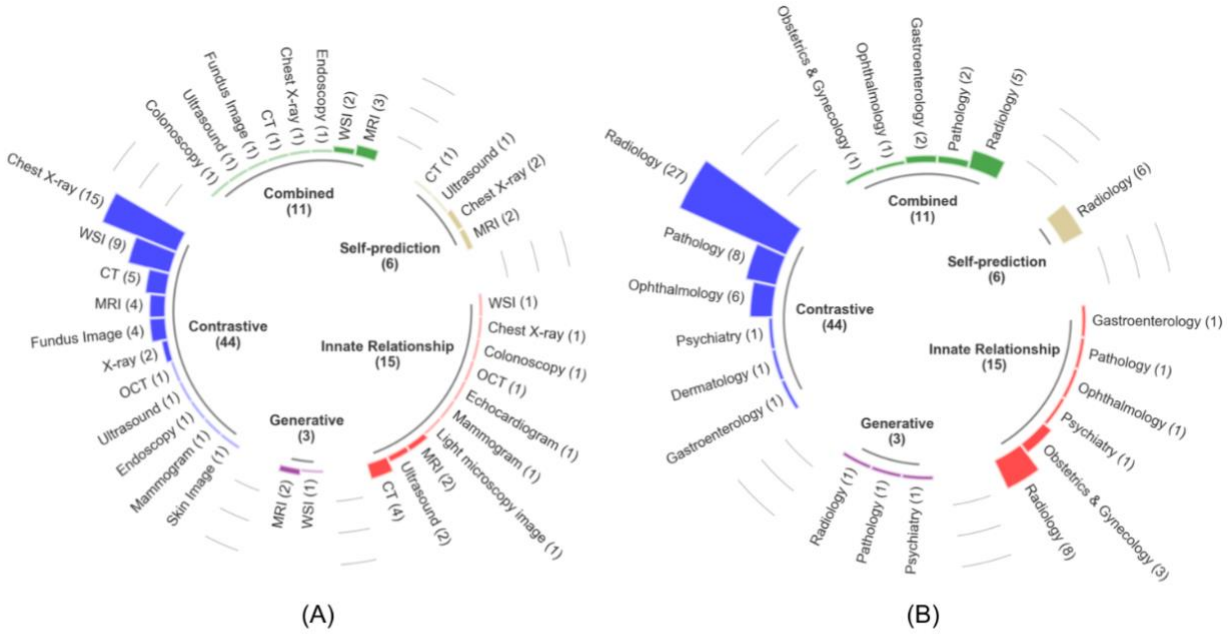
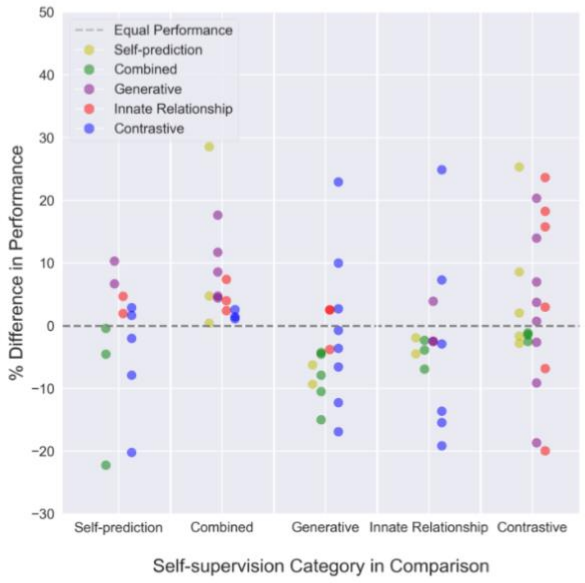
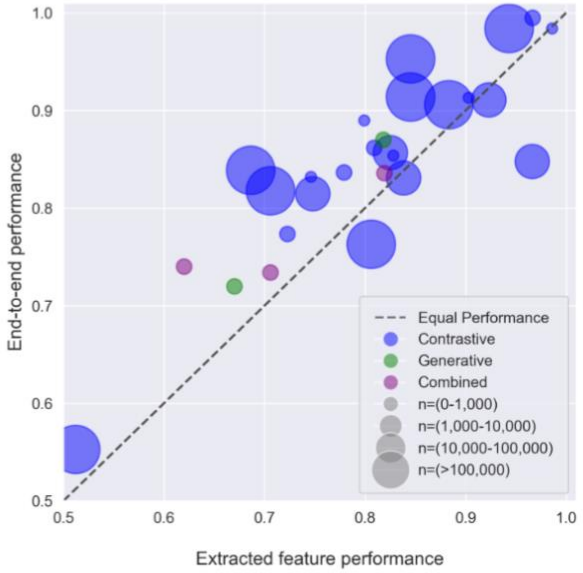


Figure 2. Authors independently screened all records for eligibility. Out of 412 studies identified from PubMed, Scopus, and ArXiv, 79 studies were included in the systematic review.

A total of 412 unique studies were identified through our systematic search. After removing duplicates and excluding studies based on title and abstract using our study selection criteria (see Methods), 148 studies remained for full-text screening. A total of 79 studies fulfilled our eligibility criteria and were included for systematic review and data extraction. Figure 2 presents a flowchart of the study screening and selection process. Table 1 displays the included studies and extracted data while Figure 3 summarizes the statistics of extracted data.


Chest X-ray (15)
WSI (9)
CT (5)
MRI (4)
Fundus Image (4)
X-ray (2)
OCT (1)
Ultrasound (1)
Endoscopy (1)
Mammogram (1)
Skin Image (1)
Contrastive (44)
Colonoscopy (1)
Ultrasound (1)
Fundus Image (1)
CT (1)
Chest X-ray (1)
Endoscopy (1)
WSI (2)
MRI (3)
Combined (11)
CT (1)
Ultrasound (1)
Chest X-ray (2)
MRI (2)
Self-prediction (6)
MRI (2)
WSI (1)
Generative (3)
WSI (1)
Chest X-ray (1)
Colonoscopy (1)
OCT (1)
Echocardiogram (1)
Mammogram (1)
Light microscopy image (1)
MRI (2)
Ultrasound (2)
CT (4)
Innate Relationship (15)
(A)
Radiology (27)
Pathology (8)
Ophthalmology (6)
Psychiatry (1)
Dermatology (1)
Gastroenterology (1)
Contrastive (44)
Obstetrics & Gynecology (1)
Ophthalmology (1)
Gastroenterology (2)
Pathology (2)
Radiology (5)
Combined (11)
Radiology (6)
Self-prediction (6)
Radiology (1)
Pathology (1)
Psychiatry (1)
Generative (3)
Gastroenterology (1)
Pathology (1)
Ophthalmology (1)
Psychiatry (1)
Obstetrics & Gynecology (3)
Radiology (8)
Innate Relationship (15)
(B)


Equal Performance
Self-prediction
Combined
Generative
Innate Relationship
Contrastive
% Difference in Performance
Self-prediction
Combined
Generative
Innate Relationship
Contrastive
Self-supervision Category in Comparison
(C)


Equal Performance
Contrastive
Generative
Combined
n=(0-1,000)
n=(1,000-10,000)
n=(10,000-100,000)
n=(>100,000)
End-to-end performance
Extracted feature performance
(D)

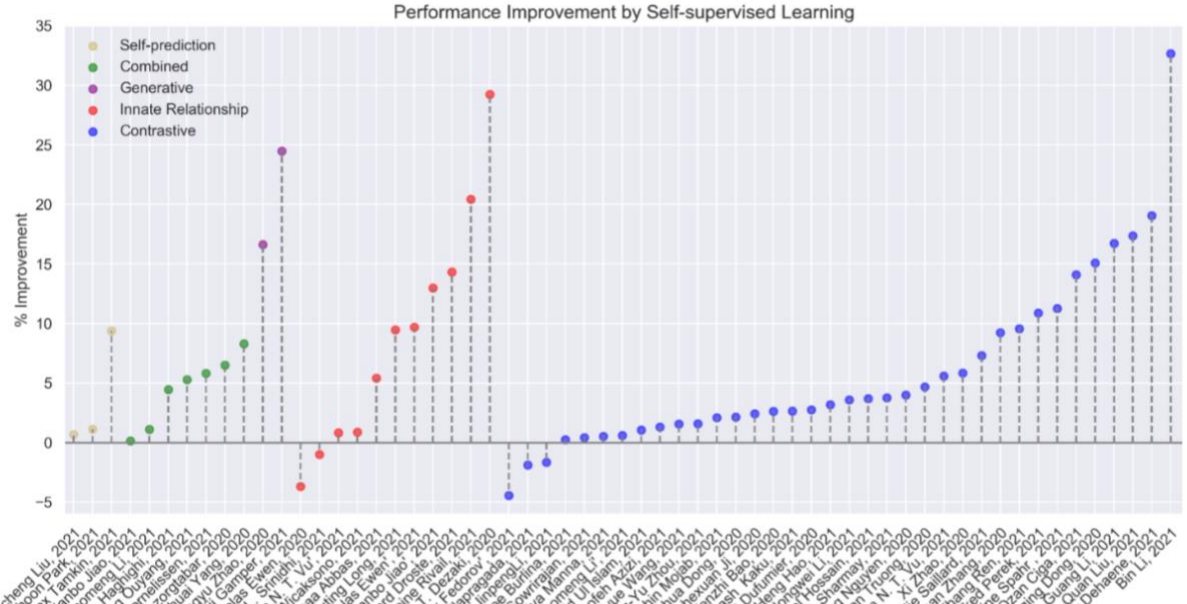

Performance Improvement by Self-supervised Learning
Self-prediction
Combined
Generative
Innate Relationship
Contrastive
% Improvement

Figure 3. Summary of extracted data from studies in our system review. A) Prevalence of different medical specialties split by self-supervised learning strategy. B) Prevalence of different medical imaging modalities split by self-supervised learning strategy. C) Relative performance difference between different types of self-supervised learning strategies on the same task. D) Performance comparison between end-to-end fine-tuning vs. training a classifier using extracted features from pretrained self-supervised models. E) Relative difference in downstream task performance between self-supervised and non-self-supervised models.

Innate Relationship

Innate relationship was used in 15 out of 79 studies (Table 1). Nine of these studies designed their innate relationship pre-text task based on different image transformations, including rotation prediction[107–110], horizontal flip prediction[103], reordering shuffled slices[105], and patch order prediction[104,109,112,113]. Notably, Jiao et al. pre-trained their models simultaneously with two innate relationship pre-text tasks (slice order prediction and geometric transformation prediction), and showed that a weight-sharing Siamese network out-performs a single disentanged model for combining the two pre-training objectives[46]. The remaining six studies designed clinically relevant pretext tasks by exploiting the unique properties of medical images. For instance, Droste et al. utilized a gaze tracking dataset and pre-trained a model to predict sonographers' gazes on ultrasound video frames with gaze-point regression[102]. Dezaki et al. employed temporal and spatial consistency to produce features for echocardiograms that are strongly correlated with the heart's inherent cyclic pattern[111]. Out of all innate relationship based studies, ten compared performance to those of supervised pre-trained models and eight of them showed improvement in performance. On average, clinically relevant pre-text tasks achieved greater improvements in performance over transformation-based pre-text tasks, when compared to purely supervised methods (13.7% vs 5.03%).

Generative

Generative self-supervised learning was used in 3 out of 79 studies (Table 1). Gamper et al. extracted histopathology images from textbooks and published papers along with the figure captions and devised an image captioning task for self-supervised pre-training, where a ResNet-18 was used for encoding images, and the representation was fed to transformers for image-captioning[100]. They were subsequently able to use the learned representations for a number of downstream histopathology tasks, including breast cancer classification. Osin et al.[98] leveraged the chronology of sequential images in brain fMRI for self-supervised pre-training. Brain fMRI scans are typically acquired with subjects alternating between a passive and an active phase, where the subject is instructed to perform some task or receives some external stimulus. During the self-supervision phase, Osin et al. trained two networks: an autoencoder to generate the active fMRI image given the passive image, and an LSTM to predict the next active image. The representations learned during the self-supervision were then used to train a classifier to predict psychiatric traits such as post-traumatic stress disorder (PTSD). Finally, Zhao et al. use a generative approach with an autoencoder with an additional constraint that explicitly associates brain age to the latent representations for longitudinally acquired brain MRIs[99]. Of the three studies, two reported comparisons with purely supervised models and showed relative improvements of 16.6%[99] and 24.5%[100] with self-supervised learning.

Contrastive

The majority of the studies that remained after our full-text screening (44/79) used contrastive learning as their self-supervised pre-training strategy (Table 1). SimCLR, MoCo and BYOL were the three most used frameworks, applied in 13, 8 and 3 papers respectively. Some papers leveraged medical domain priors to create specialized strategies for creating positive pairs. For pathology slices, Li et al. exploited that the neighborhood around a patch is likely to be similar, and used pre-clustering to find dissimilar patches[62]. In radiology, Ji et al. used multimodal contrastive learning by matching X-rays with corresponding radiology reports[76]. They extracted and fused the representations of the image and text modalities through both global image-sentence matching and local attention-based region-phrase matching. Wang et al. utilized both

radiomic features and deep features from the same image to form positive pairs[93]. They also utilized the spatial information of the patches, by mining positive pairs from proximate tumor areas and negative pairs from distant tumor areas. Dufumier et al. (2021) used patient meta-data from MRI to form positive pairs[84]. 36 studies compared contrastive SSL pre-training to supervised pre-training, and reported an average improvement in performance of 6.35%.

Self-prediction

Self-prediction was used in six out of all included studies (Table 1). We consider studies that applied local-pixel shuffling as self-prediction since the augmentation operation, which shuffles the order of pixels, is applied only to a random patch of an image. Liu et al. used a U-net model to restore Ultrasound images augmented with local-pixel shuffling, and they subsequently concatenated the outputs of the U-net encoder with featurized clinical variables (age, gender, tumor size) for the downstream prediction task[121]. Similarly, Zhong et al. designed three image restoration tasks on cine-MRI videos and used a U-net-like encoder-decoder architecture including skip connections to perform the image restoration[119]. Three different image restoration tasks were set up using local-pixel shuffling, within-frame pixel shuffling, and covering an entire video frame with random pixels. Jana et al. used an encoder-decoder architecture for image restoration of CT scans that were corrupted by swapping several small patches within a single CT slice[118]. Jung et al. created a functional connectivity matrix between pairs of region-of-interest in rs-fMRI for each subject, and created a masked auto-encoder task by randomly masking out different rows and columns of the matrix for restoration[120]. Two of the five studies compared their approach to models without self-supervised pre-training and reported slight relative improvements in performance of 1.12%[117] and 0.690%[121].

Combined Approaches

Eleven studies found creative ways to combine different self-supervised learning strategies to pretrain their medical imaging models (Table 1). Over half of these studies (6/11) combined contrastive with generative approaches. With the exception of Ke et al.'s work[53], which uses a CycleGAN for histopathology slide stain normalization, all studies utilized an autoencoder as their generative model when combined with contrastive strategies. A combination of contrastive and innate relationships was used in three studies. The innate relationship tasks range from augmentation prediction and patch positioning prediction[122], rotation prediction[50], and ultrasound video to speech correspondence prediction[46]. For the remaining two studies, Cornelissen et al. used a conditional GAN, and trained the generator network on endoscopic images of the esophagus to either recolorize, inpaint and generate super-resolution images[49]. Because their tasks consisted of both inpainting (self-prediction) and super-resolution (generative), their approach was considered combined. Haghighi et al. combined three different SSL strategies (generative, innate relationship, self-prediction) by first training an auto-encoder and group instances with similar appearances based on the latent representations from the auto-encoder[48]. Then, the authors randomly cropped image patches at a fixed coordinate for all instances in the same group, and assigned a pseudo label to the cropped patches at each coordinate. Finally, the cropped patches were randomly perturbed and a restoration autoencoder was trained simultaneously with a pseudo label classification objective. Eight of the studies that combined different strategies compared self-supervised pre-training with purely supervised approaches, all of which reported performance improvement (0.140%-8.29%).

Discussion

This review aims to aggregate the collective knowledge of prior works that applied SSL to medical classification tasks. By synthesizing the relevant literature, we provide consistent definitions for self-supervised learning techniques, categorize prior works by their pre-training strategies, and provide implementation guidelines based on lessons learned from prior works. While five studies reported a slight decrease in performance (0.980%-4.51%), the majority of self-supervised pretrained models led to relative increased performances of 0.216–32.6% AUROC, 0.440–29.2% accuracy, and 0.137-14.3% F1 score over the same model architecture without SSL pretraining, including both ImagetNet and random initialization (Figure 3E). In Figure 3C we show a comparison of different SSL strategies on the same downstream task, which suggests that a combined strategy tends to outperform other self-supervised categories. However, it is important to note that combined strategies are typically the proposed method in the manuscripts, and thus publication bias might have resulted in this trend. In Figure 3D we additionally plot the performance of the two main types of fine-tuning strategies on the same task, and the graph tends to indicate that end-to-end fine-tuning leads to better performance regardless of the dataset size. In the presence of relevant data, we recommend implementing self-supervised learning strategies for training medical image classification models since our literature review indicated that self-supervised pre-training generally results in better model performance, especially when annotations are limited (Table 1).

The types of medical images utilized for model development as well as the downstream classification task encompassed a wide range of medical domains and applications (Figure 3A&B). The vast majority of the studies explored the clinical domain of radiology (47/79), of which 27 were focused on investigating abnormalities on chest imaging such as pneumonia, COVID-19, pleural effusion and pulmonary embolism (see Table 1). The choice of this domain is likely a combination of the availability of large-scale public chest datasets such as CheXpert[123], RSPECT[124], RadFusion[125] and MIMIC-CXR[126], as well as the motivation to solve acute or emerging healthcare threats which was the case during the coronavirus pandemic[57,67,68,70,97,103,106,108,109,117,127]. The second most prevalent clinical domain was pathology (12/79). Similar to radiology, this field is centered around medical imaging in the form of whole slide images. The tasks were focused on histopathological classification, where the majority focused on colorectal cancer classification[53,54,64,91,115]. The remaining studies explored multiple other tasks and many focused on classification of malignant disorders such as breast cancer[63,93,100], skin cancer[128], and lung cancer[62]. A particularly interesting medical task that was explored was classification of psychiatric diseases or psychiatric traits using fMRI[84,98,114]. Current limited knowledge and understanding of possible imaging features arising in psychiatric diseases constitutes a major clinical challenge to making local annotations such as bounding boxes or segmentations on brain scans. In this case both Osin et al. and Hashimoto demonstrated that training a self-supervised framework could be beneficial to generate representative latent features of brain fMRIs before fine-tuning on image-level class labels[98,114].

A majority of the included studies lacked strong baselines and ablation experiments. Even though 60 out of 79 studies compared their results with purely supervised baselines, only 33 studies reported comparisons with another self-supervised learning strategy. Of the 33 studies, 26 compared with a self-supervised category that differs from their best performing model. Among the SSL baselines, SimCLR was most frequently compared (16/26), followed by autoencoders (11/26) and MoCo (9/26). Furthermore, only 18 out of 79 studies indicated use of natural image pre-trained weights, either supervised or self-supervised, to

initiate their model for subsequent in-domain self-supervised pre-training. Lastly, merely 13 studies compared performance between classification on extracted features to end-to-end fine-tuning, two of which did not report numerical results. Of the 11 studies that quantitatively reported performance, eight found end-to-end fine-tuning to outperform training a new classifier on extracting features (Figure 3d). Since self-supervised learning for medical images is a promising yet nascent research area and the optimal strategies for training these models are still to be explored, researchers should systematically investigate different categories of self-supervised learning for their medical image datasets, in addition to fine-tuning strategy and pre-trained weights. Researchers should also test their newly developed strategies on multiple datasets, ideally on different modalities and medical imaging domains.

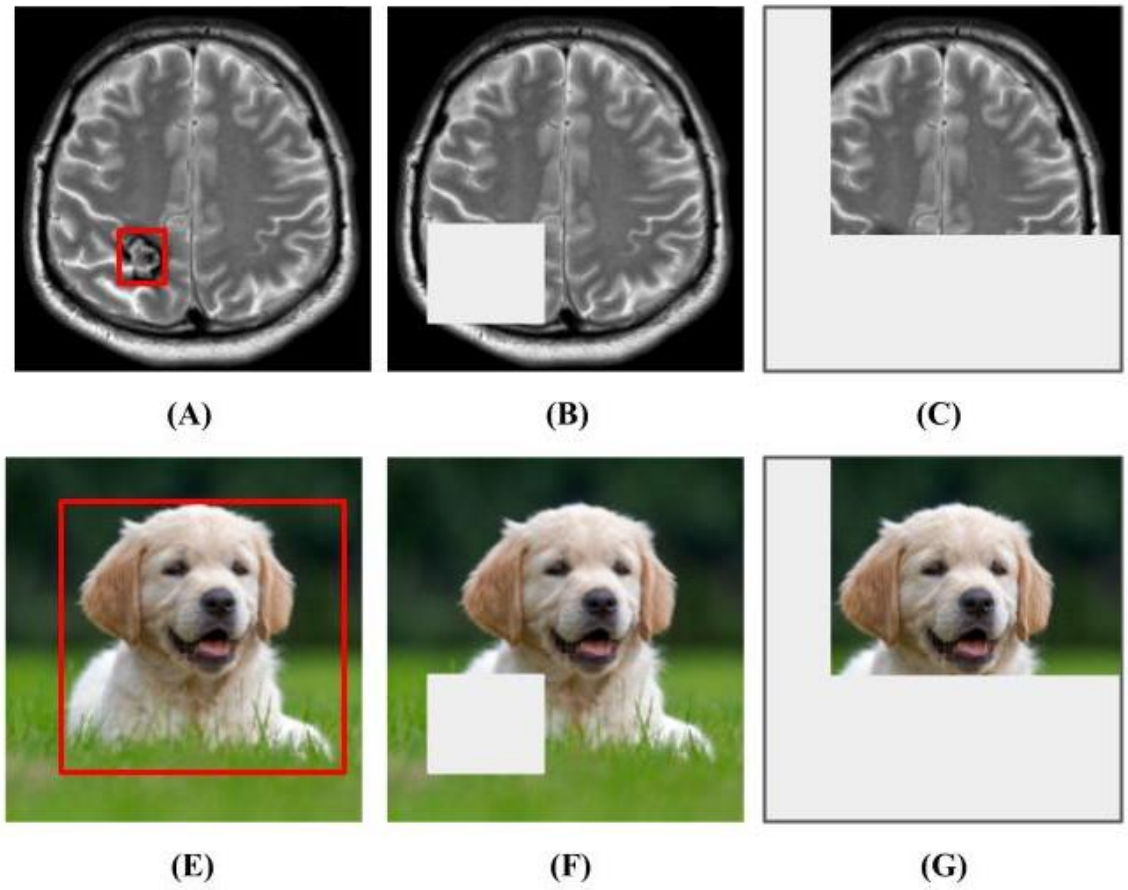


Figure 4. Examples of augmentations and transformations that alter the semantic meaning of medical images[129] but not natural images[130]. A) The image shows a T2-weighted brain MRI with a cavernoma in the right parietal lobe (bounded in red). B) and C) Masking and cropping operations can obscure the cavernoma and alter the semantic meaning of the image, as the MRI-scan no longer exhibits any abnormality. E) Image of a dog (bounded in red), F) and G) Masking and cropping operations do not obscure the dog nor alter the semantic meaning of the image.

Implementation guidelines for self-supervised learning models

Definitive conclusions on the optimal strategy for medical images cannot be made since only a subset of studies made comparisons between different types of self-supervised learning strategies. Furthermore, the optimal strategy may be dependent on a number of factors including the specific medical imaging domain, the size and complexity of the dataset, and the type of classification task[131,132]. Due to this heterogeneity, we encourage researchers to compare multiple self-supervised learning strategies for developing medical image classification models, especially in limited data regimes. Although self-supervised pre-training can be computational demanding, many models pre-trained in a self-supervised manner on large-scale natural image datasets are publicly available and should be utilized. Azizi et al. have shown that models that are SSL pre-trained using natural images tend to outperform purely supervised pre-trained models[72] for medical image classification, and continuing self-supervised pre-training with in-domain medical images leads to the best results. More recently, Azizi et al. found that using generic and large-scale *supervised* pretrained models, such as BigTransfer[133], can also benefit subsequent domain-specific self-supervised pre-training, and ultimately improve model performance and robustness for different medical imaging modalities[134]. Truong et al. have demonstrated the effectiveness of combining representations from multiple self-supervised methods to improve performance for three different medical imaging modalities[73].

It is worth noting that representations learned using certain SSL strategies can be relatively more linearly separable, while representations from other strategies can achieve better performance when more layers or the entire model are fine-tuned. For instance, for natural image datasets, MoCo outperforms MAE by training a linear model on extracted features (linear probing), while MAE achieves better performance than MoCo as the number of fine-tune layers increases[41]. Likewise, Cornelissen et al. demonstrated that using representations from earlier layers can improve downstream classification of neoplasia in Barrett's Esophagus[49]. Factors such as the degree of domain shift between SSL pre-training data and downstream task data could also affect the linear separability of the representations. Based on our aggregated results, we found that end-to-end fine-tuning generally leads to better performance for medical images (Figure 3C). However, due to the lack of ablation studies from current work, we cannot determine whether fine-tuning only later layers of the model could lead to better performance, relative to end-to-end fine-tuning. Furthermore, even though self-supervised learning strategies generate label-efficient representations, the learning process typically requires a relatively large amount of unlabeled data. For instance, reducing the number of pre-training images from 250k to 50k typically leads to more than 10.0% drop in accuracy for downstream tasks, while reducing from 1M to 250k leads to a 2.00-4.00% decrease[132]. Curating large-scale medical image datasets from multiple institutions is often challenged by the difficulty of sharing patient data while preserving patient privacy. Nevertheless, Yan et al. have demonstrated the possibility of training self-supervised models with data from multiple healthcare centers without the need for explicitly sharing data using federated learning, and shown improvement in robustness and performance over models trained using data from only one institution[135].

The field of self-supervised learning for computer vision is constantly and rapidly evolving. While many self-supervised methods have led to state-of-the-art results when fine-tuned on natural image datasets, how translatable these results are to medical datasets is unclear, mainly due to the unique properties of medical images. For instance, many contrastive based strategies have been developed based on the assumption that the class-defining object is the main focus of an image, and thus variations caused by image transformations should not alter the image's semantic meanings (Figure 4). Therefore, these methods typically apply strong

transformations to the original image and encourage the model to learn similar global representations for images with similar semantic meanings. However, the assumption made for natural images is not necessarily valid for medical images for two reasons. First, medical images have high inter-class visual similarities due to the standardized protocols for medical image acquisition and the homogeneous nature of human anatomy. Second, within the medical imaging domain, the semantic meaning of interest is rarely an object such as the anatomical organ, but is rather the presence or absence of pathological abnormalities within that organ or tissue. Many abnormalities are characterized by very subtle and localized visual cues, which can become ambiguous or obscured by augmentations (Figure 4c). The random masking operation often utilized by self-prediction self-supervised learning methods may also alter a medical image's semantic meaning by removing image regions with diseases or abnormalities (Figure 4b). Recent work has demonstrated the benefit of using learned visual word masks[136,137] or spatially constrained crops[138,139] to encourage representational invariance with semantically more similar regions of an image. In a similar vein, we believe that augmentation strategies tailored for the nature of medical images during self-supervised learning is a future research area that warrants further research and exploration.

The unique properties of medical images can be leveraged to design self-supervised learning methods more suitable for specific downstream tasks. For instance, instead of forming positive pairs with different augmented versions of the same image during contrastive learning, one can improve positive sampling according to the similarity between a patient's clinical information. In fact, several studies have shown performance improvement when constructing positive pairs with slices from the same CT series[77], images from the same imaging study[75], images from the same patient[72] and patients with similar age[84]. Future research should explore other strategies for defining positive pairs, such as leveraging patient demographics or medical history information. The unique attributes of medical images can also be utilized for creating relevant pretext tasks. Rivail et al. proposed a self-supervised approach to model disease progression by estimating the time interval between pairs of optical coherence tomography (OCT) scans from the same patient[101]. Involving additional modalities during self-supervised learning has also been shown to improve a model's performance when fine-tuned for downstream tasks. For example, Taleb et al. proposed a multimodal contrastive learning strategy between retinal fundus images and genetics data and showed improvement in performance over single modality pre-training[140]. Jiao et al. cleverly leveraged the correlation between fetal ultrasonography and the narrative speech of the sonographer to create a pretext task for self-supervision, and subsequently used the learned representations for downstream standard plane classification on sonograms[46]. Furthermore, many medical imaging modalities have corresponding radiology reports that contain detailed descriptions of the medical conditions observed by radiologists. Several studies have utilized these medical reports to provide supervision signals during self-supervised learning and shown label efficiency for downstream tasks[76,141]. By leveraging radiology reports, Huang et al. demonstrated the model's ability to localize chest abnormalities on chest x-rays without segmentation labels and revealed the possibility of zero-shot learning by converting the classification classes into textual captions and framing the image classification task as measuring the image-text similarity[142]. However, currently there are very few publicly available medical image datasets with corresponding radiology reports, largely due to the difficulties in preserving patient privacy. Therefore, these multi-modal self-supervised learning strategies are limited to implementation within a healthcare system until more datasets with medical image and report pairs are publicly released. Overall, the flexibility in creating self-supervised methods as well as the adaptability and transferability to multiple medical domains highlights the importance and utility of self-supervised techniques in clinical applications.

Limitations

For this review paper, publication bias can be a considerable limitation due to disproportionately reported positive results in the literature, which can lead to overestimation of the benefits of self-supervised learning. We also limited our search to only consider papers published after 2012, which excluded papers that applied self-supervised learning prior to the era of deep learning for computer vision[143]. Furthermore, we are unable to aggregate or statistically compare the effects of each self-supervised learning strategy on performance gain, since the included studies use different imaging modalities, report different performance metrics, and investigate different objectives. Additionally, subjectivity may have been introduced when categorizing the self-supervised learning strategy in each paper, especially for studies that implemented novel, unconventional, or a mixture of methods. Lastly, our study selection criteria only included literature for the task of medical image classification, which limits the scope of this review paper, since we recognize that self-supervised pretrained models can also be finetuned for other important medical tasks, including segmentation, regression, and registration.

Future Research

Results from this systematic review have revealed that SSL for medical image classification is a growing and promising field of research across multiple medical disciplines and imaging modalities. We found that self-supervised pre-training generally improves performance for medical imaging classification tasks over purely supervised methods. We categorized the SSL approaches used in medical imaging tasks as a framework for methodologic communication and synthesized benefits and limitations from literature to provide recommendations for future research. Future studies should include direct comparisons of different self-supervised learning strategies using shared terminology and metrics whenever applicable to facilitate the discovery of additional insights and optimal approaches.

Methods

This systematic review was conducted based on the PRISMA guidelines[144].

Search Strategy

A systematic literature search was implemented in three literature databases: PubMed, Scopus and ArXiv. The key search terms were based on a combination of two major themes: "self-supervised learning" and "medical imaging". Search terms for medical imaging were not limited to radiological imaging but were also broadly defined to include imaging from all medical fields, i.e., fundus photography, whole slide imaging, endoscopy, echocardiography, etc. Since we specifically wanted to review literature on the task of medical image classification, terms for other computer vision tasks such as segmentation, image reconstruction, denoising and object detection were used as exclusion criteria in the search. The search encompassed papers published between Janunary. 2012 and May 2021. The start date was considered appropriate due to the rising popularity of deep learning for computer vision since the 2012 ImageNet challenge. The complete search string for all three databases is provided in Supplementary Methods.

We included all research papers in English that used self-supervision techniques to develop deep learning models for medical image classification tasks. The research papers had to be original research in the form of either journal articles, conference proceedings or extended abstracts. We excluded any publications that

were not peer-reviewed. Applicable self-supervision techniques were defined according to the different types presented in the "terminology and strategies in self-supervised learning" section. We only included studies that applied the deep learning models to a downstream medical image classification task, i.e, it was not sufficient to have developed a self-supervision model on medical images. Additionally, the medical image classification task had to be a clinical task or clinically relevant task. For example, the downstream task of classifying the frame number in a temporal sequence of frames from echocardiography[145] was not considered a clinically relevant task.

We excluded studies that used semi-supervised learning or any amount of manually curated labels during the self-supervision step. We also excluded studies that only investigated regression or segmentation in their downstream tasks. Furthermore, we excluded any studies where the self-supervised pretrained model was not directly fine-tuned for classification after pretraining. Studies that used non-human medical imaging data (i.e., veterinarian medical images) were also excluded.

Study Selection

The Covidence software (www.covidence.org) was used for screening and study selection. After the removal of duplicates, studies were screened based on title and abstract, and then full texts were obtained and assessed for inclusion. Study selection was performed by three independent researchers (S.-C.H., A.P., and M.J.), and disagreements were resolved through discussion. In cases where consensus could not be achieved a forth arbitrating researcher was consulted (A.S.C.).

Data Extraction

For benchmarking the existing approaches we extracted the following data from each of the selected articles: a) self-supervised learning strategy, b) year of publication, c) first author, d) imaging modality, e) clinical domain, f) outcome/task, g) combined method, h) self-supervised framework, i) strategy for fine-tuning, j) performance metrics, k) SSL performance, l) supervised performance, and m) difference in SSL and supervised performance (Table 1). We also computed the relative difference in performance between the supervised and self-supervised model on the p) full dataset and q) subset. We classified the specific self-supervised learning strategy based on the definitions in the section "Terminology and strategies in self-supervised learning". We extracted AUROC whenever this metric was reported, otherwise we prioritize accuracy over F1 score and sensitivity. When the article contained results from multiple models (i.e. ResNet and DenseNet), metrics from the experiment with the best average performing self-supervised model were extracted. When the authors present results from several clinical tasks, we chose tasks that best corresponded to the title and objective of the manuscript. If the tasks were deemed equal, we picked the task where the chosen SSL model had the highest performance. We picked supervised baseline with the same model architecture and pre-training dataset for performance comparison. If the author did not report performance from a supervised model that uses the same pre-training dataset, preference was given to ImageNet pretrained model over a randomly initialized one. The pre-training dataset used by the self-supervised and supervised model are recorded in the Supplementary Table. When papers report results on many percentages of fine-tuning (i.e., 1%, 10%, 100%), we pick the lowest and highest to study the label-efficiency of self-supervised learning methods. We also provide a Supplementary Table 1 with additional

technical details including model architecture, dataset details, number of training samples, comparison to selected baselines and performance on subsets of data. These items were extracted to enable researchers to find and compare current self-supervised studies in their medical field or input modalities of interest.

Data Availability

The authors declare that all data supporting the findings of this study are available within the paper and its Supplementary information files.

Contributions

S.-C.H. and A.P. are co-first authors who contributed equally to this study. Concept and design: S.-C.H. and A.P.. Study selection: S.-C.H. and A.P. Data extraction: S.-C.H., A.P., and M.J. Drafting of the manuscript: S.-C.H., A.P., and M.J. Critical revision of the manuscript for important intellectual content: S.-C.H., A.P., M.J., M.P.L., S.Y. and A.S.C. Supervision: S.Y. and A.S.C.

Acknowledgement

Research reported in this publication was supported by NIH grants R01 AR077604, R01 EB002524, R01 AR079431, R01 HL155410, R01 LM012966, and P41 EB027060; NIH contracts 75N92020C00008 and 75N92020C00021. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health

Competing interests

The authors declare no competing interests.

References

1. Hong, A. S. *et al.* Trends in Diagnostic Imaging Utilization among Medicare and Commercially Insured Adults from 2003 through 2016. *Radiology* **294**, 342–350 (2020).
2. Smith-Bindman, R. *et al.* Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016. *JAMA* **322**, 843–856 (2019).
3. McDonald, R. J. *et al.* The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Acad. Radiol.* **22**, 1191–1198 (2015).
4. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. *Nat. Rev. Cancer* **18**, 500–510 (2018).
5. Dan Lantsman, C. *et al.* Trend in radiologist workload compared to number of admissions in the

emergency department. *Eur. J. Radiol.* **149**, 110195 (2022).

6. Alonso-Martínez, J. L., Sánchez, F. J. A. & Echezarreta, M. A. U. Delay and misdiagnosis in sub-massive and non-massive acute pulmonary embolism. *Eur. J. Intern. Med.* **21**, 278–282 (2010).

7. Hendriksen, J. M. T. *et al.* Clinical characteristics associated with diagnostic delay of pulmonary embolism in primary care: a retrospective observational study. *BMJ Open* **7**, e012789 (2017).

8. Dunnmon, J. A. *et al.* Assessment of Convolutional Neural Networks for Automated Classification of Chest Radiographs. *Radiology* **290**, 537–544 (2019).

9. Rajpurkar, P. *et al.* Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* **15**, e1002686 (2018).

10. Larson, D. B. *et al.* Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs. *Radiology* **287**, 313–322 (2018).

11. Park, A. *et al.* Deep Learning–Assisted Diagnosis of Cerebral Aneurysms Using the HeadXNet Model. *JAMA Netw Open* **2**, e195600–e195600 (2019).

12. Bien, N. *et al.* Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Med.* **15**, e1002699 (2018).

13. Esteva, A. *et al.* Prostate cancer therapy personalization via multi-modal deep learning on randomized phase III clinical trials. *NPJ Digit Med* **5**, 71 (2022).

14. Esteva, A. *et al.* Development and validation of a prognostic AI biomarker using multi-modal deep learning with digital histopathology in localized prostate cancer on NRG Oncology phase III clinical trials. *J. Clin. Orthod.* **40**, 222–222 (2022).

15. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).

16. Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med.* **15**, e1002683 (2018).

17. LeCun, Y. & Misra, I. Self-supervised Learning: The Dark Matter of Intelligence. Preprint at (2021).

18. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional

Transformers for Language Understanding. *arXiv [cs.CL]* (2018).

19. Brown, Mann & Ryder. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.*
20. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv [cs.LG]* (2020).
21. Krishnan, R., Rajpurkar, P. & Topol, E. J. Self-supervised learning in medicine and healthcare. *Nat Biomed Eng* (2022) doi:10.1038/s41551-022-00914-1.
22. Shurrab, S. & Duwairi, R. Self-supervised learning methods and applications in medical imaging analysis: a survey. *PeerJ Comput Sci* **8**, e1045 (2022).
23. Lilian Weng, J. W. K. Self-Supervised Learning: Self-Prediction and Contrastive Learning. (2021).
24. Gidaris, S., Singh, P. & Komodakis, N. Unsupervised Representation Learning by Predicting Image Rotations. *arXiv [cs.CV]* (2018).
25. Noroozi, M. & Favaro, P. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. *arXiv [cs.CV]* (2016).
26. Doersch, C., Gupta, A. & Efros, A. A. Unsupervised Visual Representation Learning by Context Prediction. *2015 IEEE International Conference on Computer Vision (ICCV)* Preprint at https://doi.org/10.1109/iccv.2015.167 (2015).
27. Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015).
28. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *arXiv [stat.ML]* (2013).
29. Goodfellow, I. J. *et al.* Generative Adversarial Networks. *arXiv [stat.ML]* (2014).
30. Vincent, P., Larochelle, H., Bengio, Y. & Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. in *Proceedings of the 25th international conference on Machine learning* 1096–1103 (Association for Computing Machinery, 2008).
31. Donahue & Simonyan. Large scale adversarial representation learning. *Adv. Neural Inf. Process. Syst.*
32. Donahue, J., Krähenbühl, P. & Darrell, T. Adversarial Feature Learning. *arXiv [cs.LG]* (2016).
33. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum Contrast for Unsupervised Visual

Representation Learning. *arXiv [cs.CV]* (2019).

34. Caron, M. *et al.* Emerging Properties in Self-Supervised Vision Transformers. *arXiv [cs.CV]* (2021).
35. Grill, Strub, Altché & Tallec. Bootstrap your own latent-a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.*
36. Chen, X. & He, K. Exploring Simple Siamese Representation Learning. *arXiv [cs.CV]* (2020).
37. Caron, M. *et al.* Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *arXiv [cs.CV]* (2020).
38. Asano, Y. M., Rupprecht, C. & Vedaldi, A. Self-labelling via simultaneous clustering and representation learning. *arXiv [cs.CV]* (2019).
39. Gidaris, S., Bursuc, A., Komodakis, N., Perez, P. & Cord, M. Learning Representations by Predicting Bags of Visual Words. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* Preprint at https://doi.org/10.1109/cvpr42600.2020.00696 (2020).
40. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T. & Efros, A. A. Context Encoders: Feature Learning by Inpainting. *arXiv [cs.CV]* (2016).
41. He, K. *et al.* Masked Autoencoders Are Scalable Vision Learners. *arXiv [cs.CV]* (2021).
42. Bao, H., Dong, L. & Wei, F. BEiT: BERT Pre-Training of Image Transformers. *arXiv [cs.CV]* (2021).
43. Cunningham, P. & Delany, S. J. k-Nearest Neighbour Classifiers - A Tutorial. *ACM Comput. Surv.* **54**, 1–25 (2021).
44. Bozorgtabar, B., Mahapatra, D., Vray, G. & Thiran, J.-P. SALAD: Self-supervised Aggregation Learning for Anomaly Detection on X-Rays. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* 468–478 (Springer International Publishing, 2020).
45. Hsieh, W.-T., Lefort-Besnard, J., Yang, H.-C., Kuo, L.-W. & Lee, C.-C. Behavior Score-Embedded Brain Encoder Network for Improved Classification of Alzheimer Disease Using Resting State fMRI. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2020**, 5486–5489 (2020).
46. Jiao, J. *et al.* Self-Supervised Contrastive Video-Speech Representation Learning for Ultrasound. in

Medical Image Computing and Computer Assisted Intervention – MICCAI 2020 534–543 (Springer International Publishing, 2020).

47. Tian, Y. *et al.* Constrained Contrastive Distribution Learning for Unsupervised Anomaly Detection and Localisation in Medical Images. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* 128–140 (Springer International Publishing, 2021).
48. Haghighi, F., Taher, M. R. H., Zhou, Z., Gotway, M. B. & Liang, J. Transferable Visual Words: Exploiting the Semantics of Anatomical Patterns for Self-supervised Learning. *IEEE Trans. Med. Imaging* **PP**, (2021).
49. Cornelissen & Putten. Evaluating self-supervised learning methods for downstream classification of neoplasia in barrett's esophagus. *on Image Processing …*.
50. Li, X. *et al.* Rotation-Oriented Collaborative Self-Supervised Learning for Retinal Disease Diagnosis. *IEEE Trans. Med. Imaging* **40**, 2284–2294 (2021).
51. Fedorov, A. *et al.* Tasting the cake: evaluating self-supervised generalization on out-of-distribution multimodal MRI data. in *RobustML Workshop ICLR 2021* (2021).
52. Ouyang, J. *et al.* Self-supervised Longitudinal Neighbourhood Embedding. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* 80–89 (Springer International Publishing, 2021).
53. Ke, J., Shen, Y., Liang, X. & Shen, D. Contrastive Learning Based Stain Normalization Across Multiple Tumor in Histopathology. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* 571–580 (Springer International Publishing, 2021).
54. Yang, P., Hong, Z., Yin, X., Zhu, C. & Jiang, R. Self-supervised Visual Representation Learning for Histopathological Images. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* 47–57 (Springer International Publishing, 2021).
55. Sowrirajan, H., Yang, J., Ng, A. Y. & Rajpurkar, P. MoCo-CXR: MoCo Pretraining Improves Representation and Transferability of Chest X-ray Models. in *Proceedings of Machine Learning Research* 143:727–743 (2021).

56. Zhou, H.-Y. *et al.* Comparing to Learn: Surpassing ImageNet Pretraining on Radiographs by Comparing Image Representations. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* 398–407 (Springer International Publishing, 2020).
57. Sun, L., Yu, K. & Batmanghelich, K. Context Matters: Graph-based Self-supervised Representation Learning for Medical Images. in *Proc Conf AAAI Artif Intell.* (2021).
58. Burlina, P. *et al.* Low-Shot Deep Learning of Diabetic Retinopathy With Potential Applications to Address Artificial Intelligence Bias in Retinal Diagnostics and Rare Ophthalmic Diseases. *JAMA Ophthalmol.* **138**, 1070–1077 (2020).
59. Li, X., Jia, M., Islam, M. T., Yu, L. & Xing, L. Self-Supervised Feature Learning via Exploiting Multi-Modal Data for Retinal Disease Diagnosis. *IEEE Trans. Med. Imaging* **39**, 4023–4033 (2020).
60. Fedorov, A. *et al.* On self-supervised multi-modal representation learning: An application to Alzheimer's disease. in *IEEE 18th International Symposium on Biomedical Imaging* (2021).
61. Mojab, N. *et al.* Real-World Multi-Domain Data Applications for Generalizations to Clinical Settings. in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)* 677–684 (2020).
62. Li, B., Li, Y. & Eliceiri, K. W. Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* (2021).
63. Dehaene, O., Camara, A., Moindrot, O., de Lavergne, A. & Courtiol, P. Self-Supervision Closes the Gap Between Weak and Strong Supervision in Histology. in *ML4H 2020 NeurIPS workshop* (2020).
64. Ciga, O., Xu, T. & Martel, A. L. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications* **7**, (2022).
65. Reed, C. J. *et al.* Self-Supervised Pretraining Improves Self-Supervised Pretraining. in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2022).
66. Liu, F. *et al.* Self-supervised Mean Teacher for Semi-supervised Chest X-Ray Classification. in *Machine Learning in Medical Imaging* 426–436 (Springer International Publishing, 2021).

67. Hao, H., Didari, S., Woo, J. O., Moon, H. & Bangert, P. Highly Efficient Representation and Active Learning Framework for Imbalanced Data and its Application to COVID-19 X-Ray Classification. *Conference on Neural Information Processing Systems (NeurIPS 2021)* (2021).
68. Li, J. *et al.* Multi-task contrastive learning for automatic CT and X-ray diagnosis of COVID-19. *Pattern Recognit.* **114**, 107848 (2021).
69. Gazda, M., Gazda, J., Plavka, J. & Drotar, P. Self-supervised deep convolutional neural network for chest X-ray classification. *arXiv [eess.IV]* (2021).
70. Dong, N. & Voiculescu, I. Federated Contrastive Learning for Decentralized Unlabeled Medical Images. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* 378–387 (Springer International Publishing, 2021).
71. Nguyen, N.-Q. & Le, T.-S. A Semi-Supervised Learning Method to Remedy the Lack of Labeled Data. in *2021 15th International Conference on Advanced Computing and Applications (ACOMP)* 78–84 (2021).
72. Azizi, Mustafa, Ryan & Beaver. Big self-supervised models advance medical image classification. *Proc. Estonian Acad. Sci. Biol. Ecol.*
73. Truong, T., Mohammadi, S. & Lenga, M. How Transferable are Self-supervised Features in Medical Image Classification Tasks? in *Proceedings of Machine Learning for Health* (eds. Roy, S. et al.) vol. 158 54–74 (PMLR, 2021).
74. Zhao, X. & Zhou, S. Fast Mixing of Hard Negative Samples for Contrastive Learning and Use for COVID-19. in *2021 4th International Conference on Big Data Technologies* 6–12 (Association for Computing Machinery, 2021).
75. Vu, Y. N. T. *et al.* MedAug: Contrastive learning leveraging patient metadata improves representations for chest X-ray interpretation. in *Proceedings of Machine Learning Research* 126:1–14 (2021).
76. Ji, Z. *et al.* Improving Joint Learning of Chest X-Ray and Radiology Report by Word Region Alignment. in *Machine Learning in Medical Imaging* 110–119 (Springer International Publishing,

2021).

77. Dong, H. *et al.* Case Discrimination: Self-supervised Feature Learning for the Classification of Focal Liver Lesions. in *Innovation in Medicine and Healthcare* 241–249 (Springer Singapore, 2021).

78. Islam, N. U., Gehlot, S., Zhou, Z., Gotway, M. B. & Liang, J. Seeking an Optimal Approach for Computer-Aided Pulmonary Embolism Detection. *Mach Learn Med Imaging* **12966**, 692–702 (2021).

79. Bao, W., Jin, Y., Huang, C. & Peng, W. CT Image Classification of Invasive Depth of Gastric Cancer based on 3D-DPN Structure. in *The 11th International Workshop on Computer Science and Engineering (WCSE 2021)* 115–121.

80. Jian, G.-Z., Lin, G.-S., Wang, C.-M. & Yan, S.-L. Helicobacter Pylori Infection Classification Based on Convolutional Neural Network and Self-Supervised Learning. in *2021 the 5th International Conference on Graphics and Signal Processing* 60–64 (Association for Computing Machinery, 2021).

81. Kaku, A., Upadhya, S. & Razavian, N. Intermediate layers matter in momentum contrastive self supervised learning. in *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*. (2021).

82. Yellapragada, B., Hornauer, S., Snyder, K., Yu, S. & Yiu, G. Self-Supervised Feature Learning and Phenotyping for Assessing Age-Related Macular Degeneration Using Retinal Fundus Images. *Ophthalmol Retina* **6**, 116–129 (2022).

83. Perek, S., Amit, M. & Hexter, E. Self Supervised Contrastive Learning on Multiple Breast Modalities Boosts Classification Performance. in *Predictive Intelligence in Medicine* 117–127 (Springer International Publishing, 2021).

84. Dufumier, B. *et al.* Contrastive Learning with Continuous Proxy Meta-data for 3D MRI Classification. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* 58–68 (Springer International Publishing, 2021).

85. Li, H. *et al.* Imbalance-Aware Self-supervised Learning for 3D Radiomic Representations. in

Medical Image Computing and Computer Assisted Intervention – MICCAI 2021 36–46 (Springer International Publishing, 2021).

86. Manna, S., Bhattacharya, S. & Pal, U. Interpretive self-supervised pre-training: boosting performance on visual medical data. in *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing* 1–9 (Association for Computing Machinery, 2021).

87. Roychowdhury, S., Tang, K. S., Ashok, M. & Sanka, A. SISE-PC: Semi-supervised Image Subsampling for Explainable Pathology Classification. in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)* 2806–2809 (2021).

88. Ren, Z., Guo, Y., Yu, S. X. & Whitney, D. Improve Image-based Skin Cancer Diagnosis with Generative Self-Supervised Learning. in *2021 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)* 23–34 (2021).

89. Zhao, Z. & Yang, G. Unsupervised Contrastive Learning of Radiomics and Deep Features for Label-Efficient Tumor Classification. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* 252–261 (Springer International Publishing, 2021).

90. Saillard, C. *et al.* Self supervised learning improves dMMR/MSI detection from histology slides across multiple cancers. *COMPAY@MICCAI* (2021).

91. Li, J., Lin, T. & Xu, Y. SSLP: Spatial Guided Self-supervised Learning on Pathological Images. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* 3–12 (Springer International Publishing, 2021).

92. Liu, Q. *et al.* SimTriplet: Simple Triplet Representation Learning with a Single GPU. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* 102–112 (Springer International Publishing, 2021).

93. Wang, X. *et al.* TransPath: Transformer-Based Self-supervised Learning for Histopathological Image Classification. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* 186–195 (Springer International Publishing, 2021).

94. Sharmay, Y., Ehsany, L., Syed, S. & Brown, D. E. HistoTransfer: Understanding Transfer Learning

for Histopathology. in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)* 1–4 (2021).

95. Spahr, A., Bozorgtabar, B. & Thiran, J.-P. Self-Taught Semi-Supervised Anomaly Detection On Upper Limb X-Rays. in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* 1632–1636 (2021).

96. Li, G., Togo, R., Ogawa, T. & Haseyama, M. Triplet Self-Supervised Learning for Gastritis Detection with Scarce Annotations. in *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)* 787–788 (2021).

97. Hossain, M. B., Iqbal, S. M. H. S., Islam, M. M., Akhtar, M. N. & Sarker, I. H. Transfer learning with fine-tuned deep CNN ResNet50 model for classifying COVID-19 from chest X-ray images. *Inform Med Unlocked* **30**, 100916 (2022).

98. Osin, J. *et al.* Learning Personal Representations from fMRI by Predicting Neurofeedback Performance. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* 469–478 (Springer International Publishing, 2020).

99. Zhao, Q., Liu, Z., Adeli, E. & Pohl, K. M. Longitudinal Self-Supervised Learning. *Med. Image Anal.* **71**, (2021).

100. Gamper, J. & Rajpoot, N. Multiple Instance Captioning: Learning Representations from Histopathology Textbooks and Articles. in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 16544–16554 (2021).

101. Rivail, A. *et al.* Modeling Disease Progression in Retinal OCTs with Longitudinal Self-supervised Learning. in *Predictive Intelligence in Medicine* 44–52 (Springer International Publishing, 2019).

102. Droste, R. *et al.* Ultrasound Image Representation Learning by Modeling Sonographer Visual Attention. in *International conference on information processing in medical imaging 2019* (2019).

103. Ewen, N. & Khan, N. Targeted Self Supervision For Classification On A Small Covid-19 Ct Scan Dataset. in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* 1481–1485 (2021).

104. Manna, S., Bhattacharya, S. & Pal, U. Self-Supervised Representation Learning for Detection of ACL Tear Injury in Knee MR Videos. *Pattern Recognit. Lett.* **154**, 37–43 (2022).
105. Jiao, J., Droste, R., Drukker, L., Papageorghiou, A. T. & Alison Noble, J. Self-supervised Representation Learning for Ultrasound Video. in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* (2020).
106. Abbas, A., Abdelsamea, M. M. & Gaber, M. M. 4S-DT: Self-Supervised Super Sample Decomposition for Transfer Learning With Application to COVID-19 Detection. *IEEE Trans Neural Netw Learn Syst* **32**, 2798–2808 (2021).
107. Vats, A., Pedersen, M. & Mohammed, A. A Preliminary Analysis of Self-Supervision for Wireless Capsule Endoscopy. in *2021 9th European Workshop on Visual Information Processing (EUVIP)* 1–6 (2021).
108. Ewen, N. & Khan, N. Online Unsupervised Learning For Domain Shift In Covid-19 CT Scan Datasets. in *2021 IEEE International Conference on Autonomous Systems (ICAS)* 1–5 (2021).
109. Zhu, Y. Self-supervised Learning for Small Shot COVID-19 Classification. in *2021 3rd International Conference on Information Technology and Computer Communications* 36–40 (Association for Computing Machinery, 2021).
110. Long, Y. Pneumonia Identification with Self-supervised Learning and Transfer Learning. in *Application of Intelligent Systems in Multi-modal Information Analytics* 627–635 (Springer International Publishing, 2021).
111. Dezaki, F. T. *et al.* Echo-Rhythm Net: Semi-Supervised Learning For Automatic Detection of Atrial Fibrillation in Echocardiography. in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* 110–113 (2021).
112. Wicaksono, R. S. H., Septiandri, A. A. & Jamal, A. Human Embryo Classification Using Self-Supervised Learning. in *2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS)* 1–5 (2021).
113. Vu, Y. N. T., Tsue, T., Su, J. & Singh, S. An improved mammography malignancy model with self-

supervised learning. in *Medical Imaging 2021: Computer-Aided Diagnosis* vol. 11597 210–216 (SPIE, 2021).

114. Hashimoto, Y., Ogata, Y., Honda, M. & Yamashita, Y. Deep Feature Extraction for Resting-State Functional MRI by Self-Supervised Learning and Application to Schizophrenia Diagnosis. *Front. Neurosci.* **15**, 696853 (2021).
115. Srinidhi, C. L., Kim, S. W., Chen, F.-D. & Martel, A. L. Self-supervised driven consistency training for annotation efficient histopathology image analysis. *Med. Image Anal.* **75**, (2021).
116. Tamkin, A. *et al.* DABS: A Domain-Agnostic Benchmark for Self-Supervised Learning. in *NeurIPS 2021 Datasets and Benchmarks Track* (2021).
117. Park, J., Kwak, I.-Y. & Lim, C. A Deep Learning Model with Self-Supervised Learning and Attention Mechanism for COVID-19 Diagnosis Using Chest X-ray Images. *Electronics* **10**, 1996 (2021).
118. Jana, A. *et al.* Liver Fibrosis And NAS Scoring From CT Images Using Self-Supervised Learning And Texture Encoding. in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* 1553–1557 (2021).
119. Zhong, H. *et al.* A Self-supervised Learning Based Framework for Automatic Heart Failure Classification on Cine Cardiac Magnetic Resonance Image. in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)* 2887–2890 (2021).
120. Jung, W., Heo, D.-W., Jeon, E., Lee, J. & Suk, H.-I. Inter-regional High-Level Relation Learning from Functional Connectivity via Self-supervision. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* 284–293 (Springer International Publishing, 2021).
121. Liu, C. *et al.* TN-USMA Net: Triple normalization-based gastrointestinal stromal tumors classification on multicenter EUS images with ultrasound-specific pretraining and meta attention. *Med. Phys.* **48**, 7199–7214 (2021).
122. Tian, Y. *et al.* Constrained Contrastive Distribution Learning for Unsupervised Anomaly Detection and Localisation in Medical Images. *arXiv [cs.CV]* (2021).

123. Irvin, J. *et al.* CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *arXiv [cs.CV]* (2019).
124. Colak, E. *et al.* The RSNA Pulmonary Embolism CT Dataset. *Radiol Artif Intell* **3**, e200254 (2021).
125. Zhou, Y. *et al.* RadFusion: Benchmarking Performance and Fairness for Multimodal Pulmonary Embolism Detection from CT and EHR. *arXiv [eess.IV]* (2021).
126. Johnson, A. E. W. *et al.* MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv [cs.CV]* (2019).
127. Zhang, S., Zou, B., Xu, B., Su, J. & Hu, H. An Efficient Deep Learning Framework of COVID-19 CT Scans Using Contrastive Learning and Ensemble Strategy. in *2021 IEEE International Conference on Progress in Informatics and Computing (PIC)* 388–396 (2021).
128. Liu, Q. *et al.* SimTriplet: Simple Triplet Representation Learning with a Single GPU. *arXiv [cs.CV]* (2021).
129. Hellerhoff. File:Kavernom rechts parietal 59M - MR - 001.jpg. *Wikimedia* https://commons.wikimedia.org/wiki/File:Kavernom_rechts_parietal_59M_-_MR_-_001.jpg (2022).
130. Matio, H. File:Dog Breeds.jpg. *Wikimedia Commons* https://commons.wikimedia.org/wiki/File:Dog_Breeds.jpg (2019).
131. Kotar, Ilharco & Schmidt. Contrasting contrastive self-supervised representation learning pipelines. *Proc. Estonian Acad. Sci. Biol. Ecol.*
132. Cole, E., Yang, X., Wilber, K., Aodha, O. M. & Belongie, S. When Does Contrastive Visual Representation Learning Work? *arXiv [cs.CV]* (2021).
133. Kolesnikov, A. *et al.* Big Transfer (BiT): General Visual Representation Learning. *arXiv [cs.CV]* (2019).
134. Azizi, S. *et al.* Robust and Efficient Medical Imaging with Self-Supervision. *arXiv [cs.CV]* (2022).
135. Yan, R. *et al.* Label-Efficient Self-Supervised Federated Learning for Tackling Data Heterogeneity in Medical Imaging. *arXiv [cs.CV]* (2022).
136. Shi, Y., Siddharth, N., Torr, P. H. S. & Kosiorek, A. R. Adversarial Masking for Self-Supervised

Learning. *arXiv [cs.CV]* (2022).

137. Li, G. *et al.* SemMAE: Semantic-Guided Masking for Learning Masked Autoencoders. *arXiv [cs.CV]* (2022).

138. Van Gansbeke & Vandenhende. Revisiting contrastive methods for unsupervised learning of visual representations. *Adv. Eng. Educ.*

139. Peng, Wang, Zhu & Wang. Crafting better contrastive views for siamese representation learning. *Proc. Estonian Acad. Sci. Biol. Ecol.*

140. Taleb, Kirchler & Monti. ContIG: Self-supervised Multimodal Contrastive Learning for Medical Imaging with Genetics. *Proc. IEEE*.

141. Zhang, Y., Jiang, H., Miura, Y., Manning, C. D. & Langlotz, C. P. Contrastive Learning of Medical Visual Representations from Paired Images and Text. *arXiv [cs.CV]* (2020).

142. Huang, S.-C., Shen, L., Lungren, M. P. & Yeung, S. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2021). doi:10.1109/iccv48922.2021.00391.

143. Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).

144. Moher, D. *et al.* Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst. Rev.* **4**, 1 (2015).

145. Danu, M., Ciuşdel, C. F. & Itu, L. M. Deep learning models based on automatic labeling with application in echocardiography. in *2020 24th International Conference on System Theory, Control and Computing (ICSTCC)* 373–378 (2020).