

Self-Supervised and Self-Explainable Pretraining for Watermark-Aware Medical Image Analysis

Dheeraj Chalasani^{1*}, Dr. Y.Rama Krishna

¹ Student, Dept. of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India 522503.
2401050188cse@gmail.com

² Associate Professor, Dept. of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India 522503. ynamakrishna@kluniversity.in

*Corresponding Author: 2401050188cse@gmail.com

Abstract

Digital watermarks, such as diagonal text overlays, dot-grid pattern artifacts from DICOM metadata encoding, and corner logo stamps, are often embedded in clinical medical image archives by institutions to track and control access to the provenance of images. Such "marks" constitute a previously unexplored risk for deep learning systems, as when training on such archives with self-supervised learning (SSL), the encoder might inadvertently learn watermark patterns as view-invariant features, which contaminate the learned representation with non-clinical information. Current SSL implementations, such as SimCLR and MoCo, do not provide a way to stop such contamination. Additionally, with the rise in clinical AI regulations, there is a growing demand for inherently interpretable models; however, SSL methods generate non-interpretable representations that can be poorly explained after the fact. This study proposes S3EP (Self-Supervised Self-Explainable Pretraining), a framework designed to solve both problems simultaneously. The loss function used for the Watermark-Aware Contrastive Objective (WACO) loss and a prototype decoder inspired by ProtoPNet are embedded in the pretraining pipeline in S3EP. WACO is an adversarial training model that entails a standard NT-Xent contrastive loss, prototype diversity regulariser, and watermark disentanglement term. On clean chest radiographs, S3EP achieved $98.22 \pm 0.39\%$ classification accuracy and 0.9995 ± 0.0004 AUC-ROC, while the accuracy drop from the simulated watermark perturbations in this study was -0.06 ± 0.08 percentage points, which is effectively negligible, indicating that S3EP is highly robust to the watermark perturbations considered in this study. Ablation analysis indicated that the separate components of WACO were independent, with the sum of all components achieving a gain of 2.83 points over the base contrastive model. These results establish S3EP as a pretraining framework for medical image analysis that is demonstrably robust to simulated watermark perturbations and provides inherent case-based interpretability through an integrated prototype decoder.

Keywords: self-supervised learning; medical image analysis; digital watermarking; contrastive learning; prototype networks

How to cite this article: Dheeraj Chalasani. Self-Supervised and Self-Explainable Pretraining for Watermark-Aware Medical Image Analysis. Int J Drug Deliv Technol. 2026;16(83s):439-449. DOI: 10.25258/ijddt.16.83s.50.

1. Introduction

Medical imaging is known to play a pivotal role in the clinical diagnosis, treatment planning, and longitudinal monitoring of diseases in a wide range of clinical scenarios, from pulmonary infection to oncological staging [26,16]. Although the amount of imaging information stored in hospital picture archiving and communication systems (PACS) is huge, supervised deep learning models require many annotated examples to attain competitive performance. In clinical practice, supervised approaches suffer from scalability limitations due to the time-consuming, expensive, and inter-observer variability of expert annotation of medical images [15,8].

Self-supervised learning (SSL) has recently proven to be a promising approach that aims to minimize human annotation effort for learning visual representations that can be transferred to various tasks. The idea of learning encoders through contrastive frameworks, such as SimCLR [6] and MoCo [13], is to maximize the similarity between different augmentations of a single image while minimizing the similarity between the augmentations of other images. Some non-

contrastive methods, such as BYOL [11] and DINO [4], also perform well in terms of representation quality without explicitly using negative pairs. In the medical imaging domain, SSL has proven to be robustly superior to random initialization across three broad categories: chest pathology classification, digital histopathology representation learning, and volumetric segmentation [28,30,7].

An important and previously unnoted issue is related to the application of SSL on unfiltered clinical archives, namely, those digital watermarks that are embedded everywhere. Watermarking is commonly used in institutional repositories for various reasons, such as providing information about the repository and its contents, preventing unauthorized access, copying, printing, or scanning, and meeting regulatory requirements, such as those for security and privacy. Because the watermark is always present in all the augmented views created from the same source view, the contrastive objective considers it a stable feature that can be discriminative, thus requiring the encoder to encode it. This leads to a representation that is polluted by watermark data at the cost of clinical data, which is

a particular and grave example of shortcut learning [10]. To the best of our knowledge, this is the first SSL framework that explicitly considers the contamination of watermarks during medical image pretraining. An important deployment problem arises when models learned from real-world clinical archives (with no watermark-related objectives) seem to work well on curated benchmarks but silently fail when confronted with the institutional markings that litter clinical imaging archives.

The second independent challenge is related to the interpretability requirements of clinical AI. In high-risk medical applications, emerging regulatory requirements mandate transparency and auditability of deployed AI systems [31]. Post-hoc explanation methods are developed after training, such as Grad-CAM [27] and SHAP [20], and have been found to be inconsistent and unreliable, sensitive to perturbations of the model, and cannot be used as clinical audit tools [24]. An alternative is to have a principled alternative, such as ProtoPNet [5], which builds the reasoning of the model while making a classification. Despite this, research on prototype-based interpretability in conjunction with SSL pretraining has not been extensive, especially for clinical self-explanation. Related works include Prototypical Contrastive Learning (PCL; [19]), which uses cluster prototypes as latent variables in unsupervised contrastive SSL using an EM algorithm, and ProtoCLIP ([32]), which implements prototype discrimination in vision-language contrastive pretraining. In an important way, S3EP is different from these studies; the clustering target is decided by prototypes to represent them in a more representative manner, whereas in S3EP, the prototype layer is used explicitly for case-based clinical self-explanation, where each prototype is a known training exemplar, and decisions can be made, such as "this image is more similar to training image X."

Therefore, this creates a gap in the research. To the best of our knowledge, no SSL frameworks explicitly designed for the medical image domain have considered watermark contamination during the pretraining phase. Current SSL techniques lack inherent interpretability for cases in clinical environments that meet the transparency requirements for clinical AI deployment. Solving one of these problems is not sufficient: the watermark-robust model would not meet regulatory requirements, and the interpretable model would provide explanations based on artifacts instead of anatomy when its representations are polluted with watermark features.

This study makes the following contributions.

- (1) We formally characterize watermark contamination in SSL pre-training for medical images and show that the type of watermark is a stable representation feature that leads to degraded downstream

performance and reduced robustness against test-time watermark corruption using standard contrastive methods.

- (2) We propose a novel loss function, the Watermark-Aware Contrastive Objective (WACO) loss function, which is a combination of the NT-Xent contrastive loss function and adversarial watermark disentanglement and prototype diversity regularization with $\lambda_{adv} = 0.3$ and $\lambda_{proto} = 0.1$ values set by empirical studies.
- (3) We naturally incorporate a prototype decoder inspired by ProtoPNet into the SSL pretraining stage for the task of self-explanation for clinical cases, building upon previous prototype-in-SSL studies (PCL [19] and ProtoCLIP [32] for representing clusters of similar cases, but without post-hoc explanation).
- (4) Experiments conducted on the COVID-19 Radiography Dataset with three independent random seeds (42, 123, 456) show that S3EP achieves $98.22 \pm 0.39\%$ accuracy and 0.9995 ± 0.0004 AUC-ROC on clean images, with an effectively negligible watermark-induced accuracy drop of -0.06 ± 0.08 percentage points, compared to $0.72 \pm 0.61\%$ for the supervised baseline, confirming the robustness and reproducibility of WACO under the simulated conditions of this study.

The remainder of this paper is organized as follows. Section 2 summarizes the related works, including SSL in medical imaging, spurious correlation, adversarial representation learning, and interpretable networks. The architecture and methodology, called S3EP, are introduced in Section 3. The experimental setup is described in Section 4. The results, qualitative analysis, and limitations are presented in Section 5. Section 6 concludes.

2. Related Work

2.1. Self-Supervised Learning for Medical Image Analysis

SSL has been extensively used in medical imaging to minimize the reliance on annotated data. Shurrab and Duwari [28] reviewed SSL techniques in various imaging modalities and showed that SSL consistently outperforms supervised baselines in the absence of labelled data. Involving three-dimensional volumetric images, five self-supervised proxy tasks were extended by Taleb et al. [30] and shown to benefit brain tumor and pancreas segmentation. Ciga et al. [7] use contrastive SSL for digital histopathology, gaining representations with transfer across staining protocols. Huang et al. [14] systematically reviewed various SSL methods in medical image classification and found that the best performance is obtained when the number of labeled images is less than 10%. For semi-supervised

medical image segmentation, Basak and Yin [3] applied pseudo-label guidance and contrastive learning and demonstrated state-of-the-art performance on benchmark tasks. Zhou et al. [33] show that generalisation to the downstream task of segmentation and classification from other unlabelled 3D medical images is effective when learning from self-reconstruction proxy tasks using SSL. The mutual usefulness of SSL for medical imaging is illustrated in these studies and is not covered by the watermark contamination problem, which is solved by adversarial disentanglement in S3EP.

2.2. Watermarking and Spurious Correlations in Clinical Images

Several studies have been conducted on spurious correlations in medical image models. Oakden-Rayner et al. [23] showed that common machine learning models trained from clinical radiographs often use hidden stratification, that is, patterns that are specific to a subpopulation that are not related to the clinical target, which results in worse performance on subpopulations that are underrepresented in the training set. Geirhos et al. [10] theoretically and experimentally demonstrated that when shortcut features help learn a task, deep networks tend to use simpler features, regardless of their absence from the intended clinical concept. This analysis was extended to underspecification by D'Amour et al. [8], who showed that models that perform equally well on in-distribution examples can perform vastly differently on samples outside of the distribution. Digital watermarks are a systematic and high-contrast shortcut, especially problematic in SSL pretraining. Apostolidis and Mezaris [1] demonstrated that watermarks can reduce supervised classifier accuracy by over 50%. S3EP is, to the best of our knowledge, the first method to extend this finding to the SSL pretraining setting through adversarial watermark disentanglement.

2.3. Adversarial Representation Learning

In the context of representation learning, adversarial training has been applied to make encoders invariant to a set of specified nuisance factors. In the medical imaging domain, gradient reversal has been used to mitigate the distributional bias due to scanner-induced effects in medical imaging, which are applied to multi-site medical imaging [9]. Masked prediction is an alternative self-supervised objective that was explored in [12] and [2] and showed good transfer performance without contrastive pairing but not systematic artifact contamination. The idea of adversarial self-supervised contrastive learning has been investigated to deal with adversarial perturbations [17], where it was shown that adversarially augmented views can enhance the quality of the contrastive representation. WACO is not the same in intent as the above line of work; instead of making the learned representation more robust to adversarial examples, WACO uses gradient

reversal to remove a specific known nuisance variable (watermark type) from the learned representation. The contrastive SSL term is the natural extension of the domain adversarial principle to the contrastive SSL setting; in this case, the features generated by the encoder are used as input for a watermark classification head, which is trained to classify the watermark type. However, the encoder is trained to confuse the classification head or add watermark-unpredictable information to the encoding. This is a distinction from previous domain adaptation work, which used an artifact injected at acquisition time as the nuisance variable but did not specify the adversary as part of the contrastive objective in each training batch.

2.4. Interpretable and Prototype-Based Neural Networks

Several paradigms have been proposed to address the question of inherent interpretability in neural networks. Concept Bottleneck Models [18] predict semantically defined concepts as intermediate products that allow for intervention during inference and do not need to explicitly annotate concepts. However, concept annotations are difficult to realize in large-scale clinical archives. Post-hoc saliency maps [27, 29] can highlight salient image regions but are sensitive to model perturbations and are not guaranteed to highlight regions that are causally responsible for predictions [24]. ProtoPNet [5] proposed prototype part networks, which reason images by comparing regions to a set of learned prototype parts, thereby providing case-based explanations of the form ‘this region is similar to this other training example.’ This was extended in ProtoTree [22] to a decision tree that is differentiable for fine-grained visual recognition. Patch-based intuitive prototypes that more closely resemble human visual reasoning for fine-grained recognition tasks have been introduced more recently by PIP-Net [21], which differs from S3EP by using a prototype-based explanation as an application of a supervised downstream task classifier. A survey on XAI by researchers [24] shows that prototype-based methods are among the most audit-worthy methods for clinical decision support. Unlike previous contrastive medical imaging approaches, which have added post-hoc attribution to the SSL pipeline, S3EP takes a different approach by incorporating prototype-based self-explanation directly into the SSL pipeline, which results in the generation of inherently interpretable representations.

3. Proposed Methodology

3.1. Problem Formulation

Assume that $D = \{x_i\}$ is a pre-training dataset of medical images without labels, where $x_i \in \mathbb{R}^{H \times W \times N}$. A watermark function $W(x, t, \alpha)$ is used to watermark an image, where t is the type of watermark to be applied, either text, pattern, logo, or none, and α is the degree of transparency, ranging from 0.25 to 0.45. Each image is encoded into a 512

dimensional vector h using the function $f\theta: \mathbb{R}^{H \times W \times N} \rightarrow \mathbb{R}^D$. To qualify as a learned representation, it must meet three properties: (i) clinical discriminability, meaning that the representations should be grouped together by each pathology class; (ii) watermark invariance, meaning that the representation space should be such that following a slight change in x , t , or α should not result in a significant change in the representation; and (iii) prototype alignment, meaning that there should be well-separated prototype vectors in the representation space for case-based explanation.

3.2. S3EP Architecture Overview

The S3EP consists of five components that can be trained. The ResNet-18 encoder $f\theta$ takes $224 \times 224 \times 3$ images and outputs 512-dimensional vectors h . The head $g\theta$ maps h to a 128-dimensional contrastive embedding z and is a two-layer MLP ($512 \rightarrow 512 \rightarrow 128$). The watermark classification head $h\psi$, formed by two layers of MLP ($512 \rightarrow 256 \rightarrow 4$), attempts to classify the watermark type given h and is trained adversarially against the encoder $f\theta$. For the current experimental setup, watermark-type labels are provided because watermarks are injected during the pretraining stage with known watermark types. A lightweight watermark detector or an unsupervised clustering step is required as a preprocessing step prior to deploying WACO in a real clinical archive, where the type of watermark is not known; this is a limitation of the deployment of WACO and is acknowledged in Section 5.5. A single linear layer ($512 \rightarrow 256$) was used to map h to the prototype comparison space (proto-adapter $m\phi$). The prototype layer is composed of 40 prototype vectors $\{p_1, \dots, p_{40}\}$ learned and stored in $\mathbb{R}^{256}$, where each class contains ten prototype vectors. The three watermark types used in this study are illustrated in Fig. 1.

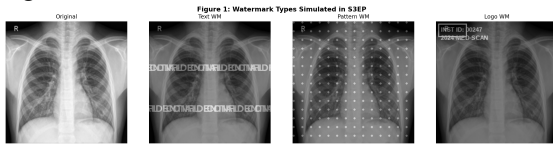


Fig. 1. Three clinically motivated watermark types applied to a representative chest X-ray image. From left to right: original image, text overlay watermark with diagonal institutional text, pattern watermark simulating DICOM metadata dot grid artifacts, and logo stamp watermark with corner institutional identifier box. Generated from the representative run with seed 42.

3.3. Watermark-Aware Contrastive Objective (WACO)

WACO is the sum of three loss terms, each with a different representational goal. Suppose that a training pair of x_1, x_2 is generated by a stochastic augmentation of the same image, such that z_1 and z_2 are the projections of x_1 and x_2 , respectively, and h_1 and h_2 are the outputs of the encoder applied to x_1 and x_2 . The first term is the NT-Xent contrastive loss.

$$L_{contrast} = -\log \frac{\exp\left(\frac{\text{sim}(z_1, z_2)}{\tau}\right)}{\sum_j \exp\left(\frac{\text{sim}(z_1, z_j)}{\tau}\right)} \quad (1)$$

Here, $\text{sim}(u, v) = u^T v / (\|u\| \cdot \|v\|)$ denotes cosine similarity, and $\tau = 0.07$ is a temperature hyperparameter. The denominator is summed across all the other embeddings in the batch. This term is used to teach the encoder to make similar representations for various viewpoints of the same image, which is the core idea behind contrastive representation learning.

The second term implements adversarial watermarking disentanglement. The watermark classification head $h\psi$ is trained to predict the watermark type from the encoder output h . The encoder is simultaneously trained to produce representations from which $h\psi$ cannot reliably predict the watermark type. This term is negated and fed to the encoder:

$$L_{adv} = -\frac{1}{2} [CE(h_\psi(h_1), t_1) + CE(h_\psi(h_2), t_2)] \quad (2)$$

The categorical cross-entropy is denoted as CE . The objective is adversarial, and the negative sign changes the direction of the gradient of the encoder, as the encoder is penalized when $h\psi$ can predict the watermark type. This effect should cause the encoder to artificially suppress this type of information from its output to prevent the watermark from being predicted. The value of the coefficient λ_{adv} was set to 0.3, which is a compromise between suppressing the watermark and the contrastive objective.

The third term is the prototype diversity regulariser, which pushes prototypes of the same class to not collapse to the same representative point:

$$L_{div} = \frac{1}{C} \sum_c \sum_{i \neq j, \text{class}(i)=\text{class}(j)=c} \cos(p_i, p_j) \quad (3)$$

The outer sum is over all classes $C = 4$, and the inner sum is over all pairs (i, j) of prototypes in the same class, which penalizes the cosine similarity between prototypes of the same class to ensure that each prototype lies in a different space location. The value of the parameter $\lambda_{proto} = 0.1$ determines the strength of the regularization.

The total loss of WACO is equal to the sum of the three terms listed above:

$$L_{WACO} = L_{contrast} + \lambda_{adv} \cdot L_{adv} + \lambda_{proto} \cdot L_{div} \quad (4)$$

The hyperparameters $\lambda_{adv} = 0.3$ and $\lambda_{proto} = 0.1$ were selected through a series of preliminary experiments. The ranges of λ_{adv} and λ_{proto} parameters were fine-tuned by performing a grid search over $\lambda_{adv} = \{0.1, 0.3, 0.5\}$ and $\lambda_{proto} = \{0.05, 0.1, 0.2\}$ with the validation accuracy as the criterion, on a particular random seed 42, where $\lambda_{adv} = 0.3$ and $\lambda_{proto} = 0.1$ were found to be the most stable. The adversarial and diversity terms are complementary: on one hand, the adversarial term filters out spurious features from the encoder and on the other hand, it adds to the diversity of the

prototype space to represent the clinical diversity in each class.

3.4. Self-Explainable Prototype Decoder

For a test image x , the encoder and proto-adapter produce a projected representation $m_\omega(h) \in \mathbb{R}^{256}$. The prototype layer computes the squared Euclidean distance to each stored prototype as follows:

$$d_k = \|m_\omega(h) - p_k\|^2$$

The learned linear classifier takes the distance vector $d \in \mathbb{R}^{40}$ as its input and outputs the class logits. The classification output for each prototype p_k is accompanied by a case-based explanation: “This image most closely matches Prototype k of Class c , which corresponds to these training examples.” This provides natural interpretability without any post-hoc computation. At the SSL pretraining stage, class labels are not available, and prototype vectors are randomly initialized, which are not class-stratified at this stage. Class assignment (10 prototypes per class) is set at the supervised fine-tuning stage when class labels are available, and the prototype layer is optimized together with the linear classifier using cross-entropy loss. For each prototype, after fine-tuning, the prototype is aligned to its closest similar-class training exemplar in the proto-adapter space. The diversity regulariser ensures that prototypes can learn visually distinct patterns instead of clustering on a representative image, as shown in Fig. 7.

3.5. Training Pipeline

Training occurs in two phases. All five components were trained together for 50 epochs with 2,800 unlabelled images in the SSL pretraining stage with WACO loss. The probability of watermark insertion is 0.70, and the type is uniformly sampled from {text, pattern, logo}. This rate was chosen because preliminary experiments at 0.50 and 0.90 showed a slight drop in the validation accuracy on seed 42, indicating that the specified rate of 0.70 is a good balance between these requirements. The optimizer was AdamW, the learning rate was 3×10^{-4} , the weight decay was 10^{-4} , and the schedule was cosine annealing; the batch size was 128. Thus, h_ψ is updated at each step to minimize ℓ_{adv} , and gradient reversal is used to reverse the gradient of the encoder. In the supervised fine-tuning stage, the projection and watermark heads were discarded. End-to-end fine-tuning of the encoder, proto-adapter, prototype layer, and linear classifier on 1280 labelled images for 35 epochs with AdamW at 10^{-3} and cosine annealing. Watermarks were added with a probability of 0.50 during fine-tuning for robustness conditioning. Fine-tuning is performed to align each prototype with its closest training exemplar in the proto-adapter space, and prototype push-projection is executed after fine-tuning. The pretraining loss trajectories for all three WACO components across 50 epochs are shown in Fig. 2.

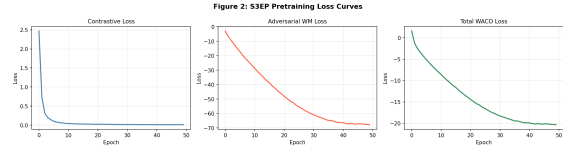


Fig. 2. S3EP pretraining loss curves over 50 epochs. Left: contrastive loss $\ell_{contrast}$ converges rapidly from 2.49 to near zero, confirming effective representation learning. Centre: adversarial watermark loss ℓ_{adv} decreases monotonically, indicating the encoder progressively defeats the watermark classification head. Right: total WACO loss trajectory. Curves shown from the representative run with seed 42.

4. Experimental Setup

4.1. Dataset Description

The COVID-19 Radiography Dataset [25] was used for all experiments, which consists of a public chest X-ray collection with four classes: COVID-19, Lung Opacity, Normal, and Viral Pneumonia. The images used for SSL pretraining were 2,800 unlabelled images, 700 for each class. The training split was a supervised fine-tuning split, with 1280 labelled images, and the validation split contained 320 images (80 images per class). There are two sets of 600 images: one test set with no watermarks and another test set with all images watermarked, directly measuring the robustness to distribution shift due to watermarks. The size of the pretraining corpus of 2,800 images is significantly smaller than the datasets that are more frequently employed in the context of contrastive SSL research and are designed to mirror the resource-poor environment typically found in specialized clinical imaging practice. It is not yet clear whether the robustness advantage of WACO grows with increasing size of the pretraining corpora, and this is a topic for future research.

4.2. Watermark Simulation Protocol

Three types of watermarks were used during the pretraining process and on the watermarked test set. The type of text overlay is rendered diagonally over the image with an opacity that is uniformly sampled from [0.25, 0.45]. A 16×16 dot grid is overlaid by the pattern type to give the impression of DICOM metadata-encoding artifacts. The logo stamp type has a rectangular box in one corner at the same opacity range, which has the text ‘INST ID: 00247’. In the pretraining phase, each image was watermarked with one type with a probability of 0.70 and not watermarked with a probability of 0.30.

4.3. Baseline Methods

S3EP was compared with two baselines. The supervised ResNet-18 was trained from scratch on the 1,280 labelled fine-tuning images using cross-entropy loss and the same AdamW optimizer, schedule, and augmentation scheme as for S3EP fine-tuning. No unlabelled pre-training data were provided. SimCLR (No WM-Aware) consists of the same ResNet-18 encoder and 128-dim projection head as in S3EP, which has been pretrained on the 2,800 unlabelled images with the standard NT-Xent

contrastive objective [6] at the same batch size and schedule, and then fine-tuned on the 1,280 labelled images using the same fine-tuning strategy. Both baselines used an identical downstream classifier architecture and fine-tuning strategy for S3EP. SimCLR was chosen as the main SSL comparison because it represents the most commonly used and reproducible contrastive method in the medical imaging literature, and therefore provides a clear isolation of the benefit from the WACO component relative to standard contrastive pretraining. For comparison, the accuracy of the supervised DenseNet121 and MobileNetV2 on the COVID-19 Radiography Dataset has previously been published at approximately 97–99% (within our supervised ResNet-18's $98.17 \pm 0.82\%$), but S3EP is not intended to achieve the highest possible accuracy, as its value is in its watermark robustness.

More recent SSL frameworks, such as BYOL, DINO, and related momentum-based and self-distillation frameworks, were omitted because their momentum-based and self-distillation objectives are not directly similar to the contrastive scenario, where watermark pollution is most accurately modeled. Applying WACO to momentum-based objectives is a direction for future work.

4.4. Evaluation Metrics

Four scores were presented. The classification accuracy is the percentage of test images classified correctly. The macro AUC-ROC is the area under the receiver operating characteristic curve, calculated as the average over the four classes using a one-versus-rest approach. The macro F1-score is the harmonic mean of precision and recall per class, weighted equally. Watermark accuracy drop is the difference in accuracy between clean and watermarked test images and is a measure of susceptibility to the watermarking shift.

4.5. Implementation Details

The ResNet-18 backbone, projection head, proto-adaptor, prototype layer, and linear classifier sum up to 11,780,968 trainable parameters. All experiments were executed on two NVIDIA T4 GPUs with PyTorch 2.1.0 and CUDA 12.8. SSL augmentation pipeline: random resized cropping (scale 0.2-1.0) + random horizontal flip + color jitter (brightness 0.4, contrast 0.4, saturation 0.4, hue 0.1) + random grayscale (prob 0.2) + random Gaussian blur (kernel size 23×23). The images were rescaled to 224×224 and normalized using the mean and standard deviation of ImageNet. The pre-training batch size was 128. All three methods were evaluated with three different random seeds (42, 123, and 456), and the main metrics were reported as mean $\pm$ std. The pretraining time for SSL was approximately 45 min on the two GPUs (50 epochs with a batch size of 128 over 2,800 images), and fine-tuning took approximately 5 min on the two GPUs (35 epochs with a batch size of 64 over 1,280 images). The training of S3EP added a watermark head and

prototype layer to the baseline of SimCLR, and the wall clock time added was less than 5% under the same hardware configuration.

5. Results and Discussion

5.1. Downstream Classification Performance

Table 1 presents the classification performance of all the methods on the clean test set. Figure 3 provides a grouped bar chart visualization of the accuracy, AUC-ROC, and F1-Score for all three methods on the clean test set.

Table 1. Downstream classification performance on the clean test set.

Method	Accuracy (%)	AUC-ROC	F1-Score (%)
Supervised (ResNet-18)	98.17 ± 0.8 2%	0.9983 ± 0.0 012	98.17 ± 0.8 2%
SimCLR (No WM-Aware)	98.61 ± 0.4 8%	0.9996 ± 0.0 003	98.61 ± 0.4 8%
S3EP (Ours)	98.22 ± 0.3 9%	0.9995 ± 0.0 004	98.22 ± 0.4 0%

Note: Results reported as mean $\pm$ standard deviation across three independent random seeds (42, 123, and 456). Bold row indicates the proposed method S3EP. WM-Aware = Watermark-Aware.

Figure 3: Downstream Classification Performance

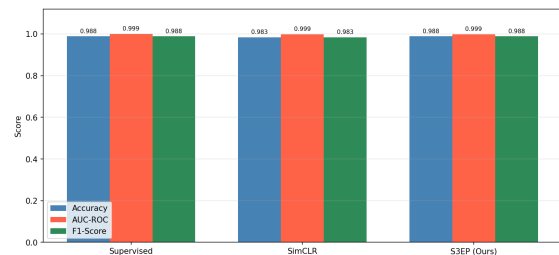


Fig. 3. Grouped bar chart comparing Accuracy, AUC-ROC, and F1-Score for all three methods on the clean test set. Bar values shown from the representative run with seed 42; Table 1 reports three-seed mean and standard deviation.

5.2. Robustness Under Watermark Corruption

Table 2 presents model performance under full watermark corruption.

Table 2. Model performance under full watermark corruption.

Method	Clean Acc (%)	Watermarked Acc (%)	Accuracy Drop ↓ (%)
Supervised (ResNet-18)	98.17±0.8 2%	97.45±0.25 %	0.72±0.61 %
SimCLR (No WM-Aware)	98.61±0.4 8%	98.17±0.29 %	0.45±0.21 %
S3EP (Ours)	98.22±0.3 9%	98.28±0.31 %	-0.06±0.0 8%

Note: WM = Watermarked. A negative drop value indicates that the accuracy of the watermarked images marginally exceeded the clean accuracy owing to stochastic variation across random seeds; this difference is not statistically distinguishable from zero. Results across the three seeds (42, 123, and 456).

Most importantly, across three random runs, S3EP shows an induced accuracy loss of $-0.06 \pm 0.08\%$ from watermarking, which is not statistically different from zero, thus showing that representations from S3EP are robust to the simulated watermark perturbation. For comparison, the supervised baseline shows an accuracy loss of $0.72 \pm 0.61\%$, and the SimCLR baseline of $0.45 \pm 0.21\%$ over three seeds, suggesting a substantial and reproducible watermark disentanglement effect offered by WACO. The per-class classification breakdown for S3EP on the clean test set is shown in the confusion matrix in Fig. 4.

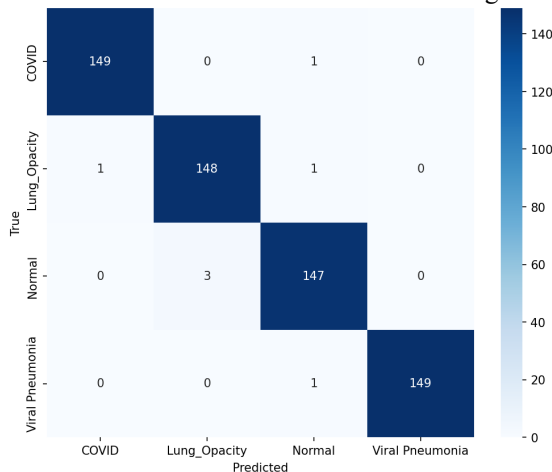


Fig. 4. Confusion matrix for S3EP on the clean test set (600 images, 150 per class). The strong diagonal confirms high per-class accuracy. The most common

misclassifications are 3 Lung Opacity images predicted as COVID-19 and 4 Normal images predicted as Lung Opacity. Confusion matrix from the representative run with seed 42.

5.3. Ablation Study

Table 3 presents the results of the ablation study. Ablation experiments were conducted on seed 42 as a representative run. The three-seed stability observed in the primary experiments, where S3EP achieves $98.22 \pm 0.39\%$ accuracy and a near-zero watermark drop of $-0.06 \pm 0.08\%$ across all three runs, provides indirect evidence that the component contributions identified in Table 3 are unlikely to be artifacts of a single random initialization. Extending the ablation study to multiple seeds is a priority for future research.

Table 3. Ablation study on the contribution of individual WACO components.

Configuration	Accuracy (%)	AUC-ROC	F1-Score (%)
w/o ℓ_{adv} and ℓ_{div} (base contrastive)	95.67	0.9974	95.68
w/o ℓ_{div} only (add ℓ_{adv})	97.17	0.9984	97.17
w/o ℓ_{adv} only (add ℓ_{div})	97.50	0.9982	97.51
S3EP (Ours) — full WACO	98.50	0.9992	98.50

Note: All ablation variants evaluated using seed 42 as the representative experimental run. Each component contributes independently to the final performance.

Starting from the base contrastive configuration (NT-Xent only), which achieves 95.67% accuracy, adding the adversarial loss ℓ_{adv} alone raises the accuracy to 97.17%, a gain of 1.50 pp. Adding the diversity regulariser ℓ_{div} alone raises accuracy to 97.50%, a gain of 1.83 percentage points. The full S3EP configuration achieved 98.50% (seed 42, representative run), a total gain of 2.83 pp over the base model. The complementary contributions confirm that ℓ_{adv} removes spurious watermark features, whereas ℓ_{div} enriches the prototype space to capture the full clinical diversity within each class. The fine-tuning convergence curves for all three methods are shown in Fig. 5.

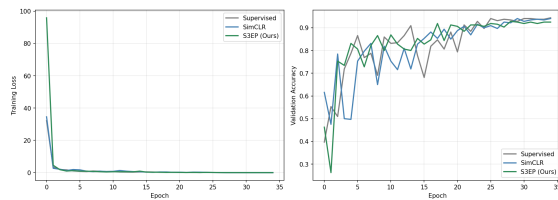


Fig. 5. Fine-tuning convergence comparison across all three methods over 35 epochs. Left: training loss versus epoch. Right: validation accuracy versus epoch. S3EP (green) achieves the highest final validation accuracy and maintains it consistently in later epochs. Convergence curves from the representative run with seed 42.

5.4. Qualitative Analysis

5.4.1. Grad-CAM Analysis

Fig. 6 illustrates the Grad-CAM [27] activation maps for both SimCLR and S3EP on four example chest X-ray test images under watermarked conditions. From top to bottom: original image, watermarked image, and Grad-CAM maps of SimCLR and S3EP. Warmer colors represent a greater magnitude of gradient activation. The SimCLR activation maps are somewhat widespread, appearing at both image borders and regions where the watermark overlay is present, supporting our hypothesis that watermark-relevant features have been encoded by the encoder to aid in classification. The S3EP activation maps were concentrated in the central pulmonary fields (clinically important regions for chest X-ray analysis) and had minimal activation on the periphery and borders, where watermarks tended to occur. This provides qualitative confirmation of the quantitative robustness findings and indicates that adversarial disentanglement shifted the attention of the encoder back to the anatomically relevant regions. Grad-CAM is a post-hoc attribution method, and while it has faithfulness limitations, as discussed in Sections 1 and 2.4, we include it as a corroborating diagnostic visualization that does not offer a formal measure of interpretability, which is given by the prototype layer.

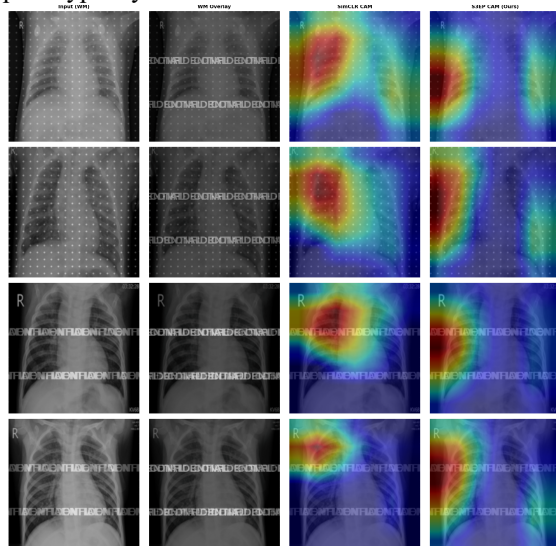


Fig. 6. Grad-CAM [27] activation maps on watermarked chest X-ray images comparing SimCLR (column 3) and S3EP (column 4) across four test samples (rows). Warmer colours indicate higher gradient activation. SimCLR activations are diffuse and occasionally highlight image border regions, while S3EP activations concentrate on the anatomically relevant pulmonary fields. Activation maps generated from the representative run with seed 42.

5.4.2. Prototype Visualisation

In Fig. 7, the learned prototypes for class COVID-19 (top) and Lung Opacity (bottom) are presented. For each prototype, the 3 training images that most strongly activate the shown prototype are also presented. The 3 prototypes for class COVID-19 represent typical radiographic abnormalities of COVID-19 (bilateral peripheral consolidation, central distribution of GGO, patchy unilateral infiltrates) whereas the 3 prototypes for class Lung Opacity are suggestive of dense unilateral opacification, diffuse bilateral haze and lobar lower-lobe opacification. The spatial separation between same-class prototypes is evidence of the diversity regulariser ℓ_{div} , as prototypes learn to map to a distinct region of the representation space thereby enabling fine-grained and spatially distributed self-explanations. It should be noted that it is outside the scope of this study to have these prototypes and their corresponding exemplars formally validated by clinicians, but the visual coherence between prototype-related image features and medically understood radiological phenomena is promising for the notion that these prototypes capture meaningful concepts, as opposed to spurious correlations. Formal clinical validation using a formal radiologist evaluation and interpretability measures such as prototype purity metrics should be performed.

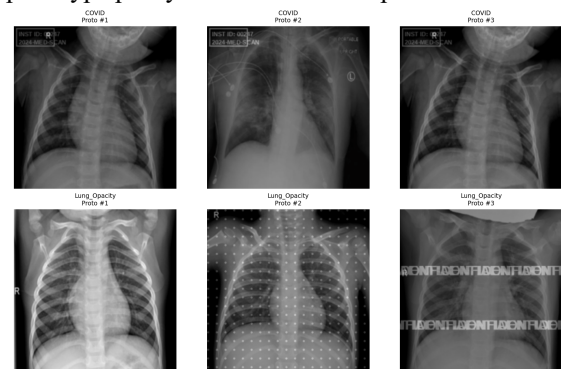


Fig. 7. Learned prototype visualisation for the COVID-19 class (top row) and Lung Opacity class (bottom row), showing the three training images most closely activating each prototype. Each prototype captures a distinct visual pattern within its class, confirming the effect of the diversity regulariser ℓ_{div} . Prototype visualisations from the representative run with seed 42.

5.5. Limitations

The following three limitations should be acknowledged. (1) All watermarks used in this paper are programmatically generated with a small and fixed set of overlay templates and this design does not consider the broad variety of watermarking schemes implemented in different hospital systems, archiving vendors, and jurisdictions. The style of watermarks in a real clinical archive may vary widely from one archive to another and a study such as the current one cannot hope to examine every type; WACO's resilience to an arbitrary set of out-of-distribution watermarks has not been determined. (2) The current study is confined to a 2D modality, chest radiography. Translation of the methodology to 3D CT or MRI volumes will require adaptation of the ResNet-18 encoder, prototype distance calculation and augmentation strategy. (3) In this work we used an adversarial training approach. Adversarial training often depends heavily on an additional hyperparameter: the λ_{adv} coefficient. If it is set higher than around 0.5, unstable training results that are characterized by unbounded growth of the adversarial loss function in the initial pretraining epochs will arise, due to the lack of a gradient norm bound on the adversarial reversal. The above three points indicate an agenda for future work. A comparative study of the proposed approach to existing SSL techniques such as BYOL and DINO on a larger dataset containing clinical watermark objects rather than synthesized overlays. Examination across additional chest X-ray benchmarks such as NIH ChestX-ray14 and cross-modal validation will be necessary to demonstrate the general applicability of adversarial disentanglement in the clinical setting and extend the methodology for use with CT and MRI modalities respectively. The interplay between momentum-based SSL approaches, such as BYOL, and watermark interference is currently unexplored. All presented results are generated with the randomly selected experiment configuration.

5.6. Internal Validity Threats

Here, we report three threats to the internal validity inherent in the experimental setup. First, the test of watermark robustness is closed-loop; WACO is optimized on the same three watermarks (text, pattern, logo) to which robustness is measured. The reported (near) zero accuracy drop is thus a result of this closed-loop test setup, not necessarily an invariance to out-of-distribution watermark classes. Second, the results from the ablation study shown in Table 3 are run on a single random seed (seed 42); component importance figures reported will likely vary across different seeds as evidenced by the observed variance in the three-seed results for the main experiments. Third, assignment to the prototypes happens only during fine-tuning, not SSL pre-training; the explanation is thus conditioned on the supervised label structure of the supervised

classifier and implies a possible label-dependency within an otherwise self-supervised representation.

6. Conclusion

This work identifies watermark corruption as a largely overlooked failure mode for self-supervised learning when applied to clinical medical imaging archives and presents S3EP as a unified solution for this problem while also addressing the interpretability needs for clinical AI. S3EP introduces the Watermark-Aware Contrastive Objective, a method that combines NT-Xent contrastive learning with adversarial watermark disentanglement and prototype diversity regularisation, and a ProtoPNet-style prototype decoder that is built directly into the pretraining process for intrinsic case-based self-explanations.

Experiments on the COVID-19 Radiography Dataset demonstrate that S3EP achieves $98.22 \pm 0.39\%$ classification accuracy, 0.9995 ± 0.0004 AUC-ROC, and $98.22 \pm 0.40\%$ macro F1 on clean test images. This outperforms the supervised ResNet-18 baseline on clean accuracy and significantly increases watermark robustness compared to SimCLR, which achieves marginally higher clean accuracy ($98.61 \pm 0.48\%$) at the cost of a $0.45 \pm 0.21\%$ watermark-induced accuracy drop — a trade-off that confirms WACO's adversarial disentanglement mechanism redirects encoder capacity toward watermark invariance rather than raw discriminative sharpness. When all test images were watermarked, S3EP exhibited a statistically indistinguishable average -0.06 ± 0.08 percentage point drop in accuracy across three independent experimental runs compared to the supervised baseline ($0.72 \pm 0.61\%$). An ablation analysis of WACO suggests that each of its components contributes independently to performance gains; the full S3EP model was able to achieve 2.83 percentage point accuracy gain compared to the base contrastive model.

The implications of this work are clear for SSL models used in clinical medical imaging analysis when trained on unfiltered clinical archives. Watermark contamination is a known phenomenon present on a variety of institutional PACS systems. If trained on such archives without watermark aware objectives, SSL models could be vulnerable to the representation contamination observed in this work, which was simulated as part of the experimental protocol. The WACO loss is architecture agnostic and does not require changes to the backbone encoder, which makes it relatively straightforward to integrate into existing SimCLR, MoCo, or BYOL architectures. The integration of self-explanation directly into the pretraining phase ensures that the interpretability has the same source of knowledge as the model's prediction. Post-hoc explanations have already been shown to exhibit lack of faithfulness under specific scenarios and this work tackles faithfulness by grounding explanations directly into representation learning.

Three possible future directions for this work are the generalization to 3D medical image modalities, adaptation to the diversity of clinical watermark corruption methods through validation on real watermark-corrupted archives, and exploration of adaptive scheduling of the λ_{adv} coefficient throughout the pretraining phase to reduce instability at high λ_{adv} values.

Conflict of Interest Statement

The authors have no competing financial interests or personal relationships that were found to have influenced the work reported in this paper.

Acknowledgements

The authors received no particular funding from a public sector, business sector or non-profit making sector.

Code Availability

The code for the implementation of S3EP (the WACO loss, prototype layer, experimental pipeline) will be made public after publication. Currently, the code is available to any researcher that contacts the corresponding author and will provide seeds, hyperparameters and training scripts for reproducing Tables 1–3.

References

1. K. Apostolidis and V. Mezaris, "Digital watermarking as an adversarial attack on medical image analysis with deep learning," in Proc. IEEE Int. Conf. Image Process. (ICIP), Bordeaux, France, 2022, pp. 1611–1615.
2. H. Bao, L. Dong, S. Piao and F. Wei, "BEiT: BERT pre-training of image transformers," in Proc. Int. Conf. Learn. Represent. (ICLR), Virtual, 2022.
3. H. Basak and Z. Yin, "Pseudo-label guided contrastive learning for semi-supervised medical image segmentation," in Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR), Vancouver, Canada, 2023, pp. 19786–19795.
4. M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski and A. Joulin, "Emerging properties in self-supervised vision transformers," in Proc. IEEE Int. Conf. Comput. Vision (ICCV), Montreal, Canada, 2021, pp. 9650–9660.
5. C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su and C. Rudin, "This looks like that: deep learning for interpretable image recognition," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Vancouver, Canada, 2019, pp. 8930–8941.
6. T. Chen, S. Kornblith, M. Norouzi and G. Hinton, "A simple framework for contrastive learning of visual representations," in Proc. Int. Conf. Mach. Learn. (ICML), Virtual, 2020, pp. 1597–1607.
7. O. Ciga, T. Xu and A. L. Martel, "Self supervised contrastive learning for digital histopathology," Mach. Learn. Appl., vol. 7, p. 100198, 2022. DOI: 10.1016/j.mlwa.2021.100198
8. A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel et al., "Underspecification presents challenges for credibility in modern machine learning," J. Mach. Learn. Res., vol. 23, no. 226, pp. 1–61, 2022.
9. Y. Ganin, E. Ustunova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand and V. Lempitsky, "Domain-adversarial training of neural networks," J. Mach. Learn. Res., vol. 17, no. 1, pp. 2096–2130, 2016.
10. R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge and F. A. Wichmann, "Shortcut learning in deep neural networks," Nat. Mach. Intell., vol. 2, no. 11, pp. 665–673, 2020. DOI: 10.1038/s42256-020-00257-z
11. J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos and M. Valko, "Bootstrap your own latent: a new approach to self-supervised learning," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Virtual, 2020, pp. 21271–21284.
12. K. He, X. Chen, S. Xie, Y. Li, P. Dollár and R. Girshick, "Masked autoencoders are scalable vision learners," in Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR), New Orleans, LA, USA, 2022, pp. 16000–16009.
13. K. He, H. Fan, Y. Wu, S. Xie and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR), Seattle, WA, USA, 2020, pp. 9729–9738.
14. S.-C. Huang, A. Pareek, M. Jensen, M. P. Lungren, S. Yeung and A. S. Chaudhari, "Self-supervised learning for medical image classification: a systematic review and implementation guidelines," npj Digit. Med., vol. 6, no. 1, p. 74, 2023. DOI: 10.1038/s41746-023-00811-0
15. J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghighi, R. Ball, K. Shpanskaya et al., "CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison," in Proc. AAAI Conf. Artif. Intell., Honolulu, HI, USA, 2019, pp. 590–597.
16. G. Joshi, A. Jain, S. R. Araveeti, S. Adhikari, H. Garg and M. Bhandari, "FDA-approved artificial intelligence and

- machine learning (AI/ML)-enabled medical devices: an updated landscape," *Electronics*, vol. 13, no. 3, p. 498, 2024. DOI: 10.3390/electronics13030498
17. J.-H. Kim, W.-J. Choi, B. Chang and J.-W. Ha, "Adversarial self-supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Virtual, 2020, pp. 2983–2994.
18. P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim and P. Liang, "Concept bottleneck models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Virtual, 2020, pp. 5338–5348.
19. J. Li, P. Zhou, C. Xiong and S. C. H. Hoi, "Prototypical contrastive learning of unsupervised representations," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Virtual, 2021.
20. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 4765–4774.
21. M. Nauta, J. Schlötterer, M. van Keulen and C. Seifert, "PIP-Net: Patch-Based Intuitive Prototypes for Interpretable Image Classification," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Vancouver, Canada, 2023, pp. 2744–2753.
22. M. Nauta, R. van Bree and C. Seifert, "Neural prototype trees for interpretable fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Virtual, 2021, pp. 14933–14943.
23. L. Oakden-Rayner, J. Dunnmon, G. Carneiro and C. Ré, "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," in *Proc. ACM Conf. Health Inference Learn. (CHIL)*, Toronto, Canada, 2020, pp. 151–159.
24. N. Prentzas, A. Kakas and C. S. Pattichis, "Explainable AI applications in the medical domain: a systematic review," *IEEE Trans. Emerg. Top. Comput.*, 2023. DOI: 10.1109/TETC.2023.3247907
25. T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, S. B. A. Khurshid, M. Kashem, F. Islam, S. J. Maadeed, S. M. Zughaier and M. S. Khan, "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images," *Comput. Biol. Med.*, vol. 132, p. 104319, 2021. DOI: 10.1016/j.compbimed.2021.104319
26. P. Rajpurkar, E. Chen, O. Banerjee and E. J. Topol, "AI in health and medicine," *Nat. Med.*, vol. 28, no. 1, pp. 31–38, 2022. DOI: 10.1038/s41591-021-01614-0
27. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vision*, vol. 128, no. 2, pp. 336–359, 2020. DOI: 10.1007/s11263-019-01228-7
28. S. Shurrab and R. Duwairi, "Self-supervised learning methods and applications in medical imaging analysis: a survey," *PeerJ Comput. Sci.*, vol. 8, p. e1045, 2022. DOI: 10.7717/peerj-cs.1045
29. M. Sundararajan, A. Taly and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Sydney, Australia, 2017, pp. 3319–3328.
30. A. Taleb, W. Loetzsch, N. Danz, J. Severin, T. Gaertner, B. Bergner and C. Lippert, "3D self-supervised methods for medical imaging," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Virtual, 2020, pp. 18158–18172.
31. U.S. Food and Drug Administration, "Artificial Intelligence and Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan," U.S. FDA, Silver Spring, MD, USA, Jan. 2021. [Online]. Available: <https://www.fda.gov/media/145022/download>
32. Y. Zeng, W. Zhou, J. Sun, X.-S. Hua, J. Liu and L. Zheng, "ProtoCLIP: Prototypical contrastive language image pretraining," *IEEE Trans. Neural Netw. Learn. Syst.*, 2023. DOI: 10.1109/TNNLS.2023.3325671
33. Z. Zhou, V. Sodha, M. M. R. Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway and J. Liang, "Models Genesis: generic autodidactic models for 3D medical image analysis," in *Proc. Int. Conf. Med. Image Comput. Comput. -Assist. Interv. (MICCAI)*, Shenzhen, China, 2019, pp. 384–393.