bioengineering

MDPI

Article

Privacy-Aware Continual Self-Supervised Learning on Multi-Window Chest Computed Tomography for Domain-Shift Robustness

Ren Tasai [1], Guang Li [2], Ren Togo [3], Takahiro Ogawa [3], Kenji Hirata [4], Minghui Tang [4], Takaaki Yoshimura [5], Hiroyuki Sugimori [5], Noriko Nishioka [4], Yukie Shimizu [4], Kohsuke Kudo [4] and Miki Haseyama [3,*]

1 Graduate School of Information Science and Technology, Hokkaido University, N-14, W-9, Kita-ku, Sapporo 060-0814, Hokkaido, Japan; tasai@lmd.ist.hokudai.ac.jp
2 Education and Research Center for Mathematical and Data Science, Hokkaido University, N-12, W-7, Kita-ku, Sapporo 060-0812, Hokkaido, Japan; guang@lmd.ist.hokudai.ac.jp
3 Faculty of Information Science and Technology, Hokkaido University, N-14, W-9, Kita-ku, Sapporo 060-0814, Hokkaido, Japan; togo@lmd.ist.hokudai.ac.jp (R.T.); ogawa@lmd.ist.hokudai.ac.jp (T.O.)
4 Faculty of Medicine, Hokkaido University, N-15 W-7, Kita-ku, Sapporo 060-8648, Hokkaido, Japan; khirata@pop.med.hokudai.ac.jp (K.H.); toumeiki@hs.hokudai.ac.jp (M.T.); nnishioka@pop.med.hokudai.ac.jp (N.N.); yshimizu7298@med.hokudai.ac.jp (Y.S.); kkudo@med.hokudai.ac.jp (K.K.)
5 Faculty of Health Sciences, Hokkaido University, N-12, W-5, Kita-ku, Sapporo 060-0812, Hokkaido, Japan; takaaki.ysm@med.hokudai.ac.jp (T.Y.); sugimori@hs.hokudai.ac.jp (H.S.)
* Correspondence: mhaseyama@lmd.ist.hokudai.ac.jp

Abstract

We propose a novel continual self-supervised learning (CSSL) framework for simultaneously learning diverse features from multi-window-obtained chest computed tomography (CT) images and ensuring data privacy. Achieving a robust and highly generalizable model in medical image diagnosis is challenging, mainly because of issues, such as the scarcity of large-scale, accurately annotated datasets and domain shifts inherent to dynamic healthcare environments. Specifically, in chest CT, these domain shifts often arise from differences in window settings, which are optimized for distinct clinical purposes. Previous CSSL frameworks often mitigated domain shift by reusing past data, a typically impractical approach owing to privacy constraints. Our approach addresses these challenges by effectively capturing the relationship between previously learned knowledge and new information across different training stages through continual pretraining on unlabeled images. Specifically, by incorporating a latent replay-based mechanism into CSSL, our method mitigates catastrophic forgetting due to domain shifts during continual pretraining while ensuring data privacy. Additionally, we introduce a feature distillation technique that integrates Wasserstein distance-based knowledge distillation and batch-knowledge ensemble, enhancing the ability of the model to learn meaningful, domain-shift-robust representations. Finally, we validate our approach using chest CT images obtained across two different window settings, demonstrating superior performance compared with other approaches.

Keywords: self-supervised learning; continual self-supervised learning; feature distillation; latent replay; chest CT image

Check for updates

Academic Editors: Carlo Cosimo Quattrocchi and Marco Parillo

Received: 17 November 2025
Revised: 24 December 2025
Accepted: 24 December 2025
Published: 27 December 2025

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Medical image analysis is crucial to clinical decision-making for diagnostic support [1,2]. Its performance has been dramatically improved by the emergence of deep learning-based supervised learning (SL) [3–5]. While automating parts of the diagnostic process can enhance the quality and efficiency of clinical judgment, models deployed in critical medical settings must be designed to simultaneously achieve high accuracy and function robustly, with generalizability across diverse datasets and conditions. However, SL is mainly limited by the significant shortage of large-scale, accurately annotated medical image datasets [6–8], and this scarcity is exacerbated because annotating medical data, which must balance privacy protection with accuracy, requires extensive expertise and considerable effort. Consequently, the model efficiency relies heavily on the availability of high-quality annotated data.

As an approach for addressing this data-scarcity issue, self-supervised learning (SSL) has garnered attention [9–12]. In SSL, a model is first pretrained using unlabeled data and then fine-tuned with a small amount of labeled data. Additionally, reports reveal that SSL achieves outstanding performance while effectively reducing labeling costs [13–15]. However, SSL still exhibits a key limitation: it lacks generalizability in real-world healthcare environments because the dynamic nature of clinical settings causes changes in the data distributions of medical images over time, resulting in domain shifts [16–18]. This shift stems from differences across medical institutions, imaging equipment, and diagnostic objectives, resulting in high diversity across medical images. A prominent example is in chest computed tomography (CT), where images often comprise multiple domains, such as the mediastinal and lung window settings, each optimized for distinct clinical-observation purposes [19,20]. In these dynamic clinical settings, data with different characteristics arrive sequentially, necessitating the continuous handling of domain shifts. However, conventional SSL relies on a joint training scenario, where a model is trained only after collecting a large amount of unlabeled data. Consequently, the model cannot flexibly adapt to new data without expensive retraining [21,22]. Furthermore, satisfying this premise in real-world scenarios is often challenging owing to the high computational costs of retraining and strict privacy constraints [23,24].

In response to these challenges, continual SSL (CSSL) was recently applied to medical imaging [25]. This approach involves allocating data with varying characteristics across multiple training stages for continual pretraining. Particularly, maintaining data-distribution diversity during this pretraining process enables the acquisition of rich feature representations that are beneficial for subsequent fine-tuning. Notably, CSSL mitigates data interference that typically occurs when integrating different modalities or diverse domains within joint SSL frameworks [26,27]. Additionally, CSSL achieves good accuracy and generalizability compared to representative supervised continual learning (SCL) paradigms [28,29]. Therefore, CSSL is projected to address labeling-cost reduction and dynamic-environment domain shifts.

Notably, the primary challenge of CSSL is catastrophic forgetting [30,31], which occurs when a model overwrites or forgets previously acquired knowledge while learning new concepts. To mitigate this issue, CSSL conventionally adopts experience-replay-based approaches [32,33] from SCL. In this approach, a portion of the original images is stored in a memory buffer (B), enabling the model to retain and revisit past knowledge during later sequential training stages. This strategy is highly versatile, exhibiting applicability in a wide range of scenarios compared with other SCL approaches. However, in the medical-data context, the retention of past datasets is often complicated by privacy concerns [34,35]. In the SCL field, latent replay (LR)-based approaches have been introduced to address catastrophic forgetting. These approaches achieve privacy preservation by storing the activations of intermediate layers in neural networks (NNs) and replaying them when

learning new knowledge [36,37]. Specifically, they store feature representations instead of preserving the original images, leveraging them in subsequent learning stages to ensure data privacy. Nevertheless, LR remains largely unexplored within the CSSL context, necessitating the exploration of a novel LR-based CSSL framework.

To satisfy the aforementioned research gap, we propose a novel CSSL framework that simultaneously addresses domain shifts in dynamic environments while effectively mitigating catastrophic forgetting under privacy-constraint conditions. The proposed framework maintains a *B* that stores only the feature representations of past data, enabling the continual pretraining of rich representations while preserving data privacy and distribution diversity. To realize this, we develop a feature distillation method that integrates Wasserstein distance (WD)-based knowledge distillation (WKD) with a batch-knowledge ensemble (BKE). While WKD enforces distributional alignment between replayed and mini-batch features, BKE aggregates feature representations to enhance consistency and reduce domain interference. This unified WKD-BKE design facilitates the learning of robust, generalizable features across multiple domains, exhibiting suitability for privacy-conscious CSSL. Further, to evaluate the effectiveness of the proposed framework, we pretrain a model using chest CT images acquired under two different window settings and evaluate its performance on two distinct public CT-image datasets. Extensive experiments demonstrate that the proposed framework consistently outperforms other approaches, achieving superior robustness and performance.

The contributions of our study are summarized below.

- We propose a novel LR-based CSSL framework to ensure data privacy and effectively address catastrophic forgetting during pretraining with chest CT images across two domains.
- We introduce a novel WKD-BKE-integrated feature distillation method to simultaneously enable robust feature-representation learning and mitigate data interference.
- Our extensive experiments reveal that our method outperforms state-of-the-art approaches on two public chest-CT-image datasets.

The remainder of this paper is organized as follows: Section 2 discusses the extant studies, Section 3 describes the details of the proposed CSSL framework, and Sections 4, 5 and 6 present the experiments, discussion, and conclusions, respectively.

2. Related Studies

2.1. Self-Supervised Learning for Addressing Domain Shifts

SSL has recently garnered significant attention in medical image analysis, which is characterized by limited annotated data. SSL has been applied across various modalities, including CT [38–40], magnetic resonance imaging (MRI) [41,42], fundus imaging [43,44], and ultrasound localization microscopy (ULM) [45,46]. However, domain shifts due to differences in imaging equipment, acquisition protocols, and diagnostic objectives typically compromise model reliability and robustness. To address this, multi-domain pretraining under the joint-training condition has been explored.

In CT-based studies, Wolf et al. [38] proposed a masked modeling-based SSL method for convolutional NNs. The authors pretrained the model on a large-scale chest-CT dataset obtained from multiple medical institutions and demonstrated its effectiveness using classification tasks. Similarly, Jiang et al. [39] investigated the robustness of SSL to domain shifts for tumor segmentation in non-small-cell lung cancer CT. Employing MRI, Fiorentino et al. [42] introduced an intensity-based self-supervised domain-adaptation approach for intervertebral disc segmentation. This approach effectively reduced annotation costs and improved generalizability across scanners with heterogeneous acquisition settings. Mojab et al. [44] employed fundus imaging to demonstrate the superior adaptability

of SSL-pretrained models, which were trained on multi-device datasets to unseen domains in glaucoma detection. Yu et al. [45] addressed domain shifts in ultrasound localization microscopy caused by differences between simulated and real ultrasound data, imaging depth, and noise characteristics, and proposed a semi-supervised approach to mitigate cross-domain discrepancies in microbubble signal representations.

Overall, these studies demonstrated that SSL-based pretraining represents an effective strategy for mitigating domain shifts in medical image analysis. However, studies revealed that representation learning across different modalities and domains can interfere with each other during pretraining, primarily because of their substantial differences, which ultimately results in data interference [26,27]. Furthermore, in the application of SSL to clinical settings that are characterized by dynamically changing data distributions, models often require retraining on the entire dataset. Such retraining requires considerable computational resources. Moreover, the often limited access to all data due to privacy constraints poses significant challenges for clinical applications.

2.2. Continual Self-Supervised Learning for Addressing Domain Shifts

In recent years, the application of CSSL has primarily focused on natural images, exploring its ability to handle incrementally arriving data. This capability is especially relevant in real-world settings where it is often impractical to assemble all data in advance. Fini et al. [23] experimented on DomainNet [47] involving sequential learning across multiple domains, demonstrating that CSSL outperformed major SSL and effectively addressed domain shifts. Furthermore, Hu et al. [29] experimentally demonstrated on DomainNet that combining CSSL with simple SCL strategies, such as experience replay or parameter regularization [48], can significantly mitigate performance degradation, even under substantial distributional shifts. This and related studies provide valuable insights, collectively indicating that CSSL operates effectively in real-world scenarios, even in the presence of domain shifts.

These advances motivated a growing interest in the application of CSSL to medical imaging analysis. For instance, Ye et al. [26] and Yao et al. [49] proposed methods for enhancing robustness and scalability in cross-modality learning. These approaches leverage experience replay and feature distillation [50,51] to efficiently integrate new modalities while suppressing the representational interference that often arises in conventional SSL. However, these studies were limited to demonstrating effectiveness across multiple modalities and did not sufficiently examine applicability to multiple domains (i.e., domain shift within a single modality). Conversely, a CSSL method was recently proposed [27] using chest CT images acquired under heterogeneous scanning conditions. The method involves the introduction of an experience-replay-based approach for balancing sample diversity and representativeness within B. This design enabled the acquisition of domain-invariant feature representations during continual pretraining. The authors deployed a COVID-19 classification task to demonstrate the effectiveness of CSSL in mitigating representational interference in conventional SSL while maintaining robustness across diverse domains.

These medical-imaging studies generally adopted experience-replay-based approaches, which store past image samples in B and reuse them in subsequent pretraining stages, thereby mitigating catastrophic forgetting. Experience replay has demonstrated improved robustness in continual pretraining across multiple modalities and domains in medical imaging. Although this strategy is highly versatile and applicable to a wide range of scenarios, retaining past datasets is often impractical owing to concerns about preserving medical data privacy. In contrast to experience replay, recent privacy-aware continual learning frameworks have investigated LR mechanisms, which store intermediate feature representations instead of original images. These approaches aim to reduce privacy risks by

avoiding the direct storage of raw data. In many existing methods, additional operations such as nonlinear transformations, compression, or quantization are applied to the stored representations in order to suppress input reconstruction via inversion models. In this study, we propose a novel CSSL framework for chest CT images spanning two domains, integrating an LR-based approach. To the best of our knowledge, the integration of LR into a CSSL framework has not been extensively explored in prior work. Accordingly, as an initial investigation, we adopt a design choice in which no additional transformations are applied to the stored feature representations. Dissimilar to conventional experience replay, our method only retains feature representations in B rather than the raw data, thereby achieving privacy-preserving continual pretraining while enabling the progressive acquisition of more expressive representations.

3. Privacy-Aware Continual Self-Supervised Learning Integrating Latent Replay and Feature Distillation

Our CSSL framework comprises a three-stage sequential process for pretraining the vision transformer (ViT) [52] encoder employed in subsequent downstream tasks. In the first stage, SSL is performed using the initial dataset, D_1, from one chest-CT-image domain. In the second stage, selected feature representations from D_1 are stored in B to preserve data diversity and privacy. In the third stage, SSL is performed again using the next (second-domain) dataset, D_2, from another domain. In this third stage, feature distillation involving WKD-BKE integration is performed using replayed features from B. Afterward, fine-tuning is performed using labeled data. Figure 1 shows an overview of the proposed CSSL framework.

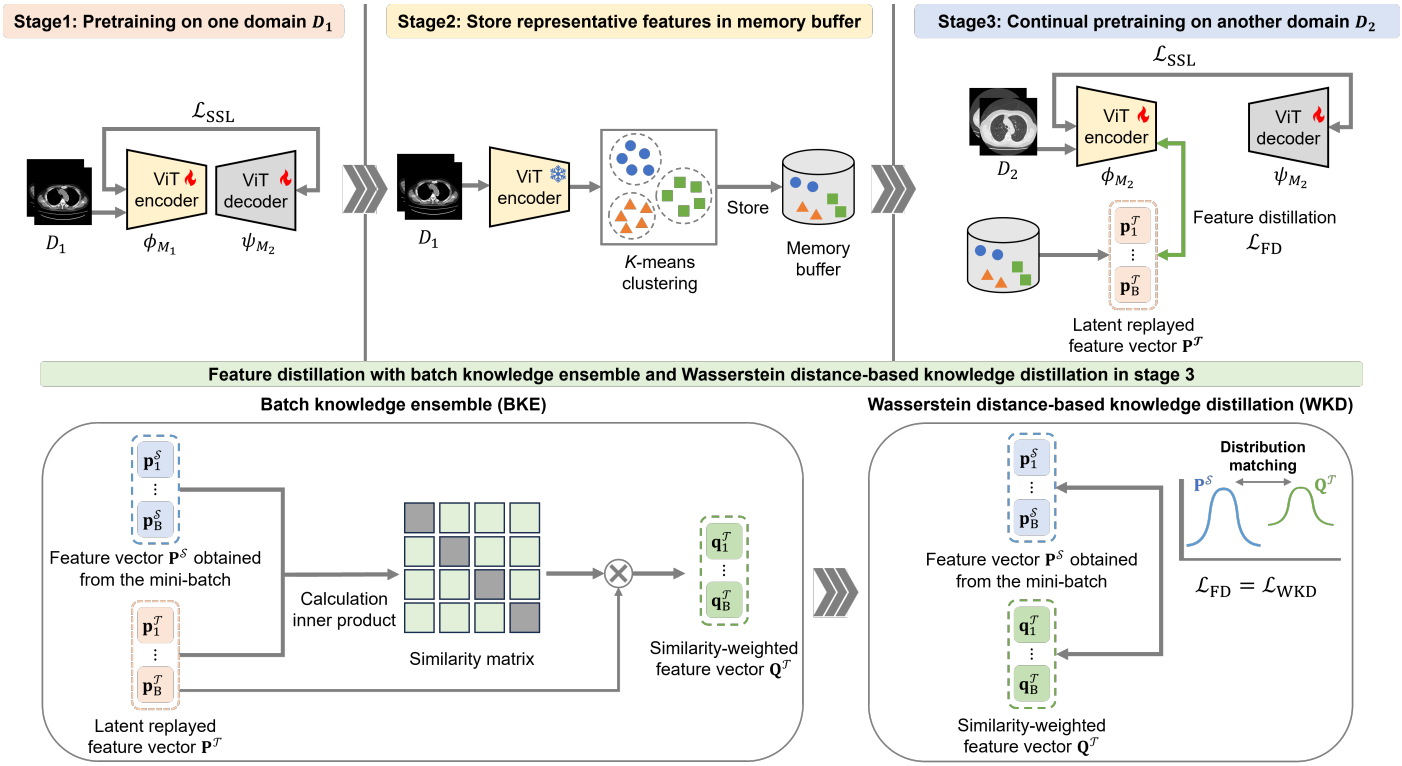


Figure 1. Overview of the proposed continual self-supervised learning (CSSL) framework.

3.1. Stage 1: Self-Supervised Learning on the First-Domain Dataset

The first pretraining stage proceeds with a model, M_1, using D_1. The masked autoencoder (MAE) method [53] is employed to learn feature representations from the input data. This pretraining task uses a reconstruction objective that compares the original patches

with their reconstructed masked counterparts, enabling the model to learn meaningful representations from unlabeled data. The network architecture consists of an encoder, ϕ_{M_1}, and its corresponding decoder, ψ_{M_1}.

In this process, each image with C channels is divided into n patches of size (V, V), which are collectively represented as $X_V \in \mathbb{R}^{n\times(V^2\times C)}$. A masking rate r is then applied to X_V, and $m = n \times r$ patches are randomly selected as the patches to be masked, which are denoted as X_m. Next, all patches are converted into a sequence of tokens using a tokenizer, $\mathcal{T}_{M_1}$. Thereafter, the tokens corresponding to the $n - m$ unmasked patches are fed into the encoder ϕ_{M_1} to generate latent feature representations. The decoder, ψ_{M_1}, reconstructs the original patch contents of X_m into reconstructed patches Y_m by predicting their pixel values using the feature representations obtained from the encoder, together with the embeddings of X_m provided by $\mathcal{T}_{M_1}$. Afterward, the model is optimized to minimize the mean squared error between X_m and Y_m as follows:

$$\mathcal{L}_{\text{SSL}} = \frac{1}{m \times V^2 \times C} ||Y_m - X_m||_2^2. \quad (1)$$

This first stage terminates with the training of M_1 to capture comprehensive feature representations from D_1. This trained (M_1) model is subsequently employed in the third CSSL stage, which integrates D_1 and D_2.

3.2. Stage 2: Sampling Features in the Memory Buffer

The second stage involves the selection of features stored in B. This process is crucial to capturing data-distribution changes across different stages and mitigating catastrophic forgetting. Additionally, the utilization of these selected features in the third stage ensures data privacy.

In this work, the output of the final encoder layer of a pretrained ViT encoder ϕ_{M_1} is used as the feature representation of D_1 in the first stage. These features integrate global contextual information across all tokens through self-attention, which can facilitate clustering and sample selection [54]. First, these feature representations are divided into $N \times \alpha$ clusters. Finally, $N \times \beta$ features closest to the cluster centers are selected and stored in B. In this algorithm, N denotes the number of images in D_1, and α and β are control parameters for the sampling ratio.

3.3. Stage 3: Continual Self-Supervised Learning with Feature Distillation Using the Second-Domain Dataset

Following the SSL pretraining of model M_1 in the first stage, another model, M_2, is pretrained in the third stage using the MAE method. Notably, M_2 is trained using D_2 and the replayed features from B in the second stage. Furthermore, WKD-BKE-integrated feature distillation enables M_2 to retain the knowledge acquired in the first stage while learning new representations for the second domain.

3.3.1. Wasserstein Distance-Based Knowledge Distillation

WKD facilitates knowledge retention by aligning the feature distributions of M_2 and M_1. Specifically, it compares the M_1-associated feature representations (replayed from B) with those generated by M_2 in the third stage.

We consider a feature map derived from the feature representations, as follows: let the spatial height, width, and channel number of this feature map be h, w, and l, respectively. Next, the feature map is transformed into a matrix, $\mathbf{F} \in \mathbb{R}^{l\times d}$, where $d = h \times w$, and the i-th column $\mathbf{f}_i \in \mathbb{R}^l$ represents the spatial features. Thereafter, we estimate the first and second moments, $\boldsymbol{\mu} = \frac{1}{d}\sum_i \mathbf{f}_i$ and $\boldsymbol{\Sigma} = \frac{1}{d}\sum_i (\mathbf{f}_i - \boldsymbol{\mu})(\mathbf{f}_i - \boldsymbol{\mu})^T$, respectively, from these

features. The feature distribution of the input images is modeled as a Gaussian distribution parameterized by the mean vector, $\boldsymbol{\mu}$, and covariance matrix, $\boldsymbol{\Sigma}$, as follows:

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{|2\pi\boldsymbol{\Sigma}|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{f}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{f}-\boldsymbol{\mu})\}, \tag{2}$$

where $|\cdot|$ is the matrix determinant. Additionally, we define the teacher's and student's feature distributions as $\mathcal{N}^{\mathcal{T}} \triangleq \mathcal{N}(\boldsymbol{\mu}^{\mathcal{T}}, \boldsymbol{\Sigma}^{\mathcal{T}})$ and $\mathcal{N}^{\mathcal{S}}$, respectively. The continuous WD between the two Gaussian distributions is expressed by the following:

$$\mathrm{D}_{\mathrm{WD}}(\mathcal{N}^{\mathcal{T}}, \mathcal{N}^{\mathcal{S}}) = \inf_{q} \int_{\mathbb{R}^l} \int_{\mathbb{R}^l} ||\mathbf{f}^{\mathcal{T}} - \mathbf{f}^{\mathcal{S}}||^2 q(\mathbf{f}^{\mathcal{T}}, \mathbf{f}^{\mathcal{S}}) d\mathbf{f}^{\mathcal{T}} d\mathbf{f}^{\mathcal{S}}, \tag{3}$$

where inf represents the infimum, which is the greatest lower bound, $\mathbf{f}^{\mathcal{T}}$ and $\mathbf{f}^{\mathcal{S}}$ are Gaussian variables, and $||\cdot||^2$ denotes the Euclidean distance. The joint distribution, q, is constrained such that its marginal distributions correspond to $\mathcal{N}^{\mathcal{T}}$ and $\mathcal{N}^{\mathcal{S}}$. Thus, to minimize this equation and following [55,56], we define the WKD loss function, $\mathcal{L}_{\mathrm{WKD}}$, as follows:

$$\mathcal{L}_{\mathrm{WKD}} = \gamma \mathrm{D}_{\mathrm{mean}}(\boldsymbol{\mu}^{\mathcal{T}}, \boldsymbol{\mu}^{\mathcal{S}}) + \mathrm{D}_{\mathrm{cov}}(\boldsymbol{\Sigma}^{\mathcal{T}}, \boldsymbol{\Sigma}^{\mathcal{S}}). \tag{4}$$

Here, $D_{\mathrm{mean}}(\boldsymbol{\mu}^{\mathcal{T}}, \boldsymbol{\mu}^{\mathcal{S}}) = ||\boldsymbol{\mu}^{\mathcal{T}} - \boldsymbol{\mu}^{\mathcal{S}}||^2$ and $\mathrm{D}_{\mathrm{cov}}(\boldsymbol{\Sigma}^{\mathcal{T}}, \boldsymbol{\Sigma}^{\mathcal{S}}) = ||\boldsymbol{\delta}^{\mathcal{T}} - \boldsymbol{\delta}^{\mathcal{S}}||^2$, where $\boldsymbol{\delta}^{\mathcal{T}}$ and $\boldsymbol{\delta}^{\mathcal{S}}$ are the standard deviation vectors formed from the square root of the diagonal elements of $\boldsymbol{\Sigma}^{\mathcal{T}}$ and $\boldsymbol{\Sigma}^{\mathcal{S}}$, respectively. Diagonal covariance matrices were employed for their robustness in estimating high-dimensional features as well as their computational efficiency [55,57]. To balance the roles of the mean and covariance, we introduce a mean–covariance ratio hyperparameter, γ. By computing $\mathcal{L}_{\mathrm{WKD}}$, we enable the feature distribution of M_2 to align with that of M_1, thereby mitigating data interference due to inter-stage domain shifts (data-distribution differences across stages).

3.3.2. Batch Knowledge Ensemble

We apply the BKE approach to enable M_2 to concurrently achieve robust learning while knowledge retention from the first stage. This approach enables feature distillation based on the similarity between B feature representations, $\mathbf{P}^{\mathcal{T}}$, randomly replayed from the memory buffer and the feature representations, $\mathbf{P}^{\mathcal{S}}$, within the mini-batch generated by the encoder, ϕ_{M_2} of M_2, in the third stage. Therefore, knowledge is propagated and ensembled via the affinity between the feature representations replayed from B and those generated by M_2 within mini-batches in the third stage.

Let the batch size, number of tokens, and embedding dimension be B, T, and E, respectively. Thus, $\mathbf{P}^{\mathcal{T}}, \mathbf{P}^{\mathcal{S}} \in \mathbb{R}^{B\times T\times E}$. First, we obtain the similarity matrix, $\mathbf{A} \in \mathbb{R}^{B\times T\times T}$, by calculating the similarities between the replayed feature representations, $\{\mathbf{p}_1^{\mathcal{T}}, \ldots, \mathbf{p}_B^{\mathcal{T}}\}$, retrieved from B, and the encoded visual features, $\{\mathbf{p}_1^{\mathcal{S}}, \ldots, \mathbf{p}_B^{\mathcal{S}}\}$, extracted from a mini-batch of B images, as follows:

$$\mathbf{A}_{i,j} = (\hat{\mathbf{p}}_i^{\mathcal{T}\top} \hat{\mathbf{p}}_j^{\mathcal{S}}). \tag{5}$$

In this equation, each feature representation is denoted as $\mathbf{p}_i^{\mathcal{T}}, \mathbf{p}_j^{\mathcal{S}} \in \mathbb{R}^{T\times E}$, where $\hat{\mathbf{p}}_i^{\mathcal{T}} = \mathbf{p}_i^{\mathcal{T}} / \|\mathbf{p}_i^{\mathcal{T}}\|_2$ and $\hat{\mathbf{p}}_j^{\mathcal{S}} = \mathbf{p}_j^{\mathcal{S}} / \|\mathbf{p}_j^{\mathcal{S}}\|_2$ represent the normalized feature representations. The indices, i and j, refer to the tokens within each mini-batch sample and replayed memory, respectively. Next, we normalize $\mathbf{A} \in \mathbb{R}^{B\times T\times T}$ as follows:

$$\hat{\mathbf{A}}_{i,j} = \frac{\exp(\mathbf{A}_{i,j})}{\sum_{j\neq i} \exp(\mathbf{A}_{i,j})}, \forall i \in \{1, \ldots, B\}. \tag{6}$$

To prevent the excessive propagation and aggregation of noisy predictions, the optimized feature representation, $\mathbf{Q}$, is generated as a weighted sum of feature representation $\mathbf{P}^{\mathcal{T}}$ and propagated probability matrix $\hat{\mathbf{A}}\mathbf{P}^{\mathcal{T}}$ as follows:

$$\mathbf{Q} = \omega\hat{\mathbf{A}}\mathbf{P}^{\mathcal{T}} + (1-\omega)\mathbf{P}^{\mathcal{T}}. \tag{7}$$

Notably, propagation can proceed multiple times to generate $\mathbf{Q}$ for feature distillation:

$$\begin{aligned}\mathbf{Q}_{(t)} &= \omega\hat{\mathbf{A}}\mathbf{Q}_{(t-1)} + (1-\omega)\mathbf{P}^{\mathcal{T}},\\ &= (\omega\hat{\mathbf{A}})^t\mathbf{P}^{\mathcal{T}} + (1-\omega)\sum_{i=0}^{t-1}(\omega\hat{\mathbf{A}})^i\mathbf{P}^{\mathcal{T}},\end{aligned} \tag{8}$$

where ω is a weight factor and t the t-th propagation and ensembling iteration. As the number of iterations approaches infinity, we obtain $\lim_{t\to\infty}(\omega\hat{\mathbf{A}})^t = 0$ and $\lim_{t\to\infty}\sum_{i=0}^{t-1}(\omega\hat{\mathbf{A}})^i = (\mathbf{I}-\omega\hat{\mathbf{A}})^{-1}$; hence an approximate inference formulation can be obtained as follows:

$$\mathbf{Q}^{\mathcal{T}} = (1-\omega)(\mathbf{I}-\omega\hat{\mathbf{A}})^{-1}\mathbf{P}^{\mathcal{T}}. \tag{9}$$

Thereafter, the optimized feature representation, $\mathbf{Q}^{\mathcal{T}}$, and $\mathbf{P}^{\mathcal{S}}$ are transformed into Equations (2) and (3) to calculate $\mathcal{L}_{\text{FD}} = \mathcal{L}_{\text{WKD}}$. By computing the feature-distillation loss $\mathcal{L}_{\text{FD}}$, we facilitate the acquisition of robust feature representations while minimizing the deviation from those learned in the first stage.

Next, our introduction of the LR-based approaches and WKD-BKE-integrated feature distillation into the CSSL framework enables the encoder to effectively capture the relationships between newly acquired data and previously learned knowledge. This integration mitigates the effects of catastrophic forgetting during pretraining as well as facilitating the learning of richer, more robust feature representations. Following the three-stage CSSL procedure, the ViT encoder ϕ_{M_1} is fine-tuned on a separate labeled dataset for downstream tasks, such as classification. During fine-tuning, ϕ_{M_1} is integrated with a randomly initialized task-specific multi-layer perceptron head and applied to the downstream tasks. Algorithm 1 summarizes the proposed CSSL framework.

Algorithm 1 Algorithm of the proposed CSSL framework.

Input: $\{D_1, D_2\}$: two subsets from different domains, B: memory buffer, $\mathcal{T}_{M_1}, \mathcal{T}_{M_2}$: tokenizers, ϕ_{M_1}, ϕ_{M_2}: encoders, ψ_{M_1}, ψ_{M_2}: model-specific decoders, $K\text{-}means(\cdot)$: k-means clustering operation, $ClusterSample(\cdot)$: operation for sampling cluster centers, $LatentReplay(\cdot)$: LR operation
Output: $\phi_{M_2}, \mathcal{T}_{M_2}$
Stage 1: SSL on D_1
Set the training dataset: $D \leftarrow D_1$
Update ϕ_{M_1}, $\mathcal{T}_{M_1}$, and ψ_{M_1} by minimizing $\mathcal{L}_{\text{SSL}}$, following Equation (1)
Stage 2: Sampling Features into the Memory Buffer
Obtain clusters: $C \leftarrow K\text{-}Means(\phi_{M_1}(D_1))$
Populate the memory buffer: $B \leftarrow ClusterSample(C)$
Stage 3: CSSL with Feature Distillation on D_2
Set the training dataset: $D \leftarrow D_2$
Extract the mini-batch feature representations: $\mathbf{P}^{\mathcal{S}} \leftarrow \phi_{M_2}(D_2)$
Retrieve replayed feature representations from B: $\mathbf{P}^{\mathcal{T}} \leftarrow LatentReplay(B)$
Obtain $\mathbf{Q}^{\mathcal{T}}$ by calculating the similarity between $\mathbf{P}^{\mathcal{S}}$ and $\mathbf{P}^{\mathcal{T}}$, following Equations (5)–(9)
Update ϕ_{M_2}, $\mathcal{T}_{M_2}$, and ψ_{M_2} by minimizing $\mathcal{L}_{\text{SSL}}$ and $\mathcal{L}_{\text{FD}}$ with $\mathbf{Q}^{\mathcal{T}}$ and $\mathbf{P}^{\mathcal{S}}$, following Equations (1) and (4), respectively.

4. Experiments

We comprehensively experiment on classification tasks to validate the effectiveness of the proposed CSSL framework. These experiments include ablation studies, hyperparame-

ter analyses, and an investigation of the impact of extending pretraining stages. The dataset and experimental settings are introduced in Section 4.1. Additionally, the classification-task performances with different pretraining datasets are discussed in Section 4.2. Furthermore, the impacts of hyperparameters on the experimental results are presented in Section 4.3. The ablation study of the proposed CSSL framework is discussed in Section 4.4. Finally, the impact of extending the pretraining stages in CSSL is demonstrated in Section 4.5.

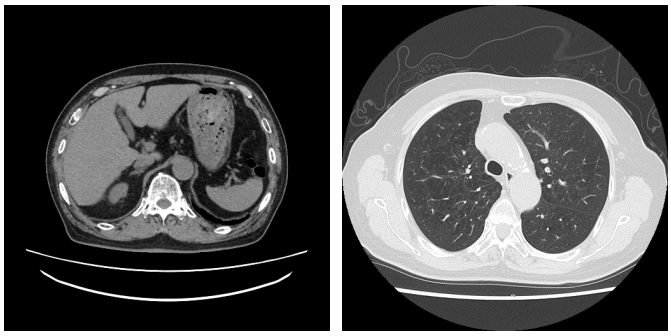
4.1. Datasets and Settings

For pretraining, we utilized a subset of the J-MID (https://www.radiology.jp/j-mid/ (accessed on 6 April 2025)) database, which contains large-scale CT scans from Japanese medical institutions, and the RICORD dataset [58], an open dataset that was developed collaboratively by the Radiological Society of North America and international partners and contains chest CT scans collected from four countries. Each dataset was constructed with two domains based on mediastinal and lung window settings in chest CT images. Both domains are denoted as D_1 and D_2, and the labels are not used during pretraining. Specifically, for the J-MID subset, D_1 (the mediastinal window) contains 31,256 CT images, and D_2 (the lung window) contains 26,403 CT images. The RICORD dataset comprises 12,897 D_1 (mediastinal window) images and 11,668 D_2 (lung window) images for pretraining. For the J-MID dataset, D_1 was generated using a window level (WL) of 40 ± 20 HU and a window width (WW) of 350 ± 50 HU, whereas D_2 was generated using a WL of -625 ± 75 HU and a WW of 1350 ± 350 HU. For the RICORD dataset, D_1 was generated with a WL of 40 ± 20 HU and a WW of 350 ± 50 HU, while D_2 was generated with a WL of -600 ± 100 HU and a WW of 1300 ± 300 HU. These parameter ranges were selected to cover clinically standard lung and mediastinal window settings while accommodating inter-scan variability in DICOM metadata. The corresponding images for each example are shown in Figures 2 and 3. For fine-tuning and evaluation, we utilized two public datasets: the SARS-CoV-2 CT-Scan Dataset [59] and the Chest CT-Scan Images Dataset (https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images (accessed on 6 April 2025)). Both datasets were used for the coronavirus disease 2019 (COVID-19) and chest cancer classification tasks, respectively. The data breakdown is as follows: the SARS-CoV-2 CT-Scan Dataset comprises 1589 training, 397 validation, and 495 test images, labeled into two (COVID-19 and Normal) classes. The Chest CT-Scan Images Dataset comprises 490 training, 123 validation, and 315 test images labeled into four (adenocarcinoma, large-cell carcinoma, squamous-cell carcinoma, and normal) classes. COVID-19 classification and lung cancer classification were selected as downstream tasks because chest CT is widely used in clinical practice for diagnosing both COVID-19 and lung cancer, allowing for an evaluation that closely reflects real-world clinical scenarios. Accordingly, to ensure reproducibility and enable fair comparisons with prior studies, we prioritized the use of publicly available datasets. The corresponding images for each example are shown in Figures 4 and 5.

In the pretraining of the MAE, the batch sizes were set to 64 and 32 in the first and third stages, respectively, and the masking ratio, r, was set to 0.75, with ViT-B [52] being deployed as the encoder. Next, augmentation techniques, such as random crop, resize, and flip, were applied to the images. Additionally, a warm-up strategy was applied during the first 40 epochs, gradually increasing the learning rate from 0 to 0.00015. Subsequently, the learning rate was reduced to 0 via a cosine schedule. For k-means sampling, the parameters, α and β, which determine the sampling ratio, were set to 0.01 and 0.05, respectively [26]. Notably, γ, which adjusts the contributions of the mean and covariance in the WKD loss, $\mathcal{L}_{\mathrm{WKD}}$, was set to 2.0 and 3.0 during the pretraining on the J-MID subset and RICORD dataset, respectively. In BKE, the hyperparameter, ω, was set to 0.5 [60–62]. The AdamW

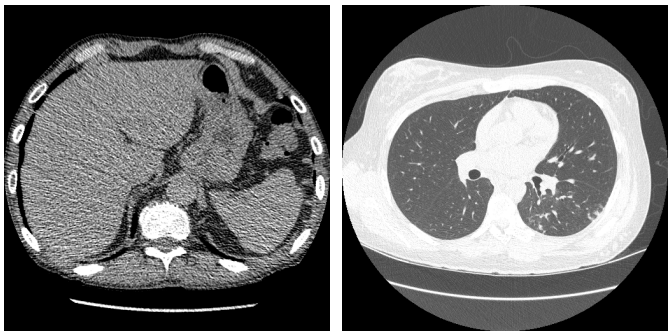
optimizer [63] was utilized, with the learning rate set to 0.00005. Pretraining and fine-tuning were conducted for 300 and 80 epochs per stage, respectively.

For the evaluation metrics, we employed three metrics: two-class classification accuracy (ACC), the area under the receiver operating characteristic curve (AUC), and the F1-score (F1). To ensure robustness, we averaged the results across three of the four random seeds (0, 10, 100, and 1000). In all tables, the best performance is highlighted in bold for each experimental result. To evaluate the effectiveness of our method, we compared it with the following approaches: the state-of-the-art CSSL method for medical imaging, MedCoSS [26], MAE [53] simultaneously pretrained on D_1 and D_2, MAE pretrained only on D_1, and MAE pretrained only on D_2. As a baseline method, we employed a model that was fine-tuned without MAE-based self-supervised pretraining.



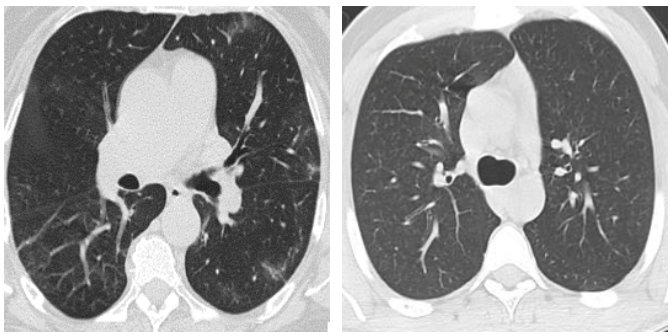
(**a**) First-domain dataset (D_1) (**b**) Second-domain dataset (D_2)

Figure 2. Examples of chest CT images on the subsets from the J-MID database: (**a**) first-domain dataset (D_1) and (**b**) second-domain dataset (D_2).



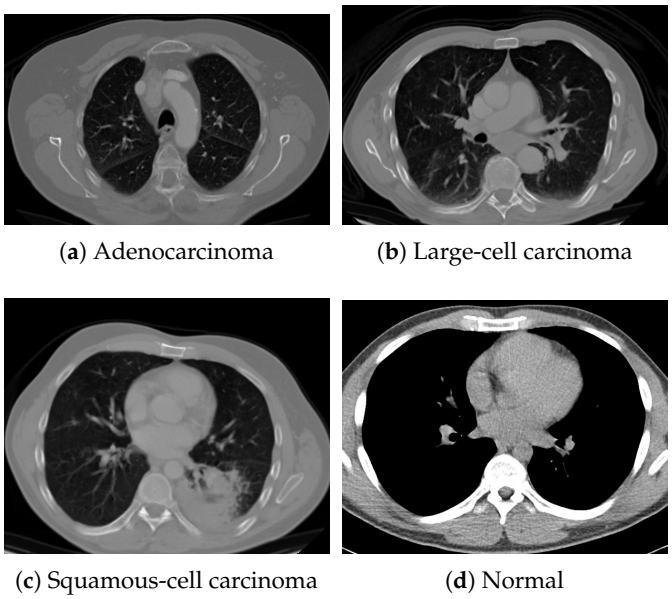
(**a**) D_1 (**b**) D_2

Figure 3. Examples of chest CT images on the subset from the RICORD dataset: (**a**) D_1 and (**b**) D_2.



(**a**) Coronavirus disease 2019 (COVID-19) (**b**) Normal

Figure 4. Examples of chest CT images on the SARS-CoV-2 CT-Scan dataset: (**a**) COVID-19 and (**b**) Normal.



(**a**) Adenocarcinoma (**b**) Large-cell carcinoma

(**c**) Squamous-cell carcinoma (**d**) Normal

Figure 5. Examples of chest CT images from the Chest CT-Scan Images dataset: (**a**) adenocarcinoma, (**b**) large-cell carcinoma, (**c**) squamous-cell carcinoma, and (**d**) normal.

4.2. Classification-Task Performance with Different Pretraining Datasets

Table 1 presents the classification results obtained after pretraining on the J-MID subset and evaluating on the SARS-CoV-2 CT-Scan dataset (for COVID-19 classification) and the Chest CT-Scan Images dataset (for lung cancer classification). Furthermore, to examine the effectiveness of the domain-pretraining order, we performed continual pretraining by interchanging domains D_1 and D_2. Notably, the highest accuracy on the SARS-CoV-2 CT-Scan dataset was achieved when continual pretraining was performed from D_1 to D_2, whereas the best performance on the Chest CT-Scan Images dataset was obtained when the pretraining order was reversed from D_2 to D_1. These results suggest that the domain-pretraining order in continual pretraining is associated with variations in downstream task performance.

Such order dependence is considered to be related to the effects of catastrophic forgetting, which is a well-known phenomenon in SCL and CSSL. Prior studies have reported that the feature space learned by deep models is influenced by parameter updates in later training stages, resulting in learned representations that tend to be biased toward data encountered more recently [64,65]. In particular, when SCL is performed without strict constraints, parameter updates in later stages can reorganize the embedding space. This reorganization can lead to the partial overwriting of representations acquired in earlier stages. Consequently, the final learned representations tend to align with the statistical characteristics of the domain learned last. In our experimental setting, similar tendencies were observed. When a domain that is semantically closer to the downstream task was presented at the final stage, the learned representations exhibited relatively higher transfer performance. In contrast, when the final domain was weakly related to the downstream task, the learned representations tended to deviate from features that are useful for the downstream task. As a result, performance degradation was observed.

Table 1. Experimental results of the proposed method and conventional state-of-the-art methods pretrained on the J-MID subset.

		SARS-CoV-2 CT-Scan Dataset			Chest CT-Scan Images Dataset		
Method	**Domain**	**ACC**	**AUC**	**F1**	**ACC**	**AUC**	**F1**
Ours	$D_1 \rightarrow D_2$	$\mathbf{0.873}_{\pm 0.030}$	$\mathbf{0.953}_{\pm 0.012}$	$\mathbf{0.873}_{\pm 0.030}$	$0.716_{\pm 0.042}$	$0.943_{\pm 0.002}$	$0.698_{\pm 0.054}$
MedCoSS [26]	$D_1 \rightarrow D_2$	$0.858_{\pm 0.020}$	$0.940_{\pm 0.016}$	$0.858_{\pm 0.020}$	$0.662_{\pm 0.020}$	$0.911_{\pm 0.008}$	$0.642_{\pm 0.015}$
Ours	$D_2 \rightarrow D_1$	$0.777_{\pm 0.010}$	$0.857_{\pm 0.016}$	$0.777_{\pm 0.010}$	$\mathbf{0.833}_{\pm 0.007}$	$\mathbf{0.965}_{\pm 0.004}$	$\mathbf{0.845}_{\pm 0.007}$
MedCoSS [26]	$D_2 \rightarrow D_1$	$0.763_{\pm 0.022}$	$0.847_{\pm 0.024}$	$0.761_{\pm 0.024}$	$0.822_{\pm 0.033}$	$0.962_{\pm 0.009}$	$0.833_{\pm 0.034}$
MAE [53]	$D_1 + D_2$	$0.784_{\pm 0.027}$	$0.871_{\pm 0.025}$	$0.783_{\pm 0.027}$	$0.807_{\pm 0.043}$	$0.964_{\pm 0.008}$	$0.804_{\pm 0.057}$
MAE [53]	D_1	$0.729_{\pm 0.015}$	$0.804_{\pm 0.014}$	$0.724_{\pm 0.015}$	$0.665_{\pm 0.061}$	$0.920_{\pm 0.003}$	$0.668_{\pm 0.063}$
MAE [53]	D_2	$0.760_{\pm 0.024}$	$0.831_{\pm 0.041}$	$0.756_{\pm 0.026}$	$0.519_{\pm 0.037}$	$0.874_{\pm 0.011}$	$0.509_{\pm 0.036}$
Baseline	None	$0.620_{\pm 0.041}$	$0.644_{\pm 0.050}$	$0.599_{\pm 0.055}$	$0.495_{\pm 0.067}$	$0.801_{\pm 0.023}$	$0.500_{\pm 0.060}$

When pretrained on the same-order domains, the proposed method consistently outperformed MedCoSS across all evaluation metrics. Although the LR mechanism and the WKD-BKE-integrated feature distillation introduced in this study do not eliminate the effects of catastrophic forgetting, the results indicate that they mitigate its negative impact. Furthermore, under the joint-learning scenario in which both domains were pretrained simultaneously, the proposed method also surpassed the MAE baseline. These findings suggest that, in CSSL, appropriately designed continual pretraining is more effective than simultaneous pretraining in reducing domain interference. In addition, Table 2 presents the results obtained after pretraining on the RICORD dataset. Except for the continual pretraining order from D_1 to D_2, these results are largely consistent with the trends observed on the J-MID dataset. Overall, these results demonstrate that the proposed LR-based CSSL framework effectively alleviates data interference during continual pretraining.

Table 2. Experimental results of the proposed method and conventional state-of-the-art methods pretrained on the RICORD dataset.

		SARS-CoV-2 CT-Scan Dataset			Chest CT-Scan Images Dataset		
Method	**Domain**	**ACC**	**AUC**	**F1**	**ACC**	**AUC**	**F1**
Ours	$D_1 \rightarrow D_2$	$\mathbf{0.844}_{\pm 0.034}$	$\mathbf{0.936}_{\pm 0.023}$	$\mathbf{0.843}_{\pm 0.034}$	$0.689_{\pm 0.103}$	$0.927_{\pm 0.031}$	$0.672_{\pm 0.148}$
MedCoSS [26]	$D_1 \rightarrow D_2$	$0.815_{\pm 0.045}$	$0.911_{\pm 0.027}$	$0.814_{\pm 0.046}$	$0.741_{\pm 0.061}$	$0.952_{\pm 0.019}$	$0.725_{\pm 0.079}$
Ours	$D_2 \rightarrow D_1$	$0.749_{\pm 0.020}$	$0.821_{\pm 0.024}$	$0.745_{\pm 0.021}$	$\mathbf{0.868}_{\pm 0.014}$	$\mathbf{0.980}_{\pm 0.002}$	$\mathbf{0.883}_{\pm 0.014}$
MedCoSS [26]	$D_2 \rightarrow D_1$	$0.748_{\pm 0.027}$	$0.814_{\pm 0.025}$	$0.744_{\pm 0.029}$	$0.818_{\pm 0.059}$	$0.973_{\pm 0.007}$	$0.827_{\pm 0.062}$
MAE [53]	$D_1 + D_2$	$0.760_{\pm 0.026}$	$0.844_{\pm 0.027}$	$0.757_{\pm 0.030}$	$0.709_{\pm 0.047}$	$0.940_{\pm 0.007}$	$0.717_{\pm 0.054}$
MAE [53]	D_1	$0.727_{\pm 0.023}$	$0.795_{\pm 0.007}$	$0.722_{\pm 0.023}$	$0.614_{\pm 0.028}$	$0.899_{\pm 0.016}$	$0.632_{\pm 0.030}$
MAE [53]	D_2	$0.768_{\pm 0.008}$	$0.843_{\pm 0.026}$	$0.765_{\pm 0.010}$	$0.474_{\pm 0.056}$	$0.854_{\pm 0.010}$	$0.465_{\pm 0.063}$
Baseline	None	$0.587_{\pm 0.019}$	$0.611_{\pm 0.017}$	$0.557_{\pm 0.030}$	$0.516_{\pm 0.037}$	$0.799_{\pm 0.026}$	$0.499_{\pm 0.061}$

In addition to performance evaluation, we analyzed the computational cost of the proposed method. During Stage 1 pretraining, the GPU memory consumption is 7731 MB, and the training time per epoch is 49.78 s. During Stage 3 continual pretraining, the GPU memory consumption is 5956 MB, and the training time per epoch is 59.94 s. Moreover, the inference time per epoch during the test phase of fine-tuning is 3.56 s. In our experiments, we used ViT-B as the backbone, resulting in a model with 86 M parameters. The batch size was set to 64 during Stage 1 pretraining and 32 during Stage 3 continual pretraining. With respect to processing requirements in hospital settings, the proposed framework can be deployed on standard GPU-equipped medical servers. In our experiments, con-

ducted using an NVIDIA GeForce RTX 4090 with approximately 24 GB of GPU memory, the proposed method maintained a moderate model size and GPU memory footprint. These results indicate that the proposed method enables efficient incremental updates without imposing excessive computational or memory overhead, making it suitable for deployment in real-world clinical environments.

4.3. Impact of Hyperparameters on the Experimental Results

To investigate feature distillation for mitigating data interference and handling distributional differences across stages, we explored optimal parameter settings to minimize deviations from the knowledge acquired in the previous stage during subsequent learning. Specifically, we examined the hyperparameter γ in the WKD loss, $\mathcal{L}_{\text{WKD}}$, and the batch size of BKE to determine their optimal values. The evaluation was on the COVID-19 classification task using the SARS-CoV-2 CT-Scan dataset.

In the proposed method, the hyperparameter in the WKD loss represents γ, which controls the relative contributions of the mean and covariance terms. Table 3 presents the classification results obtained with varying γ values when pretraining was performed on the J-MID subset. Table 4 presents the classification results with varying γ values when pretraining was performed on the RICORD dataset. For the proposed method, the optimal γ setting was 2.0 for the D_2–D_1 continual pretraining order and 3.0 for the D_1 to D_2 order. Accordingly, as γ increases, the mean term in the WKD loss exerts more significant influence, indicating that the mean plays a more crucial role than the covariance.

Table 3. Evaluation results on the SARS-CoV-2 CT-Scan dataset using the J-MID subset with varying hyperparameters, γ. Batch size was fixed at 32.

Domain	γ	ACC	AUC	F1
$D_1 \rightarrow D_2$	0.0	$0.844_{\pm 0.015}$	$0.932_{\pm 0.013}$	$0.844_{\pm 0.015}$
	1.0	$0.852_{\pm 0.012}$	$0.937_{\pm 0.017}$	$0.852_{\pm 0.012}$
	2.0	$\mathbf{0.873_{\pm 0.030}}$	$\mathbf{0.953_{\pm 0.012}}$	$\mathbf{0.873_{\pm 0.030}}$
	3.0	$0.852_{\pm 0.009}$	$0.939_{\pm 0.011}$	$0.852_{\pm 0.009}$
	4.0	$0.840_{\pm 0.023}$	$0.935_{\pm 0.019}$	$0.840_{\pm 0.023}$
$D_2 \rightarrow D_1$	0.0	$0.747_{\pm 0.016}$	$0.843_{\pm 0.012}$	$0.745_{\pm 0.016}$
	1.0	$0.752_{\pm 0.015}$	$0.844_{\pm 0.027}$	$0.751_{\pm 0.015}$
	2.0	$0.777_{\pm 0.010}$	$0.857_{\pm 0.016}$	$0.777_{\pm 0.010}$
	3.0	$0.774_{\pm 0.010}$	$0.863_{\pm 0.008}$	$0.774_{\pm 0.010}$
	4.0	$0.762_{\pm 0.021}$	$0.849_{\pm 0.021}$	$0.761_{\pm 0.015}$

Next, in the BKE of the proposed method, knowledge is propagated and ensembled based on the affinity between the feature representations replayed from the B and those generated by M_2 within mini-batches in the third learning stage. Tables 5 and 6 present the classification results with varying batch sizes when pretraining was performed on the J-MID subset and RICORD dataset, respectively. Notably, the optimal results were obtained with a batch size of 32 regardless of the utilized dataset. This behavior is closely related to the characteristics of the BKE. When the batch size is small, the diversity of feature representations jointly considered within BKE becomes limited. As a result, the estimation of batch-level feature statistics becomes unstable, potentially reducing the effectiveness of knowledge transfer. In contrast, when the batch size is large, the regularization induced by previously learned feature distributions becomes relatively stronger, which may bias the learned representations toward existing knowledge.

Table 4. Evaluation results on the SARS-CoV-2 CT-Scan dataset for the model pretrained using the RICORD dataset with varying γ. Batch size was fixed at 32.

Domain	γ	ACC	AUC	F1
$D_1 \rightarrow D_2$	0.0	$0.791_{\pm 0.041}$	$0.913_{\pm 0.030}$	$0.787_{\pm 0.042}$
	1.0	$0.804_{\pm 0.005}$	$0.904_{\pm 0.014}$	$0.803_{\pm 0.007}$
	2.0	$0.801_{\pm 0.050}$	$0.901_{\pm 0.017}$	$0.798_{\pm 0.055}$
	3.0	$\mathbf{0.844_{\pm 0.034}}$	$\mathbf{0.936_{\pm 0.023}}$	$\mathbf{0.843_{\pm 0.034}}$
	4.0	$0.832_{\pm 0.015}$	$0.916_{\pm 0.011}$	$0.831_{\pm 0.016}$
$D_2 \rightarrow D_1$	0.0	$0.749_{\pm 0.008}$	$0.822_{\pm 0.005}$	$0.746_{\pm 0.009}$
	1.0	$0.743_{\pm 0.018}$	$0.823_{\pm 0.018}$	$0.742_{\pm 0.033}$
	2.0	$0.739_{\pm 0.012}$	$0.820_{\pm 0.015}$	$0.735_{\pm 0.015}$
	3.0	$0.749_{\pm 0.020}$	$0.821_{\pm 0.024}$	$0.745_{\pm 0.021}$
	4.0	$0.757_{\pm 0.019}$	$0.833_{\pm 0.034}$	$0.755_{\pm 0.020}$

Overall, these findings indicate that the proposed method maintains robustness across a reasonable range of hyperparameter settings. The observed performance-variation trends with respect to the mean–covariance balance and batch size are consistent and interpretable, indicating that the proposed framework behaves stably and predictably under different configurations. This robustness demonstrates its practicality and reliability for continual pretraining across diverse medical-imaging domains.

Table 5. Evaluation results on the SARS-CoV-2 CT-Scan dataset for the model pretrained using the J-MID subset with varying batch sizes. γ was fixed at 2.0.

Domain	Batch Size	ACC	AUC	F1
$D_1 \rightarrow D_2$	16	$0.839_{\pm 0.019}$	$0.935_{\pm 0.008}$	$0.839_{\pm 0.019}$
	32	$\mathbf{0.873_{\pm 0.030}}$	$\mathbf{0.953_{\pm 0.012}}$	$\mathbf{0.873_{\pm 0.030}}$
	64	$0.858_{\pm 0.027}$	$0.933_{\pm 0.018}$	$0.858_{\pm 0.027}$
	128	$0.865_{\pm 0.009}$	$0.944_{\pm 0.008}$	$0.865_{\pm 0.009}$
$D_2 \rightarrow D_1$	16	$0.748_{\pm 0.009}$	$0.841_{\pm 0.006}$	$0.745_{\pm 0.030}$
	32	$0.777_{\pm 0.016}$	$0.857_{\pm 0.010}$	$0.777_{\pm 0.010}$
	64	$0.743_{\pm 0.017}$	$0.829_{\pm 0.006}$	$0.741_{\pm 0.006}$
	128	$0.739_{\pm 0.010}$	$0.825_{\pm 0.013}$	$0.737_{\pm 0.013}$

Table 6. Evaluation results on the SARS-CoV-2 CT-Scan dataset for the model pretrained on the RICORD dataset with varying batch sizes. γ was fixed at 3.0.

Domain	Batch Size	ACC	AUC	F1
$D_1 \rightarrow D_2$	16	$0.793_{\pm 0.037}$	$0.868_{\pm 0.031}$	$0.792_{\pm 0.038}$
	32	$\mathbf{0.844_{\pm 0.034}}$	$\mathbf{0.936_{\pm 0.023}}$	$\mathbf{0.843_{\pm 0.034}}$
	64	$0.832_{\pm 0.018}$	$0.930_{\pm 0.010}$	$0.832_{\pm 0.018}$
	128	$0.831_{\pm 0.031}$	$0.918_{\pm 0.020}$	$0.830_{\pm 0.031}$
$D_2 \rightarrow D_1$	16	$0.729_{\pm 0.005}$	$0.802_{\pm 0.003}$	$0.725_{\pm 0.003}$
	32	$0.749_{\pm 0.024}$	$0.821_{\pm 0.020}$	$0.745_{\pm 0.021}$
	64	$0.743_{\pm 0.013}$	$0.813_{\pm 0.014}$	$0.741_{\pm 0.013}$
	128	$0.750_{\pm 0.009}$	$0.837_{\pm 0.015}$	$0.748_{\pm 0.010}$

4.4. Ablation Studies

To evaluate the effectiveness of the proposed LR-based and WKD-BKE-integrated feature-distillation approaches, we performed an ablation study on the SARS-CoV-2 CT-Scan dataset, and Tables 7 and 8 present the results when pretraining was performed on the J-MID subset and RICORD dataset, respectively. The first row reveals the performance of

the baseline, which adopts an experience-replay-based approach with a *k*-means sampling strategy as well as performs feature distillation using only the mean-squared-error loss. The last row highlights the performance of the proposed method. We confirmed that replacing the experience-replay-based approach with LR in the proposed CSSL framework improved classification accuracy, as LR eliminated the dependence on raw image storage by replaying latent representations, thereby reducing the noise and redundancy that are inherent in pixel-level data. Consequently, the model retained informative and domain-invariant features more effectively, enhancing stability and knowledge retention during continual pretraining. Furthermore, although integrating LR with only WKD or BKE did not yield significant improvement, its incorporation with both techniques yielded substantial performance improvements. Specifically, WKD aligns the feature distributions between past and newly acquired representations to suppress domain-specific biases. Conversely, BKE exploits the similarity among feature representations within mini-batches and those replayed from the memory buffer B, facilitating feature-level knowledge propagation as well as stabilizing the optimization process. Additionally, their integration offers complementary benefits, where WKD preserves consistency across domains, and BKE promotes coherence within batches. This synergy enables the model to achieve more robust feature representations and enhanced classification accuracy.

Table 7. Results of the ablation studies on the latent replay (LR), Wasserstein distance (WD)-based knowledge distillation (WKD), and batch-knowledge ensemble (BKE) when the model was pretrained on the J-MID subset.

LR	WKD	BKE	ACC	AUC	F1
			$0.827_{\pm 0.020}$	$0.935_{\pm 0.001}$	$0.826_{\pm 0.021}$
✓			$0.833_{\pm 0.034}$	$0.936_{\pm 0.016}$	$0.832_{\pm 0.034}$
✓		✓	$0.827_{\pm 0.024}$	$0.929_{\pm 0.022}$	$0.826_{\pm 0.024}$
✓	✓		$0.845_{\pm 0.021}$	$0.934_{\pm 0.018}$	$0.845_{\pm 0.021}$
✓	✓	✓	$\mathbf{0.873_{\pm 0.030}}$	$\mathbf{0.953_{\pm 0.012}}$	$\mathbf{0.873_{\pm 0.030}}$

Table 8. Results of the ablation studies of LR, WKD, and BKE when the model was pretrained on the RICORD dataset.

LR	WKD	BKE	ACC	AUC	F1
			$0.804_{\pm 0.047}$	$0.905_{\pm 0.019}$	$0.803_{\pm 0.048}$
✓			$0.826_{\pm 0.052}$	$0.906_{\pm 0.031}$	$0.824_{\pm 0.054}$
✓		✓	$0.840_{\pm 0.035}$	$0.927_{\pm 0.023}$	$0.839_{\pm 0.037}$
✓	✓		$0.822_{\pm 0.047}$	$0.906_{\pm 0.038}$	$0.820_{\pm 0.050}$
✓	✓	✓	$\mathbf{0.844_{\pm 0.034}}$	$\mathbf{0.936_{\pm 0.023}}$	$\mathbf{0.843_{\pm 0.034}}$

Thus, the newly introduced feature distillation enhances the knowledge-retention capability of the model while maintaining its new-information adaptability. Overall, these results demonstrate the effectiveness of each component of the proposed method, underscoring their roles in mitigating data interference and improving continual-pretraining performance within the CSSL framework.

4.5. Impact of Stage Extension on Continual Pretraining

To examine the effect of progressive domain expansion during pretraining, we conducted four-stage continual pretraining on the RICORD dataset and J-MID subset, evaluating the resulting models on the SARS-CoV-2 CT-Scan dataset. The selected sequential training model was as follows: D_1–D_2 (RICORD), followed by D_1–D_2 (J-MID). This sequence was selected based on the results in Tables 1 and 2, which indicate that models pretrained on the RICORD dataset achieved lower accuracy on the SARS-CoV-2 CT-Scan

task compared with those pretrained on the J-MID subset. Therefore, we attempted to improve generalizability by first pretraining on the domains from the RICORD dataset before progressively expanding the pretraining to domains from the J-MID subset.

Table 9 summarizes the results of this extended-pretraining experiment. In Stage 1, where the model was pretrained only on RICORD D_1, it exhibited limited performance on the SARS-CoV-2 CT-Scan classification task. In Stage 2, further pretraining on RICORD D_2 significantly improved all metrics, particularly the AUC, indicating enhanced representation robustness through pretraining on multiple domains within RICORD. In Stage 3, where the model was further trained on J-MID D_1, its performance decreased slightly owing to the domain characteristics, as this D_1 is less similar to the target SARS-CoV-2 CT-Scan dataset, whereas D_2 shares more common features with the target domain. In Stage 4, following the incorporation of J-MID D_2, the model recovered and exhibited improved performance.

Table 9. Experimental results for the extended pretraining stage. Four-stage continual pretraining was conducted sequentially on RICORD and J-MID domains D_1 and D_2.

Stage	ACC	AUC	F1
Stage 1	$0.727_{\pm 0.023}$	$0.795_{\pm 0.007}$	$0.722_{\pm 0.023}$
Stage 2	$0.844_{\pm 0.034}$	$\mathbf{0.936_{\pm 0.023}}$	$0.843_{\pm 0.034}$
Stage 3	$0.754_{\pm 0.015}$	$0.849_{\pm 0.021}$	$0.754_{\pm 0.015}$
Stage 4	$\mathbf{0.857_{\pm 0.020}}$	$\mathbf{0.936_{\pm 0.020}}$	$\mathbf{0.857_{\pm 0.019}}$

Overall, these findings indicate that extending the pretraining stages enables the model to learn more diverse and transferable representations, which consequently improve its performance on the downstream task. These findings reveal that our approach can continually learn from images acquired from different domains across multiple medical institutions while preserving data privacy, thereby enabling more accurate, high-performance diagnostic capabilities.

5. Discussion

Our experimental results demonstrated the effectiveness and robustness of our method across two distinct domains of unlabeled chest CT images, namely the mediastinal and lung windows, using two publicly available datasets (J-MID and RICORD). The proposed framework outperformed existing SSL and CSSL approaches, exhibiting notable advantages. A major strength of the proposed method is its privacy-preserving design: the framework adopts LR, storing intermediate feature representations instead of original CT images. This strategy enables continual pretraining while eliminating the risk of sensitive data leakage, making the design suitable for dynamic clinical environments where imaging conditions and acquisition protocols change continually. The integration of LR with WKD-BKE was crucial to achieving an adaptation–retention balance. Our ablation studies confirmed that LR mitigated catastrophic forgetting by simultaneously retaining essential latent features and maintaining privacy protection. Moreover, the synergy between WKD-BKE exerted complementary effects. WKD aligned feature distributions across different domains, reducing domain-specific bias, whereas BKE stabilized the training process by promoting feature-level knowledge propagation among mini-batches. Collectively, these mechanisms facilitated the learning of representations that were both domain invariant and discriminative, thereby improving performance in the downstream classification task.

Despite these results, several challenges exist. A critical issue in continual pretraining is the significant dependence of model performance on the domain-exposure order: performance improves when the final domain is closely related to the evaluation data, decreasing otherwise. Therefore, for clinical applications, the relationship between the pretraining

data used and the target images must be clarified before applying the model. One direction for future work is the visualization of latent feature distributions under different domain-exposure orders. Such qualitative analyses may provide deeper insights into how the learned feature space evolves during CSSL. They may also clarify how later training stages influence the structure of the embedding space, leading to a more interpretable understanding of the observed order-dependent behavior. We also note that the current evaluation focuses on downstream tasks that are relatively close to the pretraining domains. As a next step, we plan to evaluate the proposed method on datasets exhibiting diverse domain shifts, such as inter-institutional variations across scanner vendors, reconstruction kernels, and noise levels. This evaluation will help clarify the generalizability of the proposed CSSL framework under more realistic and challenging conditions. Additionally, although the proposed method stores only representative latent features in B, memory management remains challenging in long-term continual pretraining as the number of domains increases. Future work may explore adaptive feature compression and dynamic memory-management strategies to improve efficiency. Closely related to this issue is the design choice of feature representations used for LR. In this study, LR uses features from the final encoder layer of a pretrained ViT, which capture high-level semantic information, while representations from earlier or intermediate layers tend to preserve local structural cues. The use of intermediate or multi-layer feature representations is left for future work. Furthermore, because the proposed framework explicitly stores and reuses latent feature representations, security considerations become important, especially in medical imaging. Although intermediate features are stored to retain knowledge from past domains, such representations may be vulnerable to feature inversion attacks [66]. As future work, we consider representing past-domain knowledge as statistical distributions in the latent feature space and performing LR using sampled features [67,68]. This design avoids storing raw images or sample-level feature vectors and is expected to improve robustness against feature inversion attacks.

Beyond these considerations, federated learning (FL) enables model training in distributed environments without centralizing data and has gained attention for privacy-preserving medical image analysis [69,70]. However, handling temporal changes in data distributions remains a challenge in federated settings. Recent studies have shown that integrating FL with SCL is a promising direction to address non-stationarity arising from evolving client data distributions [71,72]. Extending these insights to CSSL, we consider the integration of FL with CSSL as future work to support privacy-preserving adaptation to dynamic and heterogeneous data environments. We plan to address these limitations by further advancing the CSSL framework to improve scalability, robustness, and adaptability. Such extensions are expected to enhance generalization across diverse and evolving data distributions, ultimately enabling the development of more reliable and practical medical imaging models for real-world clinical deployment.

6. Conclusions

We proposed a privacy-aware CSSL framework to address the domain shifts in medical imaging. The method incorporates an LR mechanism with WKD-BKE-integrated feature distillation, thus effectively mitigating catastrophic forgetting and preserving data privacy simultaneously. Our experiments on multi-window chest-CT datasets demonstrated that our approach outperformed existing state-of-the-art SSL and CSSL methods, achieving superior robustness and generalizability across domains. However, several limitations persisted, including the dependence on domain-exposure order, challenges in long-term memory management, and considerations related to feature representation design and security. We plan to address these limitations by further extending the CSSL framework

toward more scalable and robust learning to establish more generalizable and privacy-preserving medical artificial intelligence systems.

Author Contributions: Conceptualization, R.T. (Ren Tasai), G.L., R.T. (Ren Togo) and T.O.; Methodology, R.T. (Ren Tasai) and G.L.; Software, R.T. (Ren Tasai); Validation, R.T. (Ren Tasai); Formal analysis, R.T. (Ren Tasai) and G.L.; Investigation, R.T. (Ren Tasai); Data curation, K.H., M.T., T.Y., H.S., N.N., Y.S., K.K. and R.T. (Ren Tasai); Writing—original draft preparation, R.T. (Ren Tasai); Writing—review and editing, R.T. (Ren Tasai), G.L., R.T. (Ren Togo), T.O., K.H., M.T., T.Y., H.S., N.N., Y.S. and K.K.; Visualization, R.T. (Ren Tasai); Supervision, G.L., R.T. (Ren Togo), T.O. and M.H.; Project administration, G.L.; Funding acquisition, G.L., R.T. (Ren Togo), T.O. and M.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly supported by JSPS KAKENHI Grant Numbers JP23K11141, JP23K21676, JP24K02942, JP24K23849, and JP25K21218, with additional support from AMED under Grant Number JP256f0137006.

Institutional Review Board Statement: This study was approved by the Research Ethics Committee, Faculty of Medicine, Juntendo University (10 September 2021).

Informed Consent Statement: The requirement for informed consent was waived by the committee because the study involved retrospective analysis of existing data and posed minimal risk to participants. All procedures were conducted in accordance with relevant ethical guidelines and regulations.

Data Availability Statement: The RICORD dataset, SARS-CoV-2 CT-Scan Dataset, and Chest CT-Scan Images Dataset used in this study are publicly available as cited in the notes and references. The J-MID database cannot be released.

Acknowledgments: We would like to thank the departments of radiology that provided the J-MID database, including Juntendo University, Kyushu University, Keio University, The University of Tokyo, Okayama University, Kyoto University, Osaka University, Hokkaido University, Ehime University, and Tokushima University.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SL	Supervised Learning
SSL	Self-Supervised Learning
CT	Computed Tomography
CSSL	Continual Self-Supervised Learning
SCL	Supervised Continual Learning
LR	Latent Replay
NNs	Neural Networks
WD	Wasserstein Distance
WKD	Wasserstein distance-based Knowledge Distillation
BKE	Batch-Knowledge Ensemble
MRI	Magnetic Resonance Imaging
ULM	Ultrasound Localization Microscopy
MAE	Masked AutoEncoder
COVID-19	Coronavirus Disease 2019
ACC	Accuracy
AUC	Area Under the receiver operating characteristic Curve
F1	F1-score
FL	Federated Learning

References

1. Pinto-Coelho, L. How artificial intelligence is shaping medical imaging technology: A survey of innovations and applications. *Bioengineering* **2023**, *10*, 1435. [CrossRef] [PubMed]
2. Maleki Varnosfaderani, S.; Forouzanfar, M. The role of AI in hospitals and clinics: Transforming healthcare in the 21st century. *Bioengineering* **2024**, *11*, 337. [CrossRef] [PubMed]
3. Zhao, Y.; Wang, X.; Che, T.; Bao, G.; Li, S. Multi-task deep learning for medical image computing and analysis: A review. *Comput. Biol. Med.* **2023**, *153*, 106496. [CrossRef]
4. Rayed, M.E.; Islam, S.S.; Niha, S.I.; Jim, J.R.; Kabir, M.M.; Mridha, M. Deep learning for medical image segmentation: State-of-the-art advancements and challenges. *Inform. Med. Unlocked* **2024**, *47*, 101504. [CrossRef]
5. Shobayo, O.; Saatchi, R. Developments in deep learning artificial neural network techniques for medical image analysis and interpretation. *Diagnostics* **2025**, *15*, 1072. [CrossRef]
6. Garcea, F.; Serra, A.; Lamberti, F.; Morra, L. Data augmentation for medical imaging: A systematic literature review. *Comput. Biol. Med.* **2023**, *152*, 106391. [CrossRef]
7. Li, Y.; Wynne, J.F.; Wu, Y.; Qiu, R.L.; Tian, S.; Wang, T.; Patel, P.R.; Yu, D.S.; Yang, X. Automatic medical imaging segmentation via self-supervising large-scale convolutional neural networks. *Radiother. Oncol.* **2025**, *204*, 110711. [CrossRef]
8. Singh, P.; Chukkapalli, R.; Chaudhari, S.; Chen, L.; Chen, M.; Pan, J.; Smuda, C.; Cirrone, J. Shifting to machine supervision: Annotation-efficient semi and self-supervised learning for automatic medical image segmentation and classification. *Sci. Rep.* **2024**, *14*, 10820. [CrossRef]
9. VanBerlo, B.; Hoey, J.; Wong, A. A survey of the impact of self-supervised pretraining for diagnostic tasks in medical X-ray, CT, MRI, and ultrasound. *BMC Med. Imaging* **2024**, *24*, 79. [CrossRef]
10. Zeng, X.; Abdullah, N.; Sumari, P. Self-supervised learning framework application for medical image analysis: A review and summary. *BioMedical Eng. OnLine* **2024**, *23*, 107. [CrossRef]
11. Li, G.; Togo, R.; Ogawa, T.; Haseyama, M. COVID-19 detection based on self-supervised transfer learning using chest X-ray images. *Int. J. Comput. Assist. Radiol. Surg.* **2022**, *18*, 715–722. [CrossRef] [PubMed]
12. Li, G.; Togo, R.; Ogawa, T.; Haseyama, M. TriBYOL: Triplet BYOL for Self-supervised Representation Learning. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 23–27 May 2022; pp. 3458–3462.
13. Zhao, Z.; Alzubaidi, L.; Zhang, J.; Duan, Y.; Naseem, U.; Gu, Y. Robust and explainable framework to address data scarcity in diagnostic imaging. *Comput. Biol. Med.* **2025**, *197*, 111052. [CrossRef] [PubMed]
14. Li, G.; Togo, R.; Ogawa, T.; Haseyama, M. Self-supervised learning for gastritis detection with gastric X-ray images. *Int. J. Comput. Assist. Radiol. Surg.* **2023**, *18*, 1841–1848. [CrossRef] [PubMed]
15. Li, G.; Togo, R.; Ogawa, T.; Haseyama, M. RGMIM: Region-Guided Masked Image Modeling for Learning Meaningful Representations from X-Ray Images. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Milan, Italy, 29 September–4 October 2024; pp. 148–157.
16. Ceccon, M.; Pezze, D.D.; Fabris, A.; Susto, G.A. Multi-label continual learning for the medical domain: A novel benchmark. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 28 February–4 March 2025; pp. 7163–7172.
17. Li, W.; Zhang, Y.; Zhou, H.; Yang, W.; Xie, Z.; He, Y. CLMS: Bridging domain gaps in medical imaging segmentation with source-free continual learning for robust knowledge transfer and adaptation. *Med. Image Anal.* **2025**, *100*, 103404. [CrossRef]
18. Liu, X.; Shih, H.A.; Xing, F.; Santarnecchi, E.; El Fakhri, G.; Woo, J. Incremental learning for heterogeneous structure segmentation in brain tumor MRI. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Vancouver, BC, Canada, 8–12 October 2023; pp. 46–56.
19. Wang, Q.; Tan, X.; Ma, L.; Liu, C. Dual windows are significant: Learning from mediastinal window and focusing on lung window. In Proceedings of the CAAI International Conference on Artificial Intelligence, Beijing, China, in 27–28 August 2022; pp. 191–203.
20. Lu, H.; Kim, J.; Qi, J.; Li, Q.; Liu, Y.; Schabath, M.B.; Ye, Z.; Gillies, R.J.; Balagurunathan, Y. Multi-window CT based radiological traits for improving early detection in lung cancer screening. *Cancer Manag. Res.* **2020**, *12*, 12225–12238. [CrossRef]
21. Mushtaq, E.; Yaldiz, D.N.; Bakman, Y.F.; Ding, J.; Tao, C.; Dimitriadis, D.; Avestimehr, S. CroMo-Mixup: Augmenting Cross-Model Representations for Continual Self-Supervised Learning. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024; pp. 311–328.
22. Purushwalkam, S.; Morgado, P.; Gupta, A. The challenges of continuous self-supervised learning. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 702–721.
23. Fini, E.; da Costa, V.G.T.; Alameda-Pineda, X.; Ricci, E.; Alahari, K.; Mairal, J. Self-supervised models are continual learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 9621–9630.

24. Cheng, H.; Wen, H.; Zhang, X.; Qiu, H.; Wang, L.; Li, H. Contrastive continuity on augmentation stability rehearsal for continual self-supervised learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 5684–5694.
25. Xinyao, W.; Zhe, X.; Raymond, Kai-yu, T. Continual learning in medical image analysis: A survey. *Comput. Biol. Med.* **2024**, *182*, 109206. [CrossRef]
26. Ye, Y.; Xie, Y.; Zhang, J.; Chen, Z.; Wu, Q.; Xia, Y. Continual self-supervised Learning: Towards universal multi-modal medical data representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 11114–11124.
27. Tasai, R.; Li, G.; Togo, R.; Tang, M.; Yoshimura, T.; Sugimori, H.; Hirata, K.; Ogawa, T.; Kudo, K.; Haseyama, M. Continual self-supervised learning considering medical domain knowledge in chest CT images. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025; pp. 1–5.
28. Madaan, D.; Yoon, J.; Li, Y.; Liu, Y.; Hwang, S.J. Representational continuity for unsupervised continual learning. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022; pp. 1–18.
29. Hu, D.; Yan, S.; Lu, Q.; Hong, L.; Hu, H.; Zhang, Y.; Li, Z.; Wang, X.; Feng, J. How well does self-supervised pre-training perform with streaming data? In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022; pp. 1–23.
30. French, R.M. Catastrophic interference in connectionist networks: Can it be predicted, can it be prevented? In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Denver, CO, USA, 29 November–2 December 1993; Volume 6, pp. 1176–1177.
31. Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A.A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 3521–3526. [CrossRef]
32. Rolnick, D.; Ahuja, A.; Schwarz, J.; Lillicrap, T.; Wayne, G. Experience replay for continual learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, USA, 8–14 December 2019; Volume 32, pp. 350–360.
33. Buzzega, P.; Boschini, M.; Porrello, A.; Calderara, S. Rethinking experience replay: A bag of tricks for continual learning. In Proceedings of the International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 2180–2187.
34. Ziller, A.; Usynin, D.; Braren, R.; Makowski, M.; Rueckert, D.; Kaissis, G. Medical imaging deep learning with differential privacy. *Sci. Rep.* **2021**, *11*, 13524. [CrossRef]
35. Sahiner, B.; Chen, W.; Samala, R.K.; Petrick, N. Data drift in medical machine learning: Implications and potential remedies. *Br. J. Radiol.* **2023**, *96*, 20220878. [CrossRef]
36. Hayes, T.L.; Kafle, K.; Shrestha, R.; Acharya, M.; Kanan, C. Remind your neural network to prevent catastrophic forgetting. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 466–483.
37. Srivastava, S.; Yaqub, M.; Nandakumar, K.; Ge, Z.; Mahapatra, D. Continual domain incremental learning for chest X-ray classification in low-resource clinical settings. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Workshop, Virtual, 27 September–1 October 2021; pp.226–238.
38. Wolf, D.; Payer, T.; Lisson, C.S.; Lisson, C.G.; Beer, M.; Götz, M.; Ropinski, T. Self-supervised pre-training with contrastive and masked autoencoder methods for dealing with small datasets in deep learning for medical imaging. *Sci. Rep.* **2023**, *13*, 20260. [CrossRef]
39. Jiang, J.; Rangnekar, A.; Veeraraghavan, H. Self-supervised learning improves robustness of deep learning lung tumor segmentation models to CT imaging differences. *Med. Phys.* **2025**, *52*, 1573–1588. [CrossRef] [PubMed]
40. Tasai, R.; Li, G.; Togo, R.; Tang, M.; Yoshimura, T.; Sugimori, H.; Hirata, K.; Ogawa, T.; Kudo, K.; Haseyama, M. Lung cancer classification using masked autoencoder pretrained on J-MID database. In Proceedings of the IEEE Global Conference on Consumer Electronics (GCCE), Kitakyushu, Japan, 29 October–1 November 2024; pp. 456–457.
41. Chang, X.; Cai, X.; Dan, Y.; Song, Y.; Lu, Q.; Yang, G.; Nie, S. Self-supervised learning for multi-center magnetic resonance imaging harmonization without traveling phantoms. *Phys. Med. Biol.* **2022**, *67*, 145004. [CrossRef] [PubMed]
42. Fiorentino, M.C.; Villani, F.P.; Benito Herce, R.; González Ballester, M.A.; Mancini, A.; López-Linares Román, K. An intensity-based self-supervised domain adaptation method for intervertebral disc segmentation in magnetic resonance imaging. *Int. J. Comput. Assist. Radiol. Surg.* **2024**, *19*, 1753–1761. [CrossRef] [PubMed]
43. Gu, R.; Wang, G.; Lu, J.; Zhang, J.; Lei, W.; Chen, Y.; Liao, W.; Zhang, S.; Li, K.; Metaxas, D.N.; et al. CDDSA: Contrastive domain disentanglement and style augmentation for generalizable medical image segmentation. *Med. Image Anal.* **2023**, *89*, 102904. [CrossRef]
44. Mojab, N.; Noroozi, V.; Yi, D.; Nallabothula, M.P.; Aleem, A.; Yu, P.S.; Hallak, J.A. Real-world multi-domain data applications for generalizations to clinical settings. In Proceedings of the IEEE International Conference on Machine Learning and Applications (ICMLA), Virtual, 14–17 December 2020; pp. 677–684.

45. Yu, X.; Luan, S.; Lei, S.; Huang, J.; Liu, Z.; Xue, X.; Ma, T.; Ding, Y.; Zhu, B. Deep learning for fast denoising filtering in ultrasound localization microscopy. *Phys. Med. Biol.* **2023**, *68*, 205002. [CrossRef]
46. Luan, S.; Yu, X.; Lei, S.; Ma, C.; Wang, X.; Xue, X.; Ding, Y.; Ma, T.; Zhu, B. Deep learning for fast super-resolution ultrasound microvessel imaging. *Phys. Med. Biol.* **2023**, *68*, 245023. [CrossRef]
47. Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; Wang, B. Moment matching for multi-source domain adaptation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1406–1415.
48. Aljundi, R.; Babiloni, F.; Elhoseiny, M.; Rohrbach, M.; Tuytelaars, T. Memory Aware Synapses: Learning What (not) to Forget. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 144–161.
49. Yao, Y.; Wu, R.; Zhou, Y.; Zhou, T. Continual Retinal Vision-Language Pre-training upon Incremental Imaging Modalities. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Daejeon, Republic of Korea, 23–27 September 2025; pp. 111–121.
50. Mansourian, A.M.; Ahmadi, R.; Ghafouri, M.; Babaei, A.M.; Golezani, E.B.; Ghamchi, Z.Y.; Ramezanian, V.; Taherian, A.; Dinashi, K.; Miri, A.; et al. A Comprehensive Survey on Knowledge Distillation. *arXiv* **2025**, arXiv:2503.12067. [CrossRef]
51. Gou, J.; Yu, B.; Maybank, S.J.; Tao, D. Knowledge distillation: A survey. *Int. J. Comput. Vis.* **2021**, *129*, 1789–1819. [CrossRef]
52. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021; pp. 1–21.
53. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked Autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009.
54. Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; Dosovitskiy, A. Do Vision Transformers See Like Convolutional Neural Networks? In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Online, 6–14 December 2021; pp. 12116–12128.
55. Lv, J.; Yang, H.; Li, P. Wasserstein distance rivals Kullback-Leibler divergence for knowledge distillation. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 10–15 December 2024; Volume 37, pp. 65445–65475.
56. Peyr'e, G.; Cuturi, M. Computational optimal transport. *Found. Trends Mach. Learn.* **2019**, *11*, 355–607. [CrossRef]
57. Yang, E.; Lozano, A.C.; Ravikumar, P. Elementary estimators for sparse covariance matrices and other structured moments. In Proceedings of the International Conference on Machine Learning (ICML), Beijing, China, 21–26 June 2014; pp. 397–405.
58. Tsai, E.B.; Simpson, S.; Lungren, M.P.; Hershman, M.; Roshkovan, L.; Colak, E.; Erickson, B.J.; Shih, G.; Stein, A.; Kalpathy-Cramer, J.; et al. The RSNA international COVID-19 open radiology database (RICORD). *Radiology* **2021**, *299*, E204–E213. [CrossRef]
59. Soares, E.; Angelov, P.; Biaso, S.; Froes, M.H.; Abe, D.K. SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification. *medRxiv* **2020**. [CrossRef]
60. Ge, Y.; Choi, C.L.; Zhang, X.; Zhao, P.; Zhu, F.; Zhao, R.; Li, H. Self-distillation with batch knowledge ensembling improves ImageNet classification. *arXiv* **2021**, arXiv:2104.13298.
61. Li, G.; Togo, R.; Ogawa, T.; Haseyama, M. Boosting automatic COVID-19 detection performance with self-supervised learning and batch knowledge ensembling. *Comput. Biol. Med.* **2023**, *158*, 106877. [CrossRef]
62. Li, G.; Togo, R.; Ogawa, T.; Haseyama, M. Self-knowledge distillation based self-supervised learning for COVID-19 detection from chest X-ray images. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 23–27 May 2022; pp. 1371–1375.
63. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019; pp. 1–18.
64. Caccia, L.; Aljundi, R.; Asadi, N.; Tuytelaars, T.; Pineau, J.; Belilovsky, E. New Insights on Reducing Abrupt Representation Change in Online Continual Learning. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022; pp. 1–27.
65. Sarfraz, F.; Arani, E.; Zonooz, B. Error Sensitivity Modulation based Experience Replay: Mitigating Abrupt Representation Drift in Continual Learning. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023; pp. 1–18.
66. Dosovitskiy, A.; Brox, T. Inverting visual representations with convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4829–4837.
67. Kumari, P.; Reisenbüchler, D.; Luttner, L.; Schaadt, N.S.; Feuerhake, F.; Merhof, D. Continual domain incremental learning for privacy-aware digital pathology. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Marrakesh, Morocco, 6–10 October 2024; pp. 34–44.

68. Lemke, N.; González, C.; Mukhopadhyay, A.; Mundt, M. Distribution-Aware Replay for Continual MRI Segmentation. In Proceedings of the International Workshop on Personalized Incremental Learning in Medicine, Marrakesh, Morocco, 10 October 2024; pp. 73–85.
69. Xia, Y.; Yu, W.; Li, Q. Byzantine-Resilient Federated Learning via Distributed Optimization. *arXiv* **2025**, arXiv:2503.10792. [CrossRef]
70. Zhu, H.; Togo, R.; Ogawa, T.; Haseyama, M. Prompt-based personalized federated learning for medical visual question answering. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 1821–1825.
71. Yang, X.; Yu, H.; Gao, X.; Wang, H.; Zhang, J.; Li, T. Federated Continual Learning via Knowledge Fusion: A Survey. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 3832–3850. [CrossRef]
72. Zhong, Z.; Bao, W.; Wang, J.; Chen, J.; Lyu, L.; Yang Bryan Lim, W. SacFL: Self-Adaptive Federated Continual Learning for Resource-Constrained End Devices. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 17169–17183. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.