Computerized Medical Imaging and Graphics 128 (2026) 102705

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

SPARSE data, rich results: Few-shot semi-supervised learning via class-conditioned image translation☆

Guido Manni [a,b,*], Clemente Lauretti [b], Loredana Zollo [b], Paolo Soda [a]

[a] Unit of Artificial Intelligence and Computer Systems, Department of Engineering, Università Campus Bio-Medico di Roma, Rome, Italy
[b] Unit of Advanced Robotics and Human-Centered Technologies, Department of Engineering, Università Campus Bio-Medico di Roma, Rome, Italy

ARTICLE INFO

Dataset link: https://github.com/GuidoManni/SPARSE, https://medmnist.com/

Keywords:
Semi-supervised learning
Few-shot learning
Medical imaging
Deep learning
GAN-based methods

ABSTRACT

Deep learning has revolutionized medical imaging, but its effectiveness is severely limited by insufficient labeled training data. This paper introduces a novel GAN-based semi-supervised learning framework specifically designed for low labeled-data regimes, evaluated across settings with 5 to 50 labeled samples per class. Our approach integrates three specialized neural networks: a generator for class-conditioned image translation, a discriminator for authenticity assessment and classification, and a dedicated classifier, within a three-phase training framework. The method alternates between supervised training on limited labeled data and unsupervised learning that leverages abundant unlabeled images through image-to-image translation rather than generation from noise. We employ ensemble-based pseudo-labeling that combines confidence-weighted predictions from the discriminator and classifier with temporal consistency through exponential moving averaging, enabling reliable label estimation for unlabeled data. Comprehensive evaluation across eleven MedMNIST datasets demonstrates that our approach achieves statistically significant improvements over six state-of-the-art GAN-based semi-supervised methods, with particularly strong performance in the extreme 5-shot setting where the scarcity of labeled data is most challenging. The framework maintains its superiority across all evaluated settings (5, 10, 20, and 50 shots per class). Our approach offers a practical solution for medical imaging applications where annotation costs are prohibitive, enabling robust classification performance even with minimal labeled data. Code is available at https://github.com/GuidoManni/SPARSE.

1. Introduction

Deep learning has demonstrated remarkable potential in revolutionizing medical imaging (Chen et al., 2022). However, insufficient labeled data for model training is one of the main challenges hindering its effectiveness that arises from several constraints such as: the stringent privacy regulations and ethical guidelines governing patient data access and distribution (Chiruvella and Guddati, 2021), and the necessity of specialized medical expertise for data annotation, a resource limited by healthcare professionals' primary commitment to patient care (Bull et al., 2015). These constraints often result in what is known as the low-data regime, i.e., a situation where the number of labeled medical images falls below the threshold needed for reliable convergence of deep networks, typically ranging from dozens to a few hundred samples depending on the complexity of the task and model architecture. To address this issue, researchers have explored unsupervised, supervised and semi-supervised learning. The approaches in the first category tackle the low-data regime through unsupervised learning, which exploits the abundant unlabeled medical images to learn meaningful representations. Models are trained to capture underlying data patterns through tasks like image reconstruction, anomaly detection, or feature learning. This approach is particularly valuable in medical imaging where unlabeled data is available, but a fundamental challenge remains: without sufficient labeled validation data, it is difficult to ensure that the extracted features are clinically relevant rather than merely statistically significant in the data distribution. In the second case of supervised learning, researchers have explored various strategies to mitigate the limited availability of labeled samples, such as transfer learning, data augmentation and synthetic data generation. Transfer learning leverages models pre-trained on large datasets by fine-tuning them for specific tasks with limited data (Ruffini et al., 2025), but its effectiveness diminishes when the target domain differs significantly from the source domain Bau et al. (2017). Data augmentation

☆ This article is part of a Special issue entitled: 'GenAI for Medical Imaging' published in Computerized Medical Imaging and Graphics.

* Corresponding author.
E-mail addresses: guido.manni@unicampus.it (G. Manni), c.lauretti@unicampus.it (C. Lauretti), l.zollo@unicampus.it (L. Zollo), p.soda@unicampus.it (P. Soda).

https://doi.org/10.1016/j.compmedimag.2026.102705
Received 8 August 2025; Received in revised form 12 December 2025; Accepted 7 January 2026
Available online 8 January 2026
0895-6111/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

techniques artificially expand training datasets through various transformations (Chen et al., 2023), but cannot introduce new information. Synthetic data generation through simulation or generative models has emerged as another strategy (Sariyildiz et al., 2023; Azizi et al., 2023), though generating fully synthetic datasets that realistically capture real-world data distributions remains challenging (Geng et al., 2025; Salmè et al., 2025). The third approach is semi-supervised learning (SSL) that simultaneously leverages labeled and unlabeled data. Traditional SSL methods relied on machine learning techniques such as self-training and co-training. However, recent advances in generative models, particularly Generative Adversarial Networks (GANs) based methods, have improved SSL performance, not by generating completely synthetic data but by learning to extract meaningful features from the real unlabeled data distribution to enhance the learning process (Sajun and Zualkernan, 2022). Despite these advances, existing SSL methods often struggle in low-data regimes as it happens in medical imaging. In this paper we tackle this issue and we introduce the following contributions:

- A novel GAN-based semi-supervised learning method specifically designed for medical image classification in low labeled-data regimes.
- A dynamic training schedule that alternates between supervised phases and unsupervised phases to optimize learning efficiency.
- An image-to-image translation mechanism employed as a secondary task that, unlike purely generative approaches that create images de novo from noise vectors, modifies existing real unlabeled images to preserve authentic anatomical features while enriching feature representations beyond what traditional generative approaches provide.
- A confidence-weighted temporal ensemble technique that combines predictions from multiple model components and previous training iterations, significantly improving pseudo-labeling reliability in low-data scenarios.
- A comprehensive empirical evaluation demonstrating competitive performance against six state-of-the-art SSL methods across eleven benchmark datasets for medical image classification tasks.

The remainder of this paper is organized as follows: Section 2 reviews the related works in the field, providing context for our research contributions. Section 3 details our proposed method. Section 4 describes the experimental configuration, including datasets, parameters, and evaluation metrics. Section 5 presents our results and provides comprehensive analysis. Finally, Section 6 concludes the paper with a summary of our findings and suggestions for future research directions.

2. Related works

Semi-supervised learning methods aim to leverage both labeled and unlabeled data to improve model performance, particularly in scenarios where labeled data is scarce or expensive to obtain. A recent survey (van Engelen and Hoos, 2020) has established a clear taxonomy of SSL approaches, distinguishing between two main classes: inductive methods, which construct classifiers that can generate predictions for any input, and transductive methods, which optimize directly over predictions for a given set of unlabeled data points. Inductive methods can be further subdivided into three categories based on how they incorporate unlabeled data: (1) wrapper methods, which iteratively train classifiers on labeled data and use their predictions to generate pseudo-labels for unlabeled samples; (2) unsupervised preprocessing methods, which extract features or determine initial parameters from unlabeled data before supervised training; and (3) intrinsically semi-supervised methods, which directly incorporate unlabeled data into the objective function or optimization procedure. Since our proposed approach falls within this third category, and it exploits GANs, the remainder of this section reviews the GAN-based SSL methods across various domains, focusing on approaches that have introduced key architectural innovations, while the interested readers can refer to Sajun and Zualkernan (2022) for a comprehensive review of approaches within this category. Semi-supervised learning is built on the fundamental assumption that the data distribution in the input space contains substantial information about label distribution in the output space. Within this context, GANs are particularly suitable candidates for SSL applications, given their inherent ability to model underlying data distributions and reveal patterns in the input space. The first significant work in this context introduced the SGAN model (Odena, 2016), which expands the traditional GAN architecture by augmenting the discriminator to perform dual functions that distinguish between real and synthetic samples while simultaneously predicting class labels for input data. This dual-purpose approach represented an important extension of the framework through pseudo-labeling, where the discriminator/classifier is trained on both labeled data and generated samples with known class labels. SGAN exemplifies what is known as a two-player model in GAN-based SSL, where the traditional generator-discriminator architecture is maintained but the discriminator is extended to perform both adversarial discrimination and classification tasks simultaneously. Building on these foundations, MatchGAN (Sun et al., 2020) introduced an innovative approach that leveraged the Wasserstein distance and conditional generation. As a semi-supervised conditional GAN, MatchGAN utilizes the label space in the target domain along with unlabeled samples to generate additional labeled training data. The framework assigns labels from the pool of labeled samples to unlabeled samples, then passes these through the generator to create synthetic versions of images based on the target labels. This work also introduces a match loss term that compares the generated images to the original labeled images from which the target labels are sampled. A breakthrough in GAN-based SSL came with the introduction of TripleGAN (Li et al., 2022), which addressed the difficulty of simultaneously optimizing both generator and discriminator performance. TripleGAN pioneered the three-player model architecture by incorporating an additional classifier that works independently from the discriminator, creating a tripartite interaction between generator, discriminator, and classifier. In this three-player setup, the classifier works in conjunction with the generator to characterize conditional distributions between images while limiting the discriminator's role to identifying fake image-label pairs. This separation of concerns allows each component to specialize in its primary task, potentially leading to improved overall performance compared to two-player models where the discriminator must balance competing objectives. Recent developments have significantly expanded upon the TripleGAN framework (Haque, 2021; Zhen et al., 2023; Xie et al., 2023). EC-GAN (Haque, 2021) proposed a mechanism where generated images are immediately processed by a classifier to produce pseudo-labels. This classifier-generator interaction is regulated through a hyperparameter-weighted loss function that precisely controls the influence of generated samples on classifier training. SEC-CGAN (Zhen et al., 2023) introduced a co-supervised learning paradigm where a conditional GAN is trained alongside the classifier, providing semantics-conditioned, confidence-aware synthesized examples during training. CISSL-GAN (Xie et al., 2023) then extended the Triple-GAN framework to address semi-supervised learning with class-imbalanced data through a dynamic class-rebalancing sampler that strategically selects pseudo-labeled samples from unlabeled data. Beyond GAN-based approaches, recent efforts in semi-supervised learning have focused on medical image segmentation tasks, introducing several innovative paradigms that offer valuable insights for SSL methodologies. VerSemi (Zeng et al., 2025a) proposes a versatile framework that unifies multiple SSL segmentation tasks within a single model using dynamic task prompts and CutMix-based synthetic task generation, demonstrating that learning across diverse tasks improves model generalization. In a complementary direction, PICK (Zeng et al., 2025b) employs masked image modeling as a pretext task, masking pseudo-label-guided attentive regions and reconstructing them to learn robust representations while mitigating the impact of erroneous pseudo-labels. Building on the

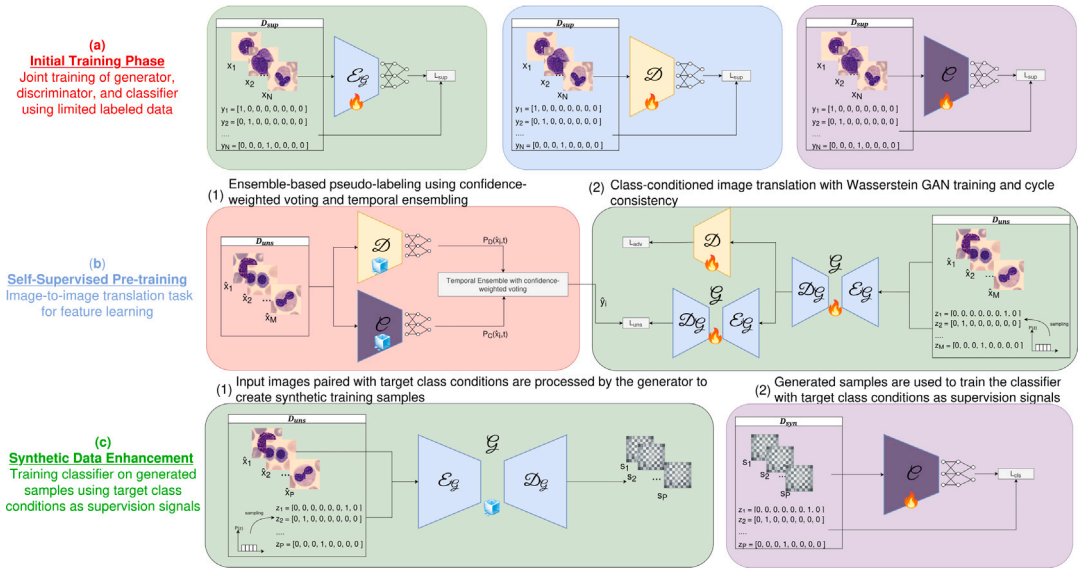

Fig. 1. Three-phase framework for semi-supervised learning with limited labeled data. Our approach integrates three specialized networks: a generator ($\mathscr{G}$) for class-conditioned image synthesis, a discriminator ($\mathscr{D}$) for authenticity assessment and classification signaling, and a dedicated classifier ($\mathscr{C}$) for the primary classification task. The generator comprises an encoder ($\mathscr{E}_{\mathscr{G}}$) and decoder ($\mathscr{D}_{\mathscr{G}}$) for image-to-image translation. (a) Initial Training Phase: Joint training of these three networks using limited labeled data. (b) Self-Supervised Pre-training: Two-part approach combining (1) ensemble-based pseudo-labeling using confidence-weighted voting and temporal ensembling, and (2) class-conditioned image translation with Wasserstein GAN training and cycle consistency. (c) Synthetic Data Enhancement Phase: Training classifier on generated samples using target class conditions as supervision signals, where (1) input images paired with target class conditions are processed by the generator to create synthetic training samples, and (2) generated samples are used to train the classifier with target class conditions as supervision signals. Fire symbols indicate trainable networks while ice symbols represent frozen weights during respective training phases.

theme of leveraging model discrepancies, the consistency-guided differential decoding approach (LeFeD) (Zeng et al., 2025c), exploits feature-level differences between multiple decoders to enhance representation learning, showing that differential features naturally emerge when decoders pursue consistent predictions. From a different perspective focused on classification, PEFAT (Zeng et al., 2023) introduces loss distribution modeling for pseudo-label quality assessment, using Gaussian Mixture Models to distinguish high-quality pseudo-labels from noisy ones while employing feature-level adversarial training to leverage low-confidence samples. These advances in segmentation demonstrate the value of ensemble mechanisms, multi-task learning, and sophisticated pseudo-labeling strategies, i.e., principles that resonate with our approach, though applied here to the distinct challenge of classification in extremely low labeled-data regimes. The analysis of the literature reported so far shows that several open issues still exist in current GAN-based SSL approaches:

- None of the existing methods specifically addresses the challenges of extremely low labeled data regimes, such as the case where only 5–10 labeled samples per class are available.
- Current approaches like SGAN (Odena, 2016), TripleGAN (Li et al., 2022), and EC-GAN (Haque, 2021) primarily rely on generation-based paradigms for unsupervised learning, lacking effective mechanisms to integrate supervised and unsupervised signals.
- Existing methods rely on a single discriminator or on a single classifier, while none has investigated possible advantages given by the use of an ensemble of models that, in other domains, has proven to provide complementary outputs that enhance model robustness (Cao et al., 2020; Guarrasi et al., 2022).

The next section introduces our methodology that addresses these limitations.

3. Methods

We propose a novel semi-supervised learning strategy called SPARSE (**S**emi-supervised **P**seudo-labeling via **A**dversarial **R**epresentation tran**S**lation **E**nhancement), designed to achieve robust classification performance in extremely low labeled-data regimes. Our approach integrates three specialized neural networks: a generator ($\mathscr{G}$) that performs class-conditioned image translation, a discriminator ($\mathscr{D}$) that assesses image authenticity while providing classification signals, and a dedicated classifier ($\mathscr{C}$) that focuses exclusively on the classification task. Our approach consists of three main phases:

1. Supervised training phase (Fig. 1a): it jointly trains the three aforementioned networks, $\mathscr{G}, \mathscr{D}$ and $\mathscr{C}$. Each model is trained with a supervised dataset $D_{sup} = \{(x_i, y_i)\}_{i=1}^{N}$, where N is the number of few-shot samples, x_i is the input sample, e.g., an image in our experiments, and y_i is the corresponding one-hot encoded ground truth vector, with $y_i \in \{0,1\}^{K\times 1}$ where K is the total number of classes. This initial phase is crucial for maintaining classification accuracy and preventing drift in the unsupervised learning process, represented in panel b of the same figure, by providing supervised signals from the limited labeled data.
2. Self-supervised pre-training phase (Fig. 1b), which consists of two components. The first is an ensemble-based pseudo-labeling block that combines the confidence-weighted scores provided by $\mathscr{D}$ and $\mathscr{C}$ given a set $D_{uns} = \{\hat{x}_i\}_{i=1}^{M}$ of M unsupervised samples, with $M \gg N$. This ensemble outputs the pseudo-labels $\{\hat{y}_i\}_{i=1}^{M}$, with $\hat{y}_i \in \{0,1\}^{K\times 1}$, assigned to each sample in D_{uns}. The second component of the self-supervised pretraining phase introduces a class-conditioned image translation task that uses randomly sampled class conditions $\{z_i\}_{i=1}^{M}$, with $z_i \in \{0,1\}^{K\times 1}$. Hence, this phase leverages the abundant unlabeled data to improve feature representations and model generalization capabilities through image-to-image translation tasks.
3. Synthetic data enhancement phase (Fig. 1c): it receives as input D_{uns} paired with one-hot encoded class vectors $\{z_i\}_{i=1}^{P}$ randomly sampled from a uniform distribution, which are processed by the generator $\mathscr{G}$ to create synthetic training samples $D_{syn} = \{s_i\}_{i=1}^{P}$, with $P \gg N$. Subsequently, D_{syn} is used to train the classifier

$\mathscr{C}$ with these same class vectors $\{\hat{z}_i\}_{i=1}^{P}$ serving as supervision signals. This final phase aims to expand the effective training set by creating synthetic samples that augment the limited labeled data.

The training schedule alternates between two phases: the supervised phase (executed at every epoch) and the combined self-supervised and synthetic data enhancement phase (executed every μ epochs, where μ is a hyperparameter), ensuring stable and effective utilization of the entire dataset.

It is worth noting that our approach addresses two primary issues in semi-supervised learning in an extremely low labeled-data regime. The first is the availability of insufficient labeled samples for effective supervised learning, which may affect panel (a) of Fig. 1: the self-supervised pretraining in panel (b) synthesizes new samples that are used in panel (c) to train the classifier with a large amount of samples. The second issue concerns the integration of supervised and unsupervised learning signals. While the paradigm in the literature (Odena, 2016; Li et al., 2022; Haque, 2021) defines pre-tasks where a generative model conditionally generates new samples that enhance the downstream classification task, our approach integrates an image-to-image translation pre-task that enriches the model with more semantic information than a standard generative step (Fig. 1b).

The rest of this section details such three phases: next Section 3.1 presents panel (a) of Fig. 1, whilst Section 3.2 describes both panels (b) and (c) of the same f

3.1. Supervised training phase

The supervised phase in (Fig. 1a) is crucial for maintaining classification accuracy and preventing drift in the unsupervised learning process. It occurs at every epoch, utilizing the limited labeled data to train simultaneously the encoder of the generator $E_{\mathscr{G}}$, $\mathscr{D}$ and $\mathscr{C}$ to perform classification. We leverage deep supervision by adding a classification tail to the bottleneck of the generator's encoder $E_{\mathscr{G}}$ denoted as a set of interconnected neurons in panel (a) of Fig. 1; it is located at the bottleneck because it serves as the information compression point between the encoder and decoder paths. By applying classification supervision at this critical juncture, we ensure that the most compact representation in the network encodes both structural information needed for generation and semantic information required for classification.

The supervised loss function used by all the three networks ($\mathcal{L}_{sup}$) combines four specialized loss terms, each addressing a specific challenge in few-shot learning and weighted by coefficients to balance their contribution.

$$\mathcal{L}_{sup} = \mathcal{L}_{prototype} + \alpha\mathcal{L}_{mutual} + \beta\mathcal{L}_{entropy} + \gamma\mathcal{L}_{mixup} \tag{1}$$

The prototype loss ($\mathcal{L}_{prototype}$) helps create discriminative class-specific features by learning robust prototypical representations for each class. The mutual learning loss ($\mathcal{L}_{mutual}$) enables knowledge sharing between the three models, leveraging their complementary perspectives on the data. The entropy minimization loss ($\mathcal{L}_{entropy}$) encourages the models to make confident predictions, helping combat the uncertainty inherent in limited-data scenarios. Finally, the mixup loss ($\mathcal{L}_{mixup}$) provides regularization through data augmentation, helping prevent overfitting which is particularly crucial when training with few samples. The rest of this section details the computation of each of these four loss functions.

The prototype loss $\mathcal{L}_{prototype}$ creates discriminative class-specific representations by comparing softmax probabilities with class prototypes:

$$\mathcal{L}_{prototype} = -\sum_{i=1}^{N} \log \frac{\exp(-d(p_T(x_i), c_{y_i}))}{\sum_{k=1}^{K} \exp(-d(p_T(x_i), c_k))} \tag{2}$$

where $p_T(x_i)$ represents the softmax probabilities of input x_i with temperature T, c_k is the prototype of class k computed as the mean of the class probabilities:

$$c_k = \frac{1}{|S_k|} \sum_{x_i \in S_k} p_T(x_i) \tag{3}$$

where S_k is the set of samples from class k. The distance function $d(\cdot,\cdot)$ is defined as the negative sum of squared differences:

$$d(p_T(x), c_k) = -\sum_{j=1}^{K} (p_{T,j}(x) - c_{k,j})^2 \tag{4}$$

where K is the total number of classes.

The mutual learning loss $\mathcal{L}_{mutual}$ facilitates knowledge transfer between models through a combination of supervised cross-entropy and KL divergence:

$$\mathcal{L}_{mutual} = \mathcal{L}_{ce} + \lambda_{kl}\mathcal{L}_{kl} \tag{5}$$

where λ_{kl} is the weight coefficient for the KL divergence term, $\mathcal{L}_{ce}$ is the sum of cross-entropy losses for each model:

$$\mathcal{L}_{ce} = \sum_{m \in \{E_{\mathscr{G}}, \mathscr{D}, \mathscr{C}\}} \mathrm{CE}(p_m(x), y) \tag{6}$$

and $\mathcal{L}_{kl}$ is the symmetric KL divergence between each model's probabilities and the average of other models' probabilities:

$$\mathcal{L}_{kl} = \sum_{m \in \{E_{\mathscr{G}}, \mathscr{D}, \mathscr{C}\}} \mathrm{KL}(p_m(x) | \frac{1}{2} \sum_{n \neq m} p_n(x)) \tag{7}$$

The entropy minimization loss $\mathcal{L}_{entropy}$ promotes confident predictions:

$$\mathcal{L}_{entropy} = -\sum_{k=1}^{K} p_k(x) \log(p_k(x)) \tag{8}$$

Finally, the mixup loss $\mathcal{L}_{mixup}$ provides regularization by training on interpolated samples and labels. For each pair of samples (x_i, y_i) and (x_j, y_j), the interpolation weight λ_{mix} is sampled from a Beta distribution:

$$\lambda_{mix} \sim \mathrm{Beta}(\alpha_{mix}, \alpha_{mix}) \tag{9}$$

The Beta distribution is particularly suitable for generating interpolation weights as it is bounded between [0,1] and can be symmetric around 0.5, ensuring a balanced mixing of samples while maintaining their relative contributions. The hyperparameter α_{mix} in the Beta distribution controls the strength of interpolation, i.e., higher values of α_{mix} lead to interpolation weights closer to 0.5, while lower values favor weights closer to 0 or 1. This weight is then used to create interpolated samples and labels:

$$\tilde{x} = \lambda_{mix} x_i + (1 - \lambda_{mix}) x_j \tag{10}$$

$$\tilde{y} = \lambda_{mix} y_i + (1 - \lambda_{mix}) y_j \tag{11}$$

The mixup loss is then computed as:

$$\mathcal{L}_{mixup} = \mathcal{L}_{ce}(p(\tilde{x}), \tilde{y}) \tag{12}$$

where $\mathcal{L}_{ce}$ is the cross-entropy loss.

3.2. Unsupervised training phase

The unsupervised training phase, illustrated in panel (b) of Fig. 1, executes every μ epochs. This phase leverages unlabeled data through an image-to-image translation framework with three distinct stages.

First, we use an ensemble-based pseudo-labeling mechanism to estimate class probabilities for unlabeled samples (Section 3.2.1). These probability estimates then guide a class-conditioned image translation process, which learns to generate class-specific variations of input images (Section 3.2.2). Finally, we employ these generated samples in a synthetic data enhancement phase. Here, the synthetic images serve as additional training data to strengthen the classifier's ability to distinguish between classes (Section 3.2.3).

3.2.1. Ensemble-based pseudo-labeling

To effectively utilize unlabeled samples, we require reliable class probability estimates. Our approach addresses three key challenges: (1) quantifying model uncertainty, (2) aggregating predictions from multiple models, and (3) maintaining temporal stability throughout training. We tackle these challenges through an ensemble mechanism (Fig. 1b) that integrates confidence-weighted voting, temporal ensembling, and adaptive thresholding.

At epoch t, each unlabeled image $\hat{x}_i$ from $D_{uns} = \{\hat{x}_i\}_{i=1}^{M}$ is processed by the models trained during the previous initial training phase. We apply temperature scaling to obtain calibrated class probabilities:

$$p_m(\hat{x}_i, t) = \text{softmax}\left(\frac{l_m(\hat{x}_i, t)}{T}\right) \tag{13}$$

where $l_m(\hat{x}_i, t)$ represents the **logits** (raw pre-softmax outputs) from model $m \in \{\mathscr{D}, \mathscr{C}\}$ for input $\hat{x}_i$ at epoch t. The temperature parameter T controls prediction sharpness: lower values ($T < 1$) yield more confident predictions, while higher values ($T > 1$) produce smoother probability distributions.

Each model's prediction reliability is quantified using an entropy-based confidence measure:

$$c_m(\hat{x}_i, t) = 1 - \frac{H(p_m(\hat{x}_i, t))}{H_{max}} \tag{14}$$

where the entropy $H(p_m(\hat{x}_i, t))$ is computed as:

$$H(p_m(\hat{x}_i, t)) = -\sum_{k=1}^{K} p_m^{(k)}(\hat{x}_i, t) \log p_m^{(k)}(\hat{x}_i, t) \tag{15}$$

and $H_{max} = \log K$ represents the maximum possible entropy for K classes. The term $p_m^{(k)}(\hat{x}_i, t)$ denotes model m's predicted probability for class k. This confidence score ranges from 0 (complete uncertainty) to 1 (complete certainty).

We combine predictions from both models by weighting each according to its confidence score:

$$p_{weighted}(\hat{x}_i, t) = \frac{\sum_{m \in \{\mathscr{D}, \mathscr{C}\}} c_m(\hat{x}_i, t) \cdot p_m(\hat{x}_i, t)}{\sum_{m \in \{\mathscr{D}, \mathscr{C}\}} c_m(\hat{x}_i, t)} \tag{16}$$

The denominator ensures proper normalization. This weighting scheme allows more confident models to contribute more strongly to the final prediction.

To enhance prediction stability, we incorporate historical information by blending current predictions with past predictions:

$$p_{ens}(\hat{x}_i, t) = \alpha \cdot p_{weighted}(\hat{x}_i, t) + (1 - \alpha) \cdot p_{ema}(\hat{x}_i, t - 1) \tag{17}$$

Here, $p_{ema}(\hat{x}_i, t-1)$ captures the temporal history through an exponential moving average (EMA) of past ensemble predictions. The parameter $\alpha \in [0, 1]$ balances current information (higher α) against historical stability (lower α). This temporal smoothing prevents abrupt prediction changes that could destabilize the training process.

After computing the current ensemble prediction, we update the EMA:

$$p_{ema}(\hat{x}_i, t) = \beta \cdot p_{ema}(\hat{x}_i, t - 1) + (1 - \beta) \cdot p_{ens}(\hat{x}_i, t) \tag{18}$$

The momentum parameter $\beta \in [0, 1]$ determines the temporal memory span. Large values (e.g., $\beta = 0.99$) maintain longer memory for stable predictions, while smaller values allow faster adaptation to recent changes.

With reliable probability estimates in hand, we select only the most confident predictions for pseudo-labeling:

$$S(t) = \{\hat{x}_i \in D_{uns} : \max_k(p_{ens}^{(k)}(\hat{x}_i, t)) > \tau(t)\} \tag{19}$$

where $\max_k(p_{ens}^{(k)}(\hat{x}_i, t))$ is the highest class probability from the ensemble prediction. The subset $S(t)$ contains selected samples at epoch t, and the adaptive threshold $\tau(t)$ is defined as:

$$\tau(t) = Q_\rho(\{\max_k(p_{ens}^{(k)}(\hat{x}_i, t)) : \hat{x}_i \in D_{uns}\}) \tag{20}$$

Here, Q_ρ denotes the ρ-th percentile of maximum probabilities across all unlabeled samples in the current batch, where $\rho \in [0, 1]$ is the percentile threshold parameter.

This percentile-based approach automatically adjusts to the model's current performance level. It selects $(1-\rho)\times 100\%$ of the most confident samples. For instance, setting $\rho = 0.8$ selects the top 20% most confident predictions. This adaptive mechanism prevents error accumulation from unreliable pseudo-labels while naturally accommodating the model's improving performance.

For each selected confident sample $\hat{x}_i \in S(t)$, we create a discrete pseudo-label:

$$\hat{y}_i = \text{one-hot}(\arg\max_k p_{ens}^{(k)}(\hat{x}_i, t)) \tag{21}$$

This produces a one-hot encoded pseudo-label $\hat{y}_i \in \{0, 1\}^{K\times 1}$ for each unlabeled sample $\hat{x}_i$. The complete set of pseudo-labels $\{\hat{y}_i\}_{i=1}^{|S(t)|}$ for all selected samples then feeds into the subsequent class-conditioned image translation process.

3.2.2. Class-conditioned image translation

Using the pseudo-labels $\{\hat{y}_i\}_{i=1}^{M}$ obtained from our ensemble mechanism, we implement a class-conditioned image translation process that leverages $\mathscr{G}$ and $\mathscr{D}$ working in tandem (Fig. 1b right). Now the generator $\mathscr{G}$, which consists of a complete U-Net architecture, not just the encoder as in the supervised phase, learns to perform class-conditioned image translation while preserving semantic features relevant to classification. Its training objective $\mathcal{L}_{uns}$ combines four components:

$$\mathcal{L}_{uns} = \mathcal{L}_{adv} + \mathcal{L}_{cls}^{\mathscr{D}} + \mathcal{L}_{cls}^{\mathscr{G}} + \lambda_{rec}\mathcal{L}_{rec} \tag{22}$$

where λ_{rec} is the weight coefficient for the reconstruction loss, and the adversarial loss $\mathcal{L}_{adv}$ uses the Wasserstein distance metric to assess image realism:

$$\mathcal{L}_{adv} = -\mathbb{E}_{\hat{x}\sim D_{uns}}[\mathscr{D}(\mathscr{G}(\hat{x}, z_{target}))], \tag{23}$$

Next, the classification losses $\mathcal{L}_{cls}^{\mathscr{D}}$ and $\mathcal{L}_{cls}^{\mathscr{G}}$ ensure accurate conditioning on target classes using cross-entropy from both the discriminator and generator classifiers:

$$\mathcal{L}_{cls}^{\mathscr{D}} = -\mathbb{E}_{\hat{x}\sim D_{uns}} \sum_{k=1}^{K} z_{target}^{k} \log(p_k^{\mathscr{D}}(\mathscr{G}(\hat{x}, z_{target}))) \tag{24}$$

$$\mathcal{L}_{cls}^{\mathscr{G}} = -\mathbb{E}_{\hat{x}\sim D_{uns}} \sum_{k=1}^{K} z_{target}^{k} \log(p_k^{\mathscr{G}}(\mathscr{G}(\hat{x}, z_{target}))) \tag{25}$$

where z_{target}^{k} is the kth element of the target class one-hot encoding, $p_k^{\mathscr{D}}$ represents the probability for class k from the discriminator, and $p_k^{\mathscr{G}}$ represents the probability for class k from the generator's classifier. The reconstruction loss $\mathcal{L}_{rec}$ maintains content consistency using L1 distance:

$$\mathcal{L}_{rec} = \mathbb{E}_{\hat{x}\sim D_{uns}}[|\mathscr{G}(\mathscr{G}(\hat{x}, z_{target}), z_{source}) - \hat{x}|_1] \tag{26}$$

where z_{source} is the one-hot encoding of the most probable class according to $p_{ens}(\hat{x}, t)$, connecting this translation process with our previous pseudo-labeling step.

The discriminator $\mathscr{D}$ serves a dual role: it assesses image realism through a Wasserstein distance metric while also providing classification signals. Its objective function reflects these dual tasks:

$$\mathcal{L}_{\mathscr{D}} = -\mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{gp}\mathcal{L}_{gp} \tag{27}$$

where λ_{cls} and λ_{gp} are weight coefficients for the classification loss and gradient penalty terms, respectively. Its adversarial component implements the Wasserstein distance as follows:

$$-\mathcal{L}_{adv} = -\mathbb{E}_{\hat{x}\sim D_{uns}}[\mathscr{D}(\hat{x})] + \mathbb{E}_{\hat{x}\sim D_{uns}}[\mathscr{D}(\mathscr{G}(\hat{x}, z_{target}))] \tag{28}$$

The gradient penalty term $\mathcal{L}_{gp}$ enforces the Lipschitz constraint:

$$\mathcal{L}_{gp} = \mathbb{E}_{\hat{x}_{interp}}[(|\nabla_{\hat{x}_{interp}}\mathscr{D}(\hat{x}_{interp})|_2 - 1)^2] \tag{29}$$

where $\hat{x}_{interp}$ is sampled uniformly along straight lines between pairs of real and generated images.

3.2.3. Synthetic data enhancement

After establishing the image translation process, we leverage translated images as a form of synthetic labeled data to enhance the classifier's performance (Fig. 1c). The key process involves taking unlabeled input images $\{\hat{x}_i\}_{i=1}^{M}$ and sampling random target classes $\{z_i\}_{i=1}^{P}$ from a uniform distribution over the K classes in one-hot format, where P is the number of synthetic samples generated. We then use these randomly sampled classes to condition the generator to translate the original images of unknown classes. The resulting translated images have known target classes $\{z_i\}_{i=1}^{P}$.

These translated images with known target classes can then be used to train the classifier in a supervised manner:

$$\mathcal{L}_{\mathscr{C}} = \mathcal{L}_{cls}(\mathscr{C}(\mathscr{G}(\hat{x}, z)), z) \tag{30}$$

where z represents the target class condition provided to the generator for synthetic data generation. This approach provides us with effectively labeled training samples, since we know exactly what class condition was used to generate each image. Importantly, as demonstrated in Dai et al. (2017), the visual quality of generated images is not critical for our classification objective: instead, we focus on ensuring that the translation process captures and preserves discriminative features that are useful for classification.

3.3. Inference configuration

Our approach in the inference phase can be deployed in different configurations that exploit the model trained as reported in Section 3. In particular, in the rest of the manuscript we consider the following two setups for inference:

- SPARSE: it uses only the classifier $\mathscr{C}$.
- SPARSE$_{\text{ens}}$: it exploits both the discriminator $\mathscr{D}$ and the classifier $\mathscr{C}$, which are combined in late fusion by averaging the estimates of posterior probabilities per class.

It is worth noting that in our ensemble configuration, we should also consider the potential use of the generator's encoder $\mathscr{E}_{\mathscr{G}}$. However, since it is trained to minimize a loss function that balances both generation and classification objectives, this results in a performance degradation, a finding we experimentally verified, though omitted from the manuscript for conciseness.

4. Experimental configuration

This section describes our experimental methodology: it details the materials (Section 4.1), followed by our training configuration (Section 4.2).

4.1. Materials

We conducted experiments using eleven datasets from MedMNIST repository (Yang et al., 2023): BloodMNIST, BreastMNIST, ChestMNIST, DermaMNIST, OCTMNIST, OrganAMNIST, OrganCMNIST, OrganSMNIST, PathMNIST, PneumoniaMNIST and TissueMNIST. As shown in Table 1, these datasets span different medical imaging modalities and classification tasks, with varying number of classes (2-11) and dataset sizes (from hundreds to hundreds of thousands of samples).

To evaluate our method's effectiveness in extremely low labeled-data regimes, we conducted experiments across four few-shot settings: 5-shot, 10-shot, 20-shot, and 50-shot per class, with 5-shot representing the most challenging scenario. For each N-shot setting, we constructed the training set using N labeled samples per class, with the remaining samples treated as unlabeled data. For data preprocessing, we applied a transformation pipeline consisting of random horizontal flipping for data augmentation and tensor conversion, with input images maintaining their original 128×128 pixel resolution. To ensure reproducibility, we utilized the original validation/test splits provided by the MedMNIST authors.

4.2. Training configuration

The training schedule alternates between supervised and unsupervised learning phases. The supervised phase occurs at every epoch, while the unsupervised phase is executed every μ epochs as already described in Section 3. All models were trained for 1000 epochs using the AdamW optimizer, maintaining identical configurations throughout all experiments which are reported in Table 2. During training, we implemented a model checkpoint strategy that saved the model state achieving the best validation accuracy, which was then used for final evaluation. For all the models, we did not investigate any hyperparameter configuration since their tuning is out of the scope of this manuscript. Nevertheless, because the 'No Free Lunch' Theorem for optimization (Wolpert and Macready, 1997) states that no universal set of hyperparameters will optimize a model's performance across all possible datasets, this approach ensures a fair comparison among all the approaches.

5. Results

This section presents the experimental results evaluating our framework's effectiveness in extremely low labeled data regimes in medical image classification. We compare our approach against the six state-of-the-art semi-supervised GAN architectures which are already presented in Section 2, using their original implementations without additional optimization or modifications to ensure fair comparison with our method. To assess model performance, we employ classification accuracy per class as our primary evaluation metric. For each dataset, we first compute the accuracy across all available classes and then the final score is obtained by averaging these accuracies across all datasets, providing a robust evaluation of the model's generalization capabilities across varying medical imaging tasks, modalities, and data distributions. Individual results for each dataset are presented in the appendix.

Table 3 presents the performance of all models across different few-shot settings, where we systematically vary the number of samples per class from extremely limited (5-shot) to more moderate (50-shot) situations. The first two rows show the results of our approach using only the classifier $\mathscr{C}$ (SPARSE) or using the ensemble between the discriminator $\mathscr{D}$ and $\mathscr{C}$ (SPARSE$_{\text{ens}}$). The following six rows display the results of the six competitors. We note that the two variants of our approach consistently outperform existing competitors across all settings, and the ensemble setting (SPARSE$_{\text{ens}}$) achieves the best performance in all scenarios. As the number of labeled samples increases from 5 to 50 shots per class, we observe a steady improvement in performance across all models, though the relative advantage of our approach remains consistent.

To further investigate these findings, the rest of this section examines our improvements in the 5-shot setting (Section 5.1), which represents the most challenging scenario where the extreme scarcity of labeled data tests the true effectiveness of semi-supervised learning. We then complement the results with a detailed analysis in the 50-shot setting (Section 5.2), which helps us understand how our method scales when more labeled data becomes available. Statistical significance testing of the performance differences is presented within these analyses. Finally, investigates how the frequency of unsupervised training affects model performance.

5.1. Performance in 5-shot

We now deepen our analysis on the most challenging 5-shot learning scenario, where only 5 labeled samples per class are available for training while the remaining samples are treated as unlabeled. To provide robust statistical evaluation, we employ the Wilcoxon signed-rank test, a non-parametric method that compares paired accuracy values across our 11 datasets, with Benjamini–Hochberg FDR correction for multiple

Table 1
Characteristics of MedMNIST datasets used in our experiments.

Dataset	Modality	Task type	# Samples		
			Total	Training	Val/Test
BloodMNIST	Blood Cell Microscope	Multi-Class (8)	17,092	11,959	1,712/3,421
BreastMNIST	Breast Ultrasound	Binary-Class (2)	780	546	78/156
ChestMNIST	Chest X-ray	Binary-Class (2)	112,120	78,468	11,219/22,433
DermaMNIST	Dermatoscope	Multi-Class (7)	10,015	7007	1,003/2,005
OCTMNIST	Retinal OCT	Multi-Class (4)	109,309	97,477	10,832/1,000
OrganAMNIST	Abdominal CT	Multi-Class (11)	58,830	34,561	6,491/17,778
OrganCMNIST	Abdominal CT	Multi-Class (11)	23,583	12,975	2,392/8,216
OrganSMNIST	Abdominal CT	Multi-Class (11)	25,211	13,932	2,452/8,827
PathMNIST	Colon Pathology	Multi-Class (9)	107,180	89,996	10,004/7,180
PneumoniaMNIST	Chest X-ray	Binary-Class (2)	5856	4708	524/624
TissueMNIST	Kidney Cortex Microscope	Multi-Class (8)	236,386	165,466	23,640/47,280

Table 2
Training hyperparameters used in our experiments.

Parameter	Value
General Training Parameters	
Training Epochs	1000
Base Learning Rate	0.0002
Optimizer	AdamW
μ (unsupervised phase frequency)	10
T (temperature parameter)	2.0
Loss Weights for Supervised Objective	
α (mutual learning loss weight)	0.1
β (entropy minimization loss weight)	0.01
γ (mixup loss weight)	0.5
λ_{kl} (KL divergence weight)	0.5
mixup parameters	
α_{mix} (beta distribution parameter)	0.2
loss weights for generator	
λ_{rec} (reconstruction loss weight)	10.0
loss weights for discriminator	
λ_{cls} (classification loss weight)	1.0
λ_{gp} (gradient penalty weight)	10.0
Ensemble Parameters	
α (temporal ensemble weight)	0.6
β (EMA momentum)	0.99
ρ (percentile threshold)	0.75

Table 3
Model performance across different few-shot settings.

Model	Average accuracy per class			
	5-shot	10-shot	20-shot	50-shot
SPARSE	63.21	68.50	73.44	77.15
$\text{SPARSE}_{\text{ens}}$	**66.22**	**70.95**	**75.71**	**78.28**
SGAN (Odena, 2016)	25.80	25.96	24.92	27.28
MatchGAN (Sun et al., 2020)	39.88	48.73	51.90	54.15
EC-GAN (Haque, 2021)	35.09	34.66	34.29	47.41
TripleGAN (Li et al., 2022)	64.23	68.79	71.40	76.25
SEC-GAN (Zhen et al., 2023)	58.73	63.79	66.95	73.30
CISSL (Xie et al., 2023)	45.84	46.80	49.19	51.20

comparisons. Table 4 presents the results of this statistical comparison, with the upper triangular part showing corrected p-values with effect sizes (r) and their interpretations, where significant results ($p < 0.05$) are highlighted in bold, and the lower triangular part displaying Win-Tie-Loss (W-T-L) statistics from the row model's perspective when compared against the column model.

The primary finding is that our ensemble model, $\text{SPARSE}_{\text{ens}}$, achieves statistically significant superiority over most competing approaches, also with a number of wins larger than losses, demonstrating robust performance across diverse medical imaging datasets.

We now discuss the effectiveness of the ensemble approach by comparing $\text{SPARSE}_{\text{ens}}$ to its base model, SPARSE. The results show that the ensemble model demonstrates a statistically significant improvement ($p = 4.1 \times 10^{-2}, r = 0.818$), securing wins on 10 out of 11 datasets with only a single loss. This marked improvement is achieved through ensemble averaging at inference time. Although both configurations share an identical training procedure, the classifier $\mathscr{C}$ and discriminator $\mathscr{D}$ networks develop complementary decision boundaries due to their architectural differences. Averaging their predictions effectively reduces prediction variance, a critical advantage in the extreme 5-shot regime. Let us now focus on how $\text{SPARSE}_{\text{ens}}$ performs with respect to the external competitors. The analysis shows that $\text{SPARSE}_{\text{ens}}$ obtains statistically significant improvements over a range of generative models, including SGAN ($p = 2.1 \times 10^{-2}, r = 1.000$), MatchGAN ($p =$

Table 4
5-Shot statistical comparison of model performance. The upper triangular part shows p-values from statistical comparison (significant values $p < 0.05$ highlighted in bold), while the lower triangular part shows Win-Tie-Loss (W-T-L) statistics. The r-value represents the effect size (Pearson's correlation coefficient) with interpretations: small ($r < 0.3$), medium ($0.3 \leq r < 0.5$), large ($0.5 \leq r < 0.7$), and very large ($r \geq 0.7$). Best performing model is highlighted in bold.

Model	Statistical comparison (5 shot)							
	SPARSE	$SPARSE_{ens}$	SGAN	MatchGAN	CISSL	SEC-GAN	TripleGAN	ECGAN
SPARSE	–	**4.1e−02** r=0.818 (very large)	**2.1e−02** r=0.818 (very large)	6.0e−02 r=0.636 (very large)	8.7e−02 r=0.455 (large)	2.4e−01 r=0.455 (large)	7.7e−01 r=0.091 (small)	**4.4e−02** r=0.636 (very large)
$SPARSE_{ens}$	(10-0-1)	–	**2.1e−02** r=1.000 (very large)	**4.4e−02** r=0.636 (very large)	**2.2e−02** r=0.818 (very large)	5.3e−02 r=0.636 (very large)	1.5e−01 r=0.455 (large)	**3.5e−02** r=0.800 (very large)
SGAN	(1-0-10)	(0-1-10)	–	**2.2e−02** r=0.636 (very large)	**2.2e−02** r=0.636 (very large)	**2.1e−02** r=0.818 (very large)	**2.1e−02** r=0.818 (very large)	**4.4e−02** r=0.600 (very large)
MatchGAN	(2-0-9)	(2-0-9)	(9-0-2)	–	2.6e−01 r=0.273 (medium)	6.0e−02 r=0.455 (large)	6.0e−02 r=0.455 (large)	1.0e−01 r=0.636 (very large)
CISSL	(3-0-8)	(1-0-10)	(9-0-2)	(7-0-4)	–	**4.1e−02** r=0.636 (very large)	**4.4e−02** r=0.455 (large)	**4.4e−02** r=0.455 (large)
SEC-GAN	(3-0-8)	(2-0-9)	(10-0-1)	(8-0-3)	(9-0-2)	–	8.7e−02 r=0.455 (large)	**2.7e−02** r=0.818 (very large)
TripleGAN	(5-0-6)	(3-0-8)	(10-0-1)	(8-0-3)	(8-0-3)	(8-0-3)	–	**4.4e−02** r=0.455 (large)
ECGAN	(2-0-9)	(1-1-9)	(8-1-2)	(2-0-9)	(3-0-8)	(1-0-10)	(3-0-8)	–

$4.4 \times 10^{-2}, r = 0.636$), CISSL ($p = 2.2 \times 10^{-2}, r = 0.818$) and EC-GAN ($p = 3.5 \times 10^{-2}, r = 0.800$). The largest performance gap is observed against SGAN, where our method was superior across all 11 datasets. The closest competition came from TripleGAN and SEC-GAN; while the differences did not reach statistical significance ($p = 1.5 \times 10^{-1}$ and $p = 5.3 \times 10^{-2}$, respectively), $SPARSE_{ens}$ still outperformed them on the majority of datasets with Win-Tie-Loss equal to 8-0-3 and 9-0-2. The consistent outperformance of $SPARSE_{ens}$ over this diverse set of competitors stems from a crucial methodological distinction. All competing methods are purely generative, creating images de novo from a noise vector. In contrast, our framework employs image-to-image translation, which modifies existing real, unlabeled images. This strategy preserves the authentic and complex anatomical features of the medical data, providing a more robust training signal. This fundamental advantage is further amplified when compared to two-player models like SGAN and MatchGAN, which suffer from an internal optimization conflict by tasking a single network with both discrimination and classification. Our three-player design avoids this issue. Even when compared to advanced three-player models like TripleGAN that also separate these tasks, our method's reliance on translating real images, rather than generating them from noise, appears to be the decisive factor for success in this data-scarce context.

5.2. Performance in 50-shot

We extend our analysis to the 50-shot setting to examine how model performance scales with increased labeled data availability, while still remaining within the low-labeled data regime. This configuration provides ten times more labeled samples per class compared to the 5-shot scenario, allowing us to assess the scaling properties of our methods. Table 5 presents the statistical comparison results using the Wilcoxon signed-rank test with Benjamini–Hochberg FDR correction. The table follows the same format as the 5-shot analysis, with p-values and effect sizes in the upper triangular portion and Win-Tie-Loss statistics in the lower triangular portion. The primary finding reveals that $SPARSE_{ens}$ maintains its statistical superiority over most competing approaches even as more labeled data becomes available. When comparing the two variants of our approach, $SPARSE_{ens}$ achieves a Win-Tie-Loss equal to 9-0-2 against the base SPARSE model. However, the p-value of 1.0×10^{-1} ($r = 0.636$) indicates no statistical significance: this narrowing gap can be attributed to the changing role of ensemble averaging in different data regimes. With only 5 shots per class, individual model predictions are inherently less reliable and more prone to overfitting, making the ensemble approach particularly valuable for reducing prediction variance. The discriminator and classifier develop highly complementary decision boundaries due to the scarcity of supervised signals. However, with 50 labeled samples per class, the classifier receives sufficient supervision to develop more stable and generalizable representations independently, reducing its reliance on the discriminator's complementary perspective. Examining $SPARSE_{ens}$'s performance against external competitors we note that it achieves statistically significant improvements over SGAN ($p = 2.0 \times 10^{-3}, r = 1.000$), MatchGAN ($p = 2.0 \times 10^{-3}, r = 1.000$), CISSL ($p = 2.0\times10^{-3}, r = 1.000$), and EC-GAN ($p = 2.0\times10^{-3}, r = 1.000$), with Win-Tie-Loss equal to 11-0-0 against each. These results represent stronger statistical evidence compared to the 5-shot setting, where p-values ranged from 2.1×10^{-2} to 4.4×10^{-2}. Furthermore $SPARSE_{ens}$ achieves statistical significance against SEC-GAN ($p = 2.0 \times 10^{-3}, r = 1.000$, W-T-L: 0-0-11), similarly to what happens in the 5-shot setting. When comparing against TripleGAN, we notice that the Win-Tie-Loss ratio is 2-0-9 in favor of our approach, similar to the 5-shot settings, but now the performance differences are not statistically significant at $p = 0.1$. These results suggest that performance differences become more consistent across datasets as labeled data increases. The patterns observed in this comparison can be attributed to differences in how methods utilize additional supervision. Our image-to-image translation approach leverages the increased labeled data to learn more accurate class-conditional transformations. With 50 labeled samples per class, the generator receives stronger supervision signals, enabling better preservation of discriminative features during translation. Additionally, the quality of pseudo-labels generated in the unsupervised phase improves due to more reliable initial classifiers.

Table 5
50-Shot statistical comparison of model performance. The upper triangular part shows p-values from statistical comparison (significant values $p < 0.05$ highlighted in bold), while the lower triangular part shows Win-Tie-Loss (W-T-L) statistics. The r-value represents the effect size (Pearson's correlation coefficient) with interpretations: small ($r < 0.3$), medium ($0.3 \leq r < 0.5$), large ($0.5 \leq r < 0.7$), and very large ($r \geq 0.7$). Best performing model is highlighted in bold.

Model	Statistical comparison (50 shot)							
	SPARSE	$SPARSE_{ens}$	SGAN	MatchGAN	CISSL	SEC-GAN	TripleGAN	ECGAN
SPARSE	–	1.0e−01 r=0.636 (very large)	**2.0e−03** r=1.000 (very large)	**2.0e−03** r=1.000 (very large)	**5.0e−03** r=0.818 (very large)	**1.5e−02** r=0.636 (very large)	3.5e−01 r=0.455 (large)	**4.0e−03** r=0.818 (very large)
$SPARSE_{ens}$	(9-0-2)	–	**2.0e−03** r=1.000 (very large)	**2.0e−03** r=1.000 (very large)	**2.0e−03** r=1.000 (very large)	**2.0e−03** r=1.000 (very large)	1.0e−01 r=0.636 (very large)	**2.0e−03** r=1.000 (very large)
SGAN	(0-0-11)	(0-0-11)	–	**8.0e−03** r=0.818 (very large)	**2.0e−03** r=1.000 (very large)	**2.0e−03** r=1.000 (very large)	**2.0e−03** r=1.000 (very large)	**1.1e−02** r=0.818 (very large)
MatchGAN	(0-0-11)	(0-0-11)	(10-0-1)	–	9.0e−01 r=0.091 (small)	**4.0e−03** r=0.818 (very large)	**2.0e−03** r=1.000 (very large)	1.7e−01 r=0.455 (large)
CISSL	(1-0-10)	(0-0-11)	(11-0-0)	(6-0-5)	–	**1.9e−02** r=0.455 (large)	**8.0e−03** r=0.636 (very large)	5.5e−01 r=0.091 (small)
SEC-GAN	(2-0-9)	(0-0-11)	(11-0-0)	(10-0-1)	(8-0-3)	–	**3.1e−02** r=0.818 (very large)	**4.0e−03** r=0.818 (very large)
TripleGAN	(3-0-8)	(2-0-9)	(11-0-0)	(11-0-0)	(9-0-2)	(10-0-1)	–	**2.0e−03** r=1.000 (very large)
ECGAN	(1-0-10)	(0-0-11)	(10-0-1)	(3-0-8)	(5-0-6)	(1-0-10)	(0-0-11)	–

5.3. Framework component analysis

To understand the source of our framework's performance gains, we conduct a systematic analysis of key design decisions and architectural components. This section examines two critical aspects: first, we analyze how the frequency of unsupervised training phases affects model performance, demonstrating the importance of properly balancing supervised and unsupervised learning signals. Second, we perform controlled ablation experiments to quantify the contribution of individual components, namely, the generator and exponential moving average mechanism, isolating their impact on classification performance across different data availability regimes.

5.3.1. Impact of unsupervised training frequency

In our approach, the hyperparameter μ controls how frequently the unsupervised training phase is executed, with the unsupervised phase running every μ epochs. We investigate how varying this parameter affects our model's performance across different few-shot settings, as it represents the balance between supervised and unsupervised learning signals in our semi-supervised framework. Fig. 2 presents the average accuracy across all datasets as a function of the number of shots per class for different values of μ. These results were obtained using the validation set to avoid any bias that could arise from hyperparameter selection on the test set. The key finding is that $\mu = 10$ consistently achieves the highest performance across all shot settings. In the 5-shot setting, $\mu = 10$ reaches 74.5% accuracy, representing an 8.5% point improvement over supervised-only learning ($\mu = 0$ at 66.0%). This performance advantage narrows as labeled data increases, reducing to approximately 2% points in the 50-shot setting (85.1% vs 83.1%). The figure reveals distinct patterns in how unsupervised learning frequency affects performance. Moving from supervised-only learning to any incorporation of unsupervised learning produces improvements, particularly evident in low-shot scenarios. Moderate frequencies ($\mu \in \{1, 10, 25\}$) achieve optimal performance, with $\mu = 10$ emerging as optimum. Very high frequencies ($\mu \in \{50, 100\}$) lead to performance degradation compared to the optimum, though they still outperform supervised-only learning in data-scarce settings. These findings validate our design choice of alternating between supervised and unsupervised learning phases, demonstrating that the optimal frequency ($\mu = 10$) remains consistent across different data availability scenarios and that properly scheduled unsupervised learning is particularly crucial in extreme low-data regimes. Additionally, these results shows the collaborative dynamics between the supervised phase and the combined self-supervised pre-training and synthetic data enhancement phases. The supervised-only baseline ($\mu = 0$) establishes the performance ceiling achievable without leveraging unlabeled data, reaching 66.0% in the 5-shot setting. The substantial improvement when incorporating the unsupervised phases ($\mu = 10$ achieving 74.5%, an 8.5 percentage point gain) demonstrates that the self-supervised pre-training and synthetic data enhancement phases contribute complementary learning signals that supervised training alone cannot provide. However, the performance degradation at very high frequencies ($\mu \in [50, 100]$) reveals that this collaboration requires careful balance: excessive unsupervised training can destabilize the classifier before supervised signals adequately anchor its learning. training can destabilize the classifier before supervised signals adequately anchor its learning. The optimal frequency ($\mu = 10$) thus creates a productive feedback loop: the supervised phase provides stable models that enable reliable pseudo-labeling in the self-supervised pre-training phase, while the synthetic data enhancement phase expands the effective training set that the supervised phase leverages in subsequent iterations. This interdependence explains why the alternating schedule proves more effective than either continuous supervised training or more frequent unsupervised updates.

5.3.2. Ablation study of core components

To isolate the contribution of individual components within our framework, we conduct controlled ablation experiments by systematically removing key architectural elements. Specifically, we evaluate two critical components: (1) the exponential moving average (EMA) mechanism in the pseudo-labeling ensemble, which provides temporal stability to pseudo-label predictions, and (2) the generator network, which performs class-conditioned image translation in the unsupervised learning phase. These ablations are evaluated across all four few-shot settings (5, 10, 20, and 50 shots per class), averaging results across

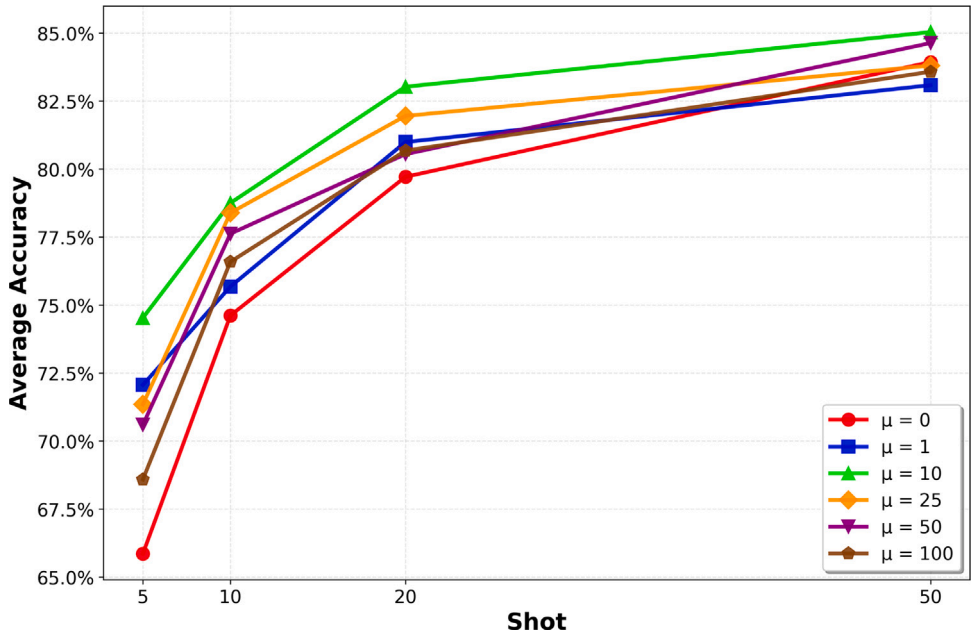

Fig. 2. Average classification accuracy as a function of the number of labeled samples per class (shots) for different unsupervised training frequencies (μ). Results are averaged across all eleven MedMNIST datasets and computed on the validation set. The parameter μ controls how frequently the unsupervised training phase is executed, with $\mu = 0$ representing supervised learning only (no unsupervised phase), $\mu = 1$ indicating unsupervised training at every epoch, and higher values ($\mu \in \{10, 25, 50, 100\}$) representing unsupervised training every μ epochs.

Table 6
Component ablation analysis across few-shot settings. Average accuracy and Cohen's d effect sizes when removing individual components from the full SPARSE framework. Results averaged over SPARSE and SPARSEens configurations across all eleven datasets. Positive d indicates full model outperforms ablated variant. Effect size interpretation: $|d| < 0.2$ (negligible), $0.2 \leq |d| < 0.5$ (small), $0.5 \leq |d| < 0.8$ (medium), $|d| \geq 0.8$ (large).

Model variant	Component removed	5-shot	10-shot	20-shot	50-shot
Full SPARSE	None	66.22	70.95	75.70	78.28
Ablation 1	EMA	65.09 d=0.21 (small)	69.73 d=0.32 (small)	75.70 d=-0.00 (neg.)	78.78 d=-0.09 (neg.)
Ablation 2	Generator	59.17 d=0.70 (medium)	65.92 d=0.66 (medium)	70.98 d=0.72 (medium)	72.48 d=0.85 (large)

both SPARSE (classifier-only) and $\text{SPARSE}_{\text{ens}}$ configurations to assess overall framework performance. To quantify the magnitude of each component's contribution, we employ Cohen's d effect sizes, which provide a standardized measure that accounts for variance across the eleven diverse datasets. This approach enables meaningful comparison of component importance through interpretable thresholds (negligible, small, medium, large effects), assessing not just whether differences exist, but whether they are practically significant.

Table 6 presents the ablation results, showing the mean performance across both inference configurations (SPARSE and $\text{SPARSE}_{\text{ens}}$) when averaged over all eleven datasets. Cohen's d effect sizes quantify the magnitude of each component's contribution. The analysis reveals that the generator component is the most critical element of the framework across all data availability regimes. Removing the generator produces medium effect sizes in the first three settings (5-shot: d = 0.702, 10-shot: d = 0.663, 20-shot: d = 0.715) and escalates to large magnitude in the 50-shot setting (d = 0.853). This increasing pattern demonstrates that the generator provides fundamental value through its image-to-image translation mechanism that enriches feature representations synergistically with supervised learning signals, rather than merely compensating for data scarcity. The modification of real unlabeled images preserves authentic anatomical features while introducing semantic variations that enhance classification performance even when moderate amounts of labeled data are available. In contrast, the EMA mechanism shows negligible to small effect sizes across all configurations (5-shot: d = 0.208, 10-shot: d = 0.318, 20-shot: d = −0.001, 50-shot: d = −0.085), with none achieving statistical significance. This indicates that the confidence-weighted ensemble mechanism already provides substantial stability by combining predictions from multiple models. Nevertheless, we retain the EMA component in our framework for two reasons: first, it introduces negligible computational overhead during training while potentially benefiting edge cases or specific datasets not fully represented in our evaluation; second, the small positive effects observed in extremely low-data regimes (5-shot and 10-shot) suggest it provides valuable stability when pseudo-label reliability is most uncertain, even if this benefit diminishes when averaged across diverse datasets and higher-shot settings. Additionally, the generator ablation reveals the specific contribution of the synthetic data enhancement phase to the overall framework collaboration. Removing the generator eliminates this phase while preserving the supervised and self-supervised pre-training phases, resulting in performance drops of 7.1 percentage points in 5-shot and 5.8 percentage points in 50-shot settings. The escalating effect size demonstrates that the synthetic data enhancement phase amplifies the benefits of supervised learning

Table 7
Computational complexity analysis of SPARSE framework components. Measurements are reported for input images of size 128× 128× 3 pixels with 8 output classes.

Model component	GFLOPs	Params (M)
Generator (image translation)	23.135	8.140
Generator (classifier mode)	6.732	4.693
Discriminator (PatchGAN)	1.368	44.787
Classifier (EfficientNet-B3)	0.667	10.709
TOTAL (Supervised phase)	**8.767**	**60.188**
TOTAL (Unsupervised phase)	**25.170**	**60.188**
TOTAL (SPARSE inference)	**0.667**	**10.709**
TOTAL ($\text{SPARSE}_{\text{ens}}$ inference)	**2.035**	**55.496**

through a specific mechanism: the self-supervised pre-training phase learns discriminative class-conditional transformations from unlabeled data, which the synthetic data enhancement phase exploits to generate high-quality samples with known labels. These synthetic samples subsequently augment the supervised phase in later training epochs, creating a three-way interaction where each phase enhances the effectiveness of the others. This collaborative mechanism, which operates iteratively across the alternating training schedule, explains why our framework substantially outperforms methods employing only supervised learning with generation or two-phase approaches that lack this complete feedback loop.

5.4. Computational analysis

The use of three specialized networks in our framework introduces computational overhead during both training and inference. This section provides a quantitative analysis of these requirements and examines the trade-offs between computational cost and classification performance.

Table 7 presents the computational complexity breakdown for each network component. During training, all three networks participate simultaneously across both learning phases. The supervised phase requires 8.767 GFLOPs per sample as the generator's encoder $E_{\mathscr{G}}$, discriminator $\mathscr{D}$ and classifier $\mathscr{C}$ jointly process the limited labeled data. The unsupervised phase is substantially more expensive at 25.170 GFLOPs per sample, with the complete generator architecture performing class-conditioned image translation accounting for 23.135 GFLOPs (92% of the total). The remaining computational cost is distributed between the discriminator (1.368 GFLOPs) and classifier (0.667 GFLOPs). The overall training cost depends critically on the frequency parameter μ. With $\mu = 10$ (the configuration that achieves optimal performance across all few-shot settings as demonstrated in) the expensive unsupervised phase occurs once every 10 epochs. This results in an average computational cost per epoch that is dominated by the unsupervised phase when considering the substantially larger unlabeled dataset ($M \gg M$). The parameter μ provides a direct mechanism to modulate computational cost: increasing μ 50 or 100 reduces the frequency of unsupervised training, thereby decreasing overall training time. However, as shown in Fig. 2, this comes at the cost of reduced classification accuracy, illustrating the fundamental trade-off between computational efficiency and model performance. The ensemble-based pseudo-labeling mechanism adds minimal overhead beyond the forward passes required for prediction. Computing confidence-weighted ensemble predictions involves averaging probability distributions and selecting samples above a percentile-based threshold, operations with complexity $O(K)$ per sample, where K is the number of classes. The temporal ensemble with exponential moving average maintains running statistics without additional forward passes, adding only $O(K)$ memory overhead per unlabeled sample.

At inference, computational requirements reduce substantially. The generator, responsible for the majority of training overhead, is discarded entirely. The SPARSE configuration requires only 0.667 GFLOPs per image using the classifier alone, while $\text{SPARSE}_{\text{ens}}$ requires 2.035 GFLOPs through ensemble averaging of the discriminator and classifier predictions. This represents a 3 times increase in inference cost for $\text{SPARSE}_{\text{ens}}$, which must be weighed against its performance gains. The empirical results demonstrate that these computational investments yield substantial performance improvements in data-scarce regimes. As shown in Table 3, $\text{SPARSE}_{\text{ens}}$ achieves 66.22% average accuracy with only 5 labeled samples per class. When compared to two-player architectures such as SGAN (25.80%), MatchGAN (39.88%), and CISSL (45.84%), the performance gap demonstrates the value of the three-network design. This advantage persists even when comparing against methods with 10 times more labeled data: SGAN reaches only 27.28% with 50 samples per class, indicating that architectural sophistication in the three-network design provides learning capabilities that simpler models cannot achieve through moderate increases in labeled data alone. The $\text{SPARSE}_{\text{ens}}$ configuration further justifies its 3 times inference cost through statistically significant improvements over SPARSE as detailed in Table 4.

6. Conclusions

This paper has introduced a novel GAN-based semi-supervised learning framework specifically designed for extremely low labeled-data regimes. Our approach addresses the fundamental challenge of insufficient labeled data through three key innovations: (1) a dynamic training schedule that alternates between supervised and unsupervised phases, (2) an image-to-image translation mechanism that enriches feature representations by learning class-conditional transformations from real unlabeled images, and (3) a confidence-weighted temporal ensemble technique for reliable pseudo-labeling. By leveraging these components within a three-player GAN architecture, our method effectively combines the complementary strengths of a generator, discriminator, and dedicated classifier. The comprehensive empirical evaluation across eleven MedMNIST datasets demonstrates the effectiveness of our approach. In the extreme 5-shot setting, our ensemble configuration achieved statistically significant improvements over six state-of-the-art semi-supervised methods, with effect sizes ranging from large to very large. The method's superiority stems from its fundamental design choice of performing image-to-image translation rather than generating images from noise, which leverages real medical images as the foundation for learning discriminative features crucial for classification tasks. This advantage is further amplified by our temporal ensemble mechanism, which aggregates predictions across training epochs to produce more reliable pseudo-labels in data-scarce scenarios. Our analysis of the unsupervised training frequency revealed that moderate alternation between supervised and unsupervised phases ($\mu = 10$) balances both learning signals. This finding provides practical guidance for deployment, as the optimal frequency remained consistent across different data availability scenarios, eliminating the need for extensive hyperparameter tuning in clinical applications.

Despite these promising results, several avenues for future research remain. The computational requirements of maintaining multiple networks may pose challenges in resource-constrained clinical settings, motivating the development of more efficient architectures. Additionally, extending the framework to incorporate domain-specific medical knowledge and multi-modal imaging data could further enhance its clinical applicability.

We acknowledge that while our evaluation on the MedMNIST benchmark provides reproducibility through standardized preprocessing and consistent evaluation protocols, and demonstrates generalizability across diverse imaging modalities (X-ray, OCT, ultrasound, CT, microscopy), anatomical regions, and classification tasks, each dataset may present specific limitations or biases inherent to its source and preparation process. Future work should validate SPARSE on additional real-world clinical datasets with varying data distributions, acquisition protocols, and patient populations to further assess its robustness and practical applicability in diverse clinical settings. Furthermore, investigating the method's performance on datasets with different types of class imbalances, rare diseases, and multi-label classification scenarios would provide valuable insights into its broader clinical utility. In conclusion, our GAN-based semi-supervised learning framework represents a significant advancement in addressing the labeled data scarcity challenge in medical imaging. By effectively leveraging both labeled and unlabeled data through image translation and ensemble techniques, our method enables robust classification performance even with as few as five labeled samples per class, offering a practical solution for medical imaging applications where annotation costs are prohibitive.

CRediT authorship contribution statement

Guido Manni: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Clemente Lauretti:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition, Conceptualization. **Loredana Zollo:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition, Conceptualization. **Paolo Soda:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Guido Manni is a Ph.D. student enrolled in the National Ph.D. in Artificial Intelligence, XXXVII cycle, course on Health and life sciences, organized by Università Campus Bio-Medico di Roma. This work was partially founded by: (i) Piano Nazionale Ripresa e Resilienza (PNRR) - HEAL ITALIA Extended Partnership - SPOKE 2 Cascade Call - "Intelligent Health" with the project BISTOURY - 3D-guided roBotIc Surgery based on advanced navigaTiOn systems and aUgmented viRtual realitY (CUP: J33C22002920006) and (ii) PNRR MUR project PE0000013-FAIR. Resources are provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) and the Swedish National Infrastructure for Computing (SNIC) at Alvis C3S.

Appendix

A.1. Dataset characteristics and performance patterns

We analyze the relationship between dataset characteristics and SPARSE performance to identify which datasets benefit most from our approach and why. To understand when and why our method achieves strong performance, we investigate two potential factors that relate to the core mechanisms of our approach. First, we examine whether discriminative feature content, the degree to which classes exhibit visually distinguishable characteristics, affects the effectiveness of our image-to-image translation mechanism. Since the generator must learn meaningful class-conditional transformations, we hypothesize that performance may depend on the presence of clear visual differences between classes. Second, we investigate how unlabeled dataset size impacts semi-supervised learning performance. Given that our method relies on pseudo-labeling and translation of unlabeled samples during the unsupervised training phase, we explore whether the quantity of available unlabeled data influences the model's ability to extract additional information beyond the limited labeled samples.

A.1.1. Discriminative features and image-to-image translation

We first investigate whether the presence of discriminative features, visual characteristics that distinguish between classes, affects SPARSE's performance, given that our method relies on image-to-image translation as its core semi-supervised learning mechanism. Unlike purely generative approaches that synthesize images from noise, our method modifies real unlabeled images to preserve authentic features while learning class-conditional transformations. We hypothesize that the generator can only learn meaningful class-conditional transformations when classes exhibit visually distinguishable characteristics. Our analysis of the results in Tables 8–11 reveals a clear performance stratification based on discriminative feature strength. Datasets where classes exhibit distinct morphological or structural characteristics achieve superior performance: BloodMNIST (96.8% in 50-shot, Table 11) contains 8 blood cell types with fundamentally different cellular structures and nuclear; PathMNIST (91.7%) comprises 9 tissue types with distinct histological patterns; OrganAMNIST (93.0%) represents 11 abdominal organs with distinct anatomical shapes and densities. In these cases, the generator appears to learn meaningful class-conditional transformations that emphasize distinctive characteristics, while the discriminator recognizes authentic class-specific features. Conversely, datasets with limited discriminative features show substantially reduced performance. TissueMNIST (47.2%) contains cells from the same anatomical region sharing similar morphologies; ChestMNIST (61.6%) presents subtle and overlapping disease manifestations. This pattern suggests that when biological similarity or imaging constraints limit discriminative features, image translation cannot create information absent in the original data. Datasets with intermediate performance present moderate discriminative features: DermaMNIST (70.1%) contains skin lesions with some overlapping visual characteristics; OCTMNIST (68.3%) captures subtle retinal tissue changes that may not present dramatically distinct patterns; PneumoniaMNIST (86.0%) shows variable pneumonia manifestations that may not always differ obviously from normal chest X-rays. These findings suggest that discriminative feature strength is a primary determinant of SPARSE performance, with the image-to-image translation mechanism requiring visually distinguishable class characteristics to be effective.

A.1.2. Unlabeled dataset size impact

We next investigate how unlabeled data quantity affects semi-supervised learning performance. The MedMNIST datasets vary significantly in training set size (546 to 165,466 samples, Table 1), providing a natural experiment for this analysis. Since SPARSE alternates between supervised and unsupervised phases, where the unsupervised phase

Table 8
Comparison of accuracy scores across different methods on medical image datasets using 5-shot learning. Best performance for each dataset is highlighted in bold.

Dataset	SPARSE	$SPARSE_{ens}$	SGAN	MatchGAN	CISSL	SEC-GAN	TripleGAN	ECGAN
bloodmnist	0.868	**0.887**	0.173	0.302	0.463	0.813	0.862	0.241
breastmnist	0.542	0.667	0.667	**0.792**	0.615	0.719	0.677	0.667
chestmnist	0.545	0.571	0.539	0.535	**0.610**	0.533	0.510	0.532
dermamnist	0.511	0.576	0.124	**0.626**	0.566	0.535	0.563	0.626
octmnist	0.488	0.465	0.263	0.333	0.254	0.363	**0.528**	0.265
organamnist	0.780	**0.801**	0.180	0.114	0.422	0.683	0.780	0.198
organcmnist	0.754	**0.792**	0.104	0.268	0.436	0.683	0.782	0.200
organsmnist	0.567	**0.588**	0.051	0.216	0.215	0.487	0.561	0.141
pathmnist	0.751	**0.778**	0.100	0.187	0.367	0.453	0.676	0.186
pneumoniamnist	0.796	0.806	0.585	0.710	0.785	**0.854**	0.762	0.767
tissuemnist	0.351	0.353	0.052	0.302	0.308	0.339	**0.363**	0.040

Table 9
Comparison of accuracy scores across different methods on medical image datasets using 10-shot learning. Best performance for each dataset is highlighted in bold.

Dataset	SPARSE	$SPARSE_{ens}$	SGAN	MatchGAN	CISSL	SEC-GAN	TripleGAN	ECGAN
bloodmnist	0.895	**0.907**	0.171	0.407	0.505	0.851	0.890	0.266
breastmnist	0.604	0.698	0.688	0.760	0.615	0.781	0.677	**0.792**
chestmnist	0.563	**0.568**	0.517	0.532	0.559	0.536	0.549	0.532
dermamnist	**0.629**	0.626	0.114	0.626	0.626	0.527	0.567	0.625
octmnist	0.490	0.515	0.246	0.240	0.316	0.481	**0.606**	0.247
organamnist	0.820	**0.834**	0.194	0.108	0.393	0.795	0.797	0.103
organcmnist	0.765	**0.808**	0.148	0.744	0.405	0.676	0.793	0.167
organsmnist	0.662	**0.694**	0.050	0.495	0.230	0.521	0.605	0.188
pathmnist	0.835	**0.873**	0.099	0.372	0.332	0.657	0.792	0.050
pneumoniamnist	0.854	0.848	0.577	0.775	0.831	0.838	**0.881**	0.802
tissuemnist	0.417	**0.434**	0.052	0.302	0.336	0.355	0.409	0.040

Table 10
Comparison of accuracy scores across different methods on medical image datasets using 20-shot learning. Best performance for each dataset is highlighted in bold.

Dataset	SPARSE	$SPARSE_{ens}$	SGAN	MatchGAN	CISSL	SEC-GAN	TripleGAN	ECGAN
bloodmnist	0.925	**0.939**	0.173	0.649	0.556	0.913	0.917	0.311
breastmnist	0.698	**0.792**	0.646	0.667	0.708	0.750	0.688	0.646
chestmnist	0.563	0.574	0.525	0.562	0.574	0.565	**0.582**	0.532
dermamnist	0.652	**0.654**	0.111	0.623	0.623	0.612	0.627	0.622
octmnist	0.609	**0.667**	0.248	0.334	0.292	0.463	0.642	0.259
organamnist	0.880	**0.885**	0.161	0.098	0.441	0.822	0.843	0.134
organcmnist	0.859	**0.881**	0.081	0.784	0.423	0.771	0.809	0.171
organsmnist	0.724	**0.737**	0.064	0.573	0.212	0.587	0.654	0.181
pathmnist	0.894	**0.909**	0.095	0.335	0.439	0.649	0.801	0.084
pneumoniamnist	**0.823**	0.817	0.585	0.781	0.815	0.792	0.821	0.756
tissuemnist	0.450	**0.472**	0.052	0.302	0.326	0.440	0.470	0.075

Table 11
Comparison of accuracy scores across different methods on medical image datasets using 50-shot learning. Best performance for each dataset is highlighted in bold.

Dataset	SPARSE	$SPARSE_{ens}$	SGAN	MatchGAN	CISSL	SEC-GAN	TripleGAN	ECGAN
bloodmnist	0.966	**0.968**	0.172	0.819	0.612	0.936	0.956	0.807
breastmnist	**0.844**	0.813	0.646	0.708	0.698	0.781	0.792	0.750
chestmnist	0.613	**0.616**	0.522	0.532	0.604	0.568	0.577	0.532
dermamnist	**0.708**	0.701	0.133	0.625	0.619	0.617	0.660	0.624
octmnist	0.648	0.683	0.241	0.278	0.339	0.643	**0.826**	0.247
organamnist	0.928	**0.930**	0.178	0.141	0.485	0.867	0.887	0.155
organcmnist	0.884	**0.888**	0.183	0.797	0.429	0.837	0.841	0.619
organsmnist	0.746	**0.763**	0.023	0.623	0.290	0.682	0.686	0.499
pathmnist	0.906	**0.917**	0.069	0.333	0.369	0.841	0.856	0.116
pneumoniamnist	0.813	**0.860**	0.783	0.798	0.850	0.842	0.821	0.815
tissuemnist	0.431	0.472	0.052	0.302	0.335	0.449	**0.485**	0.052

leverages all available unlabeled samples for pseudo-labeling and image translation, we examine whether the size of the unlabeled dataset impacts how much additional information the model can extract beyond the limited labeled samples. Our analysis reveals that unlabeled dataset size alone does not guarantee performance. Examining large unlabeled datasets, we observe that TissueMNIST (165,466 training samples) achieves only 47.2% despite having the largest unlabeled dataset, while PathMNIST (89,996 samples) reaches 91.7%. This demonstrates that unlabeled data quantity cannot compensate for absent discriminative features. However, when discriminative features are present, we find that larger unlabeled datasets amplify performance gains. PathMNIST, OrganAMNIST (34,561 samples, 93.0%), and OCTMNIST (97,477 samples, 68.3%) all show substantial improvements from 5-shot to 50-shot settings (+13.9%, +12.9%, +21.8% respectively, comparing

Tables 8 and 11), indicating effective utilization of unlabeled data for pseudo-labeling and translation. Conversely, datasets with smaller unlabeled sets demonstrate that strong discriminative features can compensate for limited unlabeled data. BloodMNIST (11,959 samples) achieves 96.8%, the highest accuracy among all datasets, while PneumoniaMNIST (4708 samples) reaches 86.0%. Both show more modest 5-to-50-shot improvements (+8.1%, +5.4%), suggesting performance depends more on labeled sample quality and discriminative features than unlabeled data quantity when the unlabeled set is small. These findings reveal that the relationship between discriminative feature strength and unlabeled dataset size is multiplicative rather than additive. Strong discriminative features enable effective learning even with limited unlabeled data (BloodMNIST, PneumoniaMNIST), while weak discriminative features limit performance regardless of unlabeled data abundance (TissueMNIST, ChestMNIST). The optimal scenario combines both strong discriminative features and large unlabeled datasets (PathMNIST, OrganAMNIST), where SPARSE achieves performance exceeding 90%.

Data availability

Code is available at https://github.com/GuidoManni/SPARSE All datasets used in this study are publicly available from the MedMNIST repository https://medmnist.com/.

References

Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., Fleet, D.J., 2023. Synthetic data from diffusion models improves ImageNet classification. arXiv:2304.08466. URL https://arxiv.org/abs/2304.08466.

Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A., 2017. Network dissection: Quantifying interpretability of deep visual representations. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 3319–3327. http://dx.doi.org/10.1109/CVPR.2017.354.

Bull, S., Cheah, P.Y., Denny, S., Jao, I., Marsh, V., Merson, L., Shah More, N., Nhan, L.T., Osrin, D., Tangseefa, D., Wassenaar, D., Parker, M., 2015. Best practices for ethical sharing of individual-level health research data from low- and middle-income settings. J. Empir. Res. Hum. Res. Ethics 10 (3), 302–313. http://dx.doi.org/10.1177/1556264615594606.

Cao, Y., Geddes, T.A., Yang, J.Y.H., Yang, P., 2020. Ensemble deep learning in bioinformatics. Nat. Mach. Intell. 2 (9), 500–508. http://dx.doi.org/10.1038/s42256-020-0217-y.

Chen, X., Wang, X., Zhang, K., Fung, K.-M., Thai, T.C., Moore, K., Mannel, R.S., Liu, H., Zheng, B., Qiu, Y., 2022. Recent advances and clinical applications of deep learning in medical image analysis. Med. Image Anal. 79, 102444. http://dx.doi.org/10.1016/j.media.2022.102444, URL https://www.sciencedirect.com/science/article/pii/S1361841522000913.

Chen, J., Yang, N., Pan, Y., Liu, H., Zhang, Z., 2023. Synchronous medical image augmentation framework for deep learning-based image segmentation. Comput. Med. Imaging Graph. 104, 102161. http://dx.doi.org/10.1016/j.compmedimag.2022.102161, URL https://www.sciencedirect.com/science/article/pii/S0895611122001318.

Chiruvella, V., Guddati, A.K., 2021. Ethical issues in patient data ownership. Interact. J. Med. Res. 10 (2), e22269. http://dx.doi.org/10.2196/22269.

Dai, Z., Yang, Z., Yang, F., Cohen, W.W., Salakhutdinov, R.R., 2017. Good semi-supervised learning that requires a bad GAN. Adv. Neural Inf. Process. Syst. 30.

Geng, S., Hsieh, C.-Y., Ramanujan, V., Wallingford, M., Li, C.-L., Koh, P.W., Krishna, R., 2025. The unmet promise of synthetic training images: Using retrieved real images performs better. arXiv:2406.05184. URL https://arxiv.org/abs/2406.05184.

Guarrasi, V., D'Amico, N.C., Sicilia, R., Cordelli, E., Soda, P., 2022. Pareto optimization of deep networks for COVID-19 diagnosis from chest X-rays. Pattern Recognit. 121, 108242.

Haque, A., 2021. EC-GAN: Low-sample classification using semi-supervised algorithms and GANs (student abstract). In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, (18), pp. 15797–15798.

Li, C., Xu, K., Zhu, J., Liu, J., Zhang, B., 2022. Triple generative adversarial networks. IEEE Trans. Pattern Anal. Mach. Intell. 44 (12), 9629–9640. http://dx.doi.org/10.1109/TPAMI.2021.3127558.

Odena, A., 2016. Semi-supervised learning with generative adversarial networks. arXiv preprint arXiv:1606.01583.

Ruffini, F., Ayllon, E.M., Shen, L., Soda, P., Guarrasi, V., 2025. Benchmarking foundation models and parameter-efficient fine-tuning for prognosis prediction in medical imaging. arXiv:2506.18434. URL https://arxiv.org/abs/2506.18434.

Sajun, A.R., Zualkernan, I., 2022. Survey on implementations of generative adversarial networks for semi-supervised learning. Appl. Sci. 12 (3), http://dx.doi.org/10.3390/app12031718, URL https://www.mdpi.com/2076-3417/12/3/1718.

Salmè, M., Tronchin, L., Sicilia, R., Soda, P., Guarrasi, V., 2025. Beyond the generative learning trilemma: Generative model assessment in data scarcity domains. arXiv:2504.10555. URL https://arxiv.org/abs/2504.10555.

Sariyildiz, M.B., Alahari, K., Larlus, D., Kalantidis, Y., 2023. Fake it till you make it: Learning transferable representations from synthetic ImageNet clones. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 8011–8021. http://dx.doi.org/10.1109/CVPR52729.2023.00774.

Sun, J., Bhattarai, B., Kim, T.-K., 2020. MatchGAN: A self-supervised semi-supervised conditional generative adversarial network. In: Proceedings of the Asian Conference on Computer Vision.

van Engelen, J.E., Hoos, H.H., 2020. A survey on semi-supervised learning. Mach. Learn. 109 (2), 373–440. http://dx.doi.org/10.1007/s10994-019-05855-6.

Wolpert, D., Macready, W., 1997. No free lunch theorems for optimization. IEEE Trans. Evol. Comput. 1 (1), 67–82. http://dx.doi.org/10.1109/4235.585893.

Xie, Y., Wan, Q., Xie, H., Xu, Y., Wang, T., Wang, S., Lei, B., 2023. Fundus image-label pairs synthesis and retinopathy screening via GANs with class-imbalanced semi-supervised learning. IEEE Trans. Med. Imaging 42 (9), 2714–2725. http://dx.doi.org/10.1109/TMI.2023.3263216.

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B., 2023. MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Sci. Data 10 (1), 41.

Zeng, Q., Lu, Z., Xie, Y., et al., 2025b. PICK: Predict and mask for semi-supervised medical image segmentation. Int. J. Comput. Vis. 133, 3296–3311. http://dx.doi.org/10.1007/s11263-024-02328-9.

Zeng, Q., Xie, Y., Lu, Z., Lu, M., Wu, Y., Xia, Y., 2025a. Segment together: A versatile paradigm for semi-supervised medical image segmentation. IEEE Trans. Med. Imaging 44 (7), 2948–2959. http://dx.doi.org/10.1109/TMI.2025.3556310.

Zeng, Q., Xie, Y., Lu, Z., Lu, M., Zhang, J., Xia, Y., 2025c. Consistency-guided differential decoding for enhancing semi-supervised medical image segmentation. IEEE Trans. Med. Imaging 44 (1), 44–56. http://dx.doi.org/10.1109/TMI.2024.3429340.

Zeng, Q., Xie, Y., Lu, Z., Xia, Y., 2023. PEFAT: Boosting semi-supervised medical image classification via pseudo-loss estimation and feature adversarial training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, Vancouver, BC, Canada, pp. 15671–15680. http://dx.doi.org/10.1109/CVPR52729.2023.01504.

Zhen, H., Shi, Y., Yang, J.J., Vehni, J.M., 2023. Co-supervised learning paradigm with conditional generative adversarial networks for sample-efficient classification. Appl. Comput. Intell. 3 (1), 13–26. http://dx.doi.org/10.3934/aci.2023002, URL https://www.aimspress.com/article/doi/10.3934/aci.2023002.