Contrastive Learning with Partial Knowledge for Medical Image Representation

by

Saba Hosseinipanah

A research project
presented to the University of Waterloo
in fulfillment of the
research paper requirement for the degree of
Master of Mathematics
in
Computational Mathematics

Waterloo, Ontario, Canada, 2025

© Saba Hosseinipanah 2025

Author's Declaration

I hereby declare that I am the sole author of this research paper. This is a true copy of the research paper, including any required final revisions, as accepted by my examiners. I understand that my research paper may be made electronically available to the public.

I understand that my thesis may be made electronically available to the public.

Abstract

Medical image classification plays a critical role in clinical decision-making, yet its performance is often limited by incomplete or uneven supervision. In practice, reliable labels may be unavailable, and comprehensive clinical knowledge can be fragmented and costly to obtain. This research project proposes a representation learning framework to augment classification under partial knowledge rather than assuming fully annotated supervision. Using data from the MedMNIST benchmarks, we simulate two realistic settings that reflect asymmetric knowledge distribution, modeled after medical students with incomplete training. In the first setting, Partial-Knowledge Contrastive Learning (PKCL), multiple students possess class-specific knowledge: each student can reliably recognize one target class but cannot consistently distinguish the others. In the second setting, Semi-Supervised Contrastive Learning (SSCL), each student can only separate normal from abnormal samples, but not the fine-grained abnormal categories. Both scenarios model real diagnostic workflows where supervision is incomplete but structured. Our approach leverages contrastive learning to integrate these partial signals into a shared latent representation. This latent embedding space allows complementary supervision sources to reinforce one another. We evaluate the learned representation using embedding analysis and downstream classification across three distinct modalities: PathMNIST, OCTMNIST, and BloodMNIST. On held-out test sets, our representation supports classifiers achieving highly competitive accuracies across all modalities—ranging from 91.0% on OCTMNIST up to 99.6% on BloodMNIST—with a strong 97.9% accuracy on the primary PathMNIST benchmark. These results indicate that partial knowledge can be effectively combined by contrastive learning to mitigate incomplete supervision, offering a highly flexible approach for real-world medical imaging environments.

Acknowledgements

I would like to thank my supervisor, Professor Mu Zhu, for his guidance, thoughtful feedback, and continuous support throughout this research project. His perspective and encouragement played an important role in shaping both the direction and clarity of this work.

This work benefited greatly from publicly available datasets, MedMNIST, including PathMNIST(NCT-CRC-HE100K), OctMNIST, and BloodMNIST which made this research possible.

Dedication

To my family, especially my parents, who endured the distance and supported me throughout this journey.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The effectiveness of every machine learning system is intrinsically dependent on the quality of the data representation it utilizes. Bengio, Courville, and Vincent [5] say that a successful representation should not only compress high-dimensional sensory inputs into a lower-dimensional manifold, but it should also capture the underlying explanatory elements that regulate the data-generating process. These factors (like the object's identity, stance, illumination, or in medical contexts, tumor shape, imaging modality, and patient demographics) should be recorded in a way that is spread out, sparse, consistent across time, and, most importantly, separate. In this context, disentanglement that each dimension or subspace of the latent representation corresponds to a single, semantically meaningful source of variation that is not related to any other factors. These kinds of representations not only make models easier to understand and more reliable, but they also make it easier to transfer learning, adjust to new situations with few examples, and reason about causes. These are all important skills in high-stakes fields like healthcare.

For much of the deep learning era, the dominant paradigm has been implicit representation learning: representations are learned end-to-end as a byproduct of optimizing a downstream task such as image classification or segmentation. While effective, this approach conflates task-specific biases with general data structure, often yielding representations that are brittle under distribution shift and opaque to human scrutiny.The shortcomings of this paradigm appeared particularly evident in cases where there was an inadequate amount of

labeled data which is a common scenario in medical imaging. In such situations a pixel-level annotations for tumor segmentation require skilled radiologists which highly expensive at scale. This challenge brought about a transition to explicit representation learning, in which the representation itself becomes the key objective of optimization, independent of any particular downstream task. Many methods have been proposed, among them Contrastive Learning (CL) has become one of the most powerful and adaptive techniques for explicit representations learning. CL is based on a simple but deep idea: learn a mapping that pulls together representations of "similar" inputs (positive pairs) and pushes apart representations of "dissimilar" inputs (negative pairs). while Generative models reconstruct inputs in pixel space which often have struggle with high-dimensional medical volumes, and discriminative models rely on human-annotated labels, Contrastive algorithms, acquire information by comparing samples while incorporating the structure of unlabeled data to create useful inductive biases. The modern instantiation of this concept originates from the InfoNCE loss [6], which describes contrastive learning as a non-parametric classification challenge: upon receiving a query representation, the model is assigned to discern the corresponding positive key over a set of negatives.This loss function establishes a lower bound on the mutual information across representations of enriched views of the identical input, therefore anchoring contrastive learning in information-theoretic principles. The efficacy of contrastive learning (CL) was solidified by studies such as SimCLR [4] and MoCo [7] , which illustrated that instance discrimination(considering each image as a distinct class) combined with robust data augmentations (e.g., random cropping, color jitter, Gaussian blur), could result in representations that outperform those achieved through supervised pretraining in downstream tasks such as object detection and semantic segmentation. However, Le-Khac et al [8] correctly point out that contrastive learning has roots that go back much further than the rise in popularity from 2018 to 2020. For example, Becker and Hinton's [9] work on maximizing mutual information across stereo views or Siamese networks for signature verification was developed by Bromley et al. [10] The shared intuition that invariance through comparison is an effective technique to learn strong features brings these different approaches into harmony. Standard contrastive learning, despite its empirical successes, is hindered by a significant structural drawback: it promotes global instance-level discrimination without offering a precise mechanism to guarantee the disentanglement of

the learnt representation. In other words, various factors that can change, such as tumor grade, MRI sequence type (T1, T2, FLAIR), scanner manufacturer, or even patient age, are often mixed up in the same latent dimensions. This entanglement results in a reliable representation in differentiating between brain scans, yet inadequately isolates the semantic attributes of interest to clinicians. For example, a model trained on multi-sequence glioma MRI might learn to connect how a tumor looks with the high contrast of post-contrast T1 (T1c) images. Suppose T1c is absent during inference, a common scenario in clinical workflows. In that case, the model's performance may dwindle, not due to a deficiency in anatomical comprehension, but because its representation is entangled with modality-specific aberrations. This fragility conflicts with a fundamental requirement of medical AI: resilience to absent modalities and acquisition protocols. Entanglement is not just a theoretical issue; it has real-world effects on how fair, understandable, and generalizable models are. An entangled representation may unintentionally incorporate erroneous correlations, such as linking disease severity with demographic factors like sex or ethnicity, resulting in biased predictions that worsen health inequities. Furthermore, without disentanglement, it is infeasible to audit the rationale behind a model's decision, thereby eroding clinician trust and regulatory compliance. On the contrary, a disentangled model would assign one subspace to tumor shape, another to imaging physics, and a third to patient-specific anatomy, facilitating clear, modular, and resilient reasoning. To overcome this limitation, a novel research avenue, Disentangled Contrastive Learning (DCL) has been developed at the confluence of self-supervised learning and structured representation theory. DCL aims to maintain contrastive objectives' scalability and empirical efficacy while openly imposing a factorized structure on the latent space. The main idea is that the encoder can be taught to give different semantic or physical aspects their own axes of variation by breaking the representation down into several low-dimensional subspaces and using contrastive learning in an organized way across these subspaces. Formally, let the encoder output a d-dimensional vector $\mathbf{r} = f_\theta(\mathbf{x})$. In DCL, this vector is partitioned into M non-overlapping subspaces:

$$\mathbf{r} = [\mathbf{r}^{(1)}; \mathbf{r}^{(2)}; \ldots; \mathbf{r}^{(M)}],$$

where each $\mathbf{r}^{(m)} \in \mathbb{R}^{d_m}$ is intended to capture a distinct aspect of the input. The DCL

objective then combines a standard contrastive loss with a disentanglement regularizer:

$$\mathcal{L}_{\text{DCL}} = \mathcal{L}_{\text{CL}} + \lambda \cdot \mathcal{R}_{\text{dis}},$$

where $\lambda > 0$ balances the two terms, and $\mathcal{R}_{\text{dis}}$ enforces statistical or geometric independence among the subspaces. Some common ways to do $\mathcal{R}_{\text{dis}}$ are to put orthogonality requirements on projection matrices [11], minimize mutual information between subspaces [12], or use factor-specific contrastive losses to create positive/negative pairings per subspace. [11] The case for DCL in medical imaging, a field characterized by multimodality, a lack of annotations, and safety-critical stakes, is very persuasive. T1 represents anatomy, T2 shows edema, FLAIR blocks cerebrospinal fluid, and T1c shows vascularized tumors. Multi-sequence magnetic resonance imaging (MRI) is one method that provides additional information. A disentangled model can extract information from a shared anatomical subspace that remains invariant to modality, while simultaneously preserving modality-specific signatures in separate subspaces. Because of this structure, it is possible to segment data even when one or more sequences are missing. DCL is capable of identifying the difference between scanner-invariant pathological attributes and acquisition-specific noise in multi-institutional settings, which makes it more generalizable across hospitals. Recent studies have confirmed this intuition: DC-Seg [11] achieved a mean Dice score of 87.54% on the BraTS 2020 brain tumor segmentation benchmark by clearly separating anatomical and modality factors. This meant that it could outperform strong baselines like nnU-Net and even large pre-trained models when a modality was missing. PLGCL [13] utilized pseudo-labels derived from a semi-supervised segmentation model to facilitate contrastive sampling. This led to fewer class collisions and a 6.3% increase in Dice scores on the ACDC cardiac MRI dataset, even though only 10% of the data was labeled. DCL not only enhances performance, but it also makes things easier to understand, which is a crucial prerequisite for clinical application. By looking at activations along disentangled axes, clinicians may be sure that the model is focusing on relevant pathologies rather than false correlations, such as scanner artifacts or patient demographics. This aligns with the European Medicines Agency's (EMA) and the U.S. Food and Drug Administration's [14] guidelines, which underscore the significance of explainable AI in healthcare. Disentangled representations also allow for fine-grained control, which makes it easier to test hypotheses and do counterfactual analysis by allowing one to manipulate a specific latent axis (such

"increase tumor grade") and observe the resulting impact on the output. Despite these advances, significant challenges remain. First, most medical datasets lack ground-truth annotations for underlying factors of variation (e.g., "tumor invasiveness" or "scanner type"), making it difficult to supervise disentanglement directly. Second, extending DCL to full 3D volumes is computationally demanding, as batch sizes are constrained by GPU memory, limiting the number of negative samples—a key ingredient for effective contrastive learning. Third, there is no consensus on the theoretical conditions under which CL plus a regularizer implies disentanglement; much of the current work is empirically driven. Finally, the evaluation of disentanglement itself is nontrivial. Although criteria such as Mutual Information Gap (MIG) and Disentanglement-Completeness-Informativeness (DCI) are well established in synthetic settings, their applicability to actual medical data is constrained. Recent research has identified significant correlations between disentanglement metrics and downstream performance. Locatello et al. [15] reported a correlation of 0.997 between DCI Disentanglement and classification accuracy across 12,000 models, indicating that disentanglement serves as a practical predictor of utility rather than merely a theoretical concept. This work therefore asks whether multiple incomplete notions of similarity can be reconciled inside a shared latent space for medical image classification. More specifically, it investigates whether fragmented expert knowledge can be organized through state-specific relational views, whether weak normal-versus-abnormal supervision can be strengthened by augmentation-based self-supervision, and whether the resulting embeddings remain both geometrically interpretable and useful for downstream classification. The scope is intentionally limited to compact 2D benchmarks rather than full 3D segmentation, missing-modality reconstruction, or formal factor-level disentanglement evaluation. The reminder of this chapter reviews the theoretical foundations of contrastive learning, its general framework, applications across domains, and recent developments in disentangled contrastive learning for medical imaging. Chapter 2 then presents the proposed methodology, including the two learning scenarios, architectural design, dataset preparation, and optimization strategy. Chapter 3 reports the experimental results and embedding analysis across the selected benchmarks. Finally, Chapter 4 discusses the main findings, limitations, and possible future directions.

1.1 Theoretical Foundations and Historical Evolution

The landscape of representation learning has undergone a profound transformation over the past three decades, evolving from handcrafted feature extractors to deep, data-driven encoders capable of capturing intricate patterns across modalities and domains. Within this evolution, contrastive learning (CL) has emerged not as a sudden innovation but as the crystallization of a long-standing intuition: that meaningful representations can be learned by comparing related and unrelated data points. As Le-Khac et al. [8] compellingly argue, the origins of CL trace back to the early 1990s, with foundational work by Becker and Hinton [9] on mutual information maximization across stereo views and Bromley et al. [10] on Siamese networks for signature verification. These early efforts established the core principle that would define CL for decades to come (learning by comparison) a paradigm that stands in contrast to both generative models, which reconstruct inputs, and discriminative models, which predict labels. This section presents the historical and theoretical foundations that informed the proposed methodology, emphasizing how comparison-based learning facilitates structured and disentangled representations in medical imaging classification. This study adheres closely to the metric-learning domain, implementing similarity via state-specific pair formation and employing a margin-based contrastive loss across several projection subspaces.

The theoretical foundation of modern contrastive learning arises from information theory, especially the concept of mutual information (MI). Becker and Hinton [9] suggested that an effective representation must encapsulate the shared structure among several perspectives of the same scene and simoultaneously eliminating view-specific noise. Formally, for two perspectives $\mathbf{x}$ and $\mathbf{x}'$ of the same underlying reality, the goal is to maximize the mutual information between their representations.

$$\mathbf{I} = I(f(x); f(x')) \tag{1.1}$$

However, direct maximizing MI is impractical since it requires estimating high-dimensional probability densities. To address this different methods have been proposed but Noise-Contrastive Estimation (NCE) [16] have gained more attention. The NCE algorithm treating density estimation as a binary classification problem: distinguishing data sam-

ples from noise. Oord, Li, and Vinyals [6] utilized NCE to formulate the InfoNCE loss, establishing a manageable lower constraint on MI[6]:

$$\mathcal{L}_{\text{InfoNCE}}^{(i)} = -\log \frac{\exp\left(\text{sim}(\mathbf{z}, \mathbf{z}^+)/\tau\right)}{\exp\left(\text{sim}(\mathbf{r}_i, \mathbf{r}_i^+)/\tau\right) + \sum_{j \neq i} \exp\left(\text{sim}(\mathbf{r}_i, \mathbf{r}_j)/\tau\right)}, \tag{1.2}$$

where $\mathbf{z}$ and $\mathbf{z}^+$ are representations of a positive pair, $\mathbf{z_k}$ includes one positive and $K-1$ negative samples,sim is a similarity function (e.g., cosine similarity), and τ is a temperature parameter. This loss function not only grounded CL in information theory but also unified disparate approaches under a common objective. Alongside the information-theoretic lineage, the metric learning community developed additional frameworks. Chopra, Hadsell, and LeCun [17] developed the contrastive pair loss for face verification, which reduces the distance between embeddings of the same identity while increasing the distance between embeddings of different identities beyond a certain margin.

$$\mathcal{L}_{\text{pair}} = \begin{cases} \|f(x_i) - f(x_j)\|^2, & \text{if } y_i = y_j, \\ \max(0, m - \|f(x_i) - f(x_j)\|)^2, & \text{if } y_i \neq y_j. \end{cases} \tag{1.3}$$

Later, this was expanded to include the triplet loss [18], [19] which adds a relative distance restriction. This means that the distance between an anchor and a positive must be at least m lower than the distance between the anchor and a negative. These early approaches worked, but they took a long time to converge because there were not enough sample interactions per batch, and "hard" negatives had to be mined carefully. Multi-sample losses, such as the lifted structured embedding loss [20] and the multi-class N pair loss [21], solved this problem by using more than one negative per query. This idea would become very important in current CL. The convergence of these resulted in the sample discrimination paradigm [22], which categorizes each image as a distinct class. This method separates representation learning from human-annotated labels leveraging the vast amount of unlabeled data available in fields like ImageNet. Wu and his colleagues [22] created a memory bank to hold embeddings, which separated the number of negatives from the batch size. MoCo [7] improved on this by using a momentum-updated encoder and a dynamic queue to keep a steady set of negatives. SimCLR[4] demonstrated that combining powerful data augmentations (such as random cropping and color distortion) with a projection head and

high batch sizes can achieve the best transfer performance without a memory bank. These works set up the contemporary CL framework: augment, encode, project, and contrast.

1.2 General Framework for Contrastive Learning

Le-Khac et al. [8] suggested a general Contrastive Representation Learning (CRL) framework that brings together different contrastive learning methods. The CRL framework has four key elements, each with several design options. In the present study, these same four elements are instantiated through state-dependent pair definitions, a pretrained visual backbone with a lightweight encoder, learned projection heads for specialized 3D subspaces, and a contrastive objective tailored to each relational state.

1.2.1 Distributions of Similarity and Dissimilarity

The core of any contrastive learning system is the delineation of what qualifies as a "similar" (positive) or "dissimilar" (negative) pair. Le-Khac et al. [8] classify definitions of similarity into five general strategies. The first strategy leverages multi-sensory signals, which utilize natural connections between various types of data, such as the correspondence between audio and video in videos [23] or the alignment between L/ab channels in color images [24]. The second strategy employs data transformation, using semantic-preserving augmentations to generate diverse representations of the same instance. Examples include cropping, rotating, and color jittering in images [4]; back-translation in text [25]; and SpecAugment in audio [26]. The third strategy relies on connections between context and instances, contrasting the global context (such as a summary vector of an image) with local instances (such as patches in a feature map), as demonstrated in Deep InfoMax [27] and Contrastive Predictive Coding [6]. The fourth strategy uses sequential coherence, employing temporal or spatial continuity to categorize neighboring frames in time or space as positive and those that are distant as negative [28]. The fifth strategy applies natural clustering, using cluster assignments (e.g., from k-means) as pseudo-labels to delineate positives within the identical cluster [29]. This taxonomy demonstrates the versatility of contrastive learning:

the same basic framework can be used for supervised, self-supervised, or even unsupervised clustering by merely changing the similarity distribution. In the framework developed later in this thesis, that definition becomes explicitly state-dependent: what counts as a positive or negative pair changes with the active relational hypothesis rather than remaining fixed throughout training.

1.2.2 Encoders

The encoder $e(.)$ maps raw inputs to a representation space. Consistent with this taxonomy, the current study uses a pretrained backbone for efficient feature extraction and a lightweight trainable encoder to map those features into a compact shared latent space. Le-Khac et al. [8] categorize encoder update strategies into three types. The first is end-to-end updating, where both query and key encoders are updated directly via backpropagation (e.g., SimCLR). While simple, this approach is memory-intensive. The second strategy is online-offline updating, where the query encoder is updated online while the key encoder is updated offline, either via a momentum average as in MoCo [7] or using a memory bank as in Wu et al. [22]. This approach decouples memory from batch size. The third strategy uses a pre-trained encoder, where one encoder is frozen, such as a teacher network in knowledge distillation [24] or a pre-trained BERT model in cross-modal learning [30].

1.2.3 Transform Heads

The transform head $h(.)$ converts the encoder output into a metric embedding space optimized for contrastive loss. This role is especially central in the present work, where separate learned projection heads define distinct 3D relational views of a shared 4D representation. Three distinct head types have been identified in the literature. Projection heads are straightforward multi-layer perceptrons (MLPs) designed to map data into a lower-dimensional space, as discussed by Chen et al. [4]. The separation of representation learning from the contrastive objective that projection heads provide frequently results in improved transfer performance. Contextualization heads consolidate data over temporal or spatial dimensions, as exemplified by Gated Recurrent Units in Contrastive Predictive

Coding [6] and pooling mechanisms in Deep InfoMax [27]. Quantization heads map continuous representations to discrete codes, such as the Gumbel-softmax utilized in wav2vec 2.0 [31] and the Sinkhorn-Knopp algorithm employed in SwAV [29].

1.2.4 Contrastive Loss Functions

Contrastive loss functions are classified according to their foundational theoretical principles. Energy-based margin losses comprise pair or triplet losses that enforce distance margins, as delineated by Chopra et al. [17]. Probabilistic noise-contrastive estimation (NCE)-based losses, such as InfoNCE and NT-Xent, conceptualize contrastive learning as a classification task, as suggested by Oord et al. [6] and Chen et al. [4]. Loss functions based on mutual information directly enhance lower bounds of mutual information, as demonstrated by Hjelm et al. [27]. This framework illustrates that the effectiveness of modern contrastive learning methods, such as SimCLR [4] and MoCo [7], arises from a synergistic integration of robust augmentations, extensive negative sets, projection heads, and the InfoNCE [6] loss function. For the present study, however, the most directly relevant branch is the energy-based formulation, since the proposed multi-state model applies a margin-based contrastive loss within each relational subspace.

1.3 Applications Across Domains

CL shows considerable flexibility when dealing with various data formats. In computer vision, CL has functioned as a principal testbed, transcending the parameters of basic instance discrimination as described by Wu et al. [22] and Chen et al. [4]. The approach has offered a more straightforward time representation, utilizing methods such as Time-Contrastive Networks [28] and Dense Predictive Coding [32], which do an exceptional job of capturing spatio-temporal dependencies from video data. Moreover, cross-modal alignment methodologies, exemplified by Audio-Visual Embedding Networks [23], facilitate the synchronization of visual and auditory information. Other methods, such as SwAV [29] and Prototypical CL [33], utilize clustering objectives to enhance the semantic

structure. In natural language processing (NLP), CL enables the creation of representations at both the word and sentence levels. The Skip-gram with Negative Sampling model [34] is an early version of CL. Models like Quick-Thought [35] and Sentence-BERT [36] train sentence-level embeddings using triplet or InfoNCE losses. A contrastive technique has been employed to examine BERT's next-sentence prediction problem, focusing on how well text segments cohere sequentially. [37] In the realm of audio and voice, Contrastive Learning (CL) facilitates the unsupervised acquisition of speech representations without the need for tagged transcripts. CPC [6] and wav2vec [38] are examples of models that learn frame-level features, which aid in phoneme classification. Wav2vec 2.0 [31] combines quantization with Transformer topologies to achieve the best performance in automated speech recognition (ASR). CL has been adapted for graph-structured data, enabling the acquisition of embeddings at both the node and graph levels. The goal of Deep Graph Infomax [39] is to maximize the mutual information between local and global representations of graphs.Graph Contrastive Coding [40], on the other hand, views subgraphs as contrastive examples and utilizes graph diffusion to augment them. This breadth of successful application also motivates the experimental design of the present study, which evaluates a single structured contrastive framework across handwritten digits, histopathology, retinal OCT, and blood-cell microscopy.

1.4 DCL in Medical Imaging

Although it was published prior to the recent wave of Disentangled Contrastive Learning (DCL) applications in medical imaging, the survey carried out by Le-Khac et al. [8] still provides a useful conceptual foundation across multiple disciplines. In medical imaging, the central motivation for disentanglement remains clear: clinical data are heterogeneous, labels are often incomplete, and meaningful variation is frequently mixed with acquisition-driven variability. In the present project, that motivation is narrowed to a more concrete experimental setting (2D medical image classification under incomplete, asymmetric, or weak supervision) where the aim is to determine whether a compact multi-state representation can recover useful class structure without access to full fine-grained labels.

1.4.1 Key Methodologies in Medical Imaging

Recently, Many papers aiming at significant methodological advancements in contrastive learning for medical image analysis have been published. In 2023, Zhou et al. presented an approach called DC-Seg, a bidirectional contrastive loss method for segmenting brain tumors.[11] The anatomical loss aligns $\mathbf{r} * \text{anat}$ across modalities for the same patient, whereas the modality loss aligns $\mathbf{r} * \text{mod}^{(M)}$ across patients for the same modality. Upon evaluation using the BraTS 2020 dataset, DC-Seg attains a Dice score of 87.54%, surpassing nnU-Net, especially in instances involving missing modalities. Pseudo-labels guided contrastive learning (PLGCL) was another method introduced by basak et al. [13], which presents a semi-supervised framework. The PLGCL utilizes pseudo-labels generated by a teacher model to guide contrastive sampling, thereby alleviating class collision problems frequently encountered in dense prediction tasks. This method gains a 6.3% enhancement in Dice score on the ACDC dataset with only 10% of the data labeled. In DDCL [41], orthogonality is maintained across distortion-invariant (DIR) and distortion-variant (DVR) subspaces by the regularization term $|\mathbf{W} * \text{DIR}^{\top}\mathbf{W} * \text{DVR}|_F^2$. This method improves the resilience of the acquired representations against scanner-induced artifacts. Ultimately, CF-SimCLR [42] employs counterfactual augmentation to replicate scanner changes, allowing the model to acquire domain-invariant features for enhanced pneumonia diagnosis across diverse imaging settings. These studies establish the broader methodological case for structured contrastive objectives in medicine. However, the present work shifts the problem from segmentation and missing-modality robustness to classification under partial supervision. Instead of relying on full volumetric MRI or dense pseudo-label guidance alone, it asks whether partial expert knowledge, augmentation-based supervision, orthogonal projection heads, and cyclic multi-state training are sufficient to produce compact 4D representations that remain interpretable in geometry and strong in downstream accuracy across MedMNIST benchmarks.

Chapter 2

Methodology

2.0.1 Motivation

The motivating idea of this work is illustrated using the analogy of several medical students, each possessing only partial knowledge about the relationships between tissue classes (e.g, Cancer, Stroma, Normal). Each student understands certain similarities and differences among tissue types, but none has access to the complete relational structure. In practice, reliable labels may be unavailable, and expert knowledge can be fragmented and costly to obtain. A fundamental challenge in many classification pipelines arises from the implicit reliance on Euclidean distance as a measure of similarity. Euclidean distance operates at the pixel level and assumes that small numerical differences correspond to semantic similarity. However, this assumption is frequently violated in medical imaging. Two pathology images may appear highly similar to the human observer, while exhibiting a large Euclidean distance due to variations in staining, illumination, or local texture. Figure 2.1 illustrates a simple example in which two binary representations appear visually similar, yet their Euclidean distance is large. Formally, consider the two vectors.

$$\mathbf{X}_1 = [1, 0, 1, 0, 1, 0], \quad \mathbf{X}_2 = [0, 1, 0, 1, 0, 1].$$

Their Euclidean distance is given by $\mathbf{D} = \|X_1 - X_2\|_2$, with

$$D^2 = (1-0)^2 + (0-1)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2 + (0-1)^2 = 6,$$

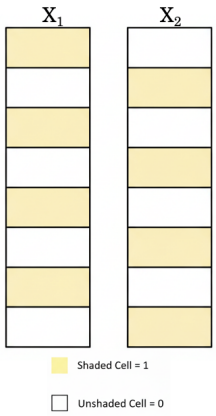

Figure 2.1: Limitation of Euclidean distance in original space: Visually similar representations with large Euclidean distance.

Their Euclidean distance is given by $D = \|X_1 - X_2\|_2 = \sqrt{6} \approx 2.45$. Despite the apparent structural similarity between the arrays, the Euclidean distance suggests they are far apart. This mismatch indicates that the original image space is not appropriate for measuring meaningful similarity or dissimilarity. Contrastive learning provides a principled framework for explicitly defining pairwise relationships, reshaping the geometry of the latent space to align with human perception. However, in this setting, a single global contrastive objective can conflate incompatible similarity signals, motivating multiple relational views. This necessitates a representation learning framework designed to leverage partial expert signals without collapsing them into a single inconsistent objective.

2.0.2 Core Research

Medical image classification plays a critical role in clinical decision-making, yet its performance is often limited by incomplete or uneven supervision. To address the lack of accessible fine-grained annotations, this work investigates a combined self-supervised visual learning framework augmented with partial expert knowledge. This approach systematically explores several benchmarks. Initially, the feasibility and behavior of the contrastive multi-state method using a controlled proof-of-concept on the MNIST Digits dataset [43]

was examined. Following this validation, the analysis was extended to realistic medical diagnostic workflows using MedMNIST datasets [2, 3], specifically PathMNIST [44], OCTMNIST [45], and BloodMNIST‘[46]. To reflect asymmetric knowledge distribution, we explicitly model two complementary learning scenarios that share the same architectural backbone but differ in the supervision available during training:

- **Scenario 1: Partial-Knowledge Contrastive Learning (PKCL)**: Multiple technicians possess class-specific expertise: each expert can reliably recognize one target class while can not consistently distinguish the others. Supervision is incomplete and structured, as some class distinctions are reliable while others are misinterpreted.
- **Scenario 2: Semi-Supervised Contrastive Learning (SSCL)**: Each annotator can only separate normal from abnormal samples, but not the fine-grained abnormal categories. To compensate for the absence of detailed annotations, augmentation-driven self-supervision is utilized.

Both scenarios use the same 4D representation framework, enabling a controlled study of how supervision structure influences representation quality and downstream classification.

2.1 Architectural Framework

The framework is designed to learn structured and disentangled representations under incomplete expert annotation. It consists of four main components: a convolutional backbone (MobileNetV2 [1]), a lightweight encoder producing a shared 4D latent representation, multiple projection heads [4] that map this representation into state-specific 3D subspaces, and a cyclic state-based training strategy. This design extends the foundational siamese learning paradigm [47] to a multi-state setting, allowing the encoder to synthesize fragmented clinical expertise within a unified embedding space.

As illustrated in Figure 2.2, the proposed architecture is shared across both learning scenarios and consists of several key components described below.

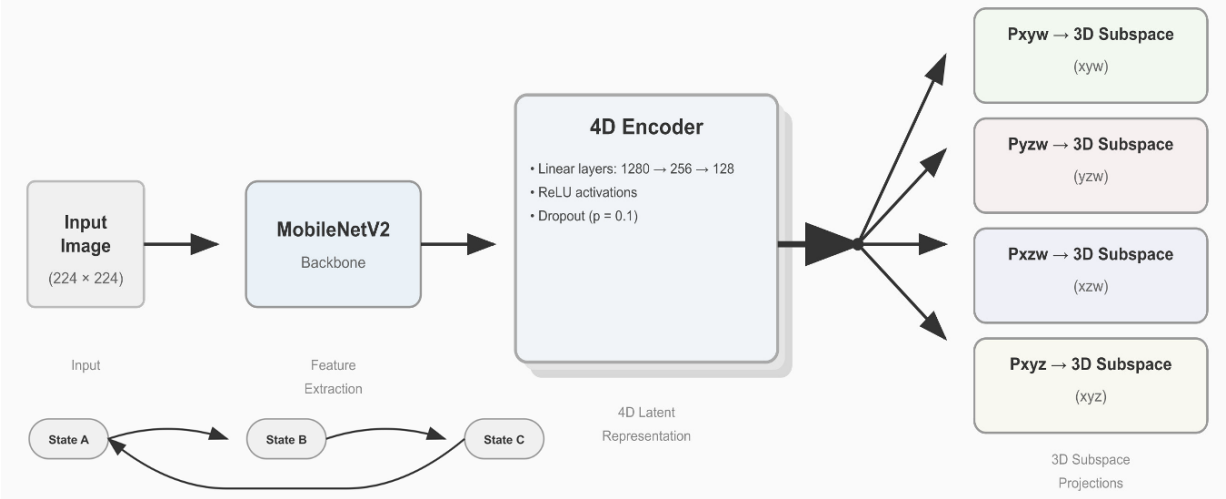

Figure 2.2: Overview of the proposed 4D representation framework. An input image is processed by a MobileNetV2 backbone [1], followed by a lightweight encoder that maps features into a shared 4D latent representation. The 4D embedding is projected into multiple 3D subspaces using learnable projection heads, each corresponding to a relational state. Training proceeds in a cyclic manner across states A, B, and C.

Backbone Network. Given an input image $x \in \mathbb{R}^{3\times224\times224}$, feature extraction is performed using a MobileNetV2 [1] backbone pretrained on ImageNet [48]. MobileNetV2 is chosen due to its lightweight design and ability to produce stable, high-level visual features while maintaining computational efficiency. The backbone outputs a feature vector of dimension 1280.

4D Encoder. A lightweight multi-layer perceptron (MLP) maps the extracted features to a shared 4D embedding, $f_\theta : \mathcal{X} \to \mathbb{R}^4$, such that $z = f_\theta(x)$. The MLP utilizes the layer structure

$$1280 \to 256 \to 128 \to 4,$$

incorporating ReLU activations after each intermediate layer and a dropout rate of $p = 0.1$ for regularization [49]. The ReLU activation function [50]introduces nonlinearity while maintaining strong and stable gradients during training. The final output is a shared 4D latent vector $z \in \mathbb{R}^4$ that serves as a common representation across all relational states.

Projection Heads and Subspaces. To represent different relational views, we learn three linear projection matrices during training:

$$P^{(k)} \in \mathbb{R}^{3\times 4}, \quad \text{where } k \in \{XYW, YZW, XZW\}.$$

Each row of $P^{(k)}$ is ℓ_2-normalized before projection to ensure numerical stability:

$$\hat{P}^{(k)}_{r,i} = \frac{P^{(k)}_{r,i}}{\|P^{(k)}_{r,i}\|_2 + \epsilon}, \quad \epsilon = 10^{-8}.$$

A fourth matrix, $P^{(XYZ)}$, is maintained as a fixed, non-learned auxiliary head used solely for t-SNE visualization [51] and does not participate in the training loss or the orthogonality regularizer.

Relational States and Cyclic Training. The system learns a unified embedding space trained under multiple relational states, where each state defines similarity through expert-specific positive and negative pairings. During training, the states use the fixed mapping $k(A) = XYW$, $k(B) = YZW$, and $k(C) = XZW$. Training across these states is performed in a cyclic manner:

$$\text{State A} \rightarrow \text{State B} \rightarrow \text{State C} \rightarrow \text{State A}.$$

The loop back to State A allows the encoder to iteratively incorporate complementary relational signals while maintaining a shared latent space.

2.2 Rationale behind Model Selection

Transfer learning (TL) is a well-established and scientifically supported method that involves employing a pretrained convolutional neural network—such as MobileNetV2 [1]—and retraining only the specialized final layers. The underlying reasoning is that deep neural networks learn hierarchical feature representations when trained on large, varied datasets, allowing these representations to be generalized across a wide range of visual tasks. The initial layers of a neural network frequently capture low- and medium-level features, such as

edges, textures, and shapes, while the deeper layers encode more complex semantic structures. Consequently, models can perform well on new tasks even with limited target data, as these learned representations are highly transferable, as demonstrated by Yosinski et al. [52]. Because the computationally expensive early stages of feature learning (including the discovery of general visual features) are completed during pretraining, only the final lightweight encoder and projection heads need to learn the mapping from these generic features to the new task labels. Retraining the last layers while keeping the backbone network fixed or partially frozen significantly reduces the risk of overfitting, especially when the target medical datasets are comparatively small. Studies on fine-tuning and transfer learning in deep models show that using pretrained backbones improves convergence speed, stability, and sample efficiency compared to training a full model from scratch [53, 54]. In particular, MobileNetV2 is designed with efficient separable depthwise convolutions and inverted residual blocks [1]. This makes it exceptionally well-suited for our transfer learning scenarios, where the goal is to achieve strong representation learning performance with low computational cost. Furthermore, the loss function is not strictly tied to the pretrained architecture; adapting it for the newly added multi-state projection layers is a standard practice that enables the model to accommodate our novel state-dependent contrastive learning objectives.

2.3 Datasets

To comprehensively evaluate our representation learning framework, a foundational proof-of-concept dataset alongside three realistic medical imaging benchmarks from the MedMNIST collection [2, 3] were utilized. Figure 2.3 illustrates representative samples from the medical datasets.

2.3.1 MNIST (Digits)

Initially, we use the MNIST dataset [43] as a controlled testbed to assess the efficiency and feasibility of the contrastive learning technique. We restrict the dataset to digits $\{0, 6, 9\}$ to simulate visually similar but distinct classes.

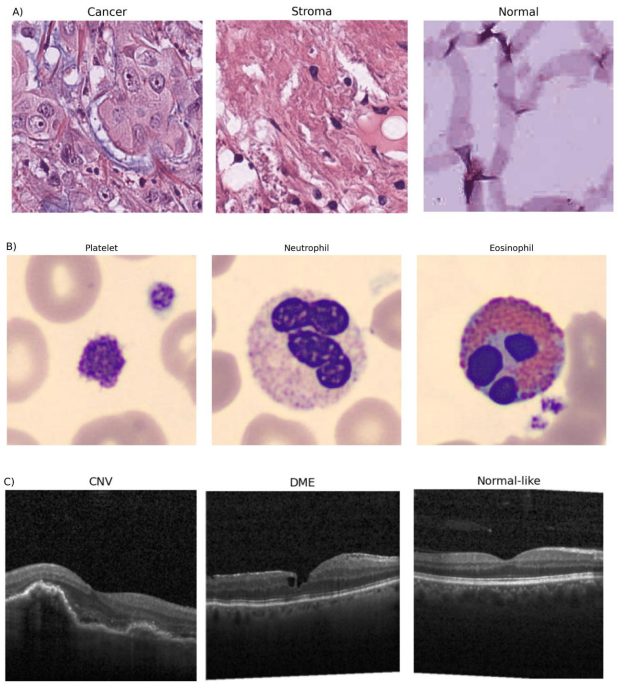

Figure 2.3: Sample images from the three medical datasets [2, 3]: (A) PathMNIST, (B) BloodMNIST, (C) OCTMNIST.

2.3.2 PathMNIST

The PathMNIST dataset [44] is the primary medical benchmark for this pipeline. It is derived from the widely-used NCT-CRC-HE-100K collection, comprising 100,000 non-overlapping image patches of hematoxylin & eosin-stained colorectal cancer (CRC) and normal colon tissue. Originally prepared as a 9-class classification task, PathMNIST serves as a valuable lightweight benchmark for computational pathology, allowing us to test models under realistic medical-image conditions.

2.3.3 BloodMNIST

Finally, we utilize BloodMNIST [46], a dataset of microscopic peripheral blood cell images, to validate the framework's ability to handle highly localized cellular structures.

2.3.4 OCTMNIST

To test the model's flexibility across different imaging modalities, we incorporate OCTMNIST [45], which consists of optical coherence tomography (OCT) images of the retina.

2.4 Data Preprocessing and Label Remapping

For all datasets, inputs are resized to 224×224 pixels to match the requirements of the MobileNetV2 [1] feature extractor. Due to the diverse nature of the datasets, specific preprocessing and label remapping strategies were applied to each.

Label Remapping: To align with clinical interpretability and diagnostic triage scenarios, we perform a semantic remapping to create consistent 3-class settings across all tasks:

- **MNIST Digits**: Labels are remapped straightforwardly as $0 \rightarrow 0$, $6 \rightarrow 1$, and $9 \rightarrow 2$.
- **PathMNIST**: We collapse the original nine classes into three clinically meaningful super-classes: *Cancer* (corresponds exclusively to colorectal adenocarcinoma epithelium), *Stroma* (cancer-associated stroma), and *Normal* (aggregates adipose, lymphocytes, mucus, smooth muscle, and normal colon mucosa). Background and debris classes are discarded.
- **BloodMNIST**: We select *Platelet*, *Neutrophil*, and *Eosinophil*, representing clinically meaningful partial expertise categories (coagulation, inflammation, and allergy).
- **OCTMNIST**: We retain *CNV*, *DME*, and *Normal*, while dropping the *DRUSEN* class.

Table 2.1 summarizes the effect of this label remapping on the overall dataset sizes. By isolating our clinically relevant super-classes and discarding extraneous data, the official MedMNIST published counts are reduced to our final filtered splits. All subsequent contrastive learning experiments are conducted exclusively on these filtered subsets.

Table 2.1: Dataset split sizes at each stage of the preprocessing pipeline. *Official* denotes the MedMNIST published counts [2, 3] across all original classes. *Filtered* represents the final counts after restricting the data to the three chosen target classes (e.g., background/debris discarded in PathMNIST; DRUSEN removed in OCTMNIST; 5 of 8 cell types removed in BloodMNIST). All experiments utilize these filtered splits.

Dataset	**Retained Classes**	**Official (Train / Val / Test)**			**Filtered (Train / Val / Test)**		
		Train	Val	Test	Train	Val	Test
PathMNIST	Cancer, Stroma, Normal	89,996	10,004	7,180	83,484	9,947	5,496
BloodMNIST	Platelet, Neutrophil, Eosinophil	11,959	1,712	3,421	11,959	1,712	1,757
OCTMNIST	CNV, DME, Normal	97,477	10,832	1,000	97,477	10,832	750

Standardization: For the Digits, PathMNIST [44], and BloodMNIST [46] datasets, inputs are normalized using standard ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$) [48]. For OCTMNIST [45], which is RGB, pixel intensities are instead scaled directly to $[-1, 1]$ to ensure stable intensity scaling.

2.4.1 Color Normalization Methodology (PathMNIST Only)

Macenko Stain Normalization [55]

[55]Due to the fact that PathMNIST comprises diverse histological slides, we apply the Macenko stain normalization algorithm exclusively to this dataset to reduce stain variability across Hematoxylin-Eosin (H&E) images. The other datasets (Digits, OCTMNIST, BloodMNIST) do not exhibit H&E stain variance and therefore skip this step. The algorithm first converts each RGB image into optical density (OD) space using the Beer-Lambert law. Given an image $I \in [0, 255]^{H \times W \times 3}$, OD values are computed as:

$$\mathrm{OD}(x) = -\log\left(\frac{I(x) + \epsilon}{255}\right),$$

where a small constant ϵ ensures numerical stability. Pixels with weak staining are removed using a luminosity threshold. Let OD_i denote the OD vector of the i-th pixel. Pixels satisfying $\|\mathrm{OD}_i\|_1 > \tau$ with $\tau = 0.8$ are retained. For the filtered set of OD vectors, singular

value decomposition (SVD) is applied to estimate the principal directions of stain variation: $M = U\Sigma V^\top$. The first two columns of V define a two-dimensional subspace capturing the dominant stain components. Angles are computed using $\phi_j = \arctan 2(V_{2j}, V_{1j})$, and the 1st and 99th percentile values determine the hematoxylin and eosin stain vectors.

Stain concentrations for each pixel are estimated via least-squares:

$$C_s = \arg\min_C \|M - H_s C\|_2,$$

where H_s is the estimated source stain matrix. To map the image into a standardized stain space, concentrations are rescaled and recombined with a reference stain matrix H_{ref}: $\text{OD}_n = H_{\text{ref}} C_n$. Finally, the normalized RGB image is reconstructed:

$$I_n(x) = 255 \exp(-\text{OD}_n(x)).$$

This pipeline effectively mitigates PathMNIST stain variability caused by slide preparation and scanner differences, improving the reliability of downstream representation learning.

2.5 State-Dependent Relational Semantics and Pair Construction

A central innovation of this work lies in the introduction of **relational states**, which redefine the notion of similarity not as a fixed binary relation but as a *contextual* and *state-contingent* construct. We define three distinct semantic states, denoted **A**, **B**, and **C**, each encoding a different clinical or biological hypothesis about class relationships.

2.5.1 Pair Construction Methodology

For a given batch of size n, the algorithm considers all unique unordered pairs (i, j) where $i \neq j$. For each state $s \in \{A, B, C\}$, we explicitly define sets of positive pairs $\mathcal{P}_s$ (to be pulled closer in embedding space) and negative pairs $\mathcal{N}_s$ (to be pushed apart).

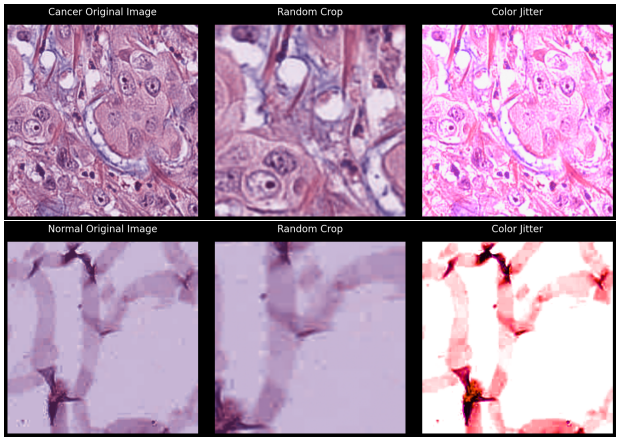

Figure 2.4: Example of data augmentation [4]used in Scenario 2 (SSCL). For each input image, two augmented views are generated using random cropping and color jitter. These augmentations preserve semantic structure while introducing appearance variability.

Scenario 1 (PKCL) Pair Construction

In the PKCL setting, each relational state simulates a specific annotator possessing asymmetric, partial knowledge of the tissue classes: Cancer (0), Stroma (1), and Normal (2). Since a single uniform similarity metric cannot capture this fragmented expertise, each state defines its own "two-vs-one" grouping to construct positive pairs ($\mathcal{P}$) that are pulled together, while negative pairs ($\mathcal{N}$) that are pushed apart in the latent space:

- **State A (Confuses Cancer and Stroma):** This expert correctly isolates Normal tissue (2) but cannot distinguish between Cancer (0) and Stroma (1). Thus, classes 0 and 1 are treated as semantically similar (forming positive pairs), while class 2 is contrasted against them (forming negative pairs).
- **State B (Confuses Stroma and Normal):** This expert reliably isolates Cancer (0) but misinterprets Stroma (1) and Normal (2) as similar. Therefore, classes 1 and 2 form the positive pairs, while class 0 is isolated as the negative contrast.
- **State C (Confuses Cancer and Normal):** This expert recognizes the cancer-associated Stroma (1) but confuses Cancer (0) and Normal (2). Consequently, classes

0 and 2 are grouped as positive pairs, and class 1 acts as the negative contrast.

Formally, for each state $s \in \{A, B, C\}$, we explicitly define the positive pair sets $\mathcal{P}_s$ and negative pair sets $\mathcal{N}_s$ as follows:

$$\mathcal{P}_A = \{(0,0), (1,1), (0,1)\}, \qquad \mathcal{N}_A = \{(0,2), (1,2), (2,2)\} \tag{2.1}$$

$$\mathcal{P}_B = \{(1,1), (2,2), (1,2)\}, \qquad \mathcal{N}_B = \{(0,0), (0,1), (0,2)\} \tag{2.2}$$

$$\mathcal{P}_C = \{(0,0), (2,2), (0,2)\}, \qquad \mathcal{N}_C = \{(1,1), (0,1), (1,2)\} \tag{2.3}$$

By cycling through these conflicting relational perspectives during training, the shared 4D encoder learns a unified latent representation that simultaneously respects the valid structural distinctions provided by all three partial experts without collapsing under their individual contradictory labels.

Scenario 2 (SSCL) Pair Construction: In the SSCL setting, we rely on a single weak expert alongside self-supervised data augmentation (illustrated in Figure 2.4):

- **State A (Partial Expert)**: We simulate an expert who only recognizes the easiest class (e.g., *Normal*). *Normal–Normal* pairs are positive, and *Normal–Non-Normal* pairs are negative. We do not construct pairs from two unknown abnormal samples.
- **State B (Augmentation-Based Self-Supervision)**: Each sample i contributes one positive pair between its base view $\mathbf{z}_{1_i}$ and strongly augmented view $\mathbf{z}_{2_i}$. Negative pairs are formed by pairing the anchor with $K = 25$ randomly sampled different images from the mini-batch.
- **State C (Hybrid)**: Combines expert constraints from State A with the augmentation-based positives from State B. If conflicts arise, the expert label takes priority.

For states defined through expert label relations, let

$$\mathcal{P}_s^\top = \{(b,a) \,:\, (a,b) \in \mathcal{P}_s\}, \qquad \mathcal{N}_s^\top = \{(b,a) \,:\, (a,b) \in \mathcal{N}_s\}.$$

A pair is then treated as positive if $(y_i, y_j) \in \mathcal{P}_s \cup \mathcal{P}_s^\top$, and as negative if $(y_i, y_j) \in \mathcal{N}_s \cup \mathcal{N}_s^\top$. During training, batches cycle through states A $\rightarrow$ B $\rightarrow$ C in a round-robin schedule, allowing the shared encoder to iteratively integrate these complementary relational signals into a unified 4D representation.

2.6 Detailed Algorithmic Optimization

2.6.1 State-Specific Contrastive Loss

For a given batch and active state $s \in \{A, B, C\}$, all valid unordered pairs (i, j) with $i \neq j$ are enumerated. If $(y_i, y_j) \in \mathcal{P}_s$, the pair is labeled positive ($\ell_{ij}^{(s)} = 1$); if it is in $\mathcal{N}_s$, it is labeled negative ($\ell_{ij}^{(s)} = 0$). Let $\mathbf{z}_i^{(s)}, \mathbf{z}_j^{(s)} \in \mathbb{R}^3$ denote the corresponding projected embeddings in the state-specific subspace, defined as $\mathbf{z}_i^{(s)} := \hat{\mathbf{P}}^{(k(s))} f_\theta(x_i)$.

The distance computation uses pairwise Euclidean distance:

$$d_{ij}^{(s)} = \left\| \mathbf{z}_i^{(s)} - \mathbf{z}_j^{(s)} \right\|_2 \tag{2.4}$$

The contrastive loss for the pair follows the margin-based formulation introduced by Hadsell et al. [47]:

$$\mathcal{L}_s^{(ij)} = \ell_{ij}^{(s)} \left(d_{ij}^{(s)} \right)^2 + \left(1 - \ell_{ij}^{(s)} \right) \max \left(0, m - d_{ij}^{(s)} \right)^2 \tag{2.5}$$

Where $m > 0$ is the margin, set to $m = 1.0$ in all experiments (selected via the validation set). The state loss $\mathcal{L}_s$ is the mean over all valid pairs (N_s) constructed under state s in the mini-batch:

$$\mathcal{L}_s = \frac{1}{N_s} \sum_{(i,j) \in \mathcal{P}_s \cup \mathcal{N}_s} \mathcal{L}_s^{(ij)}, \tag{2.6}$$

where $N_s = |\mathcal{P}_s \cup \mathcal{N}_s|$.

2.6.2 State Weighting Scheme

In the Semi-Supervised Contrastive Learning (SSCL) scenario, the integration of the three learning states employs a weighted loss formulation. At each round-robin step t, the

gradient is taken with respect to $w_{s_t}\mathcal{L}_{s_t} + \lambda\mathcal{L}_{\text{reg}}$, where s_t is the active state. The expected per-step objective is:

$$\mathcal{L}_{\text{SSCL}} = \mathbb{E}_{s\sim\text{Uniform}\{A,B,C\}}\big[w_s\,\mathcal{L}_s\big] + \lambda\,\mathcal{L}_{\text{reg}} \tag{2.7}$$

For PathMNIST, all reported SSCL results use fixed state weights of $w_A = 0.5$, $w_B = 3.0$, and $w_C = 1.5$, which yielded stable performance. For BloodMNIST and OCTMNIST, we instead adopted a learnable-weight parameterization, implemented as a softmax over learnable logits. The final learned weights and corresponding results for these datasets are reported in Appendix A.

2.6.3 Orthogonality Regularization

To encourage independence among the axes within each 3D subspace, we penalize deviations from orthogonality in the projection matrices [11]. Crucially, the regularizer is applied to $\hat{\mathbf{P}}^{(k)}$—the same row-normalized matrix used in the forward pass—ensuring consistency between training and regularization. For a given normalized projection matrix, the regularization loss is:

$$\mathcal{L}_{\text{ortho}}(\hat{\mathbf{P}}^{(k)}) = \left\|\hat{\mathbf{P}}^{(k)}(\hat{\mathbf{P}}^{(k)})^\top - \mathbf{I}_3\right\|_F^2 \tag{2.8}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\mathbf{I}_3$ is the 3×3 identity matrix. The total regularization loss aggregates over the three actively learned projections (excluding the auxiliary XYZ projection):

$$\mathcal{L}_{\text{reg}} = \sum_{k\in\{\text{XYW,YZW,XZW}\}} \mathcal{L}_{\text{ortho}}\big(\hat{\mathbf{P}}^{(k)}\big) \tag{2.9}$$

The parameter $\lambda > 0$ controls the strength of the projection regularizer.

2.7 Subspace Specialization Strategy

The framework maintains consistent subspace specialization across all implementations, projecting the 4D latent vector into specialized 3D views:

- **XYW Subspace:** Dedicated to State A learning, capturing expert-defined class distinctions (e.g., Normal versus Abnormal).
- **YZW Subspace:** Dedicated to State B learning, capturing augmentation-invariant visual features derived from self-supervision.
- **XZW Subspace:** Dedicated to State C learning, integrating the hybrid signals of expert knowledge and visual augmentations.
- **XYZ Subspace:** Maintained as a fixed, non-learned auxiliary head used strictly for multi-angle representation viewing and geometric investigation.

2.8 Downstream Classification and Evaluation Protocol

The classifier consists of a 2-layer MLP. The first layer maps the 4D embedding to a 64-dimensional hidden representation, followed by ReLU activation [50] and Dropout [49] with rate $p = 0.2$. A final linear layer produces C output logits. The classifier is optimized using the Adam optimizer [56] with standard cross-entropy loss over the TRAIN split only. Validation sets are used only for encoder checkpoint selection and not for classifier training. Final evaluation is performed on the held-out TEST set using overall accuracy, macro-averaged F1-score, and per-class precision and recall.

2.9 Visualization Strategies

Two complementary visualization techniques are employed to assess the quality and geometric disentanglement of the learned representations:

1. **t-SNE Embedding**: A 2D t-SNE [51](perplexity 50, PCA initialization, 1500 iterations) is applied to the frozen 4D latents of a balanced subset to reveal the global clustering structure across the unified representation.

2. **Multi-Angle 3D Projections**: For each of the four learned subspaces (xyw, yzw, xzw, xyz), the 3D coordinates are plotted from front, side, and top views.

For clarity, the visualizations discussed in the main text focus on PathMNIST. The corresponding scatter plots and t-SNE visualizations for BloodMNIST and OCTMNIST are presented separately in Appendix A.

Chapter 3

Result and Evaluation

3.1 Results on MNIST (Digits Proof-of-Concept)

In the first scenario, we aimed to validate the concept of "three medical students with partial knowledge" in a simple, controlled setting before advancing to complex medical images. To achieve this, we utilized three visually similar MNIST digits [43](0, 6, and 9) and treated them as distinct classes. In real medical training, a student may detect normal versus abnormal tissue but may fail to correctly separate more subtle sub-categories (like cancer vs. stroma). We simulated this asymmetric knowledge on the handwritten digits: each "student" (States A, B, and C) possesses a mix of correct and incorrect beliefs regarding class similarities. For example, a student might correctly match Digit 0 with another Digit 0, but incorrectly assume that 0 and 6 are also semantically similar. Each state defines its own pattern of positive and negative pairs, training the encoder within a different 3D projection of the shared 4D embedding. After CL pretraining, a simple linear classifier was trained on top of the frozen embeddings. As shown in Table 3.1, the results on the test set are nearly perfect, demonstrating that even with strictly incomplete and fragmented knowledge, the model successfully learns embeddings that are linearly separable. We visualized all 3D projections in a single figure (Figure 3.1). Each projection demonstrates how one student's partial knowledge physically shapes a specific view of the latent space. For instance, in the XYW projection (State A), Digits 0 and 6 are

Table 3.1: Classification Performance on MNIST (Digits 0, 6, 9)

Class	Precision	Recall	F1-score
Digit 0	0.99	0.99	0.99
Digit 6	0.99	1.00	0.99
Digit 9	1.00	0.99	1.00
Accuracy		99%	

pulled together into a single "confused" cluster, while Digit 9 is geometrically isolated. Conversely, the YZW projection (State B) isolates Digit 0 while merging 6 and 9, and the XZW projection (State C) isolates Digit 6 while merging 0 and 9. The XYZ projection provides an additional, auxiliary geometric perspective. Because each state adds a different type of partial supervision, the model is forced to resolve these conflicting views. Figure 3.2 illustrates the global t-SNE of the unified 4D embeddings. Even though the encoder was never provided with a complete set of labels separating all three digits simultaneously, the final embedding successfully disentangles them into three distinct, tight clusters. This MNIST experiment definitively confirms that the multi-state architectural setup behaves exactly as expected.

3.2 MedMNIST Evaluation: PKCL and SSCL

To assess the proposed method's clinical applicability, the 4D disentangled contrastive learning architecture was tested on three established medical benchmarks: PathMNIST [44], OCTMNIST [45], and BloodMNIST [46]. The evaluation encompasses both learning paradigms introduced in this study: the expert-driven PKCL approach and the augmentation-reliant SSCL method.

3.2.1 Partial-Knowledge Performance (PKCL)

In PKCL, each state applies a distinct partial-similarity rule via its projection subspace. As reported in Table 3.2, PathMNIST Stroma recall reaches an impressive **0.9676**, indicating

Table 3.2: Comparison of PKCL and SSCL on PathMNIST, OCTMNIST, and BloodMNIST Test Sets

Dataset	Class	PKCL			SSCL		
		Prec.	Rec.	F1	Prec.	Rec.	F1
PathMNIST	Cancer	0.9707	0.9839	0.9772	0.98	0.93	0.95
	Stroma	0.8758	0.9676	0.9194	0.84	0.91	0.87
	Normal	0.9940	0.9793	0.9866	0.98	0.99	0.99
	Acc. (N)	0.9794 (5496)			0.9700 (5496)		
OCTMNIST	CNV	0.8389	1.0000	0.9124	0.9427	0.8560	0.8973
	DME	0.9762	0.8200	0.8913	0.8681	0.9480	0.9063
	Normal	1.0000	0.9680	0.9837	1.0000	1.0000	1.0000
	Acc. (N)	0.9293 (750)			0.9347 (750)		
BloodMNIST	Platelet	0.9979	1.0000	0.9989	0.9874	1.0000	0.9937
	Neutrophil	0.9910	0.9985	0.9948	0.8869	0.8829	0.8849
	Eosinophil	1.0000	0.9904	0.9952	0.8776	0.8734	0.8755
	Acc. (N)	0.9960 (1757)			0.9108 (1760)		

that the vast majority of Stroma samples are correctly recovered despite the conflicting, partial-knowledge constraints provided during training. Overall, the unified 4D embedding supports a simple frozen-encoder classifier achieving **0.979** accuracy on PathMNIST. Furthermore, the model proves highly adaptable to other modalities, achieving **0.929** on OCTMNIST and **0.996** on BloodMNIST under the same PKCL constraints.

3.2.2 Semi-Supervised Performance (SSCL)

The SSCL evaluation investigates the network's capacity to learn without fine-grained abnormal labels, relying heavily on data augmentation alongside binary (Normal vs. Abnormal) supervision. Configured with a state weighting distribution of $w_A = 0.5$, $w_B = 3.0$,

and $w_C = 1.5$, the model achieved robust convergence. Remarkably, without any explicit labels distinguishing Cancer from Stroma, the linear probe reached a 97.0% overall accuracy on PathMNIST. Isolating the Stroma microenvironment purely through visual augmentations proved to be the most demanding task, yet the model maintained a respectable F1-score of 0.87 for this class. This self-supervised configuration also generalized effectively across other domains, securing accuracies of 93.4% on OCTMNIST and 91.0% on BloodMNIST.

3.3 Visualization of the Latent Space

To qualitatively assess the geometry of the learned representations without causing redundancy, PathMNIST is utilized as the primary representative benchmark for visual analysis in this chapter. Supplementary 3D projections and t-SNE [51] visualizations for the OCTMNIST and BloodMNIST datasets are provided in Appendix A. To maintain visual clarity and standardization throughout this study, data points are consistently color-coded: purple, orange, and cyan denote the three distinct classes.

The 3D scatter plots provided in Figure 3.3 clearly illustrate the geometrical impact of the state-dependent supervision. Under PKCL conditions, State A serves to isolate Normal tissue, State B distinctly clusters Stroma, and State C organizes Cancer cells. Conversely, in the SSCL setting, State A provides a rigid binary boundary, while States B and C generate a smoother, augmentation-induced dispersion between the subgroups.

This localized behavior is mirrored in the global t-SNE mapping (Figure 3.4). The PKCL paradigm yields three tightly consolidated clusters. In contrast, the SSCL approach exhibits a slightly higher degree of boundary overlap, primarily driven by a few cross-class anomalies resulting from the lack of explicit abnormal differentiation during training.

3.4 Statistical Robustness and Randomness in SSCL

Because the SSCL methodology relies heavily on stochastic processes—namely random negative pair sampling and image augmentations—a variance analysis was conducted to

ensure algorithmic stability. The model was trained from scratch ten separate times on the PathMNIST dataset under identical hyperparameter configurations, varying only the computational random seed.

The outcomes, depicted in Figure 3.5, demonstrate significant consistency. The model achieved a mean test accuracy of $96.74\% \pm 1.14\%$, with individual training sessions bounding the accuracy between 94.65% and 98.27%. While the overall classification accuracy proved highly resilient to stochastic variations, the specific recall metric for the Stroma class experienced wider fluctuation ($81.05\% \pm 14.08\%$). This variance underscores the biological complexity of the stromal tissue, which is intrinsically challenging to cluster effectively without deterministic, expert-provided labels.

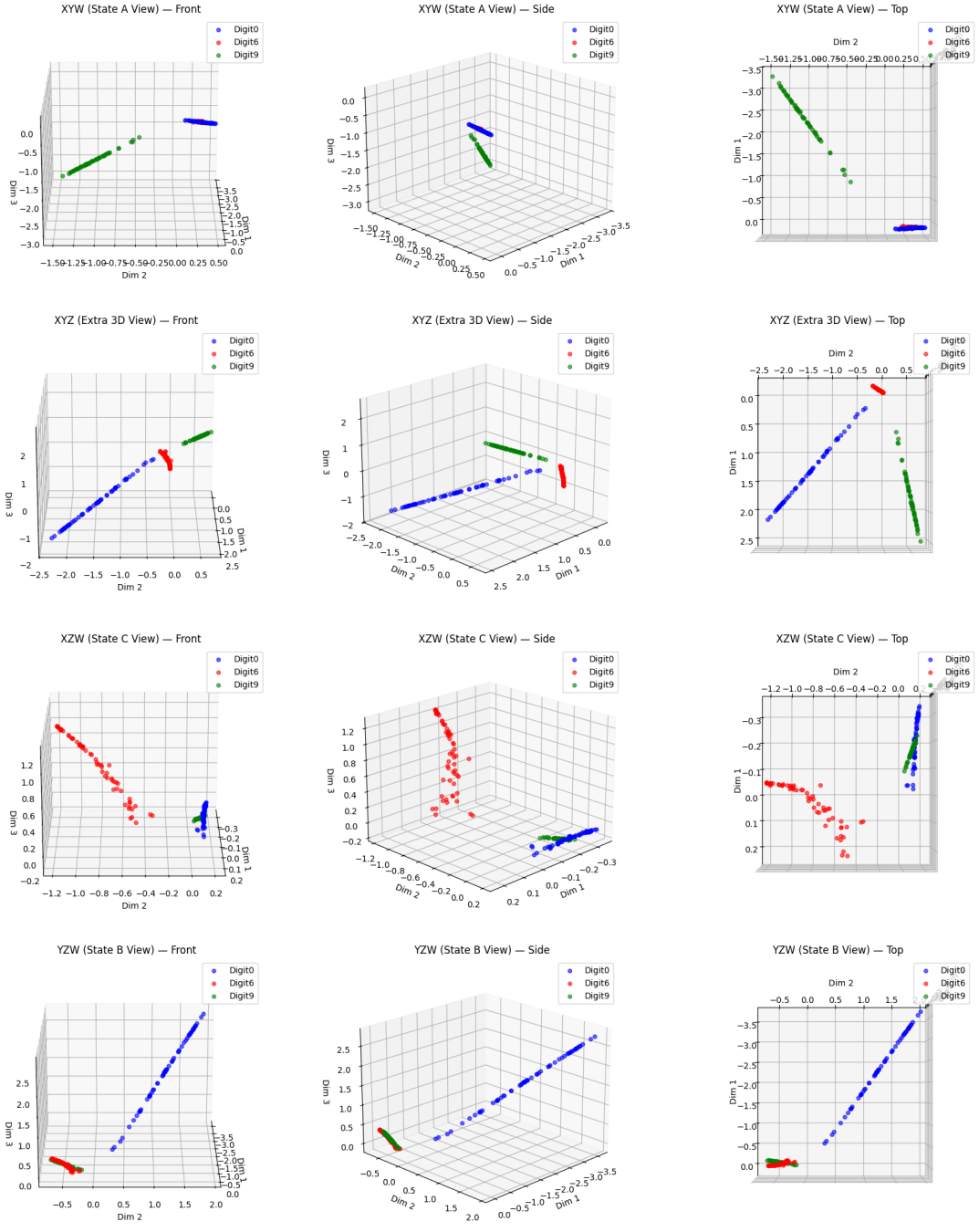

Figure 3.1: 3D projection views of MNIST embeddings using different axis combinations.

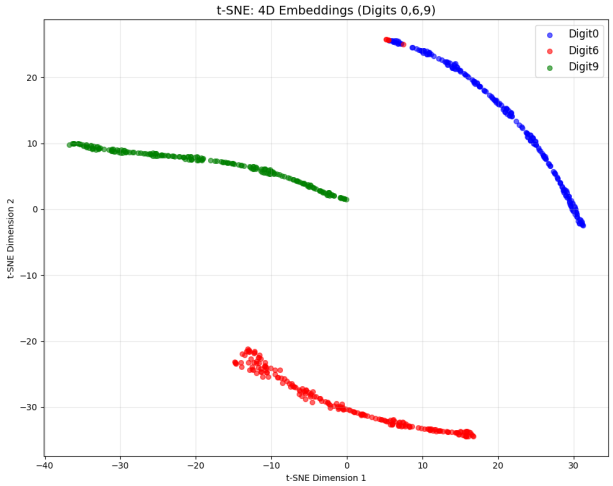

Figure 3.2: t-SNE visualization of the embeddings for digits 0, 6, and 9.

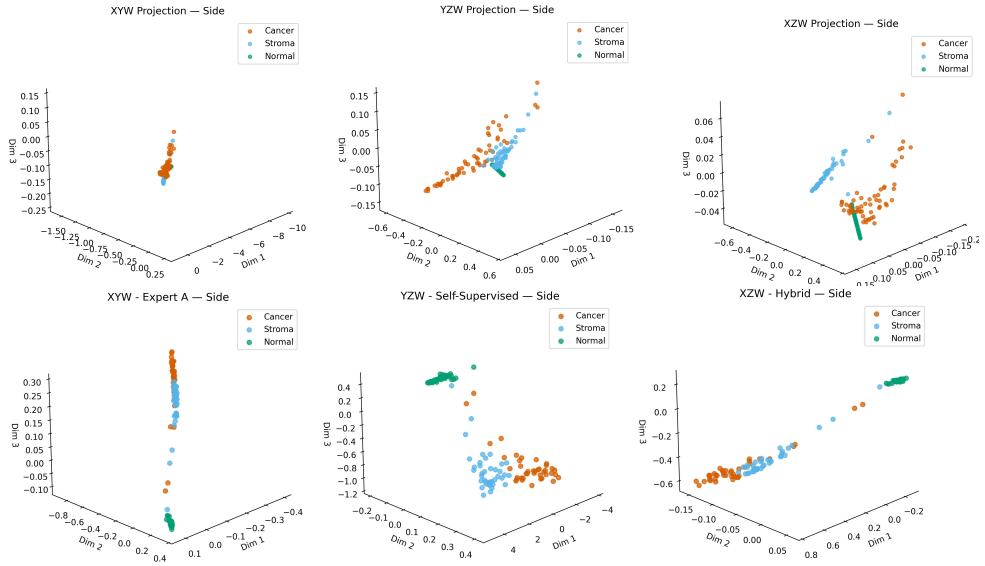

Figure 3.3: Side-view scatter plots illustrating the three projection subspaces for PathMNIST: (A) XYW, (B) YZW, and (C) XZW. The top row displays the PKCL setting, while the bottom row displays the SSCL setting. Points are colored using purple, orange, and cyan.

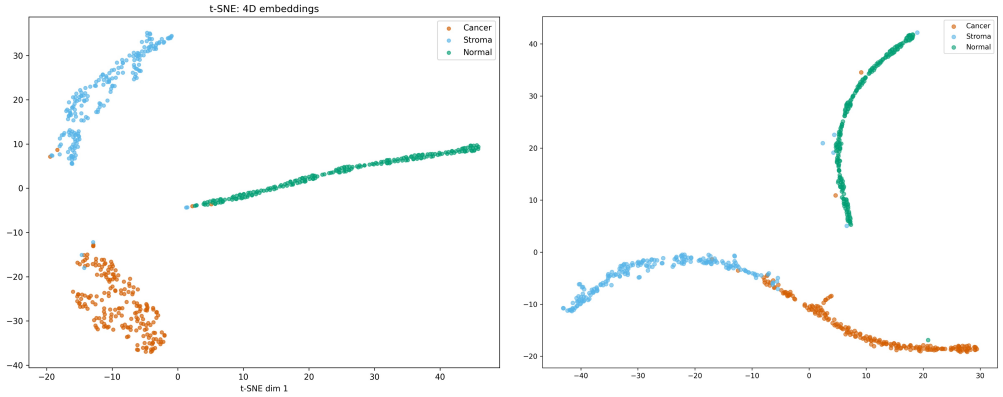

Figure 3.4: Global t-SNE visualization comparing the shared 4D embeddings on PathMNIST. PKCL is shown on the left, and SSCL on the right. The standardized purple, orange, and cyan palette is utilized.

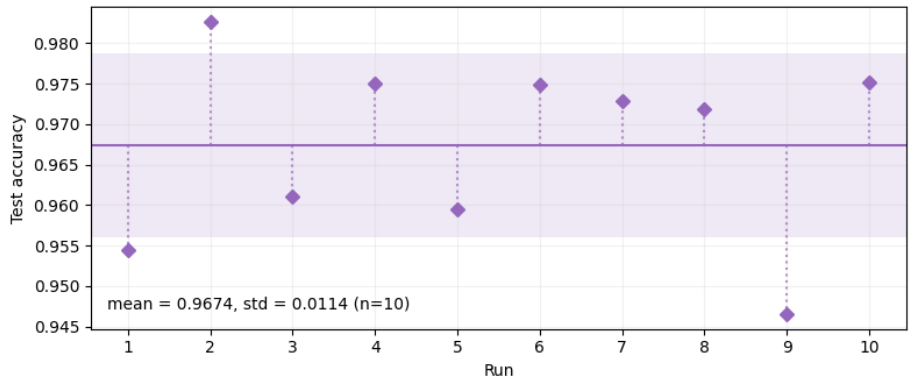

Figure 3.5: Variance in SSCL test accuracy across 10 independent training runs on PathMNIST (w_A=0.5, w_B=3, w_C=1.5). The horizontal line represents the mean, and the shaded region illustrates ±1 standard deviation (0.9674 ± 0.0114).

Chapter 4

Discussion

This project investigates how contrastive representation learning can be used in medical image classification when full supervision is not available. The core achievement of this work is the design of a framework that can learn meaningful and discriminative representations even when only partial expert knowledge is provided. In particular, the strong performance across the evaluated datasets demonstrates that the proposed framework is capable of extracting useful class structure despite incomplete annotation. Achieving high classification performance under partial knowledge suggests that the model is sufficiently robust and expressive to compensate, to a significant extent, for the absence of full expert supervision. The results also help explain the behavior of the two learning scenarios. In PKCL, the model receives guidance from multiple partial experts, but this guidance is not only incomplete; it can also contain conflicting or noisy annotations outside each expert's limited area of strength. Despite this, the framework learns a well-structured representation space and achieves consistently strong performance across datasets, with test accuracies of 97.94% on PathMNIST, 92.93% on OCTMNIST, and 99.60% on BloodMNIST. In SSCL, the setting is even more constrained. Only one expert is available, that expert has access to a narrower and less reliable form of knowledge, and the augmentation-based signal introduces additional randomness, particularly in the construction of negative pairs. Even under these conditions, the framework still maintains competitive performance, reaching 97.00% accuracy on PathMNIST, 93.47% on OCTMNIST, and 91.08%

on BloodMNIST. While PKCL performs better on PathMNIST and BloodMNIST, SSCL slightly outperforms PKCL on OCTMNIST, suggesting that the relative benefit of each supervision strategy may depend on dataset-specific visual structure. This suggests that the proposed framework is robust not only to incomplete supervision, but also to weak and partially unreliable supervisory signals. At the same time, the second scenario reveals an important limitation. The higher false negative tendency and the fluctuations across repeated runs show that the framework remains sensitive to the randomness introduced by negative pair sampling when supervision is highly constrained. Although this variability is still moderate, yielding a mean test accuracy of 96.74% ($\pm$ 1.14%) across repeated runs on PathMNIST, it indicates that representation quality can still be affected when the contrastive signal is built from weaker or partially inaccurate knowledge. This effect is especially visible in the stroma class, whose test recall varies more noticeably across runs (0.8105 ± 0.1408). Moving forward, a natural extension of this work is to apply the framework to more complex classification settings with a larger number of classes. In such cases, the problem becomes inherently more difficult because the model must capture a wider array of relationships simultaneously under partial supervision. This would likely require a higher-capacity architecture and stronger computational resources for training. If such infrastructure is available, it would be highly valuable to investigate whether the framework can still learn crisp, geometrically separable representations in these highly complex clinical settings. These findings confirm that contrastive learning under partial knowledge offers a highly practical direction for medical image representation learning. Crucially, the success of the framework across H&E histopathology (PathMNIST), optical coherence tomography (OCTMNIST), and peripheral blood smear microscopy (BloodMNIST) suggests that the approach remains adaptable across multiple imaging modalities within the evaluated benchmark setting. Because the framework organizes supervision at the level of relational structure rather than through modality-specific label engineering, it remains applicable across different imaging settings. Even when complete expert annotation is prohibitively expensive or unavailable, the framework successfully learns structured embeddings and achieves strong classification performance. This supports the broader conclusion that partial supervision, when integrated through targeted contrastive mechanisms, can provide substantial value in solving real-world medical machine learning bottlenecks.

References

[1] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] Jiancheng Yang, Rui Shi, and Bingbing Ni. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 191–195, 2021.

[3] Jiancheng Yang, Rui Shi, Donglai Wei, Zhiqiang Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[6] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018.

[7] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[8] Phong H. Le-Khac, Graham Healy, and Alan F. Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020.

[9] Suzanna Becker and Geoffrey E. Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992.

[10] Jane Bromley et al. Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1993.

[11] Tao Zhou et al. DC-Seg: Disentangled contrastive learning for multimodal brain tumor segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2023.

[12] Mingyuan Cheng, Xinru Liao, Quan Liu, Bin Ma, Jian Xu, and Bo Zheng. Learning disentangled representations for counterfactual regression via mutual information minimization. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1802–1806, 2022.

[13] Hritam Basak and Zhaozheng Yin. Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19786–19797, 2023.

[14] U.S. Food and Drug Administration. Artificial intelligence/machine learning (ai/ml)-based software as a medical device (samd) action plan. Technical report, 2021. Online.

[15] Francesco Locatello et al. A sober look at the unsupervised learning of disentangled representations and their evaluation, 2020.

[16] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[17] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[18] Kilian Q. Weinberger, John Blitzer, and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2006.

[19] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11:1109–1135, 2010.

[20] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[21] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[22] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[23] Relja Arandjelović and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 609–617, 2017.

[24] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision (ECCV)*, pages 776–794. Springer, 2020.

[25] Honglu Fang, Shu Wang, Ming Zhou, Jian Ding, and Pengtao Xie. CERT: Contrastive self-supervised learning for language understanding, 2020.

[26] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. In *INTERSPEECH*, 2019.

[27] R Devon Hjelm et al. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations (ICLR)*, 2019.

[28] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey Levine. Time-contrastive networks: Self-supervised learning from video. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[29] Mathilde Caron et al. Unsupervised learning of visual features by contrasting cluster assignments. In *European Conference on Computer Vision (ECCV)*, 2020.

[30] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer, 2019.

[31] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

[32] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *ICCV Workshops*, 2019.

[33] Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations, 2021.

[34] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality, 2013.

[35] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations, 2018.

[36] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[37] Lingpeng Kong, Cyprien de Masson d'Autume, Wang Ling, Lei Yu, Zihang Dai, and Dani Yogatama. A mutual information maximization perspective of language representation learning. In *International Conference on Learning Representations (ICLR)*, 2020.

[38] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. In *INTERSPEECH*, 2019.

[39] Petar Veličković, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. In *ICLR Workshop*, 2019.

[40] Jiezhong Qiu, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongyi Yang, Ming Ding, Kuansan Wang, and Jie Tang. GCC: Graph contrastive coding for graph neural network pre-training. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1150–1160, 2020.

[41] Yang Zhang et al. Distortion-disentangled contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[42] Amanpreet Singh, Rui Chen, and Fei Wang. Counterfactual contrastive learning for domain-invariant medical image analysis. *Medical Image Analysis*, 85:102742, 2024.

[43] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[44] Jakob Nikolas Kather et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine*, 16(1):1–22, 2019.

[45] Daniel S. Kermany, Michael Goldbaum, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9, 2018.

[46] Andrea Acevedo, Anna Merino, Sandra Alférez, Ángel Molina, Laura Boldú, and Joan Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in Brief*, 30:105474, 2020.

[47] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1735–1742. IEEE, 2006.

[48] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[49] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[50] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, pages 807–814, 2010.

[51] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[52] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[53] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *CVPR Workshops*, 2014.

[54] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better ImageNet models transfer better? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[55] Marc Macenko, Marc Niethammer, J. S. Marron, David Borland, John T. Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E. Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110. IEEE, 2009.

[56] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2015.

[57] Jean-Bastien Grill et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[58] Irina Higgins et al. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017.

[59] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

[60] Christopher Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations (ICLR)*, 2018.

[61] Ricky T. Q. Chen et al. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[62] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

[63] Liang Chen, Yuhang Wang, and Zhen Liu. Mutual information based entropy disentanglement (med): A scalable metric for high-dimensional contrastive representations, 2024.

APPENDIX A

This appendix provides the supplementary visual results for both PKCL and SSCLSSCL and scenarios applied to the retinal optical coherence tomography (OctMNIST) and peripheral blood smear (BloodMNIST) datasets.

.1 PKCL

BloodMNIST Results

The resulting the t-SNE embedding (Figure 1) reveals a well-structured representation space. Because the framework effectively prevented the collapse of incompatible similarity signals across the different experts, the latent space shows three entirely distinct and widely separated clusters. This clear separation is directly reflected in the downstream classification performance shown in the confusion matrix (Figure 2). The model achieved near-perfect accuracy, correctly classifying 100.0% of Platelets, 99.8% of Neutrophils, and 99.0% of Eosinophils. The state-specific projections (Figure 3) further illustrate how the individual experts shaped their respective 3D subspaces to separate their known classes from the unknown distributions, confirming the effectiveness of the multi-state design.

OCTMNIST Results

In the OCTMNIST PKCL experiments, the framework integrated multiple partial experts attempting to distinguish Normal retinal scans from Choroidal Neovascularization (CNV)

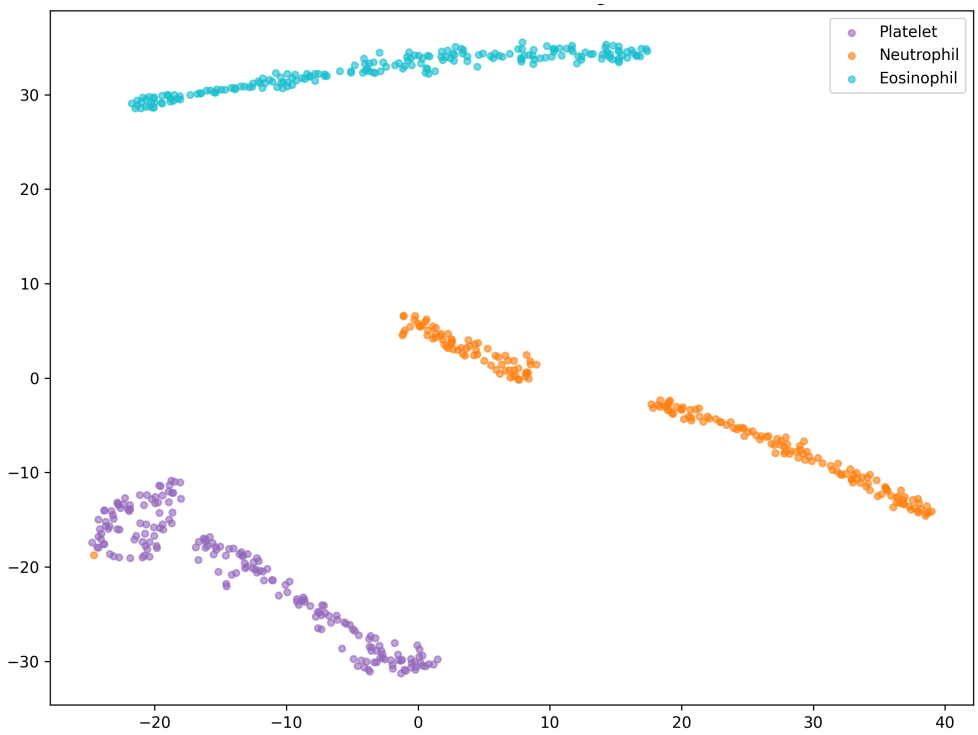

Figure 1: t-SNE visualization of BloodMNIST embeddings learned under the PKCL framework. The integration of multiple partial experts produces a structured embedding space with clear separation among all three classes.

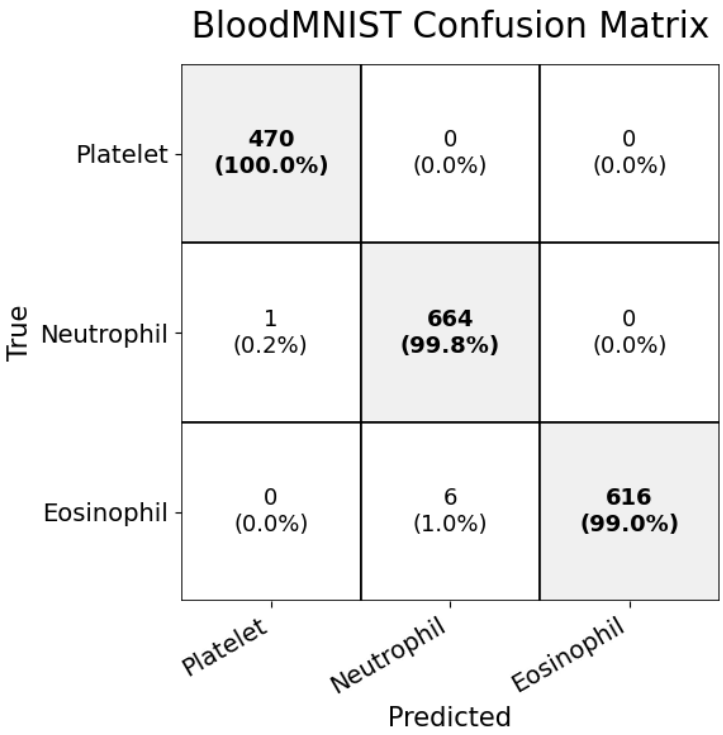

Figure 2: Confusion matrix for PKCL on the BloodMNIST test set, showing near-perfect class separation with only minimal confusion between Neutrophil and Eosinophil.

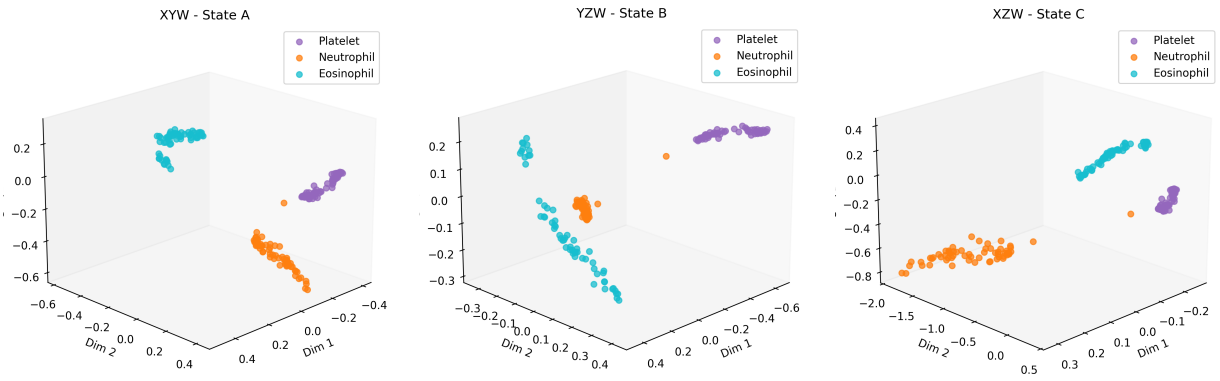

Figure 3: State-specific 3D projection subspaces for BloodMNIST (PKCL), showcasing the independent views maintained by the distinct partial experts.

and Diabetic Macular Edema (DME). This presents a significantly more challenging biological threshold, as CNV and DME share highly similar morphological features compared to healthy tissue. The t-SNE embedding (Figure 4) reflects this clinical reality. While the Normal class is cleanly separated and pushed to the right side of the manifold, the CNV and DME clusters remain spatially adjacent, forming a continuous but structured gradient on the left. The confusion matrix (Figure 5) demonstrates the model achieves perfect recall for CNV (100.0%) and strong recall for the Normal class (96.8%). The main source of error arises between the two abnormal classes, where 18.0% of DME samples are misclassified as CNV. Even with this expected overlap, the learned representation remains strongly discriminative.

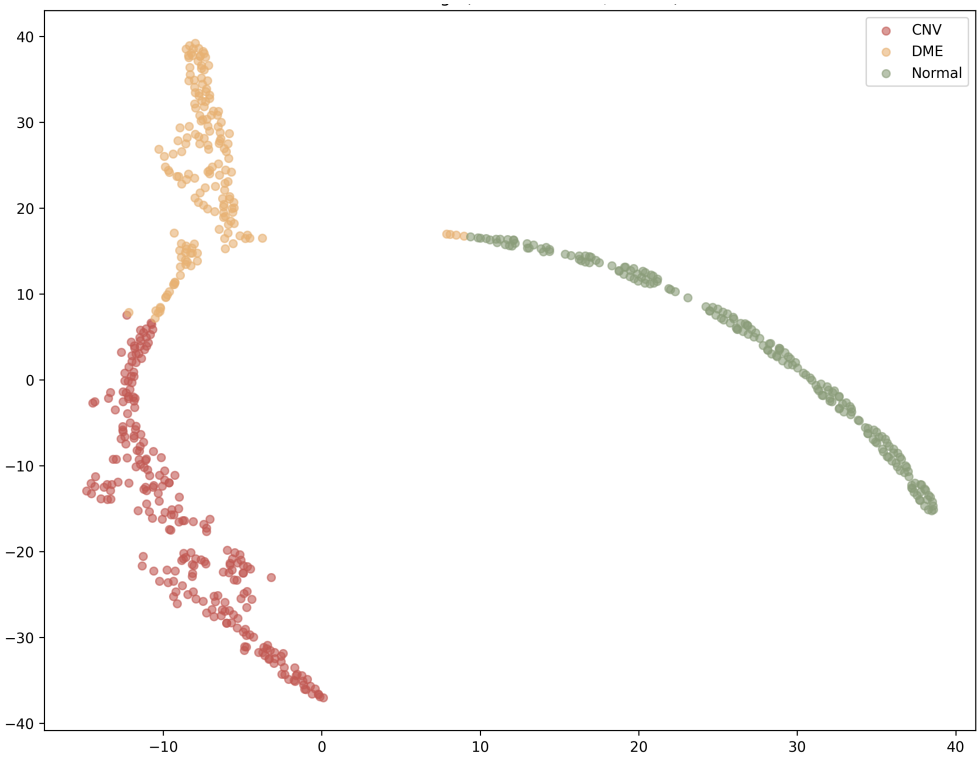

Figure 4: t-SNE embeddings for OCTMNIST under the PKCL framework. Normal tissue is clearly separated, while CNV and DME maintain distinct but adjacent topological regions due to their biological similarity.

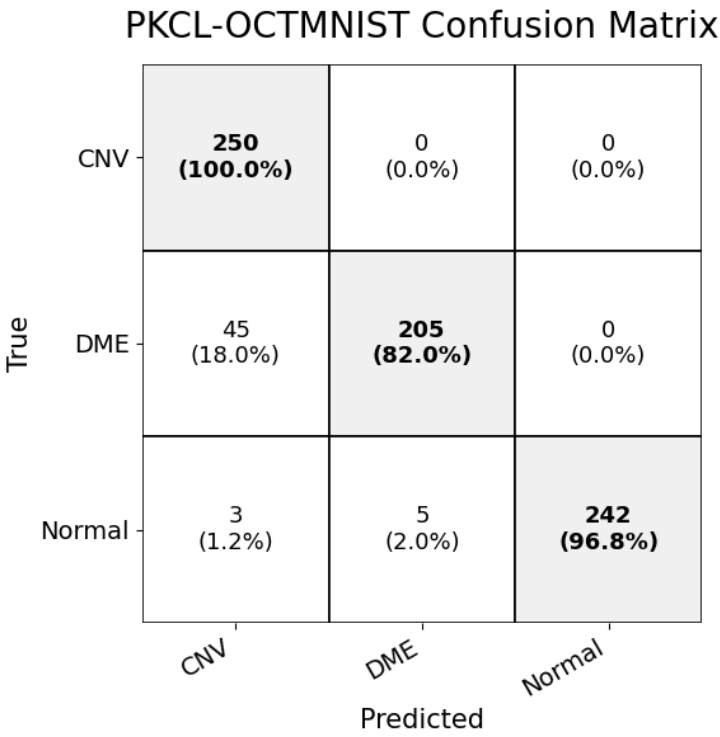

Figure 5: Confusion matrix for PKCL on the OCTMNIST test set, where the main misclassification occurs between DME and CNV, while CNV and Normal remain well separated.

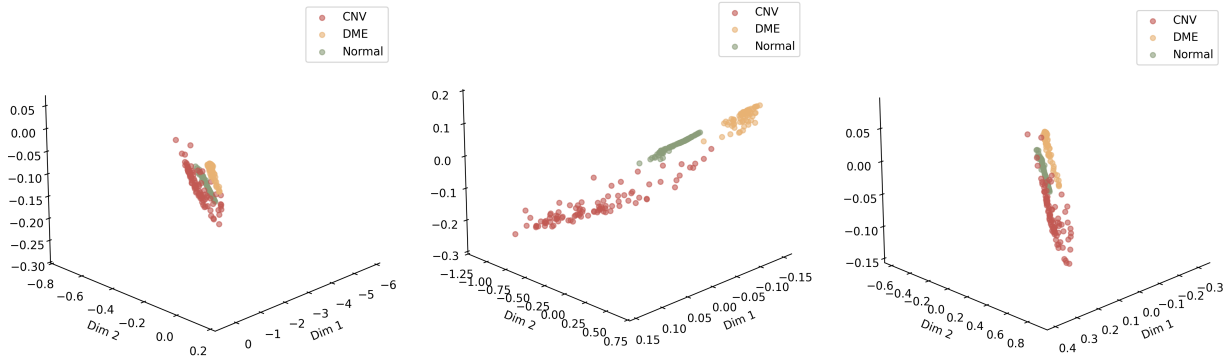

Figure 6: Multi-angle views (Front, Side, Top) of the state-specific 3D projection subspaces for OCTMNIST (PKCL). Left to Right: XYW, YZW, and XZW projections showing the geometric manipulation by each partial expert.

.2 SSCL

.2.1 BloodMNIST Results

In the BloodMNIST SSCL experiment, the framework was tasked with distinguishing Platelets, Neutrophils, and Eosinophils, where the available "expert" only possessed the knowledge to isolate Platelets (State A). The remaining two classes were initially collapsed, forcing the model to rely on self-supervised augmentation states (States B and C) to discover their latent morphological differences. As training progressed, the dynamic weighting mechanism successfully adapted to this knowledge gap. By Epoch 40, the raw loss averages and assigned state weights were as follows: State A ($AvgLoss = 0.0072, w_A = 1.27$), State B ($AvgLoss = 0.0136, w_B = 5.01$), and State C ($AvgLoss = 0.0585, w_C = 1.92$). The framework correctly placed the highest emphasis on State B to force the separation of the biologically similar white blood cells. This successful optimization is visually confirmed in the t-SNE embedding (Figure 7), which shows a massive, clean geometric margin isolating the Platelets, alongside a strong secondary separation between Neutrophils and Eosinophils. The confusion matrix (Figure 8) reflects this geometric crispness, yielding flawless Platelet recognition (100% accuracy) and minimal confusion between the remaining classes.

.2.2 OCTMNIST Results

For the OCTMNIST SSCL scenario, the model was tasked with separating Normal retinal images from two abnormal conditions: Choroidal Neovascularization (CNV) and Diabetic Macular Edema (DME). The simulated expert possessed only Normal-positive knowledge, treating CNV and DME as a single collapsed "abnormal" entity during the first relational state. Similar to the BloodMNIST trial, the adaptive weighting mechanism compensated for the missing fine-grained labels. By Epoch 40, the state distributions were: State A ($AvgLoss = 0.0178, w_A = 1.10$), State B ($AvgLoss = 0.0114, w_B = 6.05$), and State C ($AvgLoss = 0.0934, w_C = 1.06$). The severe up-weighting of State B highlights the framework's reliance on self-supervised contrastive signals to pry apart the CNV and DME

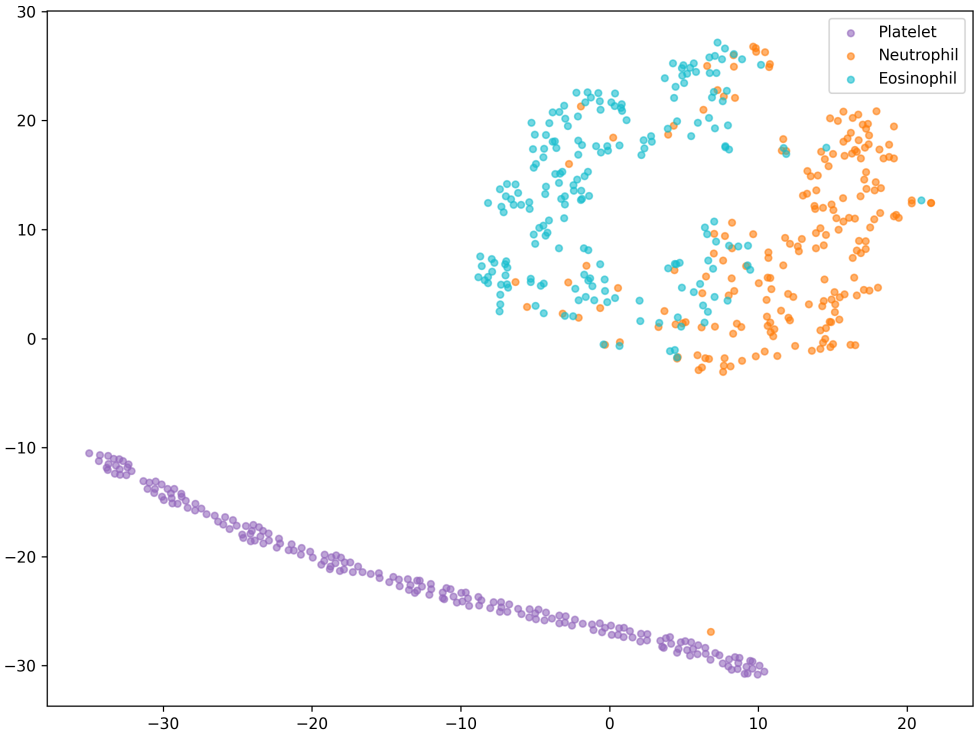

Figure 7: t-SNE embeddings for BloodMNIST under the SSCL framework. Platelets are completely isolated due to the expert signal, while self-supervision successfully separates Neutrophils and Eosinophils.

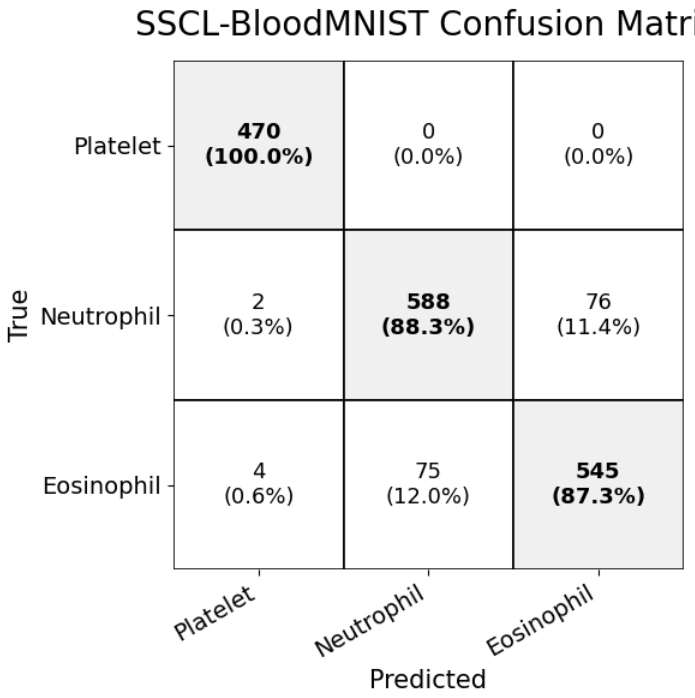

Figure 8: Confusion matrix for SSCL on the BloodMNIST test set, showing stronger confusion between Neutrophil and Eosinophil under the more constrained supervision setting.

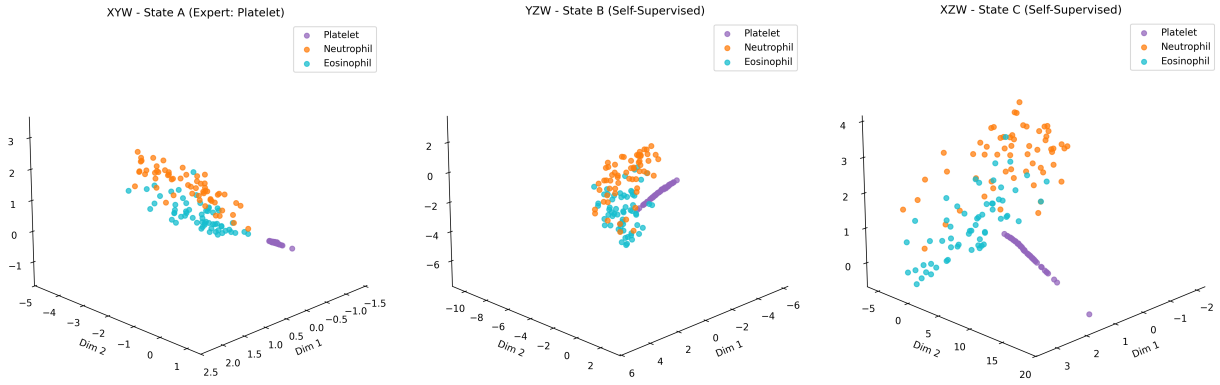

Figure 9: 3D projection subspaces for BloodMNIST. Left: State A, Middle: State B, and Right: State C

representations. The resulting t-SNE projection (Figure 10) and confusion matrix (Figure 11) illustrates the main trade-off observed in this study. The framework achieves perfect classification of the Normal class (100.0%), while a limited degree of confusion remains between CNV and DME. This suggests that under weaker supervision, the separation of visually similar abnormal classes is more challenging. Nevertheless, the overall structure of the learned representation remains strong, showing that the framework can extract clinically meaningful information from coarse and partially annotated data.

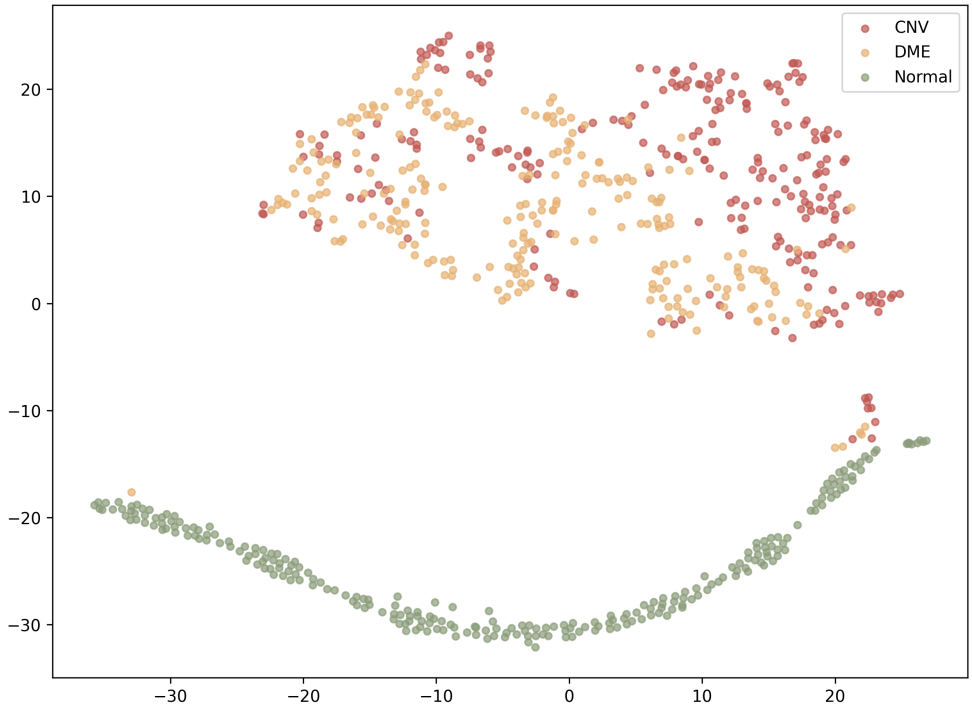

Figure 10: t-SNE embeddings for OCTMNIST under the SSCL framework. The Normal class is entirely geometrically distinct, while CNV and DME form closely adjacent but distinct clusters.

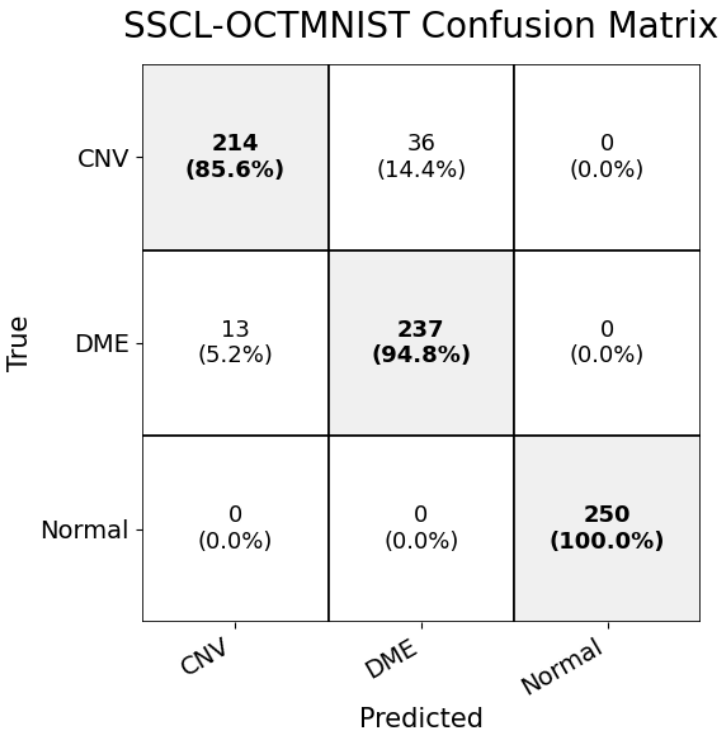

Figure 11: Test-set confusion matrix for OCTMNIST (SSCL). The model accurately isolates Normal tissue without error, with minor predictable overlap between the highly similar CNV and DME classes.

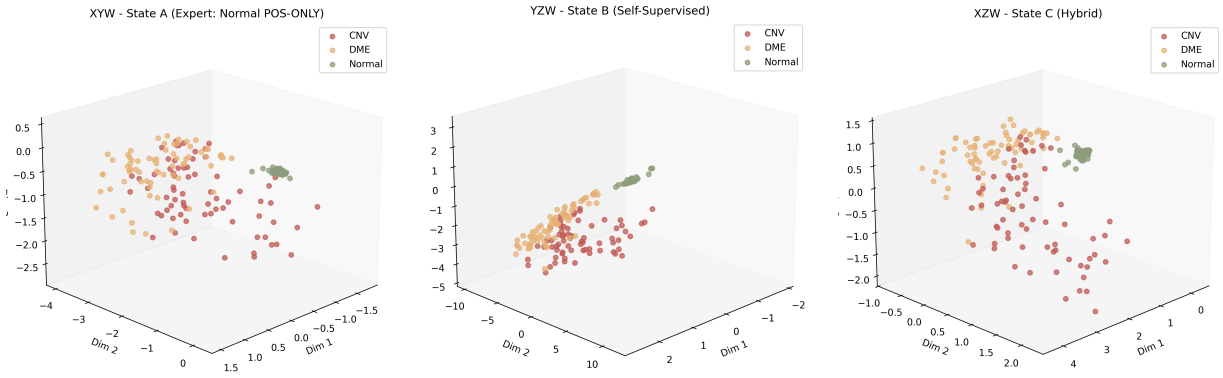

Figure 12: State-specific 3D projection subspaces for OCTMNIST. Left: State A, Middle: State B, and Right: State C.