Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Synthetic Data Augmentation in Medical Image Classification: How Architecture Determines Utility

Bachelor's Thesis

Lars Ruschak

lruschak@ethz.ch

Distributed Computing Group
Computer Engineering and Networks Laboratory
ETH Zürich

Supervisors:
Andreas Plesner, Till Aczel
Prof. Dr. Roger Wattenhofer

September 25, 2025

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Andreas Plesner and Till Aczel, for their continuous support and guidance throughout this thesis. Their expertise, constructive criticism and valuable feedback during our weekly meetings have been essential for shaping this work and for helping me grow both academically and personally. Additionally I want to thank Lorenzo Vettor for helping with the implementation of the diffusion model. Finally, I am grateful to my family and friends for their encouragement and support during this journey.

Abstract

Medical image classification requires large annotated datasets, yet medical data remains scarce due to privacy regulations, annotation costs and class imbalances. This thesis investigates whether diffusion-generated synthetic data improves medical image classification performance. We trained a class-conditional Denoising Diffusion Probabilistic Model on the PathMNIST histopathology dataset and evaluated its impact on four classifier architectures (CNN, ResNet18, DINOv2, CLIP) under three training regimes: real-only, synthetic-only and mixed data approaches. Our evaluation combined classification metrics, t-SNE feature analysis and adversarial robustness testing. Results reveal that synthetic data's utility depends critically on model architecture. Self-supervised DINOv2 achieved best performance with mixed training (92.2% accuracy), while models pretrained on distant domains showed limited benefits. Synthetic-only training consistently underperformed real-only training across all architectures. Feature space analysis confirmed strong alignment between real and synthetic samples, but adversarial robustness experiments revealed that synthetic augmentation provided no robustness benefits. These findings demonstrate that synthetic data serves as a valuable augmentation tool for specific architectures rather than a universal solution.

Contents

Chapter 1

Introduction

1.1 Motivation

A pathologist examining tissue samples under a microscope faces a familiar challenge: distinguishing between normal colon tissue and early-stage adenocarcinoma requires years of training and access to thousands of annotated examples. Yet in many medical institutions, particularly in resource-limited settings, such extensive training datasets simply don't exist. This scarcity becomes even more pronounced for rare pathological conditions, where a hospital might only see only a handful of cases per year.

This data shortage has become a critical bottleneck for deploying deep learning in medical imaging. While convolutional neural networks and transformer architectures have revolutionized computer vision on natural images, their success in medical applications remains constrained by dataset limitations. The problem extends beyond mere quantity; medical datasets face unique challenges including strict privacy regulations, expensive annotation processes requiring specialist expertise and inherent class imbalances where common conditions vastly outnumber rare but clinically relevant diseases.

1.1.1 Synthetic Data as a Solution

Traditional approaches to data scarcity have relied on basic augmentation techniques: rotating images, adjusting brightness, or flipping samples horizontally. While useful, these methods only create variations of existing data rather than expanding the underlying distribution. The emergence of generative modeling offered a more promising path: creating entirely new medical images that could expand training sets and balance underrepresented classes.

Early attempts using Generative Adversarial Networks (GANs) showed potential but revealed significant limitations. GAN training often becomes unstable, particularly problematic when working with medical images where losing specific pathological features due to mode collapse could mean missing critical diagnostic markers. For a radiologist, the difference between a benign lesion and early-stage cancer might lie in subtle texture patterns that unstable training could easily eliminate.

Denoising Diffusion Probabilistic Models (DDPMs) represent a fundamentally different approach. Instead of adversarial training, they learn to reverse a controlled noise corruption process. This framework offers more stable training and can generate diverse, high-quality samples, characteristics particularly valuable for medical imaging where both consistency and variety are essential. The question becomes: can diffusion-generated synthetic medical images improve downstream classification performance, or do they remain mere visual curiosities?

1.2 Objectives and Contribution

This thesis investigates whether synthetic data generated by diffusion models can meaningfully augment real medical image datasets for classification tasks. Rather than simply demonstrating that synthetic images look realistic, we focus on their practical utility: do classifiers trained on synthetic data perform better than those trained on real data alone?

We approach this question by training a class-conditional DDPM on PathMNIST histopathology images and evaluating its output across multiple dimensions. Our experimental design tests four different classifier architectures, from a simple CNN to sophisticated pre-trained models like DINOv2 and CLIP, under three training regimes: real-only, synthetic-only and mixed data approaches.

The contribution extends beyond performance metrics. Through representation-level analysis using t-SNE visualizations, we examine how synthetic and real images align in learned feature spaces. We also investigate adversarial robustness, a critical consideration for medical applications where small perturbations could affect diagnostic decisions.

Our findings reveal that while synthetic data cannot yet replace real medical images, it serves as a powerful augmentation tool, particularly for models prone to overfitting or those with strong pretrained representations. This work provides concrete guidance for researchers considering synthetic data in medical imaging applications, highlighting both opportunities and current limitations in this rapidly evolving field.

Chapter 2

Background and Related Work

2.1 Related Work

2.1.1 Diffusion Models in Medical Image Synthesis

Unlike GANs [Goodfellow et al., 2014] or Variational Auto-Encoders (VAEs) [Kingma and Welling, 2022], diffusion models are based on a forward-reverse denoising process [Ho et al., 2020]. A diffusion process gradually adds noise to an image over multiple time steps until it becomes almost purely noise. Conditioned on the current denoising, a neural network is then trained to learn the reverse process. Once trained, the model allows generating images, as many as needed, From pure noise.

The application of such diffusion models [Ho et al., 2020] has rapidly expanded since their introduction. Kazerouni et al. [2023] provided a comprehensive survey of diffusion models in medical imaging, covering applications across various modalities. Their survey highlighted the superior mode coverage and sample quality of diffusion models compared to GANs and VAEs, while noting computational efficiency challenges that remain for practical deployment.

Building on foundational work, Khader et al. [2023] demonstrated successful application of Denoising Diffusion Probabilistic Models (DDPMs) [Ho et al., 2020] to 3D medical image generation, specifically targeting MRI and CT synthesis. Their work showed that diffusion models could generate high-quality volumetric medical data suitable for clinical evaluation, establishing a precedent for medical diffusion applications beyond 2D image synthesis.

Pinaya et al. [2022] conducted a direct comparison between diffusion models and GANs for brain MRI synthesis, demonstrating that diffusion approaches produced more diverse and higher-quality samples while avoiding the mode collapse issues [Kossale et al., 2022] that frequently plagued GAN-based medical image generation. However, their evaluation focused primarily on visual fidelity metrics rather than downstream task utility.

2.1.2 Synthetic Data Evaluation in Medical Classification

The evaluation of synthetic medical images through downstream classification performance has received limited systematic investigation. Frid-Adar et al. [2018] provided early evidence that GAN-generated synthetic CT liver lesion images could improve classification performance when used for data augmentation, showing 2-3% accuracy improvements on small datasets. However, their evaluation was restricted to simple CNN architectures and didn't explore the interaction between synthetic data quality and different model types.

Hadžić et al. [2023] extended synthetic data evaluation to diffusion models, demonstrating that DDPM-generated medical images could improve anatomical landmark localization performance. Their work provided evidence that diffusion-based synthetic data could enhance

downstream task performance, though they focused on regression rather than classification tasks and evaluated only single architectures.

The broader machine learning community has recognized the importance of task-based evaluation for synthetic data. Lesort et al. [2019] proposed evaluating generative models through discriminative classifier performance rather than purely perceptual metrics, arguing that downstream task utility provides more meaningful assessment than visual quality alone. However, their framework was developed for natural images and hasn't been systematically applied to medical imaging domains.

2.1.3 Architecture-Specific responses to Synthetic Data

The interaction between different neural network architectures and synthetic data remains underexplored, particularly in medical imaging contexts. Most existing work evaluates synthetic data using single architectures or focuses exclusively on convolutional networks. While Radford et al. [2021] demonstrated CLIP's effectiveness for zero-shot classification across diverse natural image domains, its application to medical imaging with synthetic data hasn't been systematically investigated.

Similarly, the emergence of self-supervised learning approaches like DINOv2 [Oquab et al., 2023] has shown promise for cross-domain transfer learning, but specific interaction between self-supervised representations and synthetic medical data remains unexplored. This represents a significant gap given the potential advantages of domain-agnostic feature learning for medical applications where labeled data is scarce.

2.1.4 Pathology Image Analysis and Synthetic Generation

Histopathology presents unique challenges for synthetic data generation due to complex tissue structures and the diagnostic importance of subtle morphological features. The MedMNIST v2 collection [Yang et al., 2023], including PathMNIST, established standardized benchmarks for medical image classification but focused on evaluation frameworks rather than synthetic data augmentation strategies.

Recent advances in histopathology synthesis have primarily relied on GAN-based approaches, with limited exploration of diffusion models for tissue image generation. The complexity of histopathological features presents particular challenges for maintaining diagnostic relevance in synthetic samples.

2.1.5 Positioning of This Work

Our work addresses several critical gaps identified in the existing literature. First, we provide systematic comparison of how different neural network architectures, including CNNs, supervised pretrained models, self-supervised models and multimodal approaches, respond to synthetic histopathology data. This architecture-dependent analysis has not been systematically explored in medical imaging contexts.

Second, we employ a comprehensive evaluation framework combining task performance metrics, representation-level analysis through t-SNE [van der Maaten and Hinton, 2008] visualization and adversarial robustness testing. This multi-faceted approach provides more complete assessment than the primarily visual or single-metric evaluations in existing work.

Third, our three-regime experimental design (real-only, synthetic-only, mixed training) allows systematic isolation of synthetic data effects in ways that previous studies haven't explored. Finally, we focus specifically on practical utility of diffusion-generated pathology images for classification tasks, moving beyond visual quality assessment toward clinically relevant evaluation.

This comprehensive approach provides novel insights into when and why synthetic data benefits medical image classification, advancing understanding beyond current descriptive or single-architecture evaluations in the literature.

2.2 Beyond Visual Fidelity: Evaluating Synthetic Medical Data

If synthetic data quality depends on its ability to train effective classifiers, then understanding how different architectures respond to synthetic augmentation becomes crucial.

2.2.1 The Architecture Landscape

Convolutional Neural Networks remain the foundation of medical image analysis. Their ability to detect spatial features through convolutional layers and hierarchical feature extraction through pooling operations [O'Shea and Nash, 2015] makes them natural choices for medical imaging tasks. However, their performance on synthetic data depends heavily on whether the generated images preserve the spatial patterns that CNNs rely upon.

Residual Networks [He et al., 2015] address the vanishing gradient problem through skip connections, enabling training of very deep networks. ResNets became standard backbones in medical imaging due to their ability to capture rich hierarchical features. Yet their reliance on ImageNet [Deng et al., 2009] pretraining raises questions about domain transfer: how well do features learned on natural images adapt to medical domains, and does synthetic medical data help or hinder this adaption?

Self-Supervised Learning represents a paradigm shift particularly relevant to medical imaging. DINOv2 [Oquab et al., 2023] learns visual representations by enforcing consistency across augmented views rather than relying on labeled data. This approach seems naturally suited to medical imaging, where labeled datasets are scarce and expensive. The question becomes: do self-supervised features transfer more gracefully across the synthetic-real domain gap than supervised counterparts?

Multimodal Approaches like CLIP [Radford et al., 2021] joint-train image and text encoders on large-scale image-text pairs, enabling zero-shot classification through text prompts. While appealing for medical applications where specialized terminology matters, CLIP's pretraining on natural images and everyday text creates a substantial domain gap when applied to histopathology or radiology images.

Each of these architectural approaches brings different strengths and biases to synthetic data evaluation. Understanding how they respond to diffusion-generated images can reveal not just the quality of the synthetic data, but also the transferability of different pretraining strategies across domains.

2.3 Evaluation Beyond Accuracy: A Multi-Faceted Approach

Assessing the impact of synthetic data requires evaluation strategies that capture both representation quality and practical utility. We employ both task-level performance metrics and representation-level analysis to understand how synthetic data affects model behavior.

2.3.1 Performance Metrics in Context

Standard classification metrics take on special significance in medical imaging contexts. Accuracy, while commonly reported, can mislead when dealing with imbalanced medical datasets where rare conditions require high sensitivity. The F1 score provides a more balanced view by harmonizing precision and recall, important when false negatives could mean missing critical diagnoses. The Area Under the ROC curve (AUC) offers threshold-independent assessment of discriminative ability, valuable when diagnostic cutoffs vary across clinical settings.

However, these metrics only capture task-level performance. Understanding why synthetic data helps or hurts requires deeper investigation into learned representations.

2.3.2 Representation-Level Analysis

t-distributed Stochastic Neighbor Embedding (t-SNE) [van der Maaten and Hinton, 2008] provides a window into how models represent synthetic versus real images. By projecting high-dimensional feature embeddings into two dimensions while preserving local neighborhoods, t-SNE reveals whether synthetic images cluster near their real counterparts in feature space.

The intuition is straightforward: if a model maps synthetic tissue samples to the same region as real samples of the same type, this suggests the synthetic data captures the important features of the real dataset. Conversely, if synthetic samples form distinct clusters distant from real data, this indicates that generation hasn't fully captured the underlying data distribution.

2.4 Dataset Considerations: From ChestMNIST to PathMNIST

Our choice of evaluation dataset reflects hard-learned lessons about synthetic data assessment. Initial experiments with ChestMNIST from the MedMNIST v2 collection [Yang et al., 2023] revealed how severely imbalanced medical datasets can confound synthetic data evaluation.

ChestMNIST's extreme and undocumented class imbalance with 42.7% of images showing no pathological findings while conditions like hernia appear in only 0.2% of cases (see Figure 2.1), made it nearly impossible to assess synthetic data utility. Standard training strategies biased heavily toward majority classes, and techniques like class weighting and resampling proved insufficient for achieving stable performance.

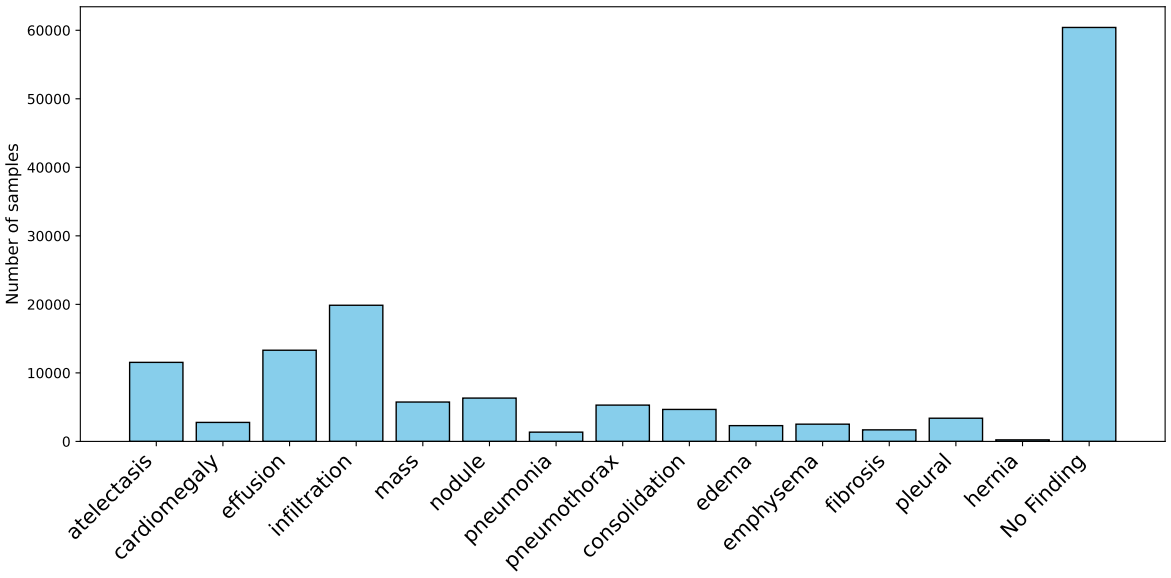

Figure 2.1: Class distribution of the ChestMNIST dataset. A large proportion of samples show no disease, while minority classes are underrepresented. This severe class imbalance hindered classifier training and motivated to switch to PathMNIST.

This experience highlighted a crucial insight: evaluating synthetic data requires datasets where the underlying classification task is well-posed. Extreme imbalances can mask or distort the effects of synthetic augmentation, making it difficult to draw meaningful conclusions about generative model quality.

PathMNIST offered a more suitable alternative. Based on colorectal cancer histopathology [Kather et al., 2019], this dataset provides tissue patches across several categories. While not perfectly balanced, PathMNIST's distribution allows for more reliable assessment of how synthetic data affects classification across different tissue types.

This foundation sets the stage for our experimental investigations: can diffusion-generated PathMNIST images meaningfully improve histopathology classification, and what does this reveal about the current state and future potential of synthetic data in medical imaging?

Chapter 3

Methodology

3.1 Dataset Description

PathMNIST [Yang et al., 2023] contains 107,180 histopathology image patches from hematoxylin and eosin-stained tissue samples, representing nine distinct tissue types: adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma and colorectal adenocarcinoma epithelium. The dataset includes 7,180 images from a different clinical center, which serve as the final test set. Figure 3.1 shows 9 images per class in a row-wise manner, labeled respectively from 0 to 8.

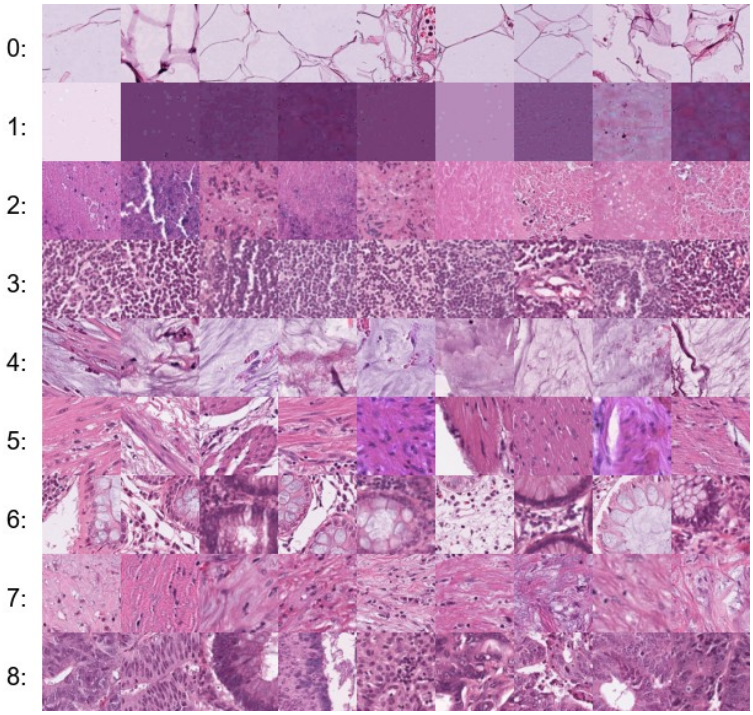

Figure 3.1: Sample images for each of the nine PathMNIST classes.

What makes PathMNIST particularly suitable for synthetic data evaluation is its reasonable class balance. Unlike ChestMNIST where extreme imbalances made it nearly impossible to assess whether synthetic augmentation helped or hurt performance, PathMNIST allows us to evaluate synthetic data effects without confounding factors from severely skewed distributions.

For training the diffusion model, we resized images to 64×64 pixels with 3 RGB channels, a

compromise between computational efficiency and preserving sufficient detail for classification. The dataset split followed a 90,000/10,000 train/validation division, with the separate clinical center test set reserved for final evaluation.

3.2 Generating Synthetic Histopathology Images

The core of our approach involves training a class-conditional diffusion model that can generate histopathology images on demand. Rather than producing random tissue samples, we wanted controlled generation, i.e. the ability to synthesize specific tissue types as needed for augmentation.

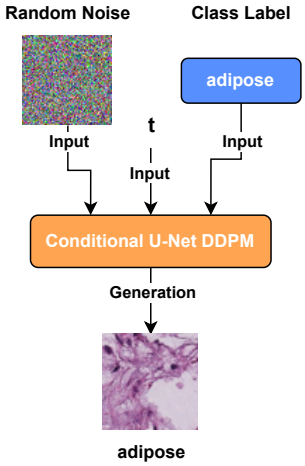

Figure 3.2: Overview of the conditional diffusion model. Random noise is combined with a class label and diffusion timestep t and passed through the conditional U-Net. Produced is an image corresponding to the desired class label.

Network Architecture and Training

We implemented the denoising network as a conditional U-Net [Ronneberger et al., 2015], chosen for its proven effectiveness in diffusion model applications. The network processes 64×64 RGB inputs with a base channel width of 64, providing sufficient capacity without excessive computational overhead. Class conditioning occurs through providing the class label, allowing the model to learn specific generation patterns. Figure 3.2 provides an overview of this setup.

The diffusion process followed a standard practice: 1000 noise steps with a linear variance schedule from $\beta_{start} = 0.0001$ to $\beta_{end} = 0.02$. At each forward step, we add Gaussian noise and during reverse generation, the U-Net predicts this added noise. We maintained an Exponential Moving Average (EMA) [Morales-Brotons et al., 2024] of model parameters throughout training, using the averaged weights at inference for more stable generation.

Training proceeded with PyTorch's [Paszke et al., 2019] AdamW optimization (learning rate 0.0002, weight decay 0.0001) using Mean Squared Error loss between predicted and true noise. PyTorch's OneCycleLR scheduler managed learning rate adaptation.

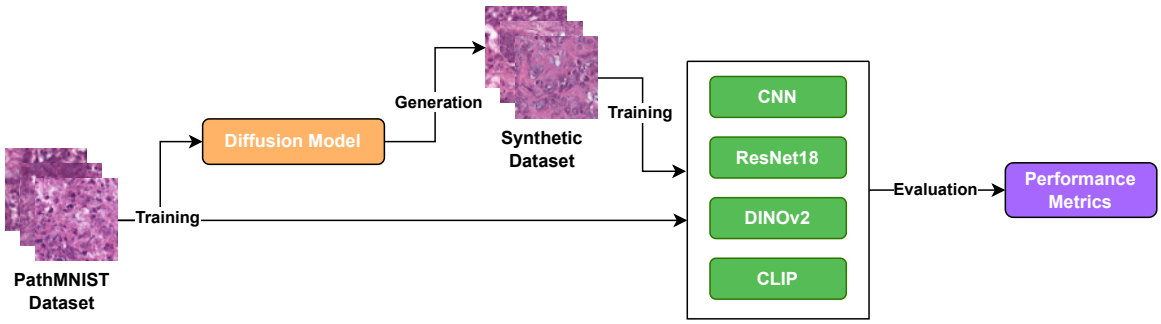

Figure 3.3: Overview of the experimental pipeline. A diffusion model is trained on the PathMNIST dataset to generate synthetic images. These are then used, alone or in combination with real data, to train classifiers whose performance is evaluated on real and synthetic test sets.

3.3 The Three-Regime Experimental Design

Our experimental strategy centers on a simple but revealing comparison: how do classifiers perform when trained on real data only, synthetic data only, or a mixture of both? This three-regime approach lets us isolate the effect of synthetic augmentation while testing whether generated images can substitute for real samples.

The logic behind this design is straightforward. If synthetic data truly captures the underlying patterns needed for classification, then synthetic-only training should match real-only performance. If synthetic data provides useful diversity without replacing real samples, then mixed training should exceed both alternatives. By testing multiple architectures under identical conditions, we determine whether synthetic data benefits depend on model characteristics.

3.3.1 Model Selection Strategy

Rather than arbitrarily choosing classifiers, our model selection reflects a deliberate experimental strategy. Each architecture tests different aspects of how synthetic data interacts with learned representations.

Simple CNN: Our lightweight baseline contains just two convolutional blocks (32 and 64 channels) with batch normalization, ReLU activation, max-pooling and dropout (p=0.25). Features feed into a linear classifier after flattening. This architecture serves as a controlled experiment; with minimal pretraining bias and limited capacity, performance changes likely stem directly from training data characteristics rather than complex feature transfer effects.

ResNet18: Pretrained on ImageNet [Deng et al., 2009] and adapted for nine-class classification, ResNet18 represents the standard supervised transfer learning approach in medical imaging. We resize inputs to 224×224 to match pretraining requirements, freeze most layers and finetune only the final classification layer plus the last ResNet block. This setup tests whether synthetic data helps the adaptation of natural features to histopathology classification.

DINOv2: This self-supervised vision transformer [Dosovitskiy et al., 2021] learned representations by enforcing consistency across augmented image views rather that using labeled

data. We use it as a frozen feature extractor (with inputs resized to 224×224) followed by a trainable linear classifier. unfreezing only the final transformer layer. DINOv2 tests whether self-supervised features handle synthetic-real domain shifts more gracefully than supervised.

CLIP: The multimodal approach [Radford et al., 2021] jointly trained image and texts encoders on web-scale data. We finetune only the vision encoder while keeping the text encoder frozen, using prompts like "This is a pathology image of [class name]". CLIP tests whether powerful multimodal representations can extract meaningful features from synthetic histopathology images despite substantial domain gaps from its natural image pretraining.

3.3.2 Training Protocol

All models used AdamW optimization with cross-entropy loss for the multi-class setting. Hyperparameters varied by architecture based on empirical performance: ResNet18 and DINOv2 used batch size 128 with learning rate 0.0002, while CNN and CLIP used batch size 64 with learning rate 0.001 and 0.0001 respectively. For each individual setup, 10 training epochs were planned. MultiStepLR scheduling managed learning rate decay, with early stopping (patience=3) monitoring validation loss to prevent overfitting. The three training regimes operated as follows:

1. **Real-only:** Standard training on authentic PathMNIST data
2. **Synthetic-only:** Training exclusively on diffusion-generated images.
3. **Mixed:** Batches containing an adjustable ratio of real and synthetic images

This controlled comparison isolates synthetic data effects while maintaining identical optimization and evaluation procedures across conditions.

3.3.3 Evaluation: Beyond Simple Accuracy

Classification performance alone doesn't tell the full story of synthetic data utility. We employed complementary evaluation approaches to understand both task-level performance and representation-level effects.

Quantitative Metrics: Accuracy provides overall performance, but F1 score offers a more balanced view of precision-recall trade-offs particularly important for medical applications. AUC gives threshold-independent discrimination assessment, valuable when diagnostic cutoffs might vary across clinical settings.

Representation Analysis: From each trained model, we extracted penultimate layer features and applied t-SNE dimensionality reduction. By jointly embedding real and synthetic samples, we can assess whether synthetic images cluster near their real counterparts in feature space, a strong indicator of distribution alignment.

Robustness Testing: Medical imaging applications demand robustness to various perturbations. We evaluated adversarial robustness using three attack types: random noise, FGSM [Goodfellow et al., 2015] and BIM [Kurakin et al., 2017], all with perturbation budget $\epsilon = 8/255$. This analysis reveals whether synthetic data improvements translate to enhanced robustness or remain superficial.

3.4 Implementation Details

The experimental infrastructure was built on PyTorch [Paszke et al., 2019] with PyTorch Lightning managing training loops and multi-model coordination. Weights & Biases [Biewald, 2020] logged training progress and metrics, while Hydra [Yadan, 2019] managed configuration files for reproducible experiments across different dataset, model and hyperparameter combinations.

All experiments ran on GPU clusters with NVIDIA GPUs: RTX 2080 Ti, TITAN Xp and TITAN RTX. To handle I/O constraints, we staged data from long-term storage to node-local scratch space during training runs. Batch sizes and model configurations adapted to available GPU memory while maintaining consistent relative comparisons across architectures.

This setup enabled systematic evaluation of synthetic data impact across multiple architectures, training regimes and evaluation criteria. The controlled design isolates synthetic data effects while providing multiple perspectives on when and why diffusion-generated images help or hurt classification performance.

Chapter 4

Experiments and Results

4.1 Synthetic Image Generation Results

After 43 epochs of training, our conditional model converged to a stable solution with training and validation losses around 0.026 and 0.025 respectively. These low loss values suggested the model had learned to effectively reverse the noise corruption process across all the PathMNIST tissue classes. Using the trained model, we generated 180,000 synthetic images across all classes. Figure 4.1 shows a side to side comparison of real and synthetic samples of each class.

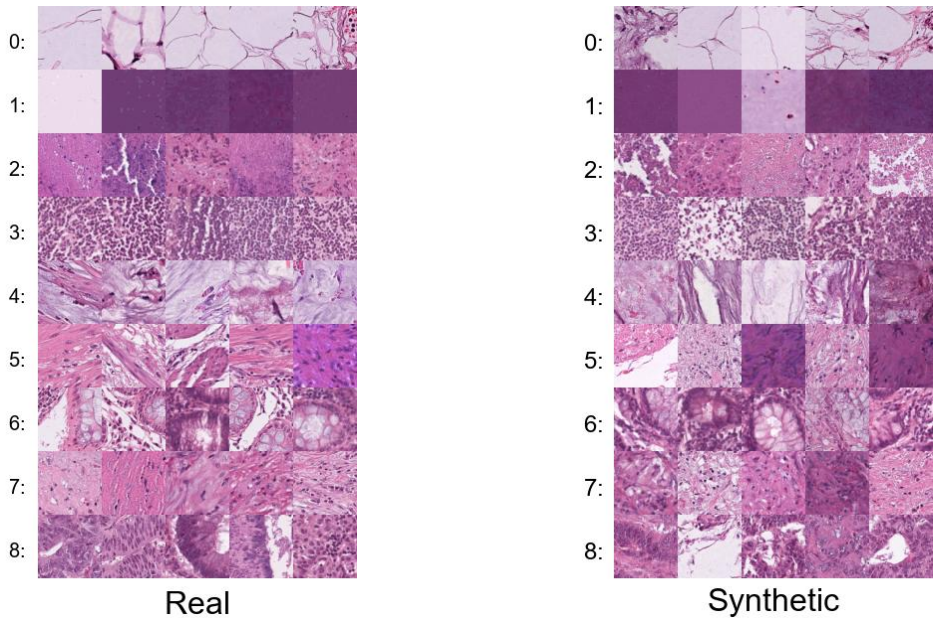

Figure 4.1: Examples of synthetic histopathology images generated by the class-conditional diffusion model. Each column corresponds to one of the nine PathMNIST classes. The samples show realistic staining patterns, though some variability in quality remains.

Visual examination shows that the model successfully learned to generate class-conditional samples. The synthetic images display the characteristic features of each tissue type: adipose tissue (label 0) shows large, clear cellular structures; adenocarcinoma epithelium (label 8) exhibits dense, irregular patterns; lymphocyte regions (label 3) contain small, darkly-stained nuclei; and background areas (label 1) appear appropriately sparse.

4.2 Classifier Performance Across Training Regimes

We trained four classifiers under three data regimes (real-only, synthetic-only, mixed) and evaluated their performance on the PathMNIST test set using accuracy, F1 score and AUC. Table 4.1 presents the complete results. The reported performances correspond to the point of convergence for each model. Validation accuracy had stabilized at this stage, and further training would not have improved results.

Model	Regime	Acc.	F1	AUC
CNN	Real	0.852	0.846	0.988
	Synth.	0.852	0.839	0.987
	Mixed	**0.882**	**0.874**	**0.993**
ResNet18	Real	**0.904**	**0.900**	**0.994**
	Synth.	0.829	0.815	0.990
	Mixed	0.895	0.891	0.992
DINOv2	Real	0.907	0.905	0.993
	Synth.	0.881	0.876	0.988
	Mixed	**0.922**	**0.921**	**0.994**
CLIP	Real	**0.888**	**0.881**	**0.972**
	Synth.	0.781	0.777	0.960
	Mixed	0.883	0.871	0.964

Table 4.1: Classifier performance of CNN, ResNet18, DINOv2 and CLIP trained on real-only, synthetic-only and mixed datasets. Metrics reported are Accuracy, F1 score and AUC on the real PathMNIST test set.

4.2.1 CNN Results

The simple CNN achieved identical accuracy (0.852) under real-only and synthetic-only training, with F1 scores of 0.846 and 0.839 respectively. Mixed training improved performance to 0.882 accuracy, 0.874 F1 and 0.993 AUC.

4.2.2 ResNet18 Results

ResNet18 performed best under real-only training (0.904 accuracy, 0.900 F1, 0.994 AUC). Synthetic-only training resulted in lower performance (0.829 accuracy, 0.815 F1, 0.990 AUC). Mixed training partially recovered performance (0.895 accuracy, 0.891 F1, 0.992 AUC) but remained below the real-only baseline.

4.2.3 DINOv2 Results

DINOv2 achieved strong performance in the real-only condition (0.907 accuracy, 0.905 F1, 0.993 AUC) and showed smallest performance drop under synthetic-only training (0.881 accuracy, 0.876 F1, 0.988 AUC). Mixed training yielded the highest overall performance across all models and conditions (0.922 accuracy, 0.921 F1, 0.994 AUC).

4.2.4 CLIP Results

CLIP demonstrated good real-only performance (0.888 accuracy, 0.881 F1, 0.972 AUC) but suffered the largest performance decline under synthetic-only training (0.781 accuracy, 0.777 F1, 0.960 AUC). Mixed training showed minimal improvement over synthetic-only (0.883 accuracy, 0.871 F1, 0.964 AUC).

4.3 Synthetic Data as Test Data

4.3.1 The Synthetic-as-Test-Data Experiment

To assess distribution alignment between real and synthetic data, we use the classifiers trained exclusively on real PathMNIST data and evaluated them on our generated synthetic dataset. The results are summarized in table 4.2.

Model	Acc.	F1	AUC
CNN	0.852 (+0)	0.852 (+0.006)	0.982 (-0.006)
ResNet18	0.866 (-0.038)	0.867 (-0.033)	**0.987** (-0.007)
DINOv2	0.868 (-0.039)	0.869 (-0.036)	0.984 (-0.009)
CLIP	**0.910** (+0.022)	**0.910** (+0.029)	0.966 (-0.006)

Table 4.2: Performance of models trained on real data and evaluated on synthetic test samples, providing insight into distributional similarity between synthetic and real images.

All models maintained reasonable performance when tested on synthetic data. The CNN showed identical accuracy (0.852) with a slight F1 improvement (+0.006) and minor AUC decrease (-0.006). ResNet18 experienced a 3.8 percentage point accuracy drop (0.866) with corresponding decreases in F1 (-0.033) and AUC (-0.007). DINOv2 showed similar patterns with 3.9 percentage point accuracy decrease (0.868) and F1 (-0.036) and AUC (-0.009) reductions.

CLIP presented an anomalous result, achieving higher performance on synthetic data than on real test data: 91.0% accuracy (+2.2 percentage points), 0.910 F1 (+0.029), with only slight AUC decrease (-0.006).

4.3.2 Feature Space Analysis

We extracted penultimate layer features from each trained model and applied t-SNE dimensionality reduction to visualize how real and synthetic samples align in learned feature spaces.

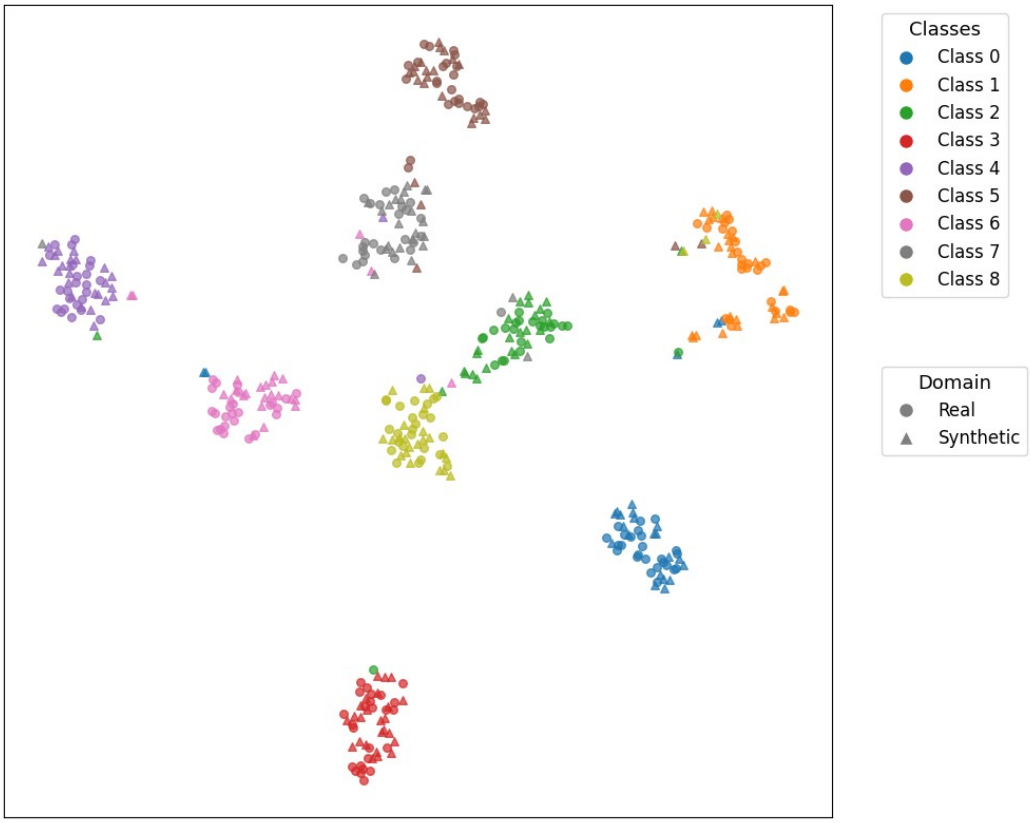

Figure 4.2: t-SNE visualization of DINOv2 embeddings from real-only training. Each point corresponds to an image, colored by label. Clusters are compact and well separated.

As seen in Figure 4.2, the t-SNE embeddings show compact clusters with clear class separation. Real and synthetic samples of the same tissue type cluster in close proximity, with synthetic images distributed throughout their corresponding real image clusters rather than forming separate synthetic-only regions.

Comparisons across architectures (see Appendix A) reveal similar attributes and the clearness of the separation matches the classifiers performance of each model. Interestingly, CLIP showed the clearest visual class separation among all models. However, this did not translate into superior classification performance.

4.4 Adversarial Robustness Evaluation

We evaluated the robustness of all trained models against three types of adversarial attacks: random noise, FGSM (Fast Gradient Sign Method) and BIM (Basic Iterative Method), all with perturbation budget $\epsilon = 8/255$.

Model	Regime	Random Noise	FGSM	BIM
CNN	Real	0.142 (-0.710)	**0.139** (-0.713)	**0.121** (-0.731)
	Synth.	**0.180** (-0.672)	0.123 (-0.729)	0.061 (-0.791)
	Mixed	0.159 (-0.723)	0.113 (-0.769)	0.102 (-0.780)
ResNet18	Real	0.283 (-0.621)	0.277 (-0.627)	**0.181** (-0.723)
	Synth.	**0.320** (-0.509)	**0.299** (-0.530)	0.085 (-0.744)
	Mixed	0.224 (-0.671)	0.214 (-0.681)	0.027 (-0.868)
DINOv2	Real	**0.736** (-0.171)	**0.653** (-0.254)	0.191 (-0.716)
	Synth.	0.566 (-0.315)	0.501 (-0.380)	0.231 (-0.650)
	Mixed	0.668 (-0.254)	0.597 (-0.325)	**0.272** (-0.650)
CLIP	Real	0.247 (-0.641)	0.218 (-0.670)	0.237 (-0.651)
	Synth.	**0.257** (-0.524)	**0.241** (-0.540)	**0.266** (-0.515)
	Mixed	0.240 (-0.643)	0.213 (-0.670)	0.167 (-0.716)

Table 4.3: Accuracy of CNN, ResNet18, DINOv2 and CLIP under adversarial perturbations using Random Noise, FGSM and BIM ($\epsilon = 8/255$). Results are reported for real-only, synthetic-only and mixed training regimes.

All models showed substantial performance degradation under adversarial attacks regardless of training regime. Under random noise, the CNN maintained 14.2-18.0% accuracy across training conditions. ResNet18 achieved 22.4-32.0% accuracy, while DINOv2 proved most resilient with 56.6-73.6% accuracy. CLIP showed 24.0-25.7% accuracy across conditions.

Gradient-based attacks (FGSM and BIM) caused even more severe degradation. Most models dropped below 30% accuracy under FGSM, with BIM attacks reducing performance to single digits for several model-regime combinations. DINOv2 maintained the highest robustness across attack types, but even its best performance (65.3% under FGSM with real-only training) represents substantial degradation from clean accuracy.

Training regimes showed no consistent pattern of improving adversarial robustness across models. In some cases, synthetic-only training showed slightly better robustness than other regimes, but these differences were small compared to the overall performance drops.

4.5 Summary of Key Findings

- Mixed training achieved the best overall performance with DINOv2 reaching 92.2% accuracy
- Synthetic-only training consistently underperformed compared to real-only training across all models
- Architecture differences were substantial: self-supervised DINOv2 showed the smallest synthetic-only performance drops, while CLIP showed the largest
- Feature space alignment showed consistent embedding of real and synthetic images across all models
- Adversarial robustness remained poor regardless of training regime, with no consistent benefits from synthetic data augmentation

Chapter 5

Discussion

5.1 Why Some Models Embraced Synthetic Data While Others Rejected It

The large performance differences across architectures reveal something fundamental about how synthetic data interacts with learned representations. DINOv2's success in mixed training, achieving our best overall performance at 92.2% accuracy, wasn't accidental. Self-supervised pretraining taught this model to extract visual features without relying on specific labeled datasets, creating representations that adapt to both real and synthetic histopathology patterns.

This contrasts sharply with ResNet18 and CLIP, both of which struggled more with synthetic data. ResNet18's ImageNet pretraining created a mismatch: feature optimized for natural images had to adapt to medical patterns, and synthetic data may have introduced artifacts that interfered with this already challenging domain transfer. CLIP faced an even larger gap, with its web-scale natural image training creating representations poorly suited to specialized medical imagery.

The simple CNN's improvement from mixed training tells a different story entirely. Without pretrained baggage, this model could learn histopathology features from scratch. The synthetic data seemed to provide additional diversity that prevented overfitting to the limited real training set. This suggests that synthetic data benefits aren't just about generation quality, they depend heavily on the receiving model's architecture and pretraining history.

5.2 The Feature Space Evidence: Where Real and Synthetic Actually Meet

The t-SNE visualizations provide crucial insight into why certain combinations worked. DINOv2's feature embeddings showed synthetic samples clustering directly alongside their real counterparts, evidence that the generated images activated similar neural pathways to authentic samples. This alignment explains why mixed training could enhance performance: synthetic images expanded the decision boundary without creating distributional conflicts.

A likely explanation for the clear separation of CLIP's feature embeddings lies in the nature of its pretraining objective: its embeddings are optimized for aligning images with natural language text, not for maximizing separability in histopathology tasks. While t-SNE highlights local embedding structure, it does not guarantee that the features are linearly separable.

5.3 Adversarial Reality Check: What Synthetic Data Can't Fix

The complete failure of synthetic data to improve adversarial robustness exposes a critical limitation. Models that showed substantial accuracy improvements from synthetic augmentation collapsed just as dramatically under adversarial attacks as their real-only counterparts. Even DINOv2's impressive 73.6% accuracy under random noise dropped to 19.1% under BIM attacks, a vulnerability that synthetic training did nothing to address.

This robustness gap suggests that our diffusion model, despite generating visually convincing images, didn't capture the complete statistical structure of real medical images. The high-level tissue patterns are there, but the low-level noise characteristics, subtle artifacts, and microscopic variations that might contribute to robustness are missing or altered. For medical applications where reliability under various image corruptions is crucial, this represents a significant limitation.

The finding also challenges a common assumption in the field: that better clean performance automatically translates to better robustness. Our results demonstrate these can be completely decoupled, with synthetic data helping one while ignoring the other entirely.

5.4 Clinical Translation: The Gap Between Lab and Hospital

From a practical standpoint, our findings raise uncomfortable questions about clinical applicability. PathMNIST, while more balanced than many medical datasets, still represents a simplified version of diagnostic reality. Real histopathology involves examining multiple tissue regions, accounting for significant inter-patient variability, and making subtle distinctions that our nine-class framework doesn't capture.

The consistent underperformance of synthetic-only training across all models suggests that current diffusion-generated images miss important aspects of real tissue complexity. In a clinical setting, these missing elements could translate to diagnostic errors or reduced confidence in edge cases, exactly the scenarios where robust performance matters most.

Moreover, the regulatory pathway for synthetic data in medical devices remains unclear. While our mixed training results look promising in academic terms, demonstrating that synthetic augmentation improves patient outcomes would require extensive clinical validation studies that our controlled experiments don't address.

5.5 The Architecture Lesson: Pretraining Strategy as Destiny

One unexpected insight concerns how different pretraining philosophies interact with synthetic medical data. DINOv2's self-supervised approach, learning visual representations through consistency rather than labeled supervision, appears fundamentally better suited to synthetic data than traditional ImageNet pretraining.

This has practical implications for medical imaging researchers. Rather than defaulting to the most powerful general-purpose models, considering pretraining domain alignment might yield better results when synthetic augmentation is planned. The medical imaging community might benefit more from developing domain-specific self-supervised pretraining than from continued adaptation of natural image features.

CLIP's poor synthetic-only performance, despite its impressive general capabilities, reinforces this point. Powerful multimodal representations trained on web-scale data don't automatically transfer to specialized medical domains, particularly when synthetic data introduces additional domain shifts.

5.6 What These Results Actually Mean for Medical AI

Our work contributes to a more nuanced understanding of synthetic data's role in medical imaging. The field appears to be moving beyond simple questions like "can we generate realistic medical images?" toward more sophisticated investigations of when, how, and for which models synthetic data provides benefits.

The success of mixed training regimes suggests synthetic data's primary value lies in augmentation rather than replacement, a finding that aligns with broader machine learning trends emphasizing data diversity over pure volume. For researchers working with small medical datasets, synthetic augmentation might provide meaningful improvements, but only when paired with appropriate architectural choices.

However, the adversarial robustness limitations suggest synthetic data alone won't solve the reliability challenges facing medical AI deployment. Real-world medical applications will likely require combining synthetic augmentation with other robustness techniques like adversarial training, uncertainty quantification, or ensemble methods.

Chapter 6

Conclusion

6.1 The Core Discovery: Architecture Determines Synthetic Data Value

This thesis set out to answer whether diffusion-generated synthetic data can meaningfully improve medical image classification. The answer, it turns out, isn't a simple yes or no, it depends on the receiving architecture and how the model learned its initial representations.

Our experiments revealed that synthetic data's utility hinges on the compatibility between generated images and pretrained features. Self-supervised models like DINOv2 could successfully leverage synthetic histopathology images to achieve better performance than either real or synthetic data alone. In contrast, models pretrained on distant domains struggled to extract useful information from synthetic samples.

This finding shifts the conversation from "how realistic are our synthetic images?" to "how compatible are they with existing model architectures". Visual fidelity, while necessary, proves insufficient for practical utility. The path forward lies in understanding and optimizing this compatibility rather than pursuing ever-more-realistic generation.

6.2 Practical Implications: When to Use Synthetic Data in Medical Imaging

For practitioners considering synthetic data augmentation, our results provide concrete guidance. Synthetic data works best as augmentation rather than replacement, particularly for models with limited capacity (prone to overfitting) or those with domain-agnostic pretraining. The combination of self-supervised backbones with synthetic augmentation emerges as the most promising approach for resource-constrained medical imaging applications.

However, researchers should temper enthusiasm with realism. Current synthetic data doesn't solve adversarial robustness problems and may not capture the full complexity of clinical imaging scenarios. The technique serves as a useful tool in the machine learning toolkit but not a panacea for medical AI's fundamental challenges.

6.3 The Research Agenda: Critical Directions

Three research priorities emerge from our work. First, exploring the use of more advanced or complex generative models, such as latent diffusion models, which may be capable of producing even higher-fidelity images.

Second, developing better evaluation frameworks for synthetic medical data that capture not just task performance but also robustness, calibration and clinical utility. Our multi-faceted

approach provides a template, but medical applications may require additional criteria specific to diagnostic reliability.

Third, investigating synthetic data's role in addressing class imbalance, one of medical imaging's most persistent challenges. Rather than generating uniform samples across classes, targeted synthesis of rare conditions might provide benefits where they're needed most.

6.4 The Final Word: Synthetic Data as Tool, not Solution

Synthetic data in medical imaging has reached a point of practical utility, but within clear boundaries. It serves as a valuable augmentation technique for specific model architectures under particular conditions, not as a universal solution to medical AI's data challenges.

The path forward requires moving beyond the question "can we generate realistic medical images?" toward "can we generate images that help specific models perform better on specific tasks?". This shift from generation realism to task utility represents the field's maturation from proof-of-concept to practical application.

For the medical imaging community, synthetic data offers genuine value when applied thoughtfully: as one tool among many for addressing data scarcity, class imbalanceand model robustness. Its limitations are real and must be acknowledged, but so are its benefits when properly matched to compatible architectures and realistic expectations.

The future of synthetic data in medical imaging may not lie in replacing clinical datasets but in strategically augmenting them in ways that align with both model capabilities and clinical needs. Our work provides a foundation for this more nuanced approach.

Bibliography

Lukas Biewald. Experiment tracking with weights and biases, 2020. URL https://www.wandb.com/. Software available from wandb.com.

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. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2018.09.013. URL https://www.sciencedirect.com/science/article/pii/S0925231218310749.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. URL https://arxiv.org/abs/1412.6572.

Arnela Hadžić, Nicola Rieke, and Shadi Albarqouni. Synthetic augmentation for anatomical landmark localization using denoising diffusion probabilistic models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 485–495. Springer, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A. Valous, Dyke Ferber, Lina Jansen, Constantino Carlos Reyes-Aldasoro, Inka Zörnig, Dirk Jäger, Hermann Brenner, Jenny Chang-Claude, Michael Hoffmeister, and Niels Halama. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine*, 16(1), 01 2019. doi: 10.1371/journal.pmed.1002730. URL https://doi.org/10.1371/journal.pmed.1002730.

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A

comprehensive survey. *Medical Image Analysis*, 88:102846, 2023. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2023.102846. URL https://www.sciencedirect.com/science/article/pii/S1361841523001068.

Firas Khader, Gustav Mueller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarburger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, Bettina Baessler, Sebastian Foersch, Johannes Stegmaier, Christiane Kuhl, Sven Nebelung, Jakob Nikolas Kather, and Daniel Truhn. Medical diffusion: Denoising diffusion probabilistic models for 3d medical image generation, 2023. URL https://arxiv.org/abs/2211.03364.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

Youssef Kossale, Mohammed Airaj, and Aziz Darouichi. Mode collapse in generative adversarial networks: An overview. In *2022 8th International Conference on Optimization and Applications (ICOA)*, pages 1–6, 2022. doi: 10.1109/ICOA55659.2022.9934291.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world, 2017. URL https://arxiv.org/abs/1607.02533.

Timothée Lesort, Andrei Stoian, Jean-François Goudou, and David Filliat. *Training Discriminative Models to Evaluate Generative Ones*, page 604–619. Springer International Publishing, 2019. ISBN 9783030305086. doi: 10.1007/978-3-030-30508-6_48. URL http://dx.doi.org/10.1007/978-3-030-30508-6_48.

Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. Exponential moving average of weights in deep learning: Dynamics and benefits, 2024. URL https://arxiv.org/abs/2411.18704.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. URL https://doi.org/10.48550/arXiv.2304.07193.

Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks, 2015. URL https://arxiv.org/abs/1511.08458.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703.

Walter H. L. Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Brain imaging generation with latent diffusion models, 2022. URL https://arxiv.org/abs/2209.07162.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Omry Yadan. Hydra - a framework for elegantly configuring complex applications. Github, 2019. URL https://github.com/facebookresearch/hydra.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan 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.

Appendix A

Additional Visualizations

A.1 t-SNE Embeddings

This section presents the complete set of t-SNE visualizations for all classifiers. Each plot shows the 2D embedding of the penultimate layer, with colors indicating class labels. These figures complement the representative DINOv2 example shown in section 4.3.2.

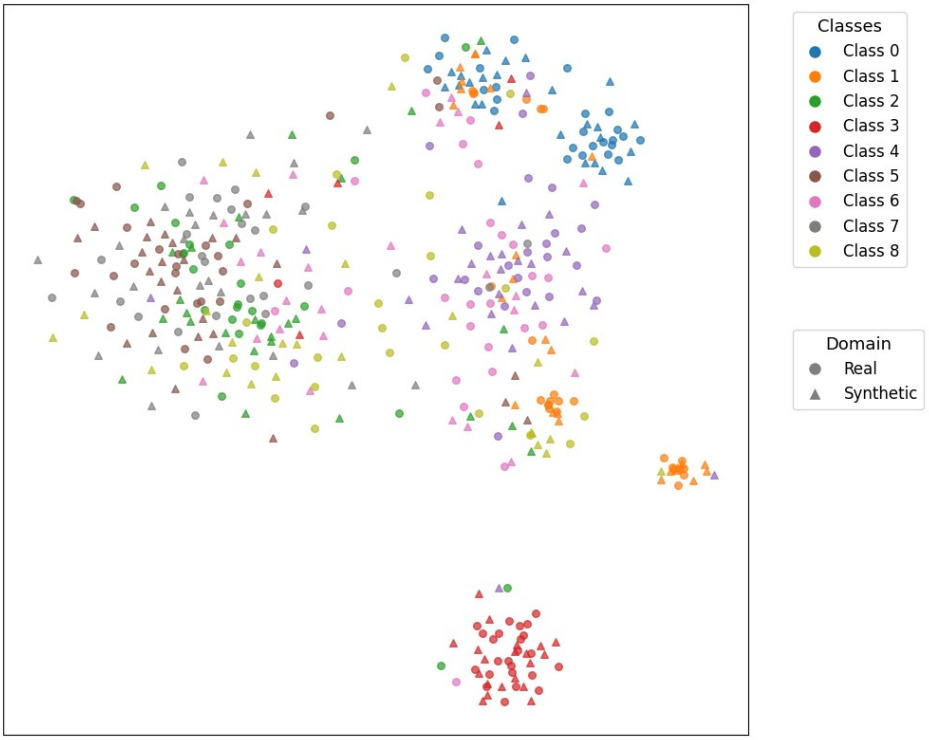

Figure A.1: t-SNE visualization of CNN embeddings, showing loose and partially overlapping clusters.

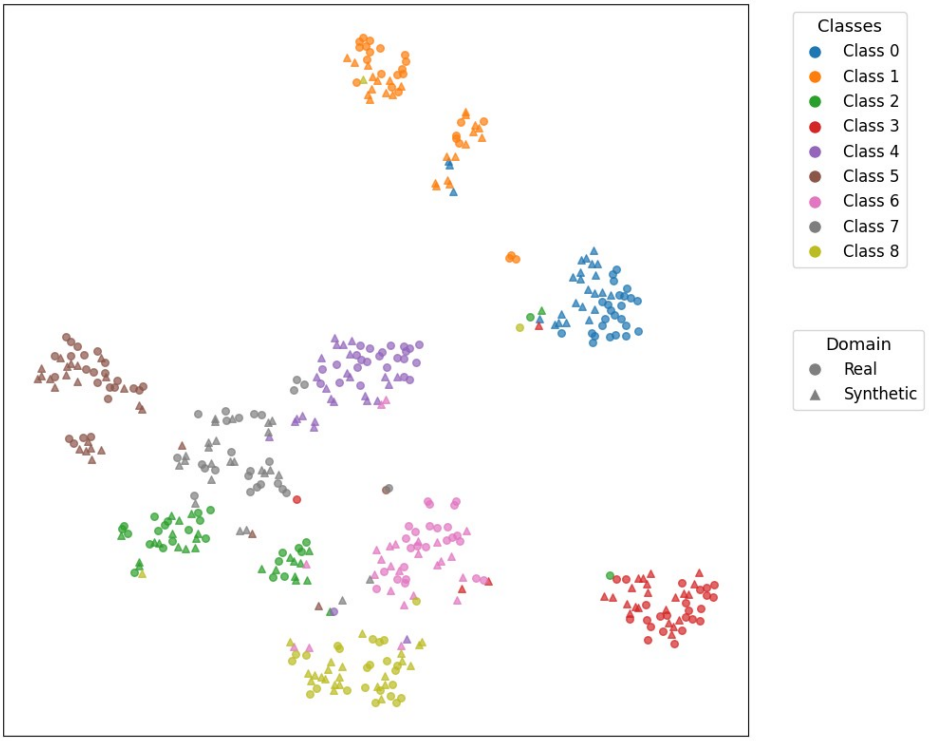

Figure A.2: t-SNE visualization of ResNet18 embeddings, showing clearer separation but still some overlaps.

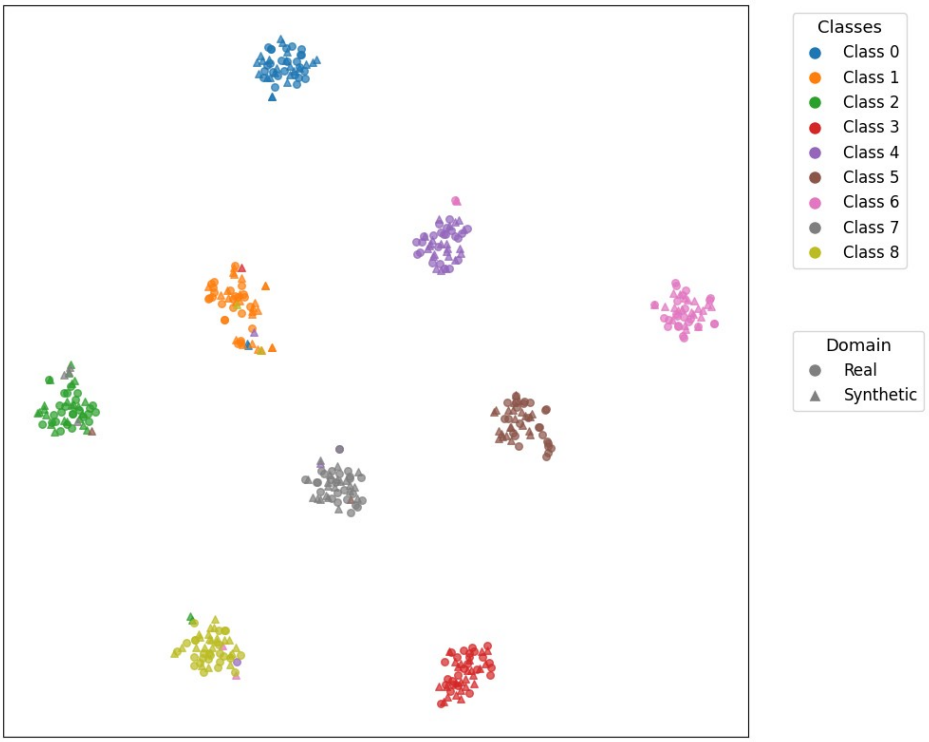

Figure A.3: t-SNE visualization of CLIP embeddings, showing clearly separated clusters, which did not translate into superior performance.