Evaluating Hyperspherical Embeddings and Targeted Augmentations for Generalizable Medical Image Analysis

Master Thesis

Master of Science in International Software Systems Science

Muhammad Tayyab Sheikh

July 2, 2025

Supervisor:

1st: Prof. Dr. Christian Ledig
2nd: Msc. Francesco Di Salvo

Chair of Explainable Machine Learning
Faculty of Information Systems and Applied Computer Sciences
Otto-Friedrich-University Bamberg

Abstract

The clinical translation of deep learning models in medical imaging is critically limited by their failure to generalize to out-of-distribution (OOD) data. Models trained in one clinical setting often underperform when encountering variations in patient populations, imaging equipment, and preparation protocols. This thesis addresses this challenge by systematically benchmarking strategies to improve OOD generalization across two distinct medical domains: inter-hospital variability in histopathology, using the **WILDS Camelyon17** dataset, and skin tone bias in dermatology, using the **Fitzpatrick17k** dataset.

Our evaluation benchmarks two fundamental approaches against the standard training method, **Empirical Risk Minimization (ERM)**. The first approach to be tested is a specialized algorithmic approach, **HYPO**, a representation learning strategy that seeks to learn domain-invariant features by actively organizing them on a hypersphere. Its goal is to make features from the same class cluster tightly together regardless of their source domain, while pushing clusters of different classes far apart. The second approach is data-centric, for which we evaluate multiple augmentation strategies. This includes leveraging **AugMix**, a powerful processing framework that creates diverse training examples by mixing multiple augmentation chains, and **MedMNIST-C**, a benchmark providing targeted corruptions that simulate realistic medical image artifacts. This culminates in our proposed novel method, **MedAugmix**, which combines these ideas by using the targeted corruptions from MedMNIST-C as the core ingredients within the robust AugMix framework.

Our findings reveal a clear hierarchy of effectiveness. While the HYPO algorithm provides a solid improvement over the baseline, confirming the value of representation learning, the most significant performance gains were achieved through our data-centric approach. The proposed **MedAugmix** strategy delivered the most substantial gains, drastically reducing the OOD performance gap. For example, on the challenging Camelyon17 benchmark, ERM with MedAugmix achieved an OOD accuracy of 0.913 compared to 0.853 for standard ERM—a substantial 6.0 percentage point improvement that exceeded all evaluated HYPO configurations. Crucially, a standard ERM model, when enhanced with MedAugmix, consistently outperformed the specialized HYPO algorithm across both the histopathology and dermatology benchmarks, demonstrating the generalizability of this core finding.

This thesis concludes that while specialized OOD algorithms are beneficial, a sophisticated, data-centric approach focused on targeted augmentation can be even more impactful for building robust medical AI. The success of the proposed MedAugmix strategy highlights that focusing on high-quality data augmentation that simulates realistic domain shifts is a crucial and highly effective path toward developing models that are not only accurate but also equitable and reliable enough for real-world clinical deployment.

[1]

[1] `https://github.com/TayyabSheikh/MedAugmix-OOD-Evaluation`

Acknowledgment

I would like to express my sincere gratitude to Prof. Dr. Christian Ledig for the opportunity to conduct my thesis under his chair. I am especially thankful to Francesco Di Salvo for his continuous support, encouragement, and exceptional supervision—his insights and feedback have been invaluable at every stage. I also extend my heartfelt thanks to my parents. To my father, who always believed in me and supported every step; to my mother, who placed my education above all else; and to my sister, whose unwavering support and presence carried me through the toughest moments. Special credits to Tahir Mamu, for his support during the hard times.

A special mention goes to my childhood friend Abdur Rehman, whose spare screen made the writing process infinitely more manageable. I'm also grateful to my friend Ahmer for keeping me entertained during the most exhausting phases of this work, and to Luqman thanks for buying that PlayStation at just the right time. Finally, a big shoutout to my university senior Ali whose quotes full of wisdom helped me trek my way forward.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation and Problem Statement

Deep learning models have achieved expert-level performance across a wide array of medical imaging tasks, holding immense promise for improving diagnostic accuracy and accelerating clinical workflows (Esteva et al., 2021; Litjens et al., 2017). However, their widespread clinical adoption is critically hindered by a lack of generalization. Models often exhibit a significant drop in performance when encountering data from new sources, a pervasive issue known as domain shift (Finlayson et al., 2021). This lack of robustness to Out-of-Distribution (OOD) data undermines clinical trust and poses significant risks to patient safety, making OOD generalization one of the most important research frontiers in medical AI (Koh et al., 2021).

This challenge manifests in diverse ways across medical specialties, highlighting the need for broadly robust solutions. To investigate this general problem, this thesis evaluates strategies against two distinct and clinically relevant OOD test cases:

- In **computational histopathology**, models must contend with significant inter-hospital variability arising from inconsistent slide staining, tissue preparation, and digital scanner hardware (Madabhushi and Lee, 2016). The **WILDS Camelyon17** benchmark encapsulates this challenge, requiring models to generalize to unseen hospitals.
- In **dermatology**, a critical OOD challenge arises from patient demographics, particularly the underrepresentation of darker skin tones in clinical datasets. The **Fitzpatrick17k** dataset provides a platform to evaluate model robustness to this skin-tone-based distribution shift (Groh et al., 2021).

Given this multifaceted problem, this study seeks to derive generalizable insights by benchmarking different approaches against these disparate challenges. This work undertakes a comprehensive evaluation of a specialized representation learning algorithm, **HYPO**, against a standard baseline. In parallel, it critically assesses the impact of data augmentation techniques, including the application of targeted corruptions sourced from the **MedMNIST-C** benchmark, the use of the **AugMix** data processing framework, and a novel hybrid strategy proposed called, **MedAugmix**, which combines these data-centric concepts.

1.2 Research Questions and Objectives

To address the general problem of Out-of-Distribution (OOD) generalization in medical imaging outlined in Section 1.1, this thesis seeks to answer a set of focused research questions through a systematic evaluation framework.

Research Questions

1. How does the performance of a specialized representation learning algorithm (HYPO) compare to a standard baseline (Empirical Risk Minimization) in improving OOD generalization across diverse medical imaging tasks, specifically histopathology and dermatology?

2. What is the comparative impact of different data-centric augmentation strategies, ranging from a basic application of targeted corruptions (Basic MedMNIST-C) to a generic advanced method (Standard AugMix) and a novel hybrid approach (MedAugmix) on model robustness?

3. Does the effectiveness of these learning algorithms and augmentation strategies generalize across different types of domain shifts, namely the inter-hospital variations in **WILDS Camelyon17** and the skin tone biases in **Fitzpatrick17k**?

4. Ultimately, which combination of learning algorithm and data augmentation strategy yields the most effective and broadly applicable solution for enhancing OOD generalization across the evaluated medical imaging benchmarks?

To answer these questions, the following concrete objectives were established for this thesis:

Objectives

1. To systematically implement and evaluate the OOD generalization performance of the ERM baseline and the HYPO algorithm on two distinct medical imaging benchmarks: WILDS Camelyon17 and Fitzpatrick17k.

2. To implement and assess the impact of three different categories of data augmentation on both learning algorithms and both datasets: (a) a basic application of targeted corruptions, (b) the standard AugMix method, and (c) the proposed novel **MedAugmix** strategy.

3. To conduct a comprehensive comparative analysis of all evaluated combinations to identify the most effective and robust strategies for improving OOD performance.

4. To analyze and interpret the findings to derive generalizable insights into the interplay between algorithmic approaches (representation learning) and data-centric approaches (structured augmentation) for building more reliable medical AI systems.

1.3 Key Contributions

This thesis makes several key contributions to the field of medical image analysis, specifically concerning the development and evaluation of robust models for Out-of-Distribution (OOD) generalization. The contributions are:

1. **Proposal and Evaluation of a Novel Targeted Augmentation Strategy (MedAugmix):** We introduce, implement, and evaluate **MedAugmix**, a novel data augmentation strategy. MedAugmix adapts the principles of the AugMix data processing pipeline to leverage **targeted** MedMNIST-C corruptions, creating a powerful, domain-aware augmentation technique specifically for medical imaging.

2. **Comprehensive Benchmarking Across Diverse Medical Domains:** This work provides a systematic and comparative evaluation of multiple OOD generalization strategies (ERM, HYPO, and various augmentations) across two distinct and challenging medical domains: histopathology (**WILDS Camelyon17**) and dermatology (**Fitzpatrick17k**). By testing these methods against different types of domain shifts inter-hospital variation and demographic (skin tone) bias this evaluation offers more generalizable insights than a single-dataset study.

3. **Empirical Demonstration of Data-Centric Superiority:** A significant discovery from our benchmarking is the empirical evidence that a standard **Empirical Risk Minimization (ERM)** algorithm, when paired with our proposed **MedAugmix** strategy, consistently outperforms other strategies tested within this study. This combination surpassed the specialized representation learning algorithm, HYPO, across both datasets and architectures.

1.4 Thesis Outline

The remainder of this thesis is organized to systematically build upon the motivation and objectives outlined above. The structure is as follows:

- **Section 2: Background and Literature Review** provides the necessary foundational knowledge. It reviews the state of deep learning in medical imaging, details the Out-of-Distribution (OOD) generalization challenge across different medical domains, and discusses existing approaches, including the HYPO algorithm, the AugMix framework, and the MedMNIST-C benchmark.

- **Section 3: Experimental Design and Methodology** details the comprehensive framework developed for this study. It introduces the two benchmark datasets, WILDS Camelyon17 and Fitzpatrick17k, describes the model architectures and evaluation protocol, and provides a detailed account of the specific strategies under evaluation: ERM, HYPO, and the various data augmentation techniques including the proposed MedAugmix.

- **Section 4: Experimental Results** factually presents the empirical findings of the study. It first details the results on the primary benchmark, Camelyon17, to establish key performance trends. It then presents a summary of findings on the secondary benchmark, Fitzpatrick17k, to assess the generalizability of these trends.
- **Section 5: Discussion** provides an in-depth interpretation and synthesis of the results. This section is structured thematically to analyze the value of representation learning (HYPO), the impact of data-centric approaches (augmentations), and the significance of the key discovery that a well-augmented ERM can surpass specialized OOD algorithms.
- **Section 6: Conclusion and Future Work** concludes the thesis by recapitulating the study, summarizing its principal findings and contributions, and discussing its strengths and limitations. Finally, it proposes several concrete directions for future research based on the insights gained from this work.

2 Background and Literature Review

2.1 Deep Learning in Medical Image Analysis: Overview

The past decade has witnessed a paradigm shift in the field of medical image analysis, largely driven by the remarkable advancements and successes of deep learning (DL) methodologies (Esteva et al., 2021; Litjens et al., 2017). Deep learning, a subfield of machine learning, employs artificial neural networks with multiple layers (hence "deep") to learn hierarchical representations of data. This capability has proven exceptionally potent for interpreting complex medical imagery, offering tools that can augment clinical decision-making, improve diagnostic accuracy, and streamline analytical workflows across various medical specialties.

At the heart of this revolution are Convolutional Neural Networks (CNNs or ConvNets), a class of deep neural networks particularly well-suited for processing grid-like data, such as images (LeCun et al., 1998; Krizhevsky et al., 2012). CNNs automatically and adaptively learn spatial hierarchies of features from images, from low-level edges and textures in the initial layers to more complex, task-specific patterns in deeper layers. This is achieved through the use of key architectural components:

- **Convolutional Layers:** These apply learnable filters to input images (or feature maps from previous layers) to create feature maps that highlight specific patterns. The filters are shared across the spatial dimensions of the input, enabling parameter efficiency and translation equivariance.
- **Pooling Layers:** These layers reduce the spatial dimensionality of the feature maps, which helps to decrease computational complexity, control overfitting, and create a degree of invariance to small translations or distortions in the input. Max pooling and average pooling are common strategies.

- **Activation Functions:** Non-linear activation functions, such as the Rectified Linear Unit (ReLU) (Nair and Hinton, 2010) and its variants, are applied element-wise to introduce non-linearity into the model, enabling it to learn more complex relationships.
- **Fully Connected Layers:** Typically found at the end of the network, these layers perform high-level reasoning by connecting every neuron from the previous layer to every neuron in the current layer, often leading to the final classification scores or regression outputs.

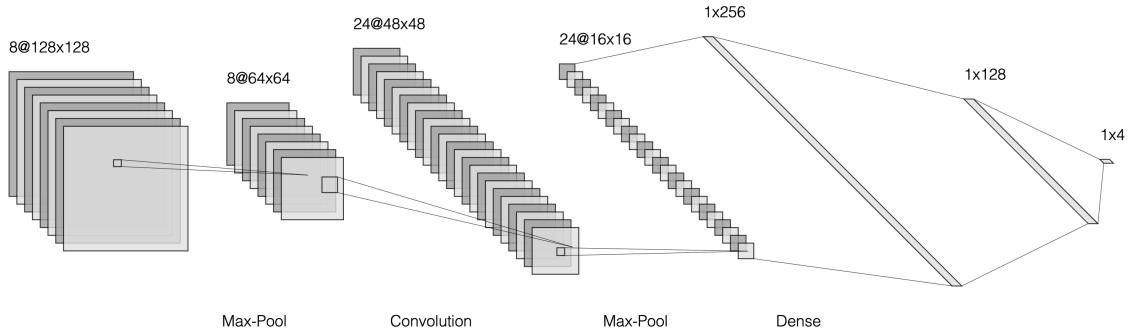
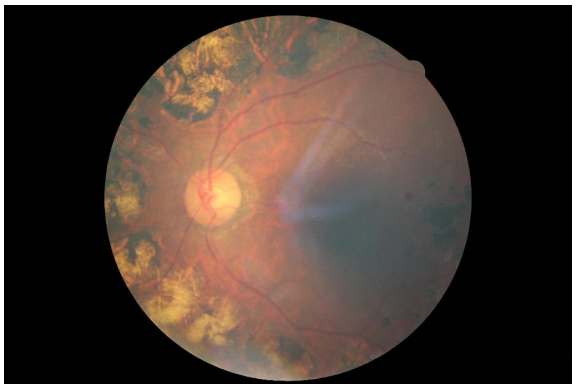


Figure 1: A simple CNN architecture showing successive convolutional and pooling layers with their channel counts and spatial dimensions. Template generated via NN-SVG.

The ability of CNNs to learn features directly from data, bypassing the need for manual, often laborious, feature engineering, has led to breakthroughs in numerous medical imaging tasks:

- **Image Classification:** Assigning a label to an entire image, such as determining the presence or absence of a disease (e.g., diabetic retinopathy detection from fundus images (Gulshan et al., 2016), classifying skin lesions (Esteva et al., 2017)), or grading tumor malignancy from histopathology slides, as is relevant to this thesis.
- **Object Detection:** Identifying and localizing specific objects or regions of interest within an image by drawing bounding boxes around them, for example, detecting nodules in chest X-rays or localizing organs (Lakhani and Sundaram, 2017).
- **Image Segmentation:** Assigning a class label to every pixel in an image, allowing for the precise delineation of anatomical structures or pathological regions. This is crucial for tasks like tumor volume quantification in MRI scans or segmenting cells in microscopy images (Ronneberger et al., 2015). The U-Net architecture (Ronneberger et al., 2015) and its variants have become a de facto standard for many medical image segmentation tasks.

(a) No DR (label 0)



(b) Proliferative DR (label 4)

Figure 2: Examples of diabetic retinopathy classification from EyePACS (Kaggle DR challenge).

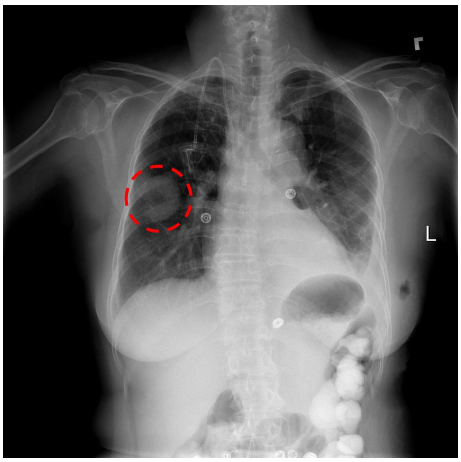


Figure 3: Example of an object detection task in medical imaging. A bounding box highlights a suspicious region in a chest X-ray, indicative of a potential abnormality such as pneumonia. Image adapted from the NIH ChestX-ray14 dataset (Wang et al., 2017).

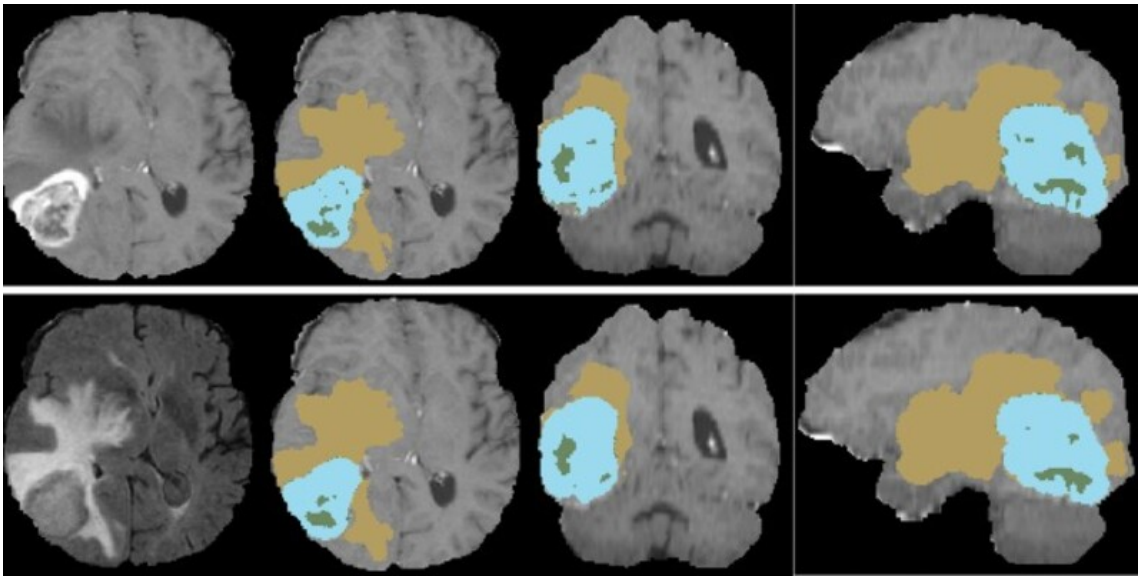


Figure 4: Example of medical image segmentation from the BraTS 2021 dataset. Brain MRI slices are shown with segmentation masks overlaid, highlighting tumor subregions: enhancing tumor (light blue), tumor core (gold), and whole tumor. Segmentation enables detailed analysis of tumor structure at the pixel level. Image credit: BraTS 2021 dataset (Baid et al., 2021).

Further advancements in CNN architectures have continued to push the boundaries of performance. Two notable architectures relevant to this work are:

- **ResNet (Residual Networks):** Introduced by He et al. (2016), ResNets addressed the degradation problem encountered when training very deep networks. They employ "residual blocks" with skip connections, which allow the network to learn an identity mapping if a layer is not beneficial, thereby enabling the training of significantly deeper and more powerful models (e.g., ResNet50, ResNet101).
- **DenseNet (Densely Connected Convolutional Networks):** Proposed by Huang et al. (2017), DenseNets connect each layer to every other layer in a feed-forward fashion within each dense block. This encourages feature reuse, strengthens feature propagation, reduces the number of parameters, and can lead to improved efficiency and performance.

These architectures, often pre-trained on large natural image datasets like ImageNet (Deng et al., 2009), serve as powerful feature extractors that can be fine-tuned for specific medical imaging tasks, leveraging transfer learning to achieve good performance even with limited medical data.

In summary, deep learning, and CNNs in particular, have become indispensable tools in medical image analysis, offering unprecedented capabilities for interpreting complex visual medical data. While their potential is vast, realizing this potential

in routine clinical practice requires addressing several challenges, most notably the ability of these models to generalize to diverse, unseen data, which will be discussed in subsequent sections.

2.2 The Domain Generalization Challenge in Medical AI

While the previous section highlighted the transformative potential of deep learning in medical image analysis, a critical impediment to its widespread and reliable clinical integration is the challenge of **domain generalization** (Wang et al., 2023). Deep learning models, despite their impressive performance on data similar to what they were trained on (in-distribution, ID data), often exhibit a significant degradation in performance when deployed in new clinical settings or applied to patient populations different from the training cohort. This phenomenon, known as **domain shift** or **dataset shift**, occurs when the statistical distribution of the test or deployment data (target domain) differs from that of the training data (Quionero-Candela et al., 2009). The ability of a model to maintain its performance when encountering such previously unseen domains, without any access to target domain data during training, is termed **Out-of-Distribution (OOD) generalization** (Koh et al., 2021). Achieving robust OOD generalization is arguably one of the most pressing research frontiers in medical AI, as failures can directly impact patient safety, diagnostic accuracy, and clinical trust (Finlayson et al., 2021; Castro et al., 2020). The research community continues to actively develop and benchmark methods to address this persistent issue, with new insights and approaches emerging regularly (Zhou et al., 2023).

Several factors contribute to domain shifts in medical imaging, creating a complex landscape of variability. These sources are often intertwined and can manifest subtly or profoundly:

1. **Patient Demographics and Biological Heterogeneity:** Patient populations naturally vary in age, sex, ethnicity, genetic predispositions, lifestyle factors, and comorbidities. These differences can lead to variations in anatomy, disease manifestation, prevalence, and image characteristics, directly impacting model performance if not accounted for (Gichoya et al., 2022). For instance, a model trained primarily on one demographic group might perform suboptimally on others, raising significant ethical concerns about fairness and equity in AI-driven healthcare (Rajpurkar et al., 2022). The Fitzpatrick17k dataset, used as a secondary benchmark in this thesis, directly encapsulates this challenge by evaluating model performance across a spectrum of skin tones.

2. **Variations in Medical Equipment and Acquisition Protocols:** This is a major source of domain shift, especially across different hospitals or even within the same institution over time.

    - *Scanner Differences:* Different manufacturers (e.g., Siemens, GE, Philips for MRI/CT; Hamamatsu, Leica, Aperio for digital slide scanners) and

models of imaging equipment possess unique hardware characteristics, internal image processing algorithms, and sensor properties, leading to systematic variations in image quality, resolution, contrast, and noise profiles (Stacke et al., 2020).

- *Acquisition Parameters:* Slight modifications in imaging protocols—such as MRI sequence parameters (e.g., echo time, repetition time), CT slice thickness or radiation dose, ultrasound probe frequency and settings, or microscope objectives and illumination in digital pathology—can substantially alter the resultant images (Glocker et al., 2019).
- *Calibration and Maintenance:* Imaging equipment can drift from its original calibration over time, or maintenance schedules might differ, introducing temporal shifts in image characteristics.

3. **Image Processing and Site-Specific Practices:** The journey from raw sensor data to a viewable medical image often involves multiple processing steps that can differ significantly.

   - *Histopathology:* This domain is notoriously affected by variability in tissue fixation methods (e.g., formalin-fixed paraffin-embedded), microtome sectioning thickness, Hematoxylin and Eosin (H&E) staining protocols leading to color and intensity variations, coverslipping techniques, and the characteristics of digital slide scanners (Madabhushi and Lee, 2016; Tizhoosh and Pantanowitz, 2018). Even subtle differences in reagent batches or technician practices can cause considerable visual shifts (Howard et al., 2021). The Camelyon17 dataset, used in this thesis, specifically captures such inter-hospital staining and scanner variations (Koh et al., 2021).
   - *Radiology and Other Modalities:* Image reconstruction algorithms for CT/MRI, post-processing filters (e.g., noise reduction, edge enhancement), image normalization techniques, and windowing settings for display can all introduce domain-specific characteristics.

4. **Geographical and Institutional Factors:** Clinical practices, disease prevalence, and even subtle environmental factors can vary by location. Data collected from a single institution or a limited set of geographically similar institutions may not capture the full spectrum of variability encountered globally or even nationally, leading to models that are geographically brittle (Zech et al., 2018). Multi-center studies frequently reveal performance drops when models are tested on external validation cohorts from unseen hospitals (DeGrave et al., 2021).

5. **Temporal Shifts (Concept Drift):** Over time, changes can occur in medical knowledge, diagnostic criteria, treatment protocols, equipment upgrades, and patient demographics. A model deployed today might see its performance degrade over months or years as the underlying data distributions gradually

evolve, a phenomenon known as concept drift or model drift (Gama et al., 2014). This necessitates continuous monitoring and periodic retraining or updating of clinical AI models.

The consequences of these domain shifts are profound for the clinical translation of medical AI. Unreliable OOD performance erodes clinician trust, poses risks of misdiagnosis or biased assessments (potentially leading to patient harm), and can exacerbate health disparities if models perform differently across demographic groups (Finlayson et al., 2021). Furthermore, the need for extensive site-specific recalibration or retraining for each new deployment environment is costly and impractical, hindering scalability. Regulatory bodies like the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) are increasingly emphasizing the need for robust validation of AI/ML-based medical devices on diverse datasets that reflect real-world variability (U.S. Food and Drug Administration, 2023).

It is important to distinguish domain generalization from the closely related problem of **domain adaptation (DA)**. In DA, some data (labeled or unlabeled) from the target domain(s) is assumed to be available during the training or fine-tuning phase, allowing the model to adapt (Wang and Deng, 2018). In contrast, domain generalization, the focus of this thesis, addresses the more challenging scenario where the model must perform well on unseen target domains without any prior exposure to their data.

Addressing the OOD generalization challenge is therefore not merely an academic exercise but a critical prerequisite for realizing the full societal and clinical benefits of AI in medicine. This thesis contributes to this endeavor by evaluating specific algorithmic and data-centric strategies designed to enhance the robustness of deep learning models against domain shifts prevalent in medical imaging..

2.3 Approaches to Out-of-Distribution Generalization

Given the critical challenges posed by domain shifts in medical AI, as detailed in Section 2.2, a significant body of research has been dedicated to developing methods that enhance Out-of-Distribution (OOD) generalization. These approaches can be broadly categorized based on where they intervene in the learning process: data-centric strategies that manipulate or augment the input data, model-centric strategies that modify the model architecture or learning objective, and learning strategy-centric approaches such as meta-learning or ensembling (Wang et al., 2023; Zhou et al., 2023). This thesis primarily investigates a model-centric approach (HYPO) and data-centric techniques (specific augmentations detailed in Section 2.4).

Empirical Risk Minimization (ERM) as a Baseline

The most fundamental approach to machine learning is Empirical Risk Minimization (ERM) (Vapnik, 1992). ERM aims to learn a model by minimizing the average loss

(empirical risk) on the observed training data, assuming that this data is representative of the test distribution. While often effective when training and test data are identically distributed, ERM can struggle with OOD generalization. Models trained via ERM may learn spurious correlations present only in the source domains, which do not hold in unseen target domains, leading to poor performance (Arjovsky et al., 2019). Consequently, ERM serves as a crucial baseline against which more sophisticated OOD generalization methods are compared.

Overview of Domain Generalization Strategies

Beyond ERM, numerous strategies have been developed to explicitly tackle OOD generalization. Some prominent categories include:

- **Domain-Invariant Feature Learning:** These methods aim to learn feature representations that are common across different domains, effectively factoring out domain-specific information while retaining task-relevant signals. Techniques include minimizing domain discrepancies using metrics like Maximum Mean Discrepancy (MMD) (Gretton et al., 2012), employing adversarial training to make features indistinguishable to a domain classifier (inspired by methods like DANN (Ganin et al., 2016)), or using contrastive learning to pull representations from the same class across different domains closer (Motiian et al., 2017).
- **Meta-Learning for Domain Generalization:** Meta-learning, or "learning to learn," approaches simulate domain shifts during training by splitting source domains into meta-train and meta-test sets. The goal is to learn an optimization procedure or model initialization that quickly adapts or generalizes to new, unseen domains (Li et al., 2018). Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017) and its variants have been explored in this context.
- **Ensemble Methods:** Combining predictions from multiple diverse models (an ensemble) can often lead to improved robustness and generalization compared to a single model. Diversity can be achieved by training models with different initializations, architectures, subsets of data, or even different learning algorithms (Lakshminarayanan et al., 2017).
- **Data Augmentation and Generation:** While ERM trains on the original data, data augmentation enriches the training set with synthetically modified samples. This can expose the model to a wider range of variations, potentially improving its robustness. General augmentation techniques are widely used, and more advanced strategies aim to generate OOD-like samples or specifically target robustness. This category, particularly targeted augmentation, is a key focus of this thesis and will be elaborated upon in Section 2.4.

Focus on Hyperspherical Embeddings: The HYPO Algorithm

Among model-centric approaches that aim to learn robust feature representations, methods leveraging hyperspherical geometry have gained attention. This thesis specifically evaluates the **HYPO** algorithm (Lee et al., 2024). The core idea behind such approaches is to constrain learned features to lie on a hypersphere, promoting desirable geometric properties in the embedding space.

The rationale for using hyperspherical embeddings for OOD generalization stems from several potential benefits:

- **Enhanced Discriminability:** By normalizing feature magnitudes, hyperspherical learning focuses on the angular separation between class features. This can lead to larger angular margins between classes, potentially making the decision boundaries more robust to OOD perturbations that might otherwise shift features across them (Liu et al., 2017).
- **Feature Compactness and Uniformity:** Some hyperspherical methods encourage features of the same class to form tight, compact clusters on the hypersphere, while ensuring that different class clusters are well-separated and that features are somewhat uniformly distributed, preventing collapse and improving representation quality (Wang et al., 2018; Deng et al., 2019).
- **Reduced Sensitivity to Magnitude Variations:** Domain shifts can sometimes manifest as changes in feature magnitudes. By projecting features onto a hypersphere, the model might become less sensitive to such variations, focusing instead on the intrinsic directional information of the features.

The HYPO algorithm, as presented by Lee et al. (2024), specifically improves OOD generalization by guiding its hyperspherical learning algorithm with principles of intra-class variation reduction and inter-class separation maximization. This ensures that features from the same class, even across different training domains, are closely aligned with their respective class prototypes on the hypersphere, while the prototypes of different classes are pushed to be maximally separated. This prototypical learning objective is designed to learn domain-invariant representations and is theoretically justified by the authors to improve the OOD generalization bound. It combines a standard classification loss with regularization terms that enforce these geometric properties in the hyperspherical feature space. The original work demonstrated HYPO's effectiveness

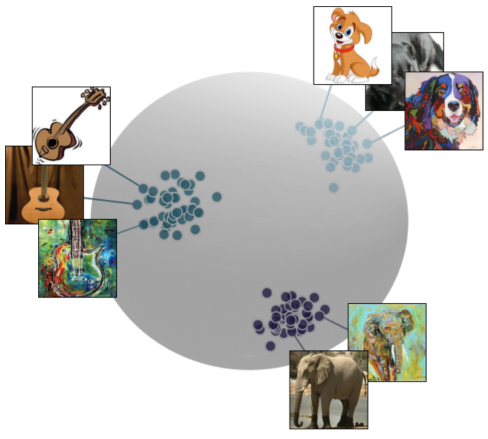


Figure 5: Illustration of hyperspherical embeddings. Reproduced fromLee et al. (2024). Subpanel images are from the PACS dataset (Li et al., 2017).

on several OOD benchmark datasets, suggesting its potential for enhancing model robustness.

Given its recent proposal and focus on learning geometrically robust representations for OOD generalization, HYPO was selected as a representative advanced model-centric technique for evaluation in this thesis. Its performance will be compared against ERM baseline and data-centric augmentation strategies to understand its relative benefits and interactions in the context of medical image analysis.

While model-centric approaches like HYPO offer promising avenues for improving intrinsic model robustness, they can often be complemented by data-centric strategies. The next section will delve into data augmentation techniques, which form another critical pillar of the investigations in this thesis.

2.3.1 The HYPO Algorithm: Hyperspherical Learning for OOD Generalization

A promising model-centric approach that has recently emerged is the HYPO (HYPerspherical Out-of-Distribution Generalization) algorithm, introduced by Lee et al. (2024). This method directly addresses the challenge of learning representations that are robust to domain shifts by leveraging the geometric properties of hyperspherical spaces. The development of HYPO is motivated by theoretical insights suggesting that OOD generalization performance is intrinsically linked to two key properties of the learned feature space: intra-class variation and inter-class separation (Lee et al., 2024).

- **Intra-class variation**, $V(\phi, E)$, quantifies the stability of a feature representation ϕ for samples of the same class across different domains e, e' within a set of environments E. Low intra-class variation implies that features for a given class remain consistent despite domain changes.
- **Inter-class separation**, $I(\phi, E)$, measures the discriminability of features from different classes within any given domain in E. High inter-class separation ensures that classes are clearly distinguishable.

While these properties were theoretically linked to OOD generalization, Lee et al. (2024) note that a practical learning algorithm directly optimizing for them with accompanying theoretical guarantees was previously an open area.

The core idea of HYPO is to learn domain-invariant feature representations by shaping the embedding space to reside on a unit hypersphere (i.e., $||z||_2 = 1$ for a feature embedding z). Within this hyperspherical space, HYPO's learning algorithm is explicitly guided by the principles of minimizing intra-class variation and maximizing inter-class separation. Specifically, it aims to ensure that feature embeddings from the same class, irrespective of their originating training domain within the available environments (E_{avail}), are closely aligned with a common class prototype on the

hypersphere. Concurrently, the prototypes representing different classes are pushed to be maximally distant from each other (Lee et al., 2024).

Methodologically, HYPO processes input samples x with corresponding labels y (for a finite set of classes Y) and learns a feature extractor ϕ that produces these normalized embeddings $z = \phi(x)$. The optimization is driven by a loss function comprising two main components:

1. $\mathcal{L}_{var}$: A variation loss term designed to reduce intra-class variation by encouraging sample embeddings to be close to their respective class prototypes within the hyperspherical space.

2. $\mathcal{L}_{sep}$: A separation loss term aimed at increasing inter-class separation by maximizing the distance between different class prototypes on the hypersphere.

By minimizing the combined loss $\mathcal{L}_{HYPO} = \mathcal{L}_{var} + \lambda\mathcal{L}_{sep}$ (where λ is a weighting factor), the algorithm directly promotes the desired geometric configuration of features. This approach of shaping the embedding geometry around class prototypes is intended to lead to smaller distribution discrepancies across domains, thereby benefiting OOD generalization (Lee et al., 2024).

A significant aspect of the HYPO framework is its theoretical justification. Lee et al. (2024) provide formal arguments demonstrating how their prototypical learning objective improves OOD generalization bounds. Their Theorem 6.1 (Appendix C.1 in their work) establishes that training with the proposed loss function, by aligning samples with class prototypes, effectively upper-bounds the supremum intra-class variation (V_{sup}). This V_{sup} is identified in their Theorem 3.1 as a key quantity determining the general upper bound on OOD generalization error. The HYPO loss directly minimizes a term ϵ that quantifies how well sample embeddings align with their class prototypes, thus tightening this bound. The theory also underscores the necessity of adequate inter-class separation for achieving OOD learnability. The relationship between training domains (E_{avail}) and the broader set of all potential test environments (E_{all}) is characterized using an "expansion function," which quantifies the potential increase in variation from E_{avail} to E_{all} and influences the learnability of an OOD problem (Lee et al., 2024).

Empirically, HYPO was evaluated on several standard OOD and domain generalization benchmarks, including corrupted versions of CIFAR-10 (e.g., CIFAR-10-C with Gaussian noise) and multi-domain datasets like PACS, Office-Home, and VLCS, often using ResNet architectures. The results reported by Lee et al. (2024) indicate that HYPO often significantly improves OOD generalization performance compared to ERM and other baselines. For instance, it demonstrated substantial accuracy gains on CIFAR-10-C and achieved competitive or superior performance on domain generalization tasks like PACS. The authors also noted that HYPO's performance could be further enhanced when combined with specialized optimization techniques such as Stochastic Weight Averaging Densely (SWAD). Importantly, their empirical analysis corroborated the theoretical claims by showing a significant reduction in

average intra-class variation in practice, and ablation studies confirmed the benefit of the inter-class separation loss component (Lee et al., 2024).

While other methods have explored hyperspherical embeddings (e.g., for face recognition (Liu et al., 2017; Deng et al., 2019)) or contrastive learning in various contexts (e.g., CIDER (Ming et al., 2023)), Lee et al. (2024) position their contribution in providing a direct theoretical linkage between their specific hyperspherical prototypical learning approach, the reduction of intra-class variation, and improved OOD generalization bounds. This distinguishes HYPO from methods that primarily aim to make feature distributions indistinguishable using discriminative classifiers or those that lack such formal OOD generalization guarantees.

However, the authors of HYPO also acknowledge limitations. Generalization to entirely arbitrary OOD scenarios remains an unsolved problem, and theoretical bounds depend on the nature of the distribution shift (e.g., the expansion function). The HYPO framework, with its reliance on class prototypes, is primarily presented for multi-class classification where the marginal label distribution $P(Y)$ is assumed to be consistent between training and testing. Significant shifts in the conditional distributions $P(X|Y)$ that drastically alter class appearance (e.g., photo-to-sketch transformations) might pose challenges for prototype-based methods (Lee et al., 2024).

Given its strong theoretical underpinnings related to intra-class variation and inter-class separation on the hypersphere, and its promising empirical performance on OOD benchmarks, HYPO is selected in this thesis as a key advanced algorithm for evaluation. Its study in the specific context of medical histopathology, with its unique domain shift characteristics, is expected to provide valuable insights.

While model-centric approaches like HYPO offer promising avenues for improving intrinsic model robustness, they can often be complemented by data-centric strategies. The next section will delve into data augmentation techniques, which form another critical pillar of the investigations in this thesis.

2.4 Data Augmentation for Enhanced Generalization and Robustness

Beyond modifying model architectures or learning objectives, data-centric approaches, particularly data augmentation, play a pivotal role in improving the generalization capabilities and robustness of deep learning models (Shorten and Khoshgoftaar, 2019). Data augmentation artificially expands the training dataset by creating modified copies of existing data or synthesizing new data instances. By exposing the model to a wider and more diverse range of training examples, augmentation helps it learn more invariant features, reduce overfitting to the original training set, and improve its performance on unseen data, including OOD samples.

Standard Data Augmentation Techniques

A variety of standard data augmentation techniques are commonly employed, especially in computer vision tasks. These can be broadly categorized:

- **Geometric Transformations:** These alter the spatial properties of images and include operations such as rotations, random scaling, translations, horizontal or vertical flipping, and random cropping. They aim to make the model invariant to changes in object position, orientation, and scale (Taylor and Nitschke, 2018).
- **Color Space Transformations:** These modify the pixel intensities and color characteristics of images. Common techniques include adjusting brightness, contrast, saturation, and hue, or adding random noise (e.g., Gaussian noise). These help the model become less sensitive to variations in lighting conditions and color distributions.
- **Other Deformations and Corruptions:** Techniques like elastic deformations, cutout/random erasing (occluding parts of an image), or applying various types of blur are also used to simulate different types of image degradations or variations.

While widely adopted and often beneficial, these generic augmentations may not always capture the full complexity or specific nature of variations encountered in specialized domains like medical imaging. Their effectiveness can be limited when domain shifts are characterized by artifacts or structural changes not well represented by simple geometric or color transformations.

Advanced Augmentation Frameworks: AugMix

To address the need for more diverse and effective augmentations that can improve robustness against unforeseen data shifts, advanced frameworks like AugMix have been proposed by Hendrycks et al. (2020). AugMix is a data processing technique designed to enhance the resilience and uncertainty estimation of image classifiers. It achieves this by creating augmented training samples that are significantly more varied and corrupted compared to standard augmentation methods, thereby pushing the model to learn more robust and invariant features.

The core technical pipeline for generating an AugMix image (x_{augmix}) from an original input image (x_{orig}) involves applying chains of simple augmentation operations, mixing the results of these chains using weights sampled from a Dirichlet distribution, and then mixing that composite image with the original using a weight sampled from a Beta distribution. This process is illustrated in Figure 6.

A critical component of the original AugMix method, beyond the data transformation itself, is the use of a **Jensen-Shannon Divergence (JSD) consistency**

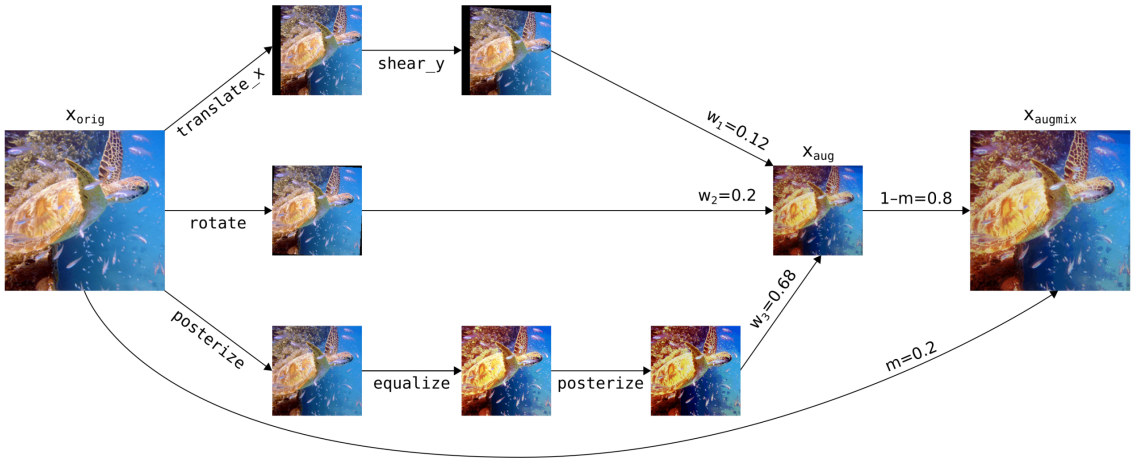


Figure 6: The AugMix data-augmentation pipeline. Briefly, AugMix generates multiple augmented variants of an image, mixes them with random convex weights, and then blends the result with the original input to improve robustness. Figure adapted directly from Hendrycks et al. (2020), Figure 4.

loss during training. This loss encourages the model to produce consistent predictions across diverse augmentations of the same input image. The total loss function is a combination of the standard classification loss and this JSD consistency loss, weighted by a hyperparameter λ:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{class}}(p_\theta(y|x_{\text{orig}}), y_{\text{true}}) + \lambda \cdot \text{JSD}\left(p_\theta(y|x_{\text{orig}}) \,||\, \frac{p_\theta(y|x_{\text{aug1}}) + p_\theta(y|x_{\text{aug2}})}{2}\right) \quad (1)$$

It is important to note that while the JSD loss is a core component of the original AugMix method, the experiments in this thesis primarily investigate the impact of the data processing and mixing pipeline itself. Therefore, this additional consistency loss term is **not** used in our implementations of AugMix-based strategies, allowing for a focused evaluation of the data transformation's effect.

While AugMix typically employs a set of general-purpose augmentation operations, its underlying framework for mixing is flexible. This flexibility is key to the MedAugmix strategy proposed in this thesis, where the set of base augmentations is replaced with **targeted** MedMNIST-C corruptions.

Targeted Augmentations and Benchmarks: MedMNIST-C

Recognizing that the nature of image variations and corruptions in medical imaging is often distinct from that in natural images, there has been a growing emphasis on developing **targeted** augmentation techniques and evaluation benchmarks.

Generic augmentations may not adequately simulate realistic medical artifacts (e.g., scanner-specific noise, patient movement, staining irregularities) or crucial biological variability, thereby limiting their effectiveness in preparing models for real-world clinical data.

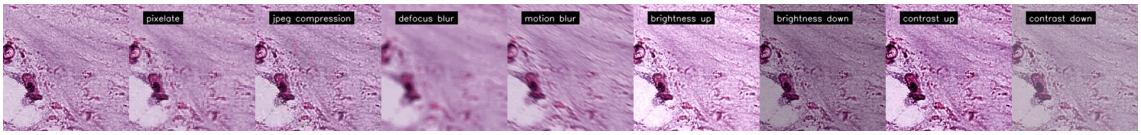
A significant step in addressing this gap was the introduction of **MedMNIST-C** by Di Salvo et al. (2024). This work presents a comprehensive benchmark and an accompanying Python library designed to simulate realistic image corruptions tailored for medical imaging tasks. The motivation behind MedMNIST-C is to fill a critical void in the medical imaging field, analogous to what ImageNet-C (Hendrycks and Dietterich, 2019) provides for the natural image domain: a standardized method to assess algorithmic robustness to common, realistic image corruptions across diverse modalities and applications. MedMNIST-C is explicitly inspired by the MedMNIST (Yang et al., 2023) APIs and the ImageNet-C repository, focusing on corruption robustness.

Technically, MedMNIST-C is built upon the MedMNIST+ collection, which comprises 12 2D datasets spanning 9 different imaging modalities. The corruptions defined in MedMNIST-C are applied to the test sets of these MedMNIST+ datasets (Di Salvo et al., 2024). A key aspect of its design is the structured categorization of corruptions and the use of domain knowledge to ensure their clinical relevance:

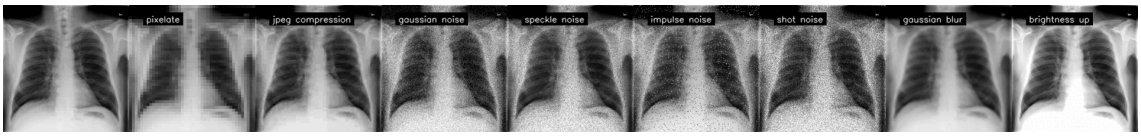
- **Corruption Categories:** The corruptions are organized into five main categories: digital, noise, blur, color, and task-specific, allowing for a nuanced evaluation of model performance against different types of potential real-world artifacts.
- **Modality-Specific Examples:**
  - *Digital Corruptions:* Applied broadly, these include JPEG compression and a `pixelate` corruption designed to mimic the effect of upsampling low-resolution images.
  - *Pathology and Blood Microscopy (Highly relevant to this thesis):* Corruptions include simulated **stain deposits** and **air bubbles**, as well as defocus and motion blur from image acquisition. Variations in brightness, contrast, and saturation due to differing illumination and scanner conditions are also incorporated.
  - *Chest X-Ray and Abdominal CT:* Includes brightness/contrast variations, Gaussian blur, various noise types common in X-ray imaging (Gaussian, speckle, impulse, shot noise), and gamma correction.
  - *Dermatoscopy:* Features noise artifacts (Gaussian, speckle, impulse, shot), blurring effects (defocus, motion, zoom), color-based artifacts (brightness, contrast), and task-specific artifacts like black corners from the dermatoscope and characters from camera overlays.

Visual examples of these corruptions applied to datasets like PathMNIST, ChestMNIST, and DermaMNIST are provided in the original work (Di Salvo et al., 2024).

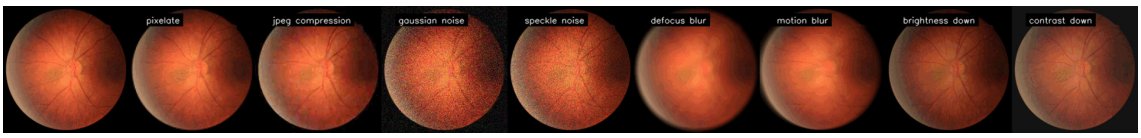
- **Severity Levels:** Crucially, following established benchmarking practices, each corruption is applied at five varying severity levels. This technical nuance allows researchers to measure the gradual degradation of model performance as corruptions intensify and to better understand model breaking points.



(a) Corruptions applied to a sample from BloodMNIST or PathMNIST (Microscopy).



(b) Corruptions applied to a sample from ChestMNIST (Chest X-Ray).



(c) Corruptions applied to a sample from RetinaMNIST (Retinal Fundus Image).

Figure 7: Visual examples of targeted medical corruptions from the MedMNIST-C framework. For each modality, an original image (left) is shown alongside several corrupted versions. Images sourced from the MedMNIST-C project GitHub repository (Di Salvo et al., 2024).

The MedMNIST-C benchmark and its associated open-source Python library facilitate the systematic assessment of algorithm robustness and, importantly, provide the tools to utilize the defined corruptions as data augmentation techniques during training. A key finding presented by Di Salvo et al. (2024) is the significant performance advantage gained by embedding domain knowledge into data augmentation strategies for improving model robustness against these specific corruptions.

Beyond benchmarking, the corruptions defined in MedMNIST-C are directly relevant to developing more robust models.

In essence, MedMNIST-C provides a much-needed, structured platform for evaluating and improving model robustness to common, realistic image corruptions. Its technical design, incorporating categorized and severity-scaled corruptions, offers a nuanced tool for researchers to both identify weaknesses in current models and rigorously test methods aimed at enhancing clinical reliability. The principles of advanced augmentation frameworks like AugMix and targeted corruption libraries like

MedMNIST-C form a crucial basis for the novel strategy proposed and evaluated in this thesis, whose detailed methodology will be presented in Section 3.

2.4.1 The WILDS Camelyon17 Dataset

A cornerstone of this thesis is the rigorous evaluation of OOD generalization strategies on a challenging and clinically relevant benchmark. For this purpose, the **Camelyon17-wilds** dataset, part of the WILDS (Workshops and Benchmarks on In-the-Wild Distribution Shifts) collection (Koh et al., 2021), was selected. The WILDS project curates datasets specifically designed to facilitate research and standardized evaluation of machine learning models under realistic distribution shifts.

The Camelyon17-wilds dataset is a patch-based variant derived from the original CAMELYON17 challenge (Bandi et al., 2019), which itself built upon data from the earlier CAMELYON16 challenge (Bejnordi et al., 2017). The original CAMELYON17 challenge focused on analyzing histological whole-slide images (WSIs) of lymph nodes from breast cancer patients, collected from five different medical centers (hospitals) in the Netherlands. The domains within the Camelyon17-wilds dataset directly correspond to these five hospitals.

The primary task in the Camelyon17-wilds benchmark is a lesion-level binary classification problem: to predict whether a given 96x96 pixel image patch, extracted from a WSI, contains tumor tissue (Koh et al., 2021). More specifically, the label pertains to whether the central 32x32 pixel region of the patch contains tumor cells. The dataset comprises approximately 455,954 such patches extracted from 50 WSIs, which were originally annotated by pathologists. A key characteristic of the dataset splits (training, validation, and test sets) provided by WILDS is that they are class-balanced, containing an equal number of positive (tumor) and negative (normal) examples (Koh et al., 2021). Metadata, including the source WSI (slide) and hospital ID (domain identifier), is available for each patch.

The critical distribution shift addressed by Camelyon17-wilds is the **inter-hospital variation**. The WILDS framework structures the dataset such that the training set includes data from a subset of these hospitals, while the OOD validation and test sets include data from hospitals not seen during training. This setup directly simulates the real-world scenario of deploying a model in a new clinical environment. The sources of this inter-hospital variation are multifaceted and include (Koh et al., 2021):

- Differences in patient populations across institutions.
- Variations in histopathological slide preparation, particularly **staining protocols** (e.g., H&E). Staining differences are noted as a major source of visual dissimilarity in such datasets and present a significant challenge for model generalization.
- Discrepancies in **image acquisition protocols** and digital slide scanning equipment used at the different medical centers.

The OOD test hospital in the WILDS setup was specifically chosen because its patches were visually distinct, emphasizing the impact of domain shift on model performance (Koh et al., 2021). Consequently, Camelyon17-wilds serves as a controlled and challenging testbed for methods aiming to learn robustness to these common variations.

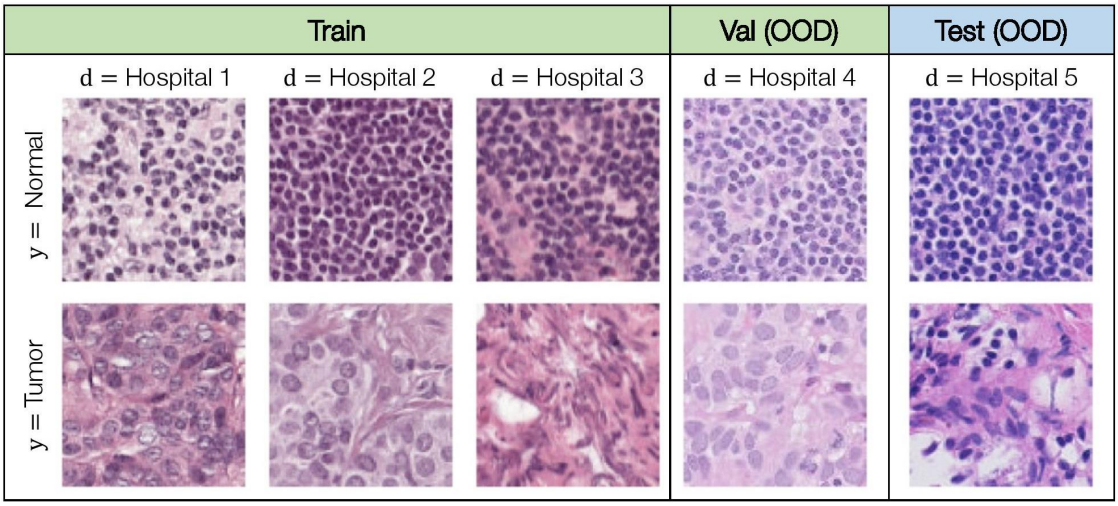


Figure 8: Example image patches from the WILDS Camelyon17 dataset, illustrating visual differences between (a) a training hospital domain and (b) an unseen OOD hospital domain. Note the variations in staining and appearance. (Figure adapted from Koh et al. (2021))

Models are evaluated on Camelyon17-wilds using **average accuracy**, which is an appropriate metric given the class-balanced nature of the dataset splits (Koh et al., 2021). Baseline Empirical Risk Minimization (ERM) models typically exhibit a substantial performance gap between in-distribution (ID) and OOD settings. For instance, Koh et al. (2021) report ERM achieving high ID accuracy on training hospital data (e.g., 98.7%) but significantly lower OOD test accuracy (e.g., 70.3%), highlighting the severity of the generalization challenge. The availability of training data from multiple distinct hospital domains within Camelyon17-wilds allows algorithms the opportunity to learn representations that are more robust to these variations.

The Camelyon17-wilds benchmark setup is distinct from some previous uses of CAMELYON data. For example, Tellez et al. (2018) explored scenarios using data from only one hospital for training, whereas Camelyon17-wilds leverages multiple training hospitals. Similarly, the PCam dataset (Veeling et al., 2018), based on the earlier CAMELYON16 challenge, primarily involved data from two hospitals. The specific multi-domain training and disjoint OOD test structure of Camelyon17-wilds makes it particularly well-suited for the domain generalization research pursued in this thesis.

Given its clinical relevance, well-defined OOD generalization problem based on real-world inter-hospital variations, and standardized evaluation protocol, the Camelyon17-wilds dataset provides an ideal platform for the investigations conducted in this work.

2.4.2 The Fitzpatrick17k Dataset

To test the generalizability of findings in a different medical domain, the **Fitzpatrick17k** dataset was used as a secondary benchmark. This dataset facilitates research into the critical challenge of model fairness and robustness in dermatology (Groh et al., 2021).

Dataset Content and Task: Fitzpatrick17k is designed for fine-grained skin disease classification. It is composed of **16,577 clinical photographs** sourced from two public dermatology atlases: DermaAmin and Atlas Dermatologico. The primary task is to classify an image into one of **114 distinct skin disease categories**, which range from common benign lesions to malignant cancers. For broader analysis, the dataset also provides labels for coarser groupings of these conditions into 9 mid-level and 3 high-level (benign, malignant, non-neoplastic) categories. Each image is annotated with a disease label and a **Fitzpatrick skin type (1–6)**, a scale used to classify human skin color based on its response to UV light.

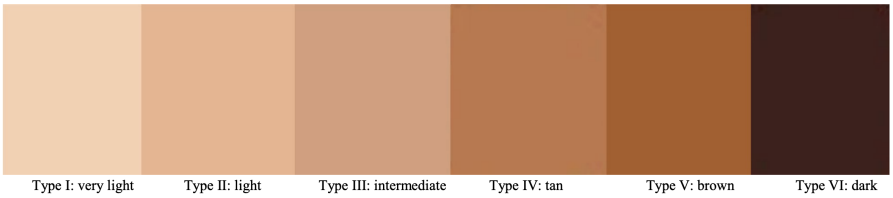


Figure 9: The Fitzpatrick scale, a widely used numerical classification scheme for human skin phototypes based on their response to sun exposure. The scale ranges from Type I (very fair) to Type VI (deeply pigmented). Image adapted from Alipour et al. (Alipour et al., 2024).

Skin-Tone Imbalance: The principal Out-of-Distribution (OOD) challenge within the Fitzpatrick17k dataset stems from a significant **skin-tone imbalance**. The dataset is skewed towards lighter skin types, with approximately 7,755 images of Fitzpatrick types 1-2 and 6,089 of types 3-4, but only 2,168 images of darker skin types 5-6. This distribution reflects a common bias in medical data collection. It creates a critical fairness and robustness challenge, where models trained predominantly on images of lighter-skinned individuals may underperform when deployed on underrepresented populations with darker skin tones.

Dataset Splits and Evaluation: For the OOD evaluation in this thesis, the dataset was split by skin tone: models were trained on images corresponding to Fitzpatrick types 1 through 4, and subsequently tested on the OOD set composed of images from types 5 and 6. The primary evaluation metric reported for this dataset is **balanced accuracy** across the 3 higher level classes, which appropriately accounts for potential class imbalance within the dataset splits.

2.4.3 Backbone Model Architectures

To evaluate the different Out-of-Distribution (OOD) generalization strategies investigated in this thesis, two widely adopted and powerful Convolutional Neural Network (CNN) architectures were selected as backbones: ResNet50 and DenseNet121. These models are frequently employed in medical image analysis tasks due to their strong performance capabilities and the common availability of weights pre-trained on large-scale natural image datasets, which facilitates transfer learning. For all experiments conducted in this study, both ResNet50 and DenseNet121 models were initialized with weights pre-trained on the ImageNet dataset (Deng et al., 2009). This standard practice allows the models to leverage foundational visual features learned from a diverse dataset, often leading to improved performance and faster convergence on more specialized medical imaging tasks with potentially smaller datasets.

- **ResNet50:** The Residual Network (ResNet) architecture, introduced by He et al. (2016), marked a significant breakthrough in deep learning by enabling the effective training of substantially deeper neural networks than was previously feasible. The core innovation of ResNets is the "residual block," which incorporates skip connections or shortcuts. These connections allow the network to learn residual functions with reference to the layer inputs, essentially allowing gradients to bypass layers if an identity mapping is optimal for that block. This mechanism mitigates the vanishing gradient problem in very deep networks. The ResNet50 variant, which comprises 50 layers, has become a de facto standard in computer vision and is frequently used as a strong baseline in medical imaging research due to its excellent balance of representational capacity and computational demand. It was chosen for this thesis to represent a well-established, high-performing architectural paradigm.
- **DenseNet121:** Densely Connected Convolutional Networks (DenseNets), proposed by Huang et al. (2017), offer an alternative architectural design aimed at improving information flow and gradient propagation throughout the network. In a DenseNet, each layer within a "dense block" receives direct connections from all preceding layers in that block, and its own feature maps are passed on to all subsequent layers. This dense connectivity encourages substantial feature reuse, reduces the number of parameters (often leading to higher parameter efficiency compared to ResNets of similar depth), and can alleviate the vanishing gradient problem. DenseNet121, a 121-layer variant, is known for its strong performance across various benchmarks. It was selected for this thesis

to provide an evaluation on a distinct architectural design, allowing for an assessment of whether the observed OOD generalization effects are consistent across different types of deep CNNs.

By employing these two different yet powerful CNN architectures, this thesis aims to draw more robust and generalizable conclusions regarding the effectiveness of the evaluated OOD generalization strategies and data augmentation techniques. The consistent performance (or lack thereof) of a method across these diverse backbones can lend greater credibility to the findings.

2.5 Evaluation Metrics for OOD Generalization

To provide a comprehensive and fair assessment of model performance, particularly for Out-of-Distribution (OOD) generalization, a carefully chosen suite of metrics was employed. The primary metric was tailored to the specific characteristics of each benchmark dataset to ensure a meaningful evaluation.

Primary Performance Metrics: For the **WILDS Camelyon17** dataset, a class-balanced binary classification task, standard **Accuracy** was used as the primary performance indicator. For the **Fitzpatrick17k** dataset, which for this study involves a high-level 3-class classification task (benign, malignant, and non-neoplastic) with inherent class imbalance, **Balanced Accuracy** was chosen as the primary metric. Balanced accuracy, defined as the average of recall obtained on each class, provides a more robust measure of performance on imbalanced multi-class problems.

OOD Evaluation Framework: A consistent framework was used to evaluate OOD generalization across both datasets. The process was designed to simulate a realistic model development and deployment scenario:

1. **Model Selection:** During training, model performance was monitored on an In-Distribution (ID) validation set. The single model checkpoint that achieved the highest performance (highest Accuracy for Camelyon17, or highest Balanced Accuracy for Fitzpatrick17k) on this ID validation set was selected as the "best" model for evaluation.

2. **OOD Performance Measurement:** This selected "best" model checkpoint was then evaluated on the completely unseen Out-of-Distribution (OOD) test set. Its performance on this OOD set is the primary metric we report for practical OOD generalization.

3. **Generalization Gap Calculation:** To quantify how well a model's performance translates from familiar to unfamiliar data, we calculated the "generalization gap." This is the simple difference between the model's peak performance on the ID validation set and its corresponding performance on the OOD test set.

Statistical Robustness and Reporting To ensure the reliability of our findings, all experiments were conducted three times using distinct random seeds (0, 1, and 2). All reported performance values in the results section are the **mean and standard deviation** across these three runs. While the discussion focuses on the key metrics described above, the detailed results tables in Section 4 also report other supporting metrics, such as the absolute best OOD performance achieved at any point during training.

3 Methodology

This chapter details the comprehensive methodology designed to systematically evaluate strategies for enhancing Out-of-Distribution (OOD) generalization in medical imaging. The experimental framework was constructed to directly address the research questions outlined in Section 1.2.

3.1 Experimental Design Overview

The core of this research is a comparative benchmark designed to assess the effectiveness of different approaches to improving OOD robustness. To ensure that our findings are not limited to a single type of domain shift, the evaluation framework is built upon two distinct and challenging medical imaging benchmarks:

1. **WILDS Camelyon17:** A histopathology dataset used to evaluate generalization across different hospitals, representing a domain shift caused by variations in lab protocols, staining, and scanners.

2. **Fitzpatrick17k:** A dermatology dataset used to evaluate generalization across different patient demographics, representing a domain shift caused by biases in skin tone representation.

Within this dual-dataset framework, we benchmarked two primary categories of OOD generalization strategies against a standard baseline. These strategies are:

- **Algorithmic Strategies:** This involves comparing the performance of standard **Empirical Risk Minimization (ERM)** against a specialized representation learning algorithm, **HYPO**, which is designed to learn a more robust feature space.

- **Data-Centric Strategies:** This involves evaluating the impact of various data augmentation techniques applied during training. These include a basic application of targeted medical corruptions, the standard (Plain) AugMix method, and our proposed novel strategy, **MedAugmix**, which tailors the AugMix pipeline with targeted medical image corruptions.

By systematically applying these strategies to both datasets and evaluating them with a consistent protocol, this thesis aims to identify which approaches offer the most significant and generalizable improvements in OOD performance. The subsequent sections of this chapter detail each component of this experimental design: the benchmark datasets and their pre-processing (Section 3.2), the model architectures and evaluation metrics (Section 2.5), the specific implementation of the training strategies (Section 3.1), and the hardware and software setup (Section 3.4).

3.2 Benchmark Datasets

The experimental framework of this thesis is built upon two distinct and publicly available medical imaging benchmarks. Each was chosen to represent a different, clinically relevant type of Out-of-Distribution (OOD) challenge. By evaluating methods against these disparate problems—one technical and process-based, the other demographic and fairness-related—this thesis aims to derive more robust and broadly applicable conclusions about OOD generalization in medical AI.

3.2.1 Primary Benchmark: WILDS Camelyon17 (Histopathology)

The primary benchmark for this study is the **WILDS Camelyon17** dataset (Koh et al., 2021). Sourced from the CAMELYON17 challenge, it consists of 96x96 pixel histopathology image patches for a binary classification task: detecting metastatic breast cancer in lymph node sections. The critical OOD challenge stems from **inter-hospital variability**, as the dataset combines images from different medical centers with variations in slide preparation, H&E staining protocols, and digital slide scanning equipment. The standard WILDS evaluation protocol involves training on a subset of hospitals and testing on data from an entirely unseen hospital. Due to the class-balanced nature of the dataset splits, standard **Accuracy** is used as the primary evaluation metric.

3.2.2 Secondary Benchmark: Fitzpatrick17k (Dermatology)

The secondary benchmark is the **Fitzpatrick17k** dataset (Groh et al., 2021), which focuses on dermatology. It is composed of 16,577 clinical photographs which can be categorized into 114 fine-grained disease classes. For the purposes of this thesis, we focus on a higher-level, 3-class classification task (benign, malignant, and non-neoplastic). The principal OOD challenge arises from a significant **skin tone imbalance**, where the dataset is skewed towards lighter skin types. For our evaluation, models were trained on images of skin types 1-4 and tested on an OOD set of underrepresented skin types 5-6. This setup creates a critical fairness and robustness challenge. Given the multi-class nature of the task, **Balanced Accuracy** is used as the primary evaluation metric.

3.3 Model Setup

To create a consistent and fair benchmark for all evaluated strategies, a common experimental setup was established for the model architectures and the protocol for measuring success. This section outlines these core methodological components.

To create a consistent and fair benchmark for all evaluated strategies, a common experimental setup was established for the model architectures. This section outlines these core methodological components.

3.3.1 Model Configuration

The experiments utilized two standard backbone architectures, **ResNet50** and **DenseNet121**, with weights pre-trained on ImageNet, to serve as powerful feature extractors. A key aspect of our experimental design was the addition of a common projection head to both backbones. This head, an MLP with two linear layers, transformed the backbone features into a **128-dimensional embedding space**.

This common architecture allowed for a direct comparison of how different learning algorithms utilize this feature space. For the **HYPO** algorithm, these 128-dimensional features were L2-normalized, as required by its hyperspherically-based loss functions. For the **ERM** baseline, in a specific configuration designed for this study, the standard Cross-Entropy loss was applied directly to the unnormalized 128-dimensional output of this same projection head. This unified model structure ensures that performance differences can be more directly attributed to the learning algorithms and augmentation strategies rather than disparate model architectures. For fine-tuning, BatchNorm layers were frozen for ResNet50, while DenseNet121 was fine-tuned end-to-end.

subsectionTraining Algorithms and Strategies

This section details the core learning algorithms and specific data augmentation strategies employed. To ensure fair comparisons, a consistent set of fundamental training hyperparameters was used for all experiments, as summarized in Table 1. Specific parameters for each algorithm are detailed in their respective subsections.

3.3.2 Empirical Risk Minimization (ERM) Baseline

Purpose The Empirical Risk Minimization (ERM) principle served as the standard baseline for all comparative evaluations.

Loss Function The training objective for ERM was to minimize the **Binary Cross-Entropy loss** (`torch.nn.CrossEntropyLoss`). As detailed in Section 3.3, this loss was applied to the 128-dimensional unnormalized output of the model's projection head.

Table 1: Common Hyperparameters for All Training Runs.

Hyperparameter	**Value**
Optimizer	SGD (Nesterov momentum = 0.9)
Weight Decay	1e-4
Batch Size	384
Total Epochs	50
Initial Learning Rate	5e-4
Learning Rate Schedule	Step decay: rate 0.1 at epochs 30 and 40
Learning Rate Warm-up	10 epochs (linear from 0.001 to 5e-4)

Optimization and Schedule The optimization process followed the common protocol detailed in Table 1.

3.3.3 HYPO Implementation

Purpose The HYPO algorithm (Lee et al., 2024) was implemented as the advanced, model-centric strategy for improving OOD generalization by learning a geometrically structured feature space.

Loss Function The HYPO training objective is a composite loss that operates on the L2-normalized 128-dimensional features from the model's projection head. It consists of two main components:

- **Compactness Loss ($\mathcal{L}_{\text{Comp}}$):** Encourages features of the same class to cluster tightly around their class-conditional prototype on the hypersphere.
- **Dispersion Loss ($\mathcal{L}_{\text{Dis}}$):** Encourages the class prototypes to be maximally separated from each other in the hyperspherical space.

The total loss for HYPO is a weighted sum of these two components:

$$\mathcal{L}_{\text{HYPO}} = w \cdot \mathcal{L}_{\text{Comp}}(f; \theta) + \mathcal{L}_{\text{Dis}}(f; \theta)$$

where f represents the features and θ the model parameters.

Optimization and Schedule Optimization followed the common protocol detailed in Table 1.

Key Hyperparameters: The HYPO loss function was governed by three key algorithm-specific hyperparameters: the compactness weight w was set to 2.0, the loss temperature was set to 0.1, and the prototype EMA momentum was set to 0.95.

3.3.4 Data Augmentation Strategies

To investigate the impact of data-centric approaches on OOD generalization, four distinct augmentation conditions were applied during model training. The strategies ranged from a baseline control to advanced mixing techniques, including the novel MedAugmix method proposed in this thesis. An overview of these strategies is provided in Table 2. For all conditions, these augmentations were applied to the training data only; validation and test sets consistently underwent only the basic pre-processing steps of tensor conversion and normalization.

Table 2: Overview of Data Augmentation Strategies Evaluated.

Strategy Name (*Execution Identifier*)	**Description**
No Advanced Augmentation (`baseline`)	Serves as the control group. Models are trained with only basic pre-processing (`ToTensor` and `Normalize`).
Basic MedMNIST-C (`medmnistc`)	Applies a single, randomly chosen targeted medical corruption with a random severity to each image to test simple targeted augmentation.
Standard AugMix (`plain_augmix`)	Applies the generic `torchvision.transforms.AugMix()` to benchmark against a standard advanced augmentation method.
MedAugmix (Proposed) (`medmnist_c_augmix`)	The novel strategy proposed in this thesis. Adapts the AugMix pipeline to use a mixture of targeted medical corruptions as its base operations.

(a) No Advanced Augmentation (Baseline Control) To establish a direct baseline performance for both ERM and HYPO without the influence of any advanced data augmentation.
Implementation: The data pipeline for this condition only included the fundamental pre-processing steps of converting images to PyTorch tensors and normalizing them with ImageNet statistics.

(b) Basic MedMNIST-C Augmentations To assess the impact of applying simple, individual targeted medical corruptions.

- **Source of Corruptions:** The transform was initialized with the set of corruptions defined for the "bloodmnist" dataset for Camelyon17 experiments, and the "dermamnist" set for Fitzpatrick17k experiments.

- **Application:** For each training image, this transform applies a single, randomly selected MedMNIST-C corruption type from the source pool at a randomly chosen severity level (from 1 to 5).

(c) Standard AugMix (Plain AugMix) To benchmark performance against a standard, off-the-shelf advanced augmentation technique that uses generic, natural-image-focused transformations.

- **Parameters:** The default parameters of the torchvision implementation were used: operational severity of 3, mixture width (k) of 3, a random chain depth of 1-3 operations, and an alpha of 1.0 for mixing.
- **JSD Loss:** Consistent with our other augmentation experiments, the Jensen-Shannon Divergence (JSD) consistency loss was not used.

(d) MedAugmix (Proposed Novel Strategy) To evaluate a novel hybrid strategy that combines the structural diversity of the AugMix pipeline with the domain relevance of targeted medical corruptions.

- **Base Operations:** The augmentation operations used within the AugMix chains were sourced from the MedMNIST-C corruption sets ("bloodmnist" for Camelyon17, "dermamnist" for Fitzpatrick17k). Specific corruptions known to cause processing issues (e.g., "MotionBlur," "ZoomBlur") were excluded.
- **Key Hyperparameters:** The implementation was configured using the experimental arguments:
  - **Severity:** All operations within a given experiment operated at a fixed severity, controlled by `args.augmix_severity` (e.g., 3 or 5).
  - **Mixture Width (k):** The number of parallel augmentation chains was controlled by `args.augmix_mixture_width` (e.g., 1, 3, or 5).
  - **Chain Depth:** A random depth of 1 to 3 MedMNIST-C operations was applied per chain.
  - **Alpha:** A value of 1.0 was used for the Dirichlet and Beta mixing distributions.
- **JSD Loss:** The JSD consistency loss was not used for this strategy.

3.4 Experimental Runs and Hardware

We detail our experimental management and compute resources to ensure full reproducibility.

3.4.1 Reproducibility Protocol

- **Configurations:** ERM and HYPO × {No Augmentation, Basic MedMNIST-C, MedCAugMix, Standard AugMix} × {ResNet50, DenseNet121}.
- **Seeds:** Each configuration executed with seeds {0, 1, 2}
- **Execution Control:** Managed via Bash scripts (e.g., `run_missing_seeds.sh`), which fix all hyperparameters except `seed`.
- **Reporting:** Results are reported as mean ± standard deviation over the three runs.

3.4.2 Hardware Specifications

Table 3: Compute Resources

GPU	NVIDIA RTX A5000 (24 GB)
CPU	Intel Core i7-13700 (13th Gen)
RAM	62 GB DDR4
Storage	1 TB NVMe SSD
OS	Ubuntu 22.04 LTS

All training and evaluation runs use a single GPU (specified via `args.gpu`).

4 Experimental Results

This section presents the empirical findings from our experimental benchmark. We first report the detailed results on the primary benchmark, WILDS Camelyon17, to establish core performance trends. Subsequently, findings from the secondary benchmark, Fitzpatrick17k, are presented to assess the generalizability of these trends.

All quantitative results are reported as the mean and standard deviation across three independent runs with different random seeds (0, 1, and 2). A detailed interpretation and discussion of these findings will follow in Section 5.

4.1 Performance on Primary Benchmark (Camelyon17)

We begin our analysis with the results on the WILDS Camelyon17 dataset. This section systematically evaluates the performance of the baseline algorithms and the incremental impact of each data augmentation strategy.

4.1.1 Baseline Performance: ERM vs. HYPO

First, we establish the foundational performance when training with no advanced data augmentation. These results, shown in Table 4, reveal the inherent difficulty of the OOD challenge and provide a direct comparison between Empirical Risk Minimization (ERM) and the HYPO algorithm.

Table 4: Baseline OOD performance on Camelyon17 (Mean ± Std over 3 seeds).

Architecture	Algorithm	ID Val Acc	OOD Test Acc	Gap
DenseNet121	ERM	0.9147 ± 0.0015	**0.8529 ± 0.0120**	0.0618
	HYPO	0.9158 ± 0.0132	**0.8626 ± 0.0172**	0.0532
ResNet50	ERM	0.9125 ± 0.0003	**0.8211 ± 0.0101**	0.0914
	HYPO	0.9067 ± 0.0048	**0.7735 ± 0.0563**	0.1333

On the DenseNet121 architecture, the HYPO algorithm achieved a slightly higher practical OOD performance than ERM, with a mean accuracy of **0.8626** compared to ERM's **0.8529**. This suggests a modest benefit from HYPO's representation learning on this backbone. Interestingly, this trend reversed for the ResNet50 architecture, where ERM outperformed HYPO with a practical OOD accuracy of **0.8211** versus **0.7735**, suggesting an architecture-dependent effect for these baseline algorithms.

These initial results highlight the severity of the OOD challenge, as all configurations exhibited substantial generalization gaps, ranging from approximately **0.05 to 0.13**. This establishes that without effective regularization, even specialized algorithms struggle to generalize. Next, we will show how each augmentation strategy impacts the performance and the generalization gaps reported above.

4.1.2 Impact of Basic MedMNIST-C Augmentations

The next set of experiments assessed the impact of a simple, targeted augmentation strategy. For this condition, models were trained with "Basic MedMNIST-C" augmentations, which involved applying a single, randomly selected corruption from the "bloodmnist" set with a random severity (1-5) to each training image, as detailed in Section 3.3.4.

Table 5 presents the performance of ERM and HYPO when trained with this augmentation strategy.

When comparing these results to the non-augmented baselines (Table 4), the effectiveness of this basic augmentation strategy appears to be highly dependent on the model architecture.

For the **DenseNet121** architecture, this approach yielded modest benefits. As shown by comparing Table 5 with the baseline results, ERM's OOD accuracy increased from 0.8529 to 0.8660. This improvement came with a reduction in the

Table 5: OOD Performance with Basic MedMNIST-C Augmentations on Camelyon17-wilds (Mean ± Std over 3 seeds).

Architecture	Algorithm	ID Val Acc	OOD Test Acc	Gap
DenseNet121	ERM	0.9175 ± 0.0042	**0.8660 ± 0.0385**	0.0515
	HYPO	0.9336 ± 0.0015	**0.8701 ± 0.0304**	0.0634
ResNet50	ERM	0.9123 ± 0.0009	**0.8205 ± 0.0065**	0.0918
	HYPO	0.9071 ± 0.0086	**0.7568 ± 0.0303**	0.1504

Table 6: Comparison of OOD Accuracy with and without Basic MedMNIST-C Augmentations.

Architecture	Algorithm	Baseline	MedMNIST-C	Change (Δ)
DenseNet121	ERM	0.8529	0.8660	+0.0131
	HYPO	0.8626	0.8701	+0.0075
ResNet50	ERM	0.8211	0.8205	-0.0006
	HYPO	0.7735	0.7568	-0.0167

generalization gap, suggesting that DenseNet121 was able to leverage these simple corruptions as a useful regularization signal. HYPO's OOD accuracy also saw a slight increase from 0.8626 to 0.8701, although its generalization gap widened, indicating it improved more on in-distribution data than out-of-distribution data with this strategy.

Conversely, for the **ResNet50** architecture, this augmentation strategy offered no benefits and was even detrimental in the case of HYPO. ERM's OOD performance remained static (0.8211 vs. 0.8205), while HYPO's performance noticeably decreased from 0.7735 to 0.7568. This suggests that for a less inherently robust architectural setup like ResNet50 in this context, the unstructured noise from single random corruptions may have disrupted the feature learning process more than it helped. It is plausible that for a sensitive algorithm like HYPO, which relies on learning a precise geometric structure, this type of random noise can make it more difficult to find stable and generalizable class prototypes.

These findings provide a key insight: simply applying random, targeted corruptions is not a universally effective strategy. Its success is conditional on the underlying model architecture's capacity to handle such variations, which motivates the need for more structured and powerful augmentation pipelines.

4.1.3 Impact of Standard AugMix (Plain AugMix)

Next, we evaluated the impact of a standard, generic advanced augmentation method, "Plain AugMix."

Table 7: OOD Performance with Standard AugMix on Camelyon17-wilds (Mean ± Std).

Architecture	Algorithm	ID Val Acc	OOD Test Acc	Gap
DenseNet121	ERM	0.9360 ± 0.0010	**0.8791 ± 0.0076**	0.0569
	HYPO	0.9427 ± 0.0042	**0.8655 ± 0.0131**	0.0771
ResNet50	ERM	0.9338 ± 0.0007	**0.8527 ± 0.0038**	0.0811
	HYPO	0.9460 ± 0.0022	**0.7924 ± 0.0157**	0.1536

To more clearly quantify the effect of this augmentation, Table 8 presents a direct comparison of the OOD performance against the non-augmented baselines.

Table 8: Comparison of OOD Accuracy with and without Standard AugMix on Camelyon17.

Architecture	Algorithm	Baseline	AugMix	Change (Δ)
DenseNet121	ERM	0.8529	0.8791	+0.0262
	HYPO	0.8626	0.8655	+0.0029
ResNet50	ERM	0.8211	0.8527	+0.0316
	HYPO	0.7735	0.7924	+0.0189

The results show that applying Standard AugMix provided a significant and positive impact, particularly for the ERM algorithm on both architectures. This suggests that the data diversification from the AugMix processing pipeline acts as a powerful regularizer, even when using generic, non-targeted augmentations.

This trend was most pronounced for ERM on the ResNet50 architecture, which saw a substantial OOD accuracy improvement of +0.0316. ERM on DenseNet121 also benefited significantly, with an increase of +0.0262. These gains highlight ERM's ability to effectively leverage the broad diversity introduced by the AugMix pipeline to learn more robust features.

In contrast, the HYPO algorithm benefited much more modestly from this generic augmentation. While its OOD accuracy did increase for both ResNet50 (+0.0189) and DenseNet121 (+0.0029), the improvements were considerably smaller than those seen with ERM. It is plausible that the non-targeted, sometimes "unrealistic" transformations from Standard AugMix do not align as well with HYPO's objective of learning a geometrically precise feature space based on class prototypes.

A key observation across both architectures is that ERM consistently outperformed HYPO when this generic Plain AugMix strategy was applied. On ResNet50, ERM's OOD accuracy of 0.8527 was substantially higher than HYPO's 0.7924. This result reverses the performance hierarchy seen in the baseline condition for DenseNet121, where ERM also surpassed HYPO (0.8791 vs. 0.8655). This further supports the finding that a simpler learning algorithm like ERM can be made more effective for

OOD generalization than a specialized algorithm, provided it is paired with a strong and diverse data augmentation pipeline.

4.1.4 Performance of Proposed MedAugmix Strategy

The final and most critical set of experiments on the Camelyon17 benchmark evaluated the performance of the novel **MedAugmix** strategy. As detailed in the methodology, this approach adapts the AugMix pipeline to use targeted MedMNIST-C corruptions sourced from "bloodmnist" as its base operations. The results for two primary configurations of MedAugmix are presented in Table 9.

Table 9: OOD Performance with Proposed MedAugmix Strategy on Camelyon17-wilds (Mean ± Std).

Arch	Algo	Config	ID Val Acc	OOD Test Acc	Gap
DenseNet121	ERM	5 / 1	0.9334 ± 0.0009	**0.9126 ± 0.0061**	0.0208
	HYPO	5 / 1	0.9341 ± 0.0031	0.8964 ± 0.0160	0.0377
	ERM	3 / 5	0.9291 ± 0.0016	0.8883 ± 0.0059	0.0408
	HYPO	3 / 5	0.9390 ± 0.0022	0.8758 ± 0.0226	0.0632
ResNet50	ERM	5 / 1	0.9265 ± 0.0022	**0.8826 ± 0.0122**	0.0439
	HYPO	5 / 1	0.9245 ± 0.0049	0.8464 ± 0.0399	0.0781
	ERM	3 / 5	0.9221 ± 0.0010	0.8528 ± 0.0088	0.0693
	HYPO	3 / 5	0.9311 ± 0.0054	0.8094 ± 0.0204	0.1218

The application of MedAugmix yielded the best OOD generalization results observed in this study. To precisely quantify this improvement, Table 10 compares the performance of the best MedAugmix configuration (Severity 5, Width 1) against the next best augmentation strategy for each algorithm (Standard AugMix for ERM, and Basic MedMNIST-C for HYPO on DenseNet121).

As the results demonstrate, the proposed MedAugmix strategy is a significant step forward. The top-performing configuration overall was **ERM combined with MedAugmix (Severity 5, Width 1) on the DenseNet121 architecture**, which achieved an OOD accuracy of 0.9126. As highlighted in Table 10, this represents a substantial improvement of +0.0335 over the next best strategy, Standard AugMix. This MedAugmix configuration also produced the smallest generalization gap observed in the entire study (0.0208), indicating a strong alignment between ID and OOD performance.

A consistent trend was observed across all experiments: ERM consistently outperformed HYPO when both were trained with MedAugmix. On ResNet50, the performance difference was particularly large, with ERM achieving an OOD accuracy of 0.8826 versus HYPO's 0.8464. This pattern strongly suggests that the powerful, targeted regularization provided by the MedAugmix data pipeline particularly benefits the simpler ERM learning objective.

Table 10: Comparison of MedAugmix against previous best augmentation strategies (OOD Accuracy).

Arch	Algo	Previous Best	MedAugmix	Change (Δ)
DenseNet121	ERM	0.8791 (AugMix)	0.9126	+0.0335
	HYPO	0.8701 (MedMNIST-C)	0.8964	+0.0263
ResNet50	ERM	0.8527 (AugMix)	0.8826	+0.0299
	HYPO	0.7924 (AugMix)	0.8464	+0.0540

Finally, comparing the MedAugmix configurations reveals that the setup with higher severity and lower mixture diversity (Severity 5, Width 1) generally outperformed the one with moderate severity and higher mixture diversity (Severity 3, Width 5). This was true for ERM on both architectures and for HYPO on ResNet50, suggesting that applying stronger, more focused targeted corruptions was more beneficial than more diverse mixing of moderately corrupted images for this task. These results factually establish MedAugmix as the most effective method evaluated in this thesis for improving OOD generalization on the Camelyon17 benchmark.

4.2 Generalization of Findings on Secondary Benchmark (Fitzpatrick17k)

To assess the broader applicability of the findings from Camelyon17, the experimental strategies were replicated on the Fitzpatrick17k dermatology dataset. This secondary benchmark tests generalization across a different type of domain shift: from well-represented lighter skin tones to underrepresented darker skin tones. As detailed in the methodology, the primary metric for this multi-class, imbalanced dataset is **Balanced Accuracy**. This section analyzes the performance trends on this distinct medical imaging challenge. All models were initialized with ImageNet pre-trained weights, and results are averaged over three random seeds.

The results for all experimental conditions on Fitzpatrick17k are summarized and ranked by OOD performance in Table 11 for ResNet50 and Table 12 for DenseNet121.

The evaluation on Fitzpatrick17k largely confirms the primary trend observed on Camelyon17: a standard ERM algorithm, when enhanced with a strong augmentation pipeline like MedAugmix or Standard AugMix, is a highly effective strategy for OOD generalization. On this benchmark, it again surpassed the specialized HYPO algorithm across all comparable conditions.

As shown in Tables 11 and 12, both MedAugmix and Standard AugMix provided a dramatic boost to ERM's OOD performance. For ResNet50, MedAugmix yielded the top result, improving OOD balanced accuracy by over 10 percentage points from its baseline (0.5009 to **0.6060**). For DenseNet121, Standard AugMix was the top performer, improving from 0.5273 to **0.5952**. The fact that both advanced augmentation methods were similarly effective on this benchmark suggests that for

Table 11: Ranked Summary of OOD Performance for **ResNet50** on Fitzpatrick17k. Sorted by OOD Balanced Accuracy (`ood_at_best_id_val_bal_acc`).

Algo	Augmentation Strategy	ID Val Bal. Acc	OOD Bal. Acc
ERM	MedAugmix (Sev 3/Width 3)	0.6937 ± 0.0119	**0.6060 ± 0.0152**
ERM	Standard AugMix (Plain)	0.6963 ± 0.0064	0.5970 ± 0.0101
ERM	No Augmentation (Baseline)	0.6029 ± 0.0241	0.5009 ± 0.0169
ERM	Basic MedMNIST-C†	0.5786 ± 0.0304	0.4964 ± 0.0266
HYPO	MedAugmix (Sev 3/Width 5)	0.5361 ± 0.0447	0.4690 ± 0.0425
HYPO	Basic MedMNIST-C†	0.5255 ± 0.0839	0.4688 ± 0.0623
HYPO	No Augmentation (Baseline)	0.5355 ± 0.0494	0.4661 ± 0.0569
HYPO	Standard AugMix (Plain)	0.5246 ± 0.0259	0.4619 ± 0.0295

†Basic MedMNIST-C (`plain_medmnistc_random`) used corruptions from "dermamnist".

Table 12: Ranked Summary of OOD Performance for **DenseNet121** on Fitzpatrick17k. Sorted by OOD Balanced Accuracy (`ood_at_best_id_val_bal_acc`).

Algo	Augmentation Strategy	ID Val Bal. Acc	OOD Bal. Acc
ERM	Standard AugMix (Plain)	0.6784 ± 0.0021	**0.5952 ± 0.0029**
ERM	MedAugmix (Sev 3/Width 5)	0.6687 ± 0.0162	0.5925 ± 0.0170
ERM	No Augmentation (Baseline)	0.6036 ± 0.0211	0.5273 ± 0.0213
ERM	Basic MedMNIST-C†	0.5993 ± 0.0139	0.4973 ± 0.0162
HYPO	MedAugmix (Sev 3/Width 5)	0.5472 ± 0.0331	0.4944 ± 0.0426
HYPO	No Augmentation (Baseline)	0.5113 ± 0.0271	0.4592 ± 0.0068
HYPO	Standard AugMix (Plain)	0.4945 ± 0.0441	0.4442 ± 0.0259
HYPO	Basic MedMNIST-C†	0.4772 ± 0.0567	0.4376 ± 0.0438

†Basic MedMNIST-C (`plain_medmnistc_random`) used corruptions from "dermamnist".

OOD shifts related to demographic attributes like skin tone, the sheer diversity and intensity of transformations from any advanced mixing pipeline may be the dominant beneficial factor, with the specific type of corruption (targeted vs. generic) being less critical than on the artifact-driven Camelyon17 task.

In contrast, the HYPO algorithm did not benefit from any of the augmentation strategies on this dataset. Its performance remained low across all conditions, indicating that neither simple nor advanced augmentations were able to help its geometric learning objective succeed on this particular fairness-related OOD challenge.

Overall, the evaluation on Fitzpatrick17k provides strong secondary evidence for the core finding of this thesis: data-centric regularization through advanced augmentation is a powerful and generalizable approach for improving OOD performance in medical imaging, and can enable a simple ERM model to be more robust than a specialized OOD algorithm.

4.3 Overall Summary of Experimental Findings

The results from the two benchmarks reveal several clear and generalizable trends. The most significant findings are summarized below and in Table 13.

- The proposed **MedAugmix** strategy, when paired with ERM, delivered the highest OOD performance on both benchmarks. It yielded a **+6.0 percentage point (pp) gain in OOD accuracy** over the baseline on Camelyon17 and a **+10.5 pp gain in OOD balanced accuracy** on Fitzpatrick17k.
- An ERM model augmented with an advanced mixing strategy (either MedAugmix or Standard AugMix) consistently matched or surpassed the performance of the specialized HYPO algorithm across all tested architectures and datasets.
- The effectiveness of targeted vs. generic augmentation appears to be context-dependent. The targeted corruptions in MedAugmix showed a clear advantage on the artifact-driven Camelyon17 benchmark, while its performance was more competitive with the generic Standard AugMix on the demographically-driven Fitzpatrick17k benchmark.

Table 13: Summary of Best OOD Performance by Augmentation Strategy for ERM.

Augmentation Strategy	**Camelyon17** (OOD Accuracy)	**Fitzpatrick17k** (OOD Balanced Accuracy)
No Augmentation (Baseline)	0.853 (DenseNet121)	0.527 (DenseNet121)
Basic MedMNIST-C	0.866 (DenseNet121)	0.497 (DenseNet121)
Standard AugMix (Plain)	0.879 (DenseNet121)	0.597 (ResNet50)
MedAugmix (Proposed)	**0.913 (DenseNet121)**	**0.606 (ResNet50)**

In summary, the results demonstrate a clear hierarchy of effectiveness. While the baseline models struggled with significant generalization gaps, the application of advanced augmentation provided substantial improvements. The data diversification from Standard AugMix improved ERM's OOD accuracy by a notable 2.6–3.2 pp on Camelyon17. However, the proposed MedAugmix strategy provided the largest and most consistent boost, increasing ERM's OOD accuracy by a full 6.0 pp on Camelyon17 (from 0.853 to 0.913 on DenseNet121) and its OOD balanced accuracy by 10.5 pp on Fitzpatrick17k (from 0.501 to 0.606 on ResNet50).

In Section 5, we will analyze why these data-centric strategies, particularly MedAugmix, appear to outshine the purely algorithmic approach of HYPO in these medical imaging contexts and discuss the broader implications for designing clinically robust models.

5 Discussion

This section provides an in-depth discussion and interpretation of the experimental findings presented in Section 4. We will synthesize the results from both the Camelyon17 and Fitzpatrick17k benchmarks to analyze the core research questions, exploring why certain strategies were more effective than others and what the broader implications are for Out-of-Distribution (OOD) generalization in medical imaging.

To facilitate a comprehensive overview, Figure 10 and Figure 11 present a visual summary of the key OOD performance metrics for all evaluated strategies on the two primary architectures. These charts will serve as a central reference for the thematic discussion that follows.

Table 14 and Table 15 present all distinct experimental configurations for Camelyon17, ranked by their OOD generalization performance. These tables serve as a central reference for the discussion that follows.

The discussion will proceed thematically. First, we will interpret the performance of the baseline algorithms (ERM vs. HYPO) in the absence of advanced augmentation. Second, we will analyze the impact of the different data augmentation strategies, with a focus on the proposed MedAugmix method. Finally, we will synthesize these findings to address the core research questions and discuss the broader significance of this work.

5.1 Analysis of Training Dynamics and Model Selection

Beyond the final performance of the selected models, analyzing their behavior over the training duration provides valuable insights into issues like overfitting to source domains and the reliability of In-Distribution (ID) validation as a proxy for Out-of-Distribution (OOD) performance. We examine this by comparing two key values: the **practical OOD performance** (achieved by the model checkpoint with the

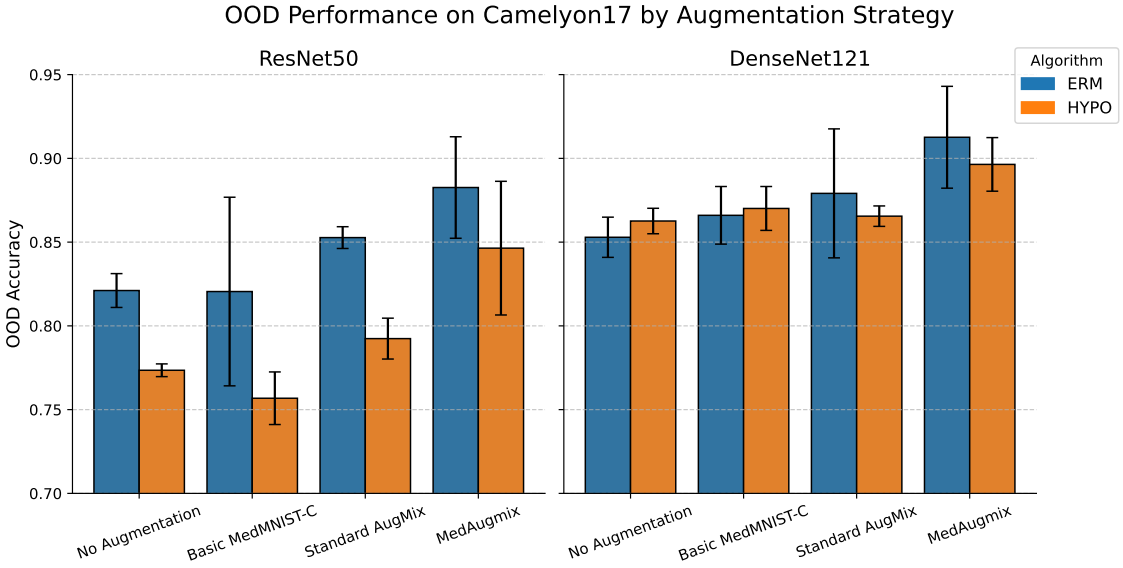


Figure 10: Summary of OOD Accuracy on the Camelyon17 benchmark across different augmentation strategies for ERM and HYPO algorithms. Error bars represent the standard deviation over 3 random seeds.

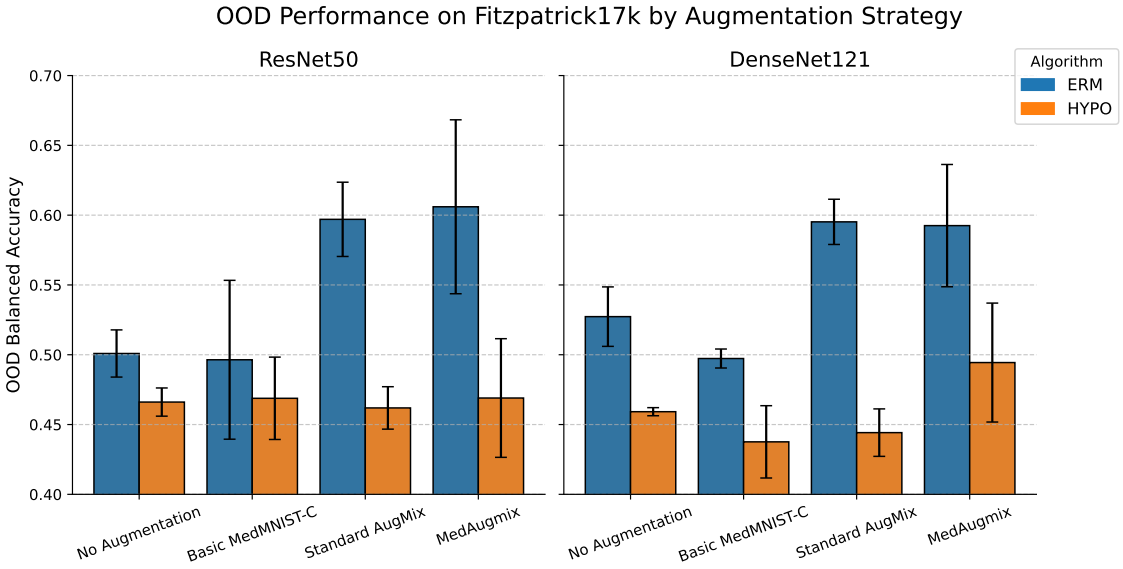


Figure 11: Summary of OOD Balanced Accuracy on the Fitzpatrick17k benchmark across different augmentation strategies for ERM and HYPO algorithms. Error bars represent the standard deviation over 3 random seeds.

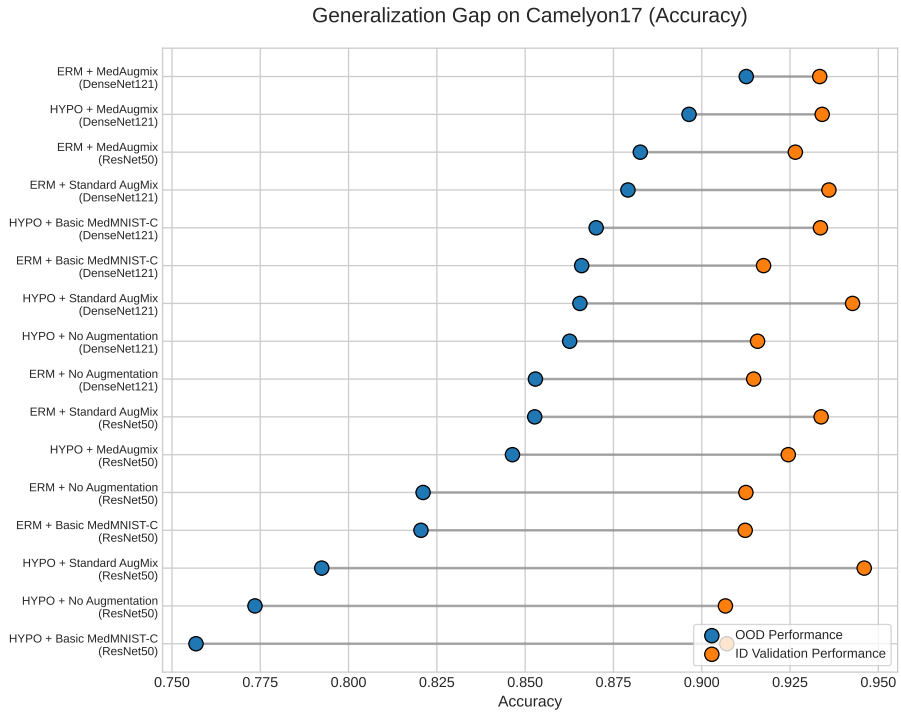


Figure 12: Generalization gap analysis for all strategies on the Camelyon17 benchmark. The plot is sorted by OOD performance (blue dots). The horizontal lines visually represent the generalization gap between ID validation performance (orange dots) and OOD performance.

Table 14: Ranked Summary of OOD Performance for **DenseNet121** Configurations on Camelyon17-wilds (Mean ± Std). Sorted by OOD Test Accuracy (`ood_at_best_id_val`).

Algorithm	Strategy	OOD Test Acc	ID Val Acc	Gap
ERM	MedAugmix (5/1)	**0.9126 ± 0.0061**	0.9334 ± 0.0009	0.0208
HYPO	MedAugmix (5/1)	0.8964 ± 0.0160	0.9341 ± 0.0031	0.0377
ERM	MedAugmix (3/5)	0.8883 ± 0.0059	0.9291 ± 0.0016	0.0408
ERM	AugMix	0.8791 ± 0.0076	0.9360 ± 0.0010	0.0569
HYPO	MedAugmix (3/5)	0.8758 ± 0.0226	0.9390 ± 0.0022	0.0632
HYPO	MedMNIST-C	0.8701 ± 0.0304	0.9336 ± 0.0015	0.0634
ERM	MedMNIST-C	0.8660 ± 0.0385	0.9175 ± 0.0042	0.0515
HYPO	AugMix	0.8655 ± 0.0131	0.9427 ± 0.0042	0.0771
HYPO	Baseline	0.8626 ± 0.0172	0.9158 ± 0.0132	0.0532
ERM	Baseline	0.8529 ± 0.0120	0.9147 ± 0.0015	0.0618

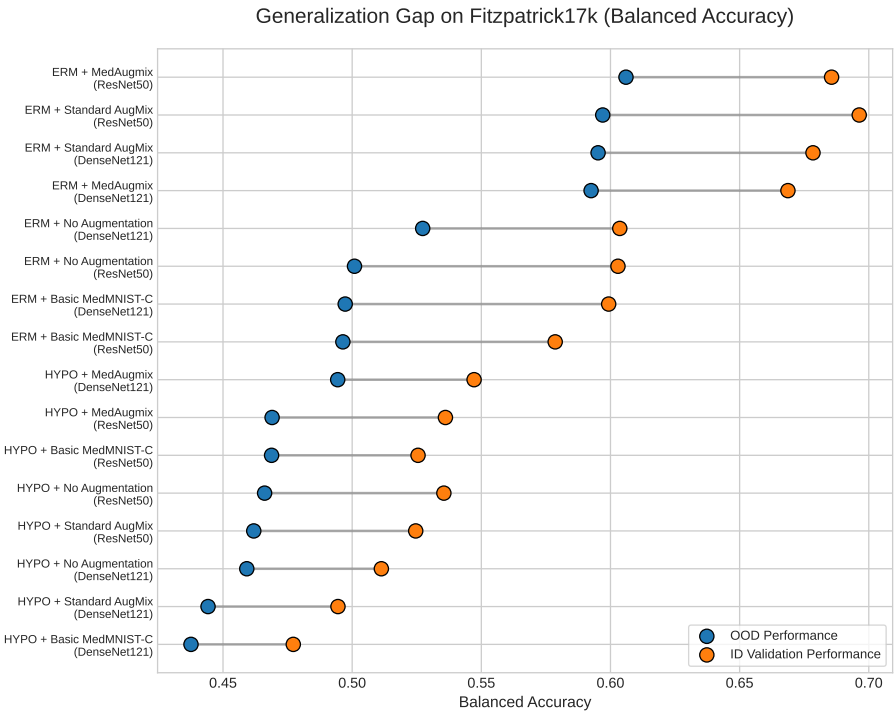


Figure 13: Generalization gap analysis for all strategies on the Fitzpatrick17k benchmark. The plot is sorted by OOD balanced accuracy (blue dots), illustrating the performance gap between ID and OOD results for each condition.

Table 15: Ranked Summary of OOD Performance for **ResNet50** Configurations on Camelyon17-wilds (Mean ± Std). Sorted by OOD Test Accuracy (`ood_at_best_id_val`).

Algorithm	**Strategy**	**OOD Test Acc**	**ID Val Acc**	**Gap**
ERM	MedAugmix (5/1)	**0.8826 ± 0.0122**	0.9265 ± 0.0022	0.0439
ERM	MedAugmix (3/5)	0.8528 ± 0.0088	0.9221 ± 0.0010	0.0693
ERM	AugMix	0.8527 ± 0.0038	0.9338 ± 0.0007	0.0811
HYPO	MedAugmix (5/1)	0.8464 ± 0.0399	0.9245 ± 0.0049	0.0781
ERM	Baseline	0.8211 ± 0.0101	0.9125 ± 0.0003	0.0914
ERM	MedMNIST-C	0.8205 ± 0.0065	0.9123 ± 0.0009	0.0918
HYPO	MedAugmix (3/5)	0.8094 ± 0.0204	0.9311 ± 0.0054	0.1218
HYPO	AugMix	0.7924 ± 0.0157	0.9460 ± 0.0022	0.1536
HYPO	Baseline	0.7735 ± 0.0563	0.9067 ± 0.0048	0.1333
HYPO	MedMNIST-C	0.7568 ± 0.0303	0.9071 ± 0.0086	0.1504

best ID validation score) and the **potential OOD performance** (the absolute best OOD score achieved at any epoch during training).

5.1.1 Training Dynamics on Camelyon17

Table 16 compares the epoch at which peak ID validation accuracy was reached versus the epoch of peak OOD test accuracy. It also shows the gap between the practical OOD performance and the potential best OOD performance.

Table 16: Training Dynamics on Camelyon17. "Potential OOD" refers to the best possible OOD accuracy achieved at any epoch. "Practical OOD" is the OOD accuracy of the model selected via best ID validation.

Architecture	Strategy	Best ID Epoch	Best OOD Epoch	Potential OOD Acc	Practical OOD Acc
ERM Algorithm					
DenseNet121	Baseline	29	5	0.8817	0.8529
ResNet50	Baseline	18	1	0.8641	0.8211
DenseNet121	MedAugmix (5/1)	7	3	0.9231	0.9126
ResNet50	MedAugmix (5/1)	6	2	0.9073	0.8826
HYPO Algorithm					
DenseNet121	Baseline	23	9	0.9008	0.8626
ResNet50	Baseline	6	11	0.8396	0.7735
DenseNet121	MedAugmix (5/1)	20	0	0.9222	0.8964
ResNet50	MedAugmix (5/1)	8	3	0.8712	0.8464

A clear trend emerges from the Camelyon17 results: for nearly all configurations, the peak OOD performance was achieved much earlier in training than the peak ID validation performance. For example, the baseline ERM model on DenseNet121 reached its best OOD accuracy at epoch 5 but continued to improve on the ID validation set until epoch 29. This divergence indicates that models quickly began to **overfit to the source domains**, and continuing to train them to maximize ID performance was actively detrimental to their ability to generalize to the unseen OOD domain.

Crucially, the application of **MedAugmix appears to mitigate this issue**. While the peaks still do not perfectly align, the gap between the epoch of best ID performance and best OOD performance is reduced. More importantly, the gap between the potential OOD accuracy and the practical OOD accuracy is much smaller for MedAugmix-trained models. For ERM on DenseNet121, the gap was reduced from 0.0288 (0.8817 - 0.8529) in the baseline to just 0.0105 (0.9231 - 0.9126) with MedAugmix. This suggests that the strong, targeted regularization provided by MedAugmix not only improves overall performance but also makes the ID validation metric a more reliable proxy for selecting a model with strong OOD capabilities.

5.1.2 Training Dynamics on Fitzpatrick17k

The training dynamics on Fitzpatrick17k, summarized in Table 17 for the key ResNet50 configurations, tell a similar story.

Table 17: Training Dynamics on Fitzpatrick17k. Comparison of the epoch of best ID performance vs. best OOD performance, and the gap between potential and practical OOD balanced accuracy.

Architecture	Strategy	Best ID Epoch	Best OOD Epoch	Potential OOD Bal	Practical OOD Bal
ERM Algorithm					
ResNet50	Baseline	10	26	0.5630	0.5009
DenseNet121	Baseline	11	28	0.5718	0.5273
ResNet50	MedAugmix (Best)	38	43	0.6210	0.6060
DenseNet121	MedAugmix (Best)	30	34	0.6126	0.5925
HYPO Algorithm					
ResNet50	Baseline	23	31	0.5410	0.4661
DenseNet121	Baseline	36	14	0.5245	0.4592
ResNet50	MedAugmix (Best)	21	17	0.5347	0.4690
DenseNet121	MedAugmix (Best)	39	29	0.5393	0.4944

Again, for all configurations, the potential OOD performance was higher than the practical performance achieved by selecting the model based on ID validation. This reinforces the finding that standard model selection can be suboptimal for OOD generalization.

A key observation from Table 17 is that, similar to Camelyon17, the potential OOD performance is consistently higher than the practical performance achieved by selecting the model based on ID validation. This again reinforces the finding that standard model selection protocols can be suboptimal for OOD generalization.

However, unlike on Camelyon17 where the best OOD performance typically occurred very early, the dynamics on Fitzpatrick17k are different. For the baseline ERM models, the peak OOD balanced accuracy was reached much later in training (e.g., epoch 26 for ResNet50) than the peak ID balanced accuracy (epoch 10). This suggests that for this specific demographic shift, longer training was required for the model to learn features that had some utility on the OOD (darker skin tone) domain, even as performance on the ID (lighter skin tone) domain may have saturated earlier.

Crucially, the application of **MedAugmix once again appears to improve the reliability of the model selection process**. For both ERM and HYPO on both architectures, the gap between the potential and practical OOD balanced accuracy was significantly reduced when MedAugmix was applied. For instance, with ERM on ResNet50, this gap shrank from a substantial 0.0621 (0.5630 - 0.5009) in the baseline condition to just 0.0150 (0.6210 - 0.6060) with MedAugmix.

This confirms the central trend seen on Camelyon17: **advanced, targeted augmentation not only boosts the overall performance ceiling but also makes the model's performance on the ID validation set a more reliable indicator of its performance on the OOD test set** . This alignment is a significant practical benefit, as it increases confidence in the model development and selection process.

5.2 Analysis of End-of-Training Performance and Overfitting

A crucial aspect of OOD generalization is understanding how models behave over the full course of training. By examining performance at the final training epoch (epoch 50), we can gain insights into whether models continue to generalize or begin to overfit to the source domains. This subsection analyzes the gap between In-Distribution (ID) and Out-of-Distribution (OOD) performance at the end of training for key experimental conditions.

5.2.1 Final Epoch Performance on Camelyon17

Table 18 presents the ID validation and OOD test accuracies at the final training epoch for the baseline and best-performing MedAugmix strategies on Camelyon17.

Table 18: End-of-Training Performance on Camelyon17 (Epoch 50). The gap represents the final difference between ID and OOD accuracy.

Architecture	Strategy	Final ID Val Acc	Final OOD Test Acc	Final Gap
ERM Algorithm				
DenseNet121	Baseline	0.9104	0.8620	0.0484
ResNet50	Baseline	0.9076	0.8300	0.0777
DenseNet121	MedAugmix (5/1)	0.9262	0.8852	0.0410
ResNet50	MedAugmix (5/1)	0.9198	0.8311	0.0887
HYPO Algorithm				
DenseNet121	Baseline	0.9091	0.8510	0.0580
ResNet50	Baseline	0.8969	0.8136	0.0833
DenseNet121	MedAugmix (5/1)	0.9272	0.8786	0.0486
ResNet50	MedAugmix (5/1)	0.9048	0.8042	0.1007

The results on Camelyon17 show that for most baseline configurations, the final OOD accuracy is similar to or slightly higher than the "practical" OOD accuracy selected earlier in training. However, when MedAugmix is applied, the models generally continue to improve on both ID and OOD metrics through to the final epoch, suggesting that the strong regularization provided by MedAugmix helps to prevent

catastrophic overfitting to the source domains. For example, the ERM DenseNet121 model with MedAugmix ends with a high OOD accuracy of 0.8852, maintaining a relatively small generalization gap. This indicates that MedAugmix not only boosts peak performance but also contributes to more stable training dynamics over a longer duration.

5.2.2 Final Epoch Performance on Fitzpatrick17k

A similar analysis for the Fitzpatrick17k benchmark reveals the profound regularizing effect of advanced augmentation. Table 19 shows the performance at epoch 50 for key ResNet50 configurations.

Table 19: End-of-Training Performance on Fitzpatrick17k (ResNet50). The gap represents the final difference between ID and OOD balanced accuracy.

Algorithm	**Strategy**	**Final ID Bal. Acc**	**Final OOD Bal. Acc**	**Final Gap**
ERM	Baseline	0.6404	0.5533	0.0871
ERM	MedAugmix (Best)	0.6930	0.6106	0.0824
HYPO	Baseline	0.5236	0.4928	0.0308
HYPO	MedAugmix (Best)	0.5550	0.4605	0.0945

For the baseline ERM model, the final OOD balanced accuracy (0.5533) is significantly higher than its practical OOD performance selected earlier (0.5009). This aligns with the epoch data from the previous section, which showed that peak OOD performance on Fitzpatrick17k tended to occur later in training. When MedAugmix is applied to ERM, it not only achieves a much higher final OOD balanced accuracy (0.6106) but also maintains a similar generalization gap, indicating that it scales performance on both ID and OOD domains effectively throughout training.

In contrast, the HYPO algorithm shows a different trend. While its final OOD performance with MedAugmix (0.4605) is similar to its practical OOD performance (0.4690), it does not reach the levels of the augmented ERM model.

Collectively, the analysis of end-of-training performance across both datasets reinforces a key finding: advanced, targeted augmentation like MedAugmix acts as a powerful regularizer. It not only elevates the ceiling of OOD performance but also helps to maintain or improve that performance through to the end of training, mitigating the kind of source-domain overfitting that can plague less-regularized models.

5.3 Interpretation of Key Findings

The experimental results reveal a clear and consistent narrative across both the Camelyon17 and Fitzpatrick17k benchmarks. This section interprets these findings,

focusing first on the baseline comparison between **ERM** and **HYPO**, and then on the transformative impact of the data augmentation strategies.

Baseline Performance: An Architecture-Dependent Advantage for HYPO The baseline experiments, conducted without advanced augmentation, highlight the inherent difficulty of the OOD tasks and reveal that the effectiveness of the specialized **HYPO** algorithm is highly contextual. On the DenseNet121 architecture for Camelyon17, **HYPO** demonstrated a modest but clear advantage over **ERM**. This aligns with the core premise of the **HYPO** paper (Lee et al., 2024), suggesting its geometric constraints on the feature space can indeed provide a benefit. However, this advantage was not universal; **ERM** surprisingly outperformed **HYPO** on the ResNet50 architecture for the same task and across both architectures on the more challenging Fitzpatrick17k benchmark. This indicates that while specialized algorithms can help, their success is not guaranteed and can depend on a favorable interaction with the backbone's learned feature space.

Impact of Augmentation: From Ineffective to Transformative The evaluation of data-centric strategies demonstrated a clear hierarchy of effectiveness. The application of **Basic MedMNIST-C** (a single random targeted corruption) proved to be an unreliable strategy, offering only marginal gains in the best-case scenario (DenseNet121 on Camelyon17) and proving detrimental in others, particularly for **HYPO** on ResNet50.

In contrast, the advanced mixing strategies had a profound impact, especially on **ERM**. **Standard AugMix**, with its generic transformations, provided a substantial boost to **ERM**'s OOD performance on both datasets, consistently making it superior to **HYPO**. This highlights the power of the AugMix processing pipeline itself as a strong regularizer.

The proposed **MedAugmix** strategy represents the pinnacle of performance in this study. By combining the structural diversity of the AugMix pipeline with the domain relevance of targeted MedMNIST-C corruptions, it delivered the highest OOD performance for **ERM** on both benchmarks. Its clear superiority over **Standard AugMix** on the artifact-driven Camelyon17 dataset underscores the benefit of using targeted corruptions that mimic realistic domain shifts. On the demographically-driven Fitzpatrick17k dataset, its performance was highly competitive with **Standard AugMix**, suggesting that for such shifts, the sheer diversity of any advanced mixing pipeline may be the dominant beneficial factor.

Training Dynamics: Augmentation as a Stabilizer Beyond elevating peak performance, the training dynamics reveal another crucial benefit of advanced augmentation. Analysis of the per-epoch metrics shows that for baseline models, the best OOD performance was often achieved very early in training, long before the model achieved its best performance on the in-distribution validation set. This divergence indicates that models were quickly starting to overfit to the source domains.

Crucially, **MedAugmix mitigated this issue significantly**. On both datasets, the gap between the model's *potential* best OOD performance and the *practical* performance achieved by selecting the model via ID validation was much smaller for **MedAugmix**-trained models. Furthermore, models trained with **MedAugmix** maintained strong OOD performance through to the final epoch. This indicates that **MedAugmix** not only boosts performance but also acts as a powerful regularizer that stabilizes training and makes the standard model selection process more reliable.

5.4 Overall Comparative Analysis and Synthesis

The preceding discussion has interpreted the performance of individual algorithms and augmentation strategies. This section now synthesizes these findings into a holistic view, drawing overarching conclusions from the ranked summary tables (Table 14 and Table 15) and addressing the core research questions that motivated this thesis.

Synthesis of Key Performance Trends

Across two distinct medical OOD benchmarks, a clear hierarchy of strategies emerged. While specialized algorithms like HYPO offered modest baseline improvements in specific contexts, the most significant and reliable gains in OOD generalization were consistently driven by advanced data augmentation. The proposed MedAugmix strategy, when paired with ERM, systematically produced the best or near-best results. Standard AugMix also proved to be a powerful regularizer for ERM, while the simple application of Basic MedMNIST-C corruptions had a limited and unreliable impact.

A central finding is the dynamic between the learning algorithm and the augmentation strategy. In non-augmented or simply-augmented settings, the relative performance of ERM and HYPO was inconsistent and architecture-dependent. However, when a strong and diverse augmentation pipeline (Standard AugMix or MedAugmix) was introduced, ERM consistently matched or surpassed HYPO's OOD performance. This suggests that the regularization effect of a powerful data-centric approach can be the dominant factor in achieving OOD robustness. Architecturally, DenseNet121 also generally proved to be a more robust backbone for this task than ResNet50 across most experimental conditions.

Addressing the Research Questions

These synthesized findings provide clear answers to the initial research questions of this study:

RQ1: How does the performance of HYPO compare against ERM? Our evaluation shows that the performance of HYPO relative to ERM is highly contextual. In baseline (non-augmented) conditions, HYPO provided a modest OOD

advantage on DenseNet121 for Camelyon17 but was outperformed by ERM on ResNet50 and on the Fitzpatrick17k benchmark. When strong augmentation strategies like Standard AugMix or MedAugmix were applied, the simpler ERM algorithm consistently achieved superior OOD performance on both datasets. Therefore, HYPO is not a universally better solution than ERM; its benefit is most apparent in the absence of strong data-centric regularization and can be superseded by a well-augmented ERM baseline.

RQ2: What is the comparative impact of different data augmentation strategies? The data-centric strategies exhibited a clear performance hierarchy. MedAugmix was the most effective strategy overall, delivering the highest OOD scores in the majority of top-performing configurations. Standard AugMix also provided substantial benefits, particularly for ERM, proving more effective than the "Basic MedMNIST-C" approach. The Basic MedMNIST-C strategy (a single random corruption) was largely ineffective, yielding only marginal gains in the best case and proving detrimental in others. This demonstrates that not just the use of augmentation, but its structural sophistication (e.g., mixing multiple operations as in AugMix and MedAugmix) is critical for significant performance gains.

RQ3: Does the effectiveness of these strategies generalize across different domain shifts? Yes, the primary findings demonstrated strong generalizability across the two different OOD challenges. The core conclusion—that a simple ERM algorithm augmented with an advanced mixing strategy (Standard AugMix or MedAugmix) is a top-performing approach—held true for both the inter-hospital variations in Camelyon17 and the skin tone biases in Fitzpatrick17k. However, a nuance was observed: the targeted nature of MedAugmix showed a clear advantage over the generic Standard AugMix on the artifact-driven Camelyon17 dataset, whereas their performance was more competitive on the demographically-driven Fitzpatrick17k dataset.

RQ4: Which combination yields the most robust OOD performance? Across all evaluated configurations, the combination that yielded the most robust and effective OOD performance was Empirical Risk Minimization (ERM) combined with the proposed MedAugmix strategy. Specifically for Camelyon17, the top performer was ERM on a DenseNet121 backbone with MedAugmix (Severity 5, Width 1). For Fitzpatrick17k, it was ERM on a ResNet50 backbone with MedAugmix (Severity 3, Width 3). This consistently highlights the power of pairing a simple learning objective with a strong, diverse, and domain-aware data augmentation pipeline.

5.5 Significance of Findings and Relation to Prior Work

The comprehensive evaluation conducted in this thesis offers several significant findings that contribute to the broader understanding of Out-of-Distribution (OOD) generalization in medical imaging. The primary significance lies not in proposing a fundamentally new learning algorithm, but in the systematic benchmarking of existing ones and the introduction of a novel, data-centric strategy, MedAugmix. This evaluation provides practical insights into the relative merits and interplay of algorithmic versus data-centric approaches to improving robustness.

A key finding of this work is the demonstrated power of a well-augmented **Empirical Risk Minimization (ERM)** baseline. Our results consistently show that ERM, when paired with an advanced, targeted augmentation strategy like MedAugmix, can achieve state-of-the-art OOD performance that surpasses a specialized OOD algorithm. This contributes a crucial perspective to the field: for certain challenging OOD problems, the most effective path to robustness may lie in sophisticated data augmentation rather than solely in the development of more complex algorithmic loss functions. This has significant practical implications, suggesting that research and engineering efforts focused on data-centric solutions can yield substantial returns.

The findings also provide valuable context for the specific methods evaluated:

- **Relation to HYPO (Lee et al., 2024):** The HYPO algorithm was proposed as a method to learn domain-invariant representations by enforcing geometric constraints on a hypersphere. Our evaluation on two large-scale medical benchmarks provides new evidence on its applicability. While HYPO did show a modest advantage over baseline ERM in some contexts (e.g., with DenseNet121 on Camelyon17), its performance was not uniformly superior and was consistently surpassed by the augmented ERM models. This does not refute the principles of HYPO, but rather contextualizes its utility, suggesting that its benefits may be most pronounced in low-data or low-augmentation regimes and can be overshadowed by powerful data-centric regularization.

- **Relation to AugMix (Hendrycks et al., 2020):** The original AugMix framework demonstrated the power of combining diverse augmentations with a consistency loss. Our evaluation contributes in two ways. First, by showing that Standard (Plain) AugMix without the JSD loss is still a potent regularizer for ERM, we highlight the inherent strength of its data mixing pipeline. Second, and more importantly, the success of our proposed **MedAugmix** demonstrates that the AugMix framework is highly adaptable. By replacing its generic operations with targeted MedMNIST-C corruptions, we show that its effectiveness can be significantly enhanced for specific domains like medical imaging.

- **Relation to MedMNIST-C (Di Salvo et al., 2024):** The MedMNIST-C paper introduced its corruptions as a vital tool for benchmarking model robustness and suggested their potential use for data augmentation. This thesis

acts directly on that suggestion and provides a key validation of their utility. Our evaluation shows that while a simple application ("Basic MedMNIST-C") has limited effect, a structured application within our MedAugmix pipeline is highly effective. This work therefore serves as a successful case study, demonstrating a concrete and powerful method for leveraging MedMNIST-C corruptions not just for testing models, but for actively *training* more robust and generalizable ones.

In summary, by conducting a rigorous evaluation across two distinct medical OOD challenges, this thesis provides a nuanced perspective on building robust AI. It confirms the value of specialized algorithms like HYPO, but ultimately champions a data-centric viewpoint, demonstrating that a novel, targeted augmentation strategy like MedAugmix enables a simple ERM baseline to achieve excellent, and in this study superior, OOD generalization.

6 Conclusion and Future Work

This final section consolidates the research presented in this thesis. It begins with a recapitulation of the study's objectives and approach, followed by a summary of the principal findings and contributions. The strengths and limitations of the work are then discussed, leading to suggestions for promising avenues for future research.

6.1 Recapitulation of Thesis Work

The clinical translation of deep learning models is critically limited by their inability to generalize to Out-of-Distribution (OOD) data. This thesis addressed this widespread challenge in medical imaging by designing and executing a systematic benchmark of OOD generalization strategies. The evaluation was conducted across two distinct medical domains to ensure the broad relevance of the findings: inter-hospital variability in histopathology, using the WILDS Camelyon17 dataset, and demographic (skin tone) bias in dermatology, using the Fitzpatrick17k dataset.

The primary objectives were to: (1) evaluate a standard Empirical Risk Minimization (ERM) baseline against a specialized representation learning algorithm, HYPO; (2) assess the impact of various data augmentation strategies, from basic targeted corruptions to standard AugMix; and (3) propose and evaluate a novel hybrid strategy, MedAugmix, which adapts the AugMix pipeline to use targeted, clinically-relevant image corruptions sourced from MedMNIST-C.

6.2 Summary of Principal Findings and Contributions

The comprehensive experimental evaluation yielded several key findings:

- The proposed **MedAugmix** strategy, when combined with ERM, consistently produced the top-performing models across both the Camelyon17 and Fitzpatrick17k benchmarks, demonstrating its effectiveness across different medical domains and OOD challenges.
- A standard ERM model, when enhanced with an advanced augmentation pipeline like MedAugmix or Standard AugMix, consistently matched or surpassed the OOD performance of the specialized HYPO algorithm.
- The benefit of using targeted MedMNIST-C corruptions (in MedAugmix) over generic corruptions (in Standard AugMix) was most pronounced on the artifact-driven Camelyon17 dataset, while both were highly competitive on the demographically-driven Fitzpatrick17k dataset.
- The HYPO algorithm's effectiveness relative to ERM was found to be context-dependent and was generally overshadowed by the significant performance gains provided by strong data-centric regularization.

The principal contributions of this thesis are:

1. A **systematic evaluation** of ERM versus HYPO across two distinct medical OOD benchmarks, providing insights into their relative strengths and weaknesses under various augmentation conditions.
2. The **proposal, implementation, and successful evaluation of MedAugmix**, a novel data augmentation strategy that tailors the AugMix framework with targeted medical corruptions, demonstrating state-of-the-art performance within this study.
3. Empirical evidence that a strong, data-centric approach can be a more effective path to OOD robustness than relying solely on the specialized learning algorithms evaluated.
4. Practical insights into the effectiveness of different augmentation approaches for improving OOD generalization in medical imaging.

6.3 Strengths of the Study

This research possesses several strengths that bolster the validity of its findings:

- **Dual-Dataset Evaluation:** By benchmarking strategies on both Camelyon17 and Fitzpatrick17k, the study tests its conclusions against two fundamentally different types of domain shifts, enhancing the generalizability of its insights.
- **Systematic Comparison:** The work provides a rigorous comparison of algorithmic versus data-centric approaches across two architectures and multiple well-defined augmentation strategies.

- **Novel Augmentation Strategy:** A key strength is the proposal and evaluation of MedAugmix, a thoughtfully designed approach that synergistically combines existing concepts.
- **Methodological Rigor:** Experiments were conducted with multiple random seeds (three for primary results) and evaluated using a comprehensive suite of metrics appropriate for each dataset.
- **Practical Implications:** The findings offer clear, practical guidance for developing more robust medical AI models, emphasizing data-centric solutions.

6.4 Limitations of the Study

While this thesis provides valuable insights, it is important to acknowledge its limitations:

- **Scope of Evaluation:** The study was limited to two datasets, two backbone architectures, and two primary learning algorithms. The findings may not generalize to all medical imaging tasks, modalities, or OOD methods.
- **MedAugmix Implementation Details:** The evaluated MedAugmix strategy did not incorporate the JSD consistency loss from the original AugMix paper, and it used a fixed operational severity within each experimental run. Exploring these variations could yield different results.
- **ERM Baseline Configuration:** The ERM baseline applied CrossEntropyLoss to the 128-dimensional output of a projection head. While kept consistent for fair comparison, this is an unconventional setup, and its performance might differ from an ERM model with a standard final classification layer.
- **Hyperparameter Tuning:** Exhaustive hyperparameter optimization for every experimental condition was beyond the scope of this work due to computational constraints.

6.5 Future Work

The findings and limitations of this thesis open several promising avenues for future research:

- **Broader Evaluation of MedAugmix:** Systematically evaluate MedAugmix on a wider variety of medical imaging datasets (including 3D data), clinical tasks (e.g., segmentation), and against a broader array of state-of-the-art OOD generalization algorithms.
- **Enhancements to MedAugmix:**

  - Investigate the impact of incorporating the JSD consistency loss into the MedAugmix training framework.
  - Explore the use of different MedMNIST-C corruption sets (e.g., "pathmnist" or a mix of sets) as source operations.
  - Experiment with adaptive policies for selecting operations and severities within the MedAugmix chains.

- **Feature Space Analysis:** Conduct in-depth analyses of the learned feature representations (e.g., using t-SNE/UMAP) to better understand how MedAugmix and HYPO shape the embedding space and contribute to OOD robustness.
- **Fairness and Demographic Shifts:** Further investigate the use of MedAugmix and other augmentation strategies specifically for mitigating demographic bias in medical AI, building on the findings from the Fitzpatrick17k benchmark.
- **Theoretical Understanding:** Explore the theoretical underpinnings of why a well-augmented ERM can achieve strong OOD generalization, potentially rivaling specialized algorithms.

Addressing these future research directions could further advance the development of robust, reliable, and equitable deep learning models for widespread clinical application.

Bibliography

Neda Alipour, Ted Burke, and Jane Courtney. Skin type diversity in skin lesion datasets: A review. *Current Dermatology Reports*, 13:1–13, aug 2024. doi: 10.1007/s13671-024-00440-0.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2019.

Ujjwal Baid, Satyam Ghodasara, Michel Bilello, et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification, 2021.

Peter Bandi, Oscar Geessink, Quirine Manson, Michael Van Veldhuizen, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, et al. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2019. doi: 10.1109/TMI.2018.2867350.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017. doi: 10.1001/jama.2017.14585.

Daniel C. Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3675, 2020. doi: 10.1038/s41467-020-17478-w.

Alex John DeGrave, Joseph D. Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts and learn spurious correlations. *Nature Machine Intelligence*, 3(11):1024–1032, 2021. doi: 10.1038/s42256-021-00338-7.

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 (CVPR)*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019.

Francesco Di Salvo, Sebastian Doerrich, and Christian Ledig. MedMNIST-C: Comprehensive benchmark and improved classifier robustness by simulating realistic image corruptions, 2024.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yang Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *NPJ Digital Medicine*, 4(1):5, 2021.

Samuel G. Finlayson, Adarsh Subbaswamy, Karandeep Singh, John D. Miller, Judy Wawira Gichoya, and Pranav Rajpurkar. The clinician and dataset shift in artificial intelligence. *The New England Journal of Medicine*, 385(3):283–286, 2021. doi: 10.1056/NEJMc2103202.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1126–1135, 2017.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46 (4):1–37, 2014. doi: 10.1145/2523813.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. volume 17, pages 1–35, 2016.

Judy Wawira Gichoya, Imon Banerjee, Anant R. Bhimireddy, John L. Burns, Leo Anthony Celi, Laila Chen, et al. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414, 2022.

Ben Glocker, Ender Konukoglu, Sotirios A. Tsaftaris, et al. Machine learning with medical imaging: The road to clinical translation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019 (Tutorial)*, 2019.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. volume 13, pages 723–773, 2012.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1820–1828, 2021.

Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22): 2402–2410, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations (ICLR)*, 2020.

Frazer M. Howard, Jack Dolezal, Sarah Kochanny, Michael Jaber, Ausaf Sayeed, Thomas T. Lee, et al. The impact of site-specific digital histology slide image domain shifts on deep learning model accuracy and bias. *Nature Communications*, 12(1):3398, 2021. doi: 10.1038/s41467-021-23722-4.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, S. M. Xie, Aditi Raghunathan, David Lovell, Jiaxin Ma, Sara Beery, Jure Leskovec, Emma Pierson, et al. WILDS: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 2021.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, pages 1097–1105, 2012.

Paras Lakhani and Baskaran Sundaram. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2):574–582, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 6402–6413, 2017.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Ik Jin Lee, Sungwoong Kim, KyuGwn Lee, and MyungJoo Kang. HYPO: HYPERSPHERICAL out-of-distribution generalization, 2024.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier: A challenge for fine-grained texture recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, pages 1260–1268, 2017.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. doi: 10.1016/j.media.2017.07.005.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 212–220, 2017.

Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33:170–175, 2016. doi: 10.1016/j.media.2016.06.037.

Yifei Ming, Haoyue Bai, and Yixuan Li. CIDER: Contrastive pre-training with hyperspherical embeddings for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16374–16384, 2023.

Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5715–5725, 2017. doi: 10.1109/ICCV.2017.611.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, 2010.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Bradley J. Erickson. Ai in health and medicine. *Nature Medicine*, 28(1):31–38, 2022. doi: 10.1038/s41591-021-01614-0.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer International Publishing, 2015.

Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019. doi: 10.1186/s40537-019-0197-0.

Kerstin Stacke, Gabriele Eilertsen, Ewert Källhammer, Lasse Mårtensson, and Claes Lundström. Measuring and reducing the effect of scanner variability in medical imaging studies using combat. *Medical Image Analysis*, 66:101794, 2020. doi: 10.1016/j.media.2020.101794.

Luke Taylor and Geoff Nitschke. Improving deep learning using generic data augmentation. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1542–1547, 2018. doi: 10.1109/SSCI.2018.8628742.

David Tellez, Maschenka Balkenhol, Irene Otte-Holler, Jeroen van der Laak, and Francesco Ciompi. Whole-slide mitosis detection in h&e breast histology using cnns: A deep transfer learning approach. *Medical Image Analysis*, 49:32–44, 2018. doi: 10.1016/j.media.2018.07.001.

Hamid R. Tizhoosh and Liron Pantanowitz. Artificial intelligence and digital pathology: An overview. *Human Pathology*, 79:43–53, 2018. doi: 10.1016/j.humpath.2018.06.017.

U.S. Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence/machine learning (ai/ml)-enabled device software functions. Technical report, U.S. Department of Health and Human Services, oct 2023. FDA Guidance Document.

Vladimir N. Vapnik. Principles of risk minimization for learning theory. In John E. Moody, Stephen J. Hanson, and Richard P. Lippmann, editors, *Advances in Neural Information Processing Systems 4 (NIPS 1991)*, pages 831–838, 1992.

Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant CNNs for digital pathology. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, volume 11071 of *Lecture Notes in Computer Science*, pages 210–218. Springer, Cham, 2018. doi: 10.1007/978-3-030-00934-2_24.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5265–5274, 2018.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Dacheng Tao. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):7925–7947, 2023. doi: 10.1109/TKDE.2022.3160602.

Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. doi: 10.1016/j.neucom.2018.05.083.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2097–2106, 2017.

Jiancheng Yang, Rui Pang, Zheyuan Zhang, Zixuan Lin, Michael P. H. Le, Liqun Li, Bingcheng Liu, Bokai Wang, Po-Choi Chen, et al. MedMNIST v2 – a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023. doi: 10.1038/s41597-022-01721-8.

John R. Zech, Michael A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric K. Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, 15(11):e1002683, 2018. doi: 10.1371/journal.pmed.1002683.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4649–4668, 2023. doi: 10.1109/TPAMI.2022.3195549.

Declaration of Authorship

Ich erkläre hiermit gemäß §9 Abs. 12 APO, dass ich die vorstehende Abschlussarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Des Weiteren erkläre ich, dass die digitale Fassung der gedruckten Ausfertigung der Abschlussarbeit ausnahmslos in Inhalt und Wortlaut entspricht und zur Kenntnis genommen wurde, dass diese digitale Fassung einer durch Software unterstützten, anonymisierten Prüfung auf Plagiate unterzogen werden kann.

Place, Date

Signature