Establishing a Hierarchy of Training Strategies for Data-Scarce Medical Imaging

Arturas Aleksandraus[1,2] ARTURAS.ALEKSANDRAUS@CONTEXTVISION.SE
Anneke Meyer[2] ANNEKE.MEYER@CONTEXTVISION.SE
Milda Poceviciute[2] MILDA.POCEVICIUTE@CONTEXTVISION.SE
Daniel Jönsson[1] DANIEL.JONSSON@LIU.SE
Gabriel Eilertsen[1] GABRIEL.EILERTSEN@LIU.SE
Björn Norell[2] BJORN.NORELL@CONTEXTVISION.SE
Jonas Unger[1] JONAS.UNGER@LIU.SE
[1] *Department of Science and Technology, Linköping University, Norrköping, Sweden*
[2] *ContextVision AB, Linköping, Sweden*

Editors: Under Review for MIDL 2026

Abstract

Deep learning in medical imaging faces fundamental constraints from data scarcity, including the inherent lack of data for rare diseases or events, and class imbalance. These challenges, compounded by privacy regulations and high expert annotation costs, make acquiring large-scale annotated datasets difficult. Although numerous training strategies aim to mitigate these issues, their comparative effectiveness in generalizing across diverse datasets remains poorly understood, providing practitioners with little guidance on prioritization.

In this paper, we investigate the effect of five training strategies: Data Augmentation, Hard Example Mining, Hard Adversarial Mining, Balancing and Reweighting, and Robustness-Oriented Training to establish a structured strategy selection for robust generalization under data scarcity.

We implement representative techniques for these families of methods and conduct 3,000+ experiments on four datasets (CIFAR-10, RetinaMNIST, OrganCMNIST, PathMNIST) with controlled scarcity. To enable fair comparison across datasets and scarcity conditions, we introduce a Normalized Potential Score (NPS) that measures strategy effectiveness relative to the achievable improvement range, where 0.0 indicates baseline performance, 1.0 represents best achieved performance, and negative values indicate performance below baseline.

Our findings establish a hierarchy of generalization capabilities: Data Augmentation yields the largest average improvements (0.30–0.60 NPS), but still leaves a lot of performance to gain from other strategies. Combining it with Hard Adversarial Mining provides further gains (0.02–0.37 NPS). Balancing strategies enhances rare-class performance (0.08–0.10 NPS) but reduces frequent-class accuracy. We observe that Exponential Moving Average (EMA) can substantially improve training ($\pm$0.30 NPS) in some domains and has low overhead, making it a useful addition to any training pipeline. These results provide a hierarchy of strategies to consider for improving generalization in medical imaging and other data-constrained scenarios.

Keywords: Data Scarcity, Class Imbalance, Training Strategies, Data Augmentation, Hard Adversarial Mining, Medical Image Classification, Empirical Evaluation

1. Introduction

Data scarcity and class imbalance constrain deep learning performance in safety-critical domains such as medical imaging, where reliable generalization is essential. One of the main causes of data scarcity is the high cost of medical data annotation, which often requires expert readers and is further limited by strict privacy regulations (Shorten and Khoshgoftaar, 2019; Cossio, 2023). Furthermore, clinically significant cases are also rare compared to healthy examples, leading to severe class imbalance (Chawla et al., 2002; Lin et al., 2017). Together, limited annotated data and class imbalance encourage overfitting: models trained repeatedly on small datasets tend to memorize specific cases rather than learn robust decision boundaries, which degrades performance on unseen data (Zhang et al., 2021).

Numerous training strategies have been developed to mitigate data scarcity and class imbalance in medical imaging, yet their interactions remain difficult to quantify. Strategies are typically evaluated in isolation: data augmentation expands the effective dataset size(Shorten and Khoshgoftaar, 2019; Mikołajczyk and Grochowski, 2018; Cossio, 2023; Zhong et al., 2025), hard example mining focuses learning on difficult cases(Shrivastava et al., 2016; Schmidt-Mengin et al., 2022; Kumar and Srivastava, 2018), balancing methods address distributional skew (Lin et al., 2017; Chawla et al., 2002), and robustness techniques stabilize training (Hendrycks et al., 2019; Tarvainen Antti, 2017). This emphasis on optimizing individual components is exemplified by works that independently study augmentation catalogues and generative augmentation, imbalance-aware losses, or hard mining schemes in medical imaging, rather than their joint behavior under data scarcity and class imbalance (Cossio, 2023; Sizikova et al., 2024; Vyver et al., 2025; Kumar and Srivastava, 2018; Schmidt-Mengin et al., 2022). As a result, practitioners lack quantitative guidance on how different strategies relate and which combinations are most effective when labeled data are limited, and must often resort to extensive trial-and-error over many possible configurations, which wastes computational resources and may be prohibitively expensive within the constraints of typical medical imaging studies.

We study interactions of five strategy families: Data Augmentation, Hard Example Mining, Hard Adversarial Mining, Balancing and Reweighting, and Robustness-Oriented Training, mapped to training pipeline stages (Fig. 1). Using representative techniques, we evaluate their individual and combined effects under data scarcity and imbalance across datasets with the Normalized Potential Score (NPS), which measures relative improvement between the best observed performance and the baseline with no techniques enabled. **Contributions**: (i) large-scale empirical study across four datasets, and (ii) practical guidance for effective training configurations without exhaustive trial-and-error.

2. Related works: Families of Training Strategies

To enable systematic comparison of training strategies under data scarcity, we organize them into five families based on algorithmic principles: **Data Augmentation** (expanding diversity), **Hard Example Mining** (prioritizing challenging samples), **Hard Adversarial Mining** (selection of confident misclassifications), **Balancing and Reweighting** (addressing imbalance), and **Robustness-Oriented Training** (model stability). These families operate at distinct levels of the training pipeline: data space, batch composition, sample selection, gradient computation, and weight updates (Fig. 1).

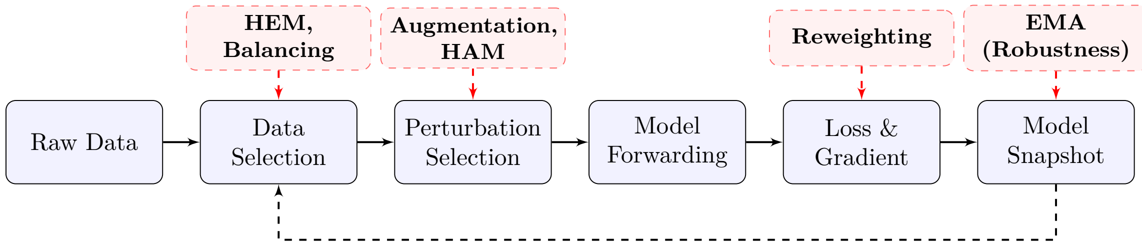

Figure 1: Visual taxonomy of training strategy families mapped to the deep learning training pipeline. The process flows from Raw Data to Data Selection (incorporating Hard Example Mining (HEM) and Balancing), then to Perturbation Selection (applying Augmentation and Hard Adversarial Mining (HAM)), followed by Model Forwarding, Loss & Gradient computation (where Reweighting applies), and finally Model Snapshot updates (using EMA for robustness). Our experiments quantify these interactions to guide effective combinations under data scarcity.

For each family we implement a representative module that captures the intended behavior, enabling a systematic comparison of family-level effectiveness. The following subsections describe these families in more detail.

2.1. Data Augmentation

Data augmentation operates at the *data level*, expanding the training distribution through label-preserving transformations. Deterministic or stochastic alterations increase diversity and improve generalization (Shorten and Khoshgoftaar, 2019; Zhong et al., 2025; Cossio, 2023).

Techniques range from geometric transformations (rotations, flips) to intensity adjustments (Shorten and Khoshgoftaar, 2019; Cossio, 2023; Mikołajczyk and Grochowski, 2018). Mixing methods such as Mixup (Zhang et al., 2018) and interpolation-based strategies have been shown to improve generalization and reduce overfitting, while consistency-focused approaches like AugMix improve robustness to corruptions (Hendrycks et al., 2019). Automated policy search (AutoAugment) and lighter-weight alternatives (RandAugment, TrivialAugment) provide practical recipes for tuning augmentation at scale (Cubuk et al., 2018, 2019; mul, 2021).

Empirical prior work highlights two practical points: (1) automated and mixing-based augmentations often yield large gains on natural-image benchmarks, but their benefit can be dataset- and label-dependent (Shorten and Khoshgoftaar, 2019; Cubuk et al., 2019); (2) in medical imaging, geometric transforms tend to be most reliable while aggressive intensity or mixing operations can distort relevant features unless carefully constrained (Mikołajczyk and Grochowski, 2018; Cossio, 2023). Recent studies also explore combining discriminative augmentations with generative synthesis (GANs, diffusion models) to increase rare-class support, although trade-offs between diversity and realism remain an open research question (Li et al., 2023; Vyver et al., 2025).

2.2. Hard Example Mining

Hard Example Mining (HEM) operates at the *batch sampling level*, prioritizing challenging data points. HEM identifies samples with high loss or low confidence and increases their representation in batches (Shrivastava et al., 2016).

Implementations include Online Hard Example Mining (OHEM) (Shrivastava et al., 2016) and curriculum learning (Liang et al., 2021). In medical applications, HEM has aided lesion segmentation (Schmidt-Mengin et al., 2022) and imbalance handling (Tang et al., 2023).

Noise and mislabeled data are inherently challenging; thus, focusing excessively on hard examples can lead to poor generalization.

2.3. Hard Adversarial Mining

Hard Adversarial Mining (HAM) targets confident misclassifications at decision boundaries. Techniques include gradient-based methods, adversarial networks, and frequency-domain approaches. Gradient-based approaches use adversarial gradients to maximize loss, while adversarial networks employ a minimax game to synthesize challenging variants. Frequency-domain methods manipulate spectral amplitude to alter domain statistics while preserving semantic content(Zhong et al., 2025).

Computational overhead is the main drawback of adversarial methods. Furthermore, aggressive adversarial perturbations can lead to data that deviates from the true distribution, reducing relevance.

2.4. Balancing and Reweighting

To counteract the majority class bias, common techniques include loss-based reweighting, such as Focal Loss (Lin et al., 2017), which down-weights easy examples, and sampling-based approaches like SMOTE (Chawla et al., 2002) that oversample minority instances. These interventions are particularly valuable for long-tailed distributions where standard training ignores rare classes. However, these methods often improve the performance of the minority class at the expense of the accuracy of the majority class. Aggressive reweighting can also lead to overfitting on rare examples, making the optimal balance dependent on specific application requirements for sensitivity versus specificity.

2.5. Robustness-Oriented Training

Robustness-Oriented Training operates at the *weight update level* to stabilize model parameters and reduce variance. By enforcing consistency in predictions or smoothing weight updates, these methods mitigate the risk of overfitting to noise in small datasets (Tarvainen Antti, 2017; Hendrycks et al., 2019).

Key techniques include Exponential Moving Average (EMA) (Tarvainen Antti, 2017; Morales-Brotons et al., 2024), which maintains a slowly updating shadow model to provide stable evaluation results with reduced fluctuations, and consistency regularization methods like AugMix (Hendrycks et al., 2019) that penalize divergence between predictions on augmented views. These approaches are particularly effective under distribution shifts.

3. Method

We evaluate training strategy families using representative technique that implement the core mechanisms of each family. Running an exhaustive experimental grid that spans 1344 unique runs across the three dimensions:

- **Domain Context:** Four datasets (CIFAR-10, RetinaMNIST, OrganCMNIST, PathMNIST).
- **Data Constraints:** Three scarcity profiles per dataset (Full, Long-tailed, Extreme Long-tailed).
- **Strategy Combinations:** A combinatorial exploration of strategy technique:
  - **Augmentation:** 4 tiers (None, Classical, RandAugment, TrivialAugment).
  - **Hard Example Mining:** 2 states (Off, Loss-based Reweighting).
  - **Hard Adversarial Mining:** 2 states (Off, Augmentation-based Selection) [1].
  - **Balancing:** 2 states (None, Batch Balancing).
  - **Loss Reweighting:** 2 states (None, Inverse-Frequency Loss).
  - **Robustness:** 2 states (None, EMA).

This design identifies individual family effects and cross-family interactions.

3.1. Dataset Protocols and Scarcity Injection

To systematically evaluate training strategies under data constraints typical of medical imaging, we induce controlled scarcity and class imbalance into standard benchmarks. We use CIFAR-10 (natural images) as a non-medical control dataset alongside three MedMNIST subsets: RetinaMNIST, OrganCMNIST, and PathMNIST. CIFAR-10 helps distinguish general algorithmic effects from medical-domain challenges while confirming our methods compete with established baselines.

Three fixed scarcity profiles are applied per dataset: the Full Profile retains the complete training set; the Long-tailed Profile preserves head (frequent) classes fully while subsampling tail classes following the methodology in (Ding et al., 2024); and the Extreme Long-tailed Profile further reduces the 50% smallest classes to 1/10th of their long-tailed amounts. These profiles progressively intensify scarcity to reveal strategy robustness, from balanced full data to severe tail-class underrepresentation.

4. Implementation

We implement one representative technique from each family with the aim of understanding which mechanisms provide the most robust starting point for practitioners. This creates a practical guideline on what families to prioritize and further explore if optimal performance is desirable.

1. Hard Adversarial Mining is only active if data augmentation is active

4.0.1. Data Augmentation

We evaluate four intensity levels:

- **Baseline (None):** Identity preprocessing with standard normalization.
- **Basic Classical:** Conservative geometric transformations (rotations $\pm 15°$, flips, mild jitter).
- **RandAugment:** Automated policy sampling (Cubuk et al., 2019) ($N = 2, M = 9$).
- **TrivialAugment:** Parameter-free automated augmentation (mul, 2021).

4.0.2. Hard Example Mining (HEM)

We implement **Online Hard Example Mining (OHEM)** (Shrivastava et al., 2016) using exponentially smoothed loss scores ($\alpha = 0.9$). Batches are sampled proportional to utility.

4.0.3. Hard Adversarial Mining (HAM) Probe

We adopt an **Augmentation-Based HAM** approach (Lin et al., 2025; Hua et al., 2021). For each step, the model generates $K = 4$ augmented variants and selects the one yielding the highest loss. This approximates adversarial training without incurring gradient-based perturbation costs or risking distortion of relevant features.

4.0.4. Balancing and Reweighting

We address class imbalance through two mechanisms:

- **Inverse-Frequency Reweighting:** Modifies loss with class weights $w_i \propto 1/f_i$, where f_i is the frequency of class i.
- **Batch Balancing:** Enforces equal class representation in each batch.

4.0.5. Robustness-Oriented

We employ **Exponential Moving Average (EMA)** (Tarvainen Antti, 2017) of model weights ($\alpha = 0.99$) as a teacher model.

4.1. Training and Evaluation Framework

Experiments use a ResNet-18 architecture adapted for small input resolutions (32×32 or 28×28). We train all models using the Adam optimizer with a learning rate of 5×10^{-4} and a batch size of 64.

To ensure fair comparison across varying dataset sizes, we define training duration based on the number of samples seen rather than epochs. All models are trained for a minimum of 1 million samples. Training continues until convergence or a maximum of 16 million samples. Convergence is defined as no improvement in the best validation score over the latest 50 steps.

Evaluation is performed on vendor-provided validation splits with an adaptive frequency schedule: every 10k samples for the first 100k samples (to capture early dynamics), every 25k samples up to 1M samples, and every 200k samples thereafter.

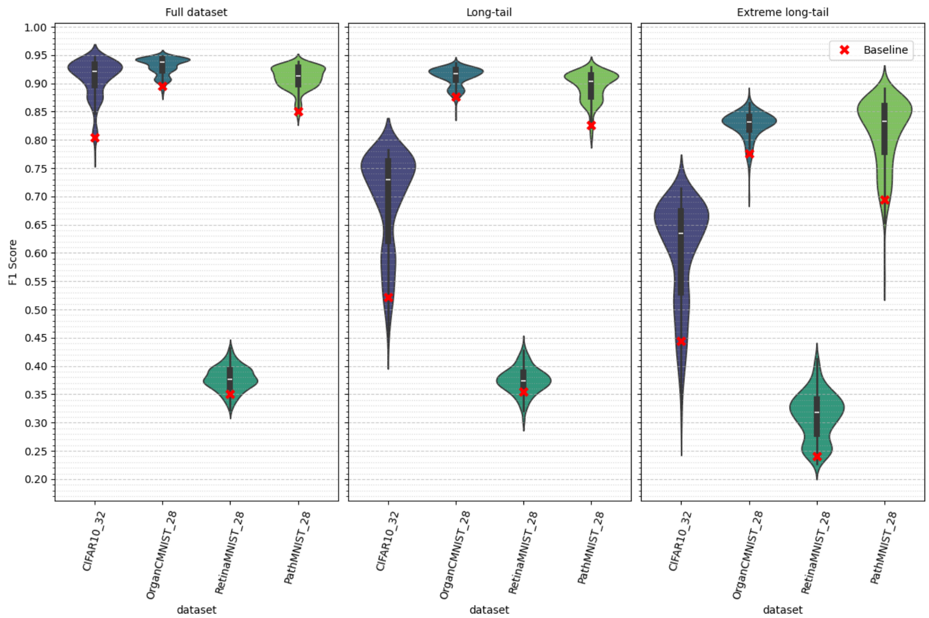

Figure 2: Performance range for all strategy combinations broken down by dataset and scarcity profile. The variability highlights the differing "headroom" for improvement available in each domain.

4.1.1. Normalized Potential Score (NPS)

We use the Normalized Potential Score (NPS) to compare effectiveness across datasets with varying baselines:

$$\text{NPS} = \frac{F1_x - F1_{baseline}}{F1_{best} - F1_{baseline} + \varepsilon} \tag{1}$$

where $F1_x$ is the F1 score of the strategy configuration, $F1_{baseline}$ is the baseline F1 (no strategies), $F1_{best}$ is the best observed F1 score across all configurations for that dataset and scarcity profile, and ε is a small constant to prevent division by zero. The NPS metric quantifies the fraction of potential improvement achieved, where 0.0 is baseline performance, 1.0 represents the best achieved performance, and negative values indicate performance below baseline. The corresponding baselines and ranges are visualized in Figure 2.

5. Empirical Observations: Quantifying Generalization Behavior

The 3,051 experiments reveal a hierarchy of strategy effectiveness. We use the Normalized Potential Score (NPS) to compare performance across datasets. Specific families provide foundational improvements, while others offer targeted refinements.

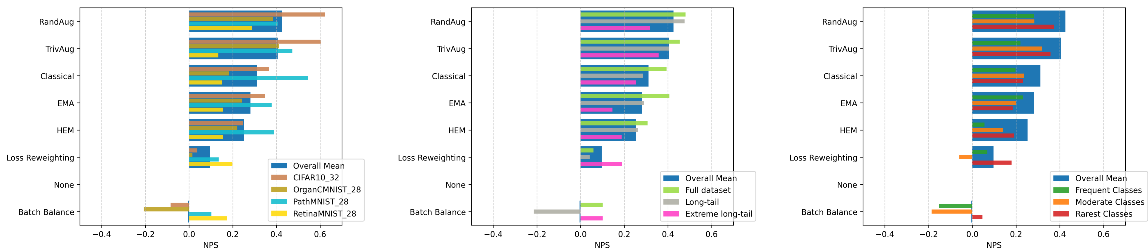

Figure 3: Normalized Potential Score (NPS) for single-strategy interventions. Data Augmentation strategies (RandAugment, TrivialAugment) consistently outperform other families when applied in isolation.

5.1. Importance of Data Augmentation

Data Augmentation provides the biggest boost, yielding 0.30–0.60 NPS gains. Aggregated results (Figure 3) show augmentation-based interventions (TrivialAugment, RandAugment, Classical) outperform other single strategies.

No single augmentation strategy dominates all datasets, i.e. the choice is dataset-dependent. RandAugment performs best on average (0.45 NPS), yet PathMNIST benefits most from Classical augmentations (0.56 NPS). RandAugment and TrivialAugment were designed for CIFAR-10, so their high performance (0.62 NPS on CIFAR) may represent an upper bound. Augmentation importance correlates with dataset imbalance: as scarcity increases, expanding the effective training distribution becomes more critical. In full data profiles, augmentation is less dominant. HEM (0.31 NPS) and EMA (0.41 NPS) perform comparably to classical augmentation (0.39 NPS) but trailing behind automated policies (0.45-0.48 NPS).

Augmentation strategies yield the largest isolated gains (Figure 3). Expanding the training distribution support is a prerequisite for optimization. Without it, other strategies are less effective.

5.2. Synergies: The "Augmentation+" Standard

Augmentation is necessary but insufficient for peak performance. Top configurations combine augmentation with complementary families.

Data Augmentation and Hard Adversarial Mining (HAM) show strong synergy. HAM appears in 77% of top configurations (Table 1). Augmentation expands the data space; HAM targets decision boundaries. This combination improves performance by 0.30 NPS over augmentation alone. It reaches 0.90 NPS in full profiles but drops to 0.77 NPS in extreme long-tail profiles, indicating limitations in handling severe imbalance.

5.3. Targeted Interventions for Imbalance

Balancing and Reweighting strategies target specific classes. Unlike augmentation, they redistribute error contributions. In Extreme Long-tail profiles, Loss Reweighting appears in 92% of top runs (Table 2), improving minority class NPS by 0.08–0.10 (Table 3). This

Table 1: Detailed breakdown of average potential for strategy combinations across all datasets, scarcity profiles, and class frequencies groups.

Strategies								Overall Mean Score		Class Size Grouping						Datasets							
Batch Balance	Loss Reweighting	Classical	RandAug	TrivAug	HEM	HAM	EMA	All		Rarest Cls		Moderate Cls		Frequent Cls		CIFAR10		OrganCMNIST		PathMNIST		RetinaMNIST	
								$\mu \pm \sigma$	n	$\mu \pm \sigma$	n	$\mu \pm \sigma$	n	$\mu \pm \sigma$	n	$\mu \pm \sigma$	n	$\mu \pm \sigma$	n	$\mu \pm \sigma$	n	$\mu \pm \sigma$	n
Full dataset																							
Best																							
-	-	-	-	x	-	x	-	**.90 ± .03**	12	.74 ± .04	20	**.68 ± .19**	75	.54 ± .07	20	.91 ± .08	5	.92 ± .02	3	.83 ± .02	3	**.94 ± .00**	1
-	x	-	-	x	-	x	x	.89 ± .01	6	**.87 ± .03**	9	.62 ± .16	32	.34 ± .16	9	.99 ± .00	2	.91 ± .00	1	.90 ± .00	1	.74 ± .05	2
x	-	-	x	-	-	x	x	.87 ± .01	8	.81 ± .07	12	.61 ± .14	47	.46 ± .20	12	.91 ± .00	1	.91 ± .00	3	**.96 ± .01**	2	.71 ± .04	2
-	x	-	-	x	x	x	x	.81 ± .05	10	.69 ± .13	14	.65 ± .15	59	.32 ± .41	14	**.99 ± .00**	3	.93 ± .00	1	.90 ± .06	4	.42 ± .12	2
x	-	-	-	x	-	x	-	.80 ± .02	12	.74 ± .06	19	.57 ± .31	78	**.70 ± .05**	19	.96 ± .00	2	.90 ± .02	5	.87 ± .05	4	.46 ± .00	1
-	-	-	-	x	x	-	x	.60 ± .08	9	.62 ± .13	14	.47 ± .44	53	.28 ± .30	14	.96 ± .01	2	**.95 ± .03**	3	.20 ± .25	2	.31 ± .01	2
Worst																							
-	x	-	-	-	x	-	-	.22 ± .02	7	.34 ± .07	12	.07 ± .38	42	**−.65 ± .06**	12	.33 ± .06	3	.18 ± .01	2	.50 ± .00	1	−.14 ± .00	1
x	x	-	-	-	x	-	-	.21 ± .01	7	.31 ± .09	12	.13 ± .18	42	−.53 ± .06	12	.39 ± .01	3	.19 ± .03	2	.60 ± .00	1	**−.35 ± .00**	1
x	-	-	-	-	-	-	-	.10 ± .08	11	.12 ± .07	18	−.06 ± .37	64	−.06 ± .13	18	.04 ± .04	5	**−.16 ± .09**	2	.09 ± .06	2	.44 ± .15	2
-	-	-	-	-	-	-	-	−.00 ± .11	15	**.00 ± .15**	24	−.00 ± .18	87	−.00 ± .37	24	**−.00 ± .10**	6	−.00 ± .16	3	.00 ± .06	3	.00 ± .12	3
x	x	-	-	-	-	-	-	**−.05 ± .06**	10	.03 ± .12	15	**−.10 ± .34**	55	−.58 ± .29	15	.05 ± .04	3	−.07 ± .03	2	**−.01 ± .12**	2	−.16 ± .06	3
Long-tail																							
Best																							
x	x	-	-	x	-	x	-	**.83 ± .01**	6	.67 ± .10	9	**.66 ± .13**	37	.59 ± .07	9	.88 ± .00	1	.74 ± .05	2	.85 ± .01	2	**.86 ± .00**	1
x	-	-	-	x	x	x	-	.79 ± .05	10	.63 ± .16	15	.47 ± .58	52	.58 ± .17	15	.88 ± .01	2	.74 ± .07	3	**.97 ± .00**	1	.56 ± .11	4
-	x	-	-	x	-	x	x	.77 ± .02	6	**.75 ± .06**	9	.62 ± .16	37	.72 ± .08	9	.96 ± .00	1	.84 ± .03	2	.91 ± .04	2	.38 ± .00	1
-	x	-	x	-	-	x	x	.76 ± .01	6	.61 ± .07	9	.36 ± .55	37	**.77 ± .06**	9	.96 ± .00	1	.82 ± .01	2	.83 ± .01	2	.42 ± .00	1
-	x	-	-	x	x	-	x	.63 ± .10	9	.63 ± .19	12	.38 ± .53	46	.46 ± .29	12	**.99 ± .00**	1	**.96 ± .02**	2	.21 ± .21	2	.38 ± .15	4
Worst																							
x	-	-	-	-	-	-	-	−.21 ± .09	11	**−.15 ± .21**	18	**−.42 ± .83**	64	−.34 ± .21	18	−.07 ± .07	5	**−.35 ± .17**	2	−.06 ± .07	2	−.37 ± .03	2
x	x	-	-	-	-	-	-	**−.25 ± .05**	11	−.12 ± .11	18	−.40 ± .64	64	**−.35 ± .31**	18	**−.17 ± .08**	5	−.09 ± .02	2	**−.22 ± .00**	2	**−.52 ± .11**	2
Extreme long-tail																							
Best																							
-	x	-	-	x	x	x	-	**.77 ± .06**	10	.68 ± .18	15	.57 ± .21	56	.71 ± .08	15	.86 ± .00	2	.77 ± .07	3	.76 ± .00	2	.68 ± .17	3
-	x	x	-	-	-	x	x	.76 ± .02	6	.61 ± .14	9	**.62 ± .20**	36	.58 ± .05	9	.78 ± .05	2	.71 ± .00	1	.85 ± .02	2	.68 ± .00	1
-	x	-	-	x	-	-	x	.73 ± .03	5	.58 ± .10	8	.51 ± .43	30	.66 ± .08	8	.93 ± .00	1	**.89 ± .11**	2	.36 ± .00	1	.73 ± .00	1
-	x	-	-	x	x	x	x	.72 ± .02	7	**.70 ± .14**	11	.58 ± .17	44	**.75 ± .04**	11	.86 ± .00	1	.69 ± .06	3	.87 ± .00	2	.47 ± .00	1
-	x	-	x	-	x	x	x	.68 ± .03	10	.52 ± .17	15	.55 ± .36	55	.70 ± .08	15	**.94 ± .05**	3	.63 ± .02	2	.78 ± .02	2	.39 ± .04	3
x	-	x	-	-	-	x	-	.68 ± .05	10	.48 ± .15	15	.55 ± .22	56	.65 ± .09	15	.66 ± .03	2	.58 ± .06	3	**.97 ± .03**	2	.50 ± .09	3
x	x	-	-	x	-	-	x	.64 ± .00	4	.53 ± .23	6	.50 ± .37	23	.46 ± .09	6	.66 ± .00	1	.59 ± .00	1	.32 ± .00	1	**1.00 ± .00**	1
Worst																							
-	-	-	x	-	-	-	x	.48 ± .01	5	.30 ± .11	8	.50 ± .19	29	.48 ± .04	8	.87 ± .02	2	.62 ± .00	1	.49 ± .00	1	**−.05 ± .00**	1
x	x	-	x	-	-	-	-	.24 ± .03	6	.28 ± .10	10	.15 ± .26	36	−.23 ± .09	10	.29 ± .06	2	.41 ± .04	2	**−.07 ± .00**	1	.32 ± .00	1
x	-	-	-	-	-	-	x	.06 ± .04	11	**−.21 ± .17**	18	.11 ± .27	64	.11 ± .12	18	.10 ± .04	5	.11 ± .09	2	−.01 ± .03	2	.03 ± .01	2
x	x	-	-	-	-	-	-	**−.09 ± .17**	11	−.10 ± .39	18	**−.13 ± .40**	64	**−.67 ± .39**	18	**−.35 ± .12**	5	**−.44 ± .32**	2	−.02 ± .17	2	.44 ± .06	2

sensitivity reduces the precision of the majority class. These strategies are appropriate only when prioritizing rare-class performance.

5.4. Ablation Analysis: Quantifying Marginal Gains

Ablation of top combinations (Table 3) confirms this structure. Removing HAM drops performance by 0.16 NPS on average (up to 0.37 NPS). Removing loss reweighting primarily degrades tail-class performance (approx. 0.10 NPS).

EMA impact is domain-dependent. In OrganCMNIST, removing EMA causes a 0.40–0.50 NPS drop; elsewhere, effects are negligible. Removing HEM has both positive and negative effects, indicating dataset-specific utility.

The tables Table 3 and Table 1 provide a lot of domain-specific insights that can be further analyzed if one is interested in a specific dataset or scarcity profile.

Table 2: Strategy usage in top-performing combinations broken down by class frequency and scarcity profile. Note the shift from Robustness strategies (HAM, EMA) for Frequent classes to Balancing strategies (LossWeight) for Rarest classes.

Group \Reduction	Full Dataset	Long-tail	Extreme Long-tail
Rarest	EMA: 75% Loss Reweighting: 67% HAM: 67% TrivAug: 58% Batch Balance: 50% HEM: 50% RandAug: 33% Classical: 8%	Loss Reweighting: 92% EMA: 83% HAM: 58% TrivAug: 50% HEM: 50% Batch Balance: 33% Classical: 25% RandAug: 25%	Loss Reweighting: 92% EMA: 75% TrivAug: 67% Batch Balance: 42% HEM: 33% Classical: 25% HAM: 25% RandAug: 8%
Moderate	EMA: 83% HAM: 67% TrivAug: 58% RandAug: 42% Loss Reweighting: 33% HEM: 33% Batch Balance: 25%	TrivAug: 83% EMA: 67% HAM: 50% Loss Reweighting: 42% HEM: 42% Batch Balance: 25% Classical: 8% RandAug: 8%	EMA: 75% Loss Reweighting: 67% TrivAug: 67% HAM: 50% Batch Balance: 42% HEM: 33% RandAug: 25% Classical: 8%
Frequent	HAM: 75% Batch Balance: 58% TrivAug: 58% EMA: 58% Loss Reweighting: 50% RandAug: 42% HEM: 42%	Loss Reweighting: 83% EMA: 75% HAM: 50% RandAug: 42% Batch Balance: 33% Classical: 33% HEM: 33% TrivAug: 25%	Loss Reweighting: 58% HAM: 58% HEM: 50% RandAug: 42% TrivAug: 33% Classical: 25% EMA: 25% Batch Balance: 17%

6. Discussion

The results indicate that training under data scarcity requires a hierarchy of complementary mechanisms rather than selection among competing techniques. We identify three principles for strategy composition.

6.1. A Hierarchy of Training Needs

Data Augmentation (0.30–0.60 NPS) is the foundational layer. Once data expansion is established, Hard Adversarial Mining (HAM) (0.02–0.37 NPS) emerges as the next most effective strategy, ensuring that selected samples induce high-magnitude gradient updates. However, large gradient updates can lead to training oscillations; thus, it is beneficial to incorporate robustness-oriented strategies like EMA to stabilize training. Since EMA maintains both teacher and student models, one can retrospectively evaluate which model yields preferable performance. Finally, Balancing and Reweighting (0.08–0.10 NPS) serve as targeted interventions to adjust the precision-recall trade-off for specific priorities, with reweighting identified as the preferable strategy based on our results.

6.2. The ”Expand-and-Refine” Paradigm

The synergy between Augmentation and HAM supports an ”Expand-and-Refine” model. Augmentation expands the training distribution, while HAM refines it by targeting decision boundaries, as confident misclassifications often indicate areas of uncertainty. In our experiments, we simplified HAM using augmentation-based selection, which guarantees that selected samples remain within the domain distribution. While adversarial perturbations could be explored in future work, they risk generating out-of-distribution samples that

Table 3: Impact of removing individual strategies from the best-performing combinations. Values indicate the change in Normalized Potential Score (NPS).

Impact of EMA

Strategies								Overall Mean Score	Class Size Grouping			Data Scarcity			Datasets			
Batch Balance	Loss Reweighting	Classical	RandAug	TrivAug	HEM	HAM	EMA	Average Δ	Rarest Δ	Moderate Δ	Frequent Δ	Full Δ	Long Δ	Extr Δ	CIFAR10 Δ	OrganCMN Δ	PathMN Δ	RetinaMN Δ
x	x	-	-	x	-	-	I	+.269	+.184	+.290	+.218	+.185	+.300	+.323	+.409	+.426	+.130	+.113
-	x	x	-	-	-	-	I	+.261	+.188	+.229	+.278	+.176	+.338	+.268	+.362	+.478	+.155	+.047
-	x	-	-	x	-	-	I	+.177	+.097	+.191	+.064	+.152	+.139	+.239	+.354	+.518	−.225	+.059
-	-	-	-	x	-	-	I	+.128	+.104	+.137	+.037	+.094	+.099	+.191	+.342	+.437	−.108	−.158
-	x	x	-	-	-	x	I	+.073	+.124	−.000	−.046	+.029	+.150	+.040	+.123	+.050	−.052	+.171
x	-	-	-	x	-	x	I	+.051	+.046	+.053	−.076	+.102	+.051	+.001	+.036	−.044	−.010	+.222
-	x	-	x	-	-	x	I	+.027	+.058	−.026	+.089	−.003	+.058	+.025	+.041	+.141	−.005	−.069
-	x	-	-	x	-	x	I	+.012	+.077	+.074	−.004	+.050	−.011	−.003	+.027	+.025	+.053	−.058
-	x	-	x	-	x	-	I	+.009	+.105	+.041	+.174	+.053	−.046	+.021	+.058	+.062	−.049	−.033
x	-	x	-	-	-	x	I	+.007	+.040	−.071	+.215	+.026	−.001	−.003	+.083	+.027	−.072	−.008
-	x	-	-	x	x	x	I	−.011	−.017	+.018	+.017	−.022	+.029	−.041	+.036	−.018	+.070	−.133
x	x	-	-	x	-	x	I	−.032	+.085	−.099	−.027	−.076	−.060	+.041	+.027	+.035	+.007	−.196
x	-	-	-	x	x	x	I	−.049	−.059	+.018	+.048	+.008	−.058	−.097	+.027	−.032	+.013	−.203
-	x	-	-	x	x	-	I	−.054	+.045	−.082	−.023	−.097	−.001	−.063	+.073	+.104	−.328	−.063
Average per column								+.062	+.077	+.055	+.069	+.048	+.071	+.067	+.143	**+.158**	−.030	−.022

Impact of HAM

Strategies								Overall Mean Score	Class Size Grouping			Data Scarcity			Datasets			
BB	LR	C	RA	TA	HEM	HAM	EMA	Average Δ	Rarest Δ	Moderate Δ	Frequent Δ	Full Δ	Long Δ	Extr Δ	CIFAR10 Δ	OrganCMN Δ	PathMN Δ	RetinaMN Δ
x	x	-	-	x	-	I	-	+.374	+.193	+.395	+.405	+.377	+.408	+.337	+.397	+.347	+.468	+.284
-	-	-	-	x	-	I	-	+.318	+.217	+.249	+.299	+.449	+.279	+.226	+.311	+.366	+.338	+.258
-	-	-	x	-	-	I	-	+.293	+.156	+.307	+.216	+.321	+.235	+.323	+.292	+.343	+.407	+.129
x	-	-	-	x	-	I	-	+.289	+.168	+.194	+.499	+.281	+.391	+.195	+.379	+.301	+.447	+.029
-	x	-	-	x	-	I	-	+.268	+.145	+.255	+.173	+.358	+.196	+.251	+.326	+.422	+.252	+.073
x	-	-	-	x	-	I	x	+.169	+.089	+.083	−.010	+.202	+.236	+.070	+.026	−.140	+.419	+.374
-	x	-	-	x	x	I	x	+.148	+.039	+.181	+.107	+.170	+.166	+.108	−.018	−.059	+.587	+.081
x	-	-	-	x	x	I	-	+.144	+.167	+.029	+.023	+.178	+.128	+.125	+.028	+.018	+.227	+.301
-	x	-	x	-	-	I	x	+.133	+.151	+.033	+.263	+.152	+.109	+.138	+.005	+.000	+.393	+.134
x	-	-	-	x	x	I	x	+.127	+.083	+.106	+.068	+.148	+.150	+.083	+.011	−.114	+.557	+.054
x	x	-	-	x	x	I	-	+.113	+.013	+.087	+.231	+.261	+.107	−.029	+.024	−.099	+.258	+.268
-	x	-	-	x	x	I	-	+.105	+.101	+.081	+.068	+.095	+.135	+.086	+.020	+.062	+.189	+.151
-	x	-	-	x	-	I	x	+.103	+.125	+.138	+.105	+.256	+.046	+.008	−.001	−.071	+.529	−.044
-	x	-	x	-	x	I	-	+.095	+.055	+.123	+.331	+.168	+.129	−.013	+.029	−.067	+.300	+.117
x	x	-	-	x	-	I	x	+.073	+.094	+.006	+.161	+.115	+.048	+.056	+.016	−.044	+.346	−.025
-	x	x	-	-	-	I	x	+.021	+.054	+.022	−.059	+.016	+.017	+.031	+.093	+.021	−.051	+.022
x	-	x	-	-	-	I	x	−.018	−.050	+.019	+.024	+.066	−.086	−.034	+.086	+.001	−.014	−.146
Average per column								+.162	+.106	+.136	+.171	+.212	+.158	+.115	+.119	+.076	**+.332**	+.121

Impact of HEM

Strategies								Overall Mean Score	Class Size Grouping			Data Scarcity			Datasets			
BB	LR	C	RA	TA	HEM	HAM	EMA	Average Δ	Rarest Δ	Moderate Δ	Frequent Δ	Full Δ	Long Δ	Extr Δ	CIFAR10 Δ	OrganCMN Δ	PathMN Δ	RetinaMN Δ
-	-	-	-	x	I	-	-	+.244	+.173	+.220	+.169	+.217	+.259	+.258	+.299	+.443	+.194	+.042
x	-	-	-	x	I	x	-	+.045	+.013	+.030	−.116	+.030	+.022	+.083	−.013	−.004	+.020	+.177
-	x	-	x	-	I	-	x	+.028	+.114	+.020	+.077	+.047	−.059	+.097	−.004	+.025	+.040	+.051
-	x	-	x	-	I	x	-	+.007	−.029	+.043	+.060	+.007	+.066	−.050	+.002	+.038	−.009	−.001
-	x	-	-	x	I	x	-	+.002	+.045	+.044	−.077	−.004	−.015	+.026	−.012	+.070	−.028	−.019
-	x	-	-	x	I	x	x	−.021	−.049	−.011	−.056	−.076	+.026	−.011	−.004	+.027	−.011	−.095
x	x	-	-	x	I	x	-	−.053	−.019	−.098	−.061	−.012	−.087	−.060	−.002	−.136	+.039	−.113
x	-	-	-	x	I	x	x	−.055	−.092	−.005	+.008	−.064	−.087	−.015	−.023	+.007	+.043	−.248
-	x	-	-	x	I	-	x	−.065	+.038	−.054	−.058	+.010	−.094	−.111	+.013	+.015	−.069	−.220
Average per column								+.015	+.022	+.021	−.006	+.017	+.003	+.024	+.028	**+.054**	+.024	−.047

Impact of Loss Reweighting

Strategies								Overall Mean Score	Class Size Grouping			Data Scarcity			Datasets			
BB	LR	C	RA	TA	HEM	HAM	EMA	Average Δ	Rarest Δ	Moderate Δ	Frequent Δ	Full Δ	Long Δ	Extr Δ	CIFAR10 Δ	OrganCMN Δ	PathMN Δ	RetinaMN Δ
-	I	-	-	x	-	-	x	+.161	+.160	+.011	+.224	+.082	+.222	+.180	+.006	+.030	+.014	+.596
-	I	x	-	-	-	x	x	+.137	+.241	−.018	+.084	+.031	+.159	+.222	+.014	+.045	+.067	+.424
-	I	-	-	x	-	x	x	+.131	+.192	+.069	+.158	+.108	+.110	+.175	+.027	+.089	+.115	+.294
-	I	x	-	-	-	-	x	+.131	+.161	+.035	+.191	+.116	+.157	+.119	+.003	+.025	+.043	+.453
-	I	-	-	x	x	x	-	+.120	+.149	+.045	+.133	+.126	+.049	+.185	+.012	+.150	−.020	+.337
-	I	-	x	-	x	-	x	+.106	+.183	+.010	+.194	+.098	+.036	+.186	+.017	+.085	−.052	+.376
-	I	-	-	x	x	x	x	+.094	+.122	+.032	+.135	+.012	+.076	+.192	+.030	+.090	+.032	+.223
-	I	-	x	-	-	x	x	+.085	+.075	−.023	+.224	−.012	+.081	+.186	−.004	+.078	+.028	+.238
-	I	-	x	-	x	x	-	+.083	+.018	+.099	+.200	+.086	+.078	+.086	+.000	+.052	+.051	+.229
x	I	-	-	x	-	-	x	+.064	+.026	+.044	−.246	+.006	+.142	+.043	−.052	−.072	+.061	+.317
-	I	-	-	x	-	x	-	+.063	+.095	−.036	+.071	−.067	+.099	+.156	+.008	+.005	+.045	+.193
-	I	-	-	x	x	-	x	+.055	+.217	−.109	−.051	+.034	+.042	+.088	+.027	+.064	−.038	+.167
x	I	-	-	x	-	x	-	+.050	−.008	+.119	−.124	+.098	+.064	−.012	−.053	−.054	−.028	+.336
x	I	-	-	x	-	x	x	−.033	+.031	−.033	−.076	−.080	−.046	+.029	−.063	+.024	−.012	−.081
x	I	-	-	x	x	x	-	−.048	−.040	−.008	−.069	+.056	−.045	−.154	−.042	−.186	−.009	+.047
Average per column								+.080	+.108	+.016	+.070	+.046	+.082	+.112	−.005	+.028	+.020	**+.277**

Impact of Batch Balance

Strategies								Overall Mean Score	Class Size Grouping			Data Scarcity			Datasets			
BB	LR	C	RA	TA	HEM	HAM	EMA	Average Δ	Rarest Δ	Moderate Δ	Frequent Δ	Full Δ	Long Δ	Extr Δ	CIFAR10 Δ	OrganCMN Δ	PathMN Δ	RetinaMN Δ
I	-	-	-	x	-	x	x	+.128	+.116	+.024	+.139	+.121	+.154	+.108	+.006	−.034	+.089	+.449
I	-	-	-	x	x	x	-	+.119	+.071	+.007	+.146	+.120	+.067	+.170	−.019	+.022	+.065	+.408
I	-	x	-	-	-	x	x	+.074	+.083	−.042	+.158	+.022	+.060	+.139	−.024	−.038	+.115	+.241
I	-	-	-	x	x	x	x	+.055	+.002	−.007	+.180	+.037	+.007	+.122	−.010	−.053	+.059	+.225
I	-	-	-	x	-	x	-	+.020	+.049	−.060	+.123	−.107	+.081	+.084	−.023	−.049	+.081	+.069
I	x	-	-	x	-	x	-	+.007	−.053	+.095	−.072	+.058	+.046	−.083	−.084	−.108	+.008	+.212
I	x	-	-	x	-	-	x	−.006	−.014	+.054	−.150	+.072	−.005	−.086	−.101	−.125	+.146	+.055
I	x	-	-	x	-	x	x	−.036	−.045	−.078	−.095	−.068	−.003	−.038	−.084	−.098	−.037	+.074
I	x	-	-	x	x	x	-	−.048	−.117	−.047	−.056	+.050	−.026	−.169	−.074	−.314	+.076	+.118
Average per column								+.035	+.010	−.006	+.042	+.034	+.042	+.027	−.046	−.089	+.067	**+.206**

may not be relevant. Our results suggest that the sheer amount of data is not the most critical factor; rather, domain coverage and sample informativeness are the key factors. We observe that both PathMNIST and OrganCMNIST exhibit similar best-case performance across scarcity profiles, whereas the baseline and worst-case performance degrade significantly with less data (Figure 2). This indicates potential redundancy in the data, highlighting the importance of selecting informative samples over merely increasing dataset size.

6.3. The Precision-Recall Trade-off in Balancing

Balancing and Reweighting strategies function as targeted interventions rather than general performance boosters. While our experiments confirm they enhance rare-class performance (0.08–0.10 NPS), this gain often comes at the expense of majority class accuracy. These techniques effectively trade overall precision for minority class sensitivity. Our results show that Balancing and Reweighting strategies are present in both the best and worst-performing configurations, highlighting their volatility. Therefore, they should be viewed as specialized tools rather than default components of a general-purpose training pipeline. Additionally, while we employed a fixed inverse-frequency reweighting scheme in this study, practitioners are advised to explore alternative parameterizations to shift the performance trade-off according to their specific needs.

7. Limitations

Several limitations constrain the generalizability of these findings.

Hyperparameter Configuration: We prioritized breadth over depth in hyperparameter exploration to ensure a comprehensive comparison across strategy families. Consequently, we employed fixed hyperparameters (e.g., $\alpha = 0.99$ for EMA, 5×10^{-4} learning rate) rather than optimizing for each specific dataset-strategy combination. While extensive per-dataset tuning was computationally infeasible given the combinatorial search space, our selected hyperparameters achieve near state-of-the-art results on CIFAR-10 and outperform the configurations reported in (Ding et al., 2024). This suggests that our findings reflect robust algorithmic properties rather than hyperparameter overfitting, though further performance gains could likely be realized through granular tuning.

Architecture and Domain Scope: The analysis is limited to ResNet-18 and image classification. Consequently, results may not generalize to other architectures (e.g., Vision Transformers) or tasks (segmentation, detection). While MedMNIST datasets provide a standardized benchmark, they represent simplified classification scenarios. Future work should validate findings on more complex medical imaging tasks and architectures.

Evaluation Methodology: The Normalized Potential Score (NPS) depends on the performance range within the experimental grid, which means that strategies that excel under different regimes may be undervalued. For instance, if the model capacity is insufficient relative to data complexity, a performance ceiling may mask the benefits of advanced training strategies. Thus, while NPS facilitates fair cross-dataset comparison by normalizing for difficulty, it can not capture absolute performance gains.

Computational Constraints: The experimental design is coarse due to scale constraints. Granular analysis of interactions, curriculum schedules, or adaptive hyperparameters requires significantly more resources.

These findings guide strategy selection under data scarcity, though fine-tuning will yield additional gains.

8. Open Problems and Future Directions

This investigation identifies areas for future research in data-constrained training.

8.1. Adaptive Strategy Selection

Representative technique show effectiveness under default configurations. Adaptive approaches adjusting strategy intensity based on dataset characteristics can be further explored. Automatic selection and tuning based on scarcity and learning dynamics is a promising direction.

8.2. Scalability Beyond Classification

While this study focuses on image classification, many workflows often rely on segmentation and detection tasks which may exhibit different training dynamics. Future work should extend this analysis to dense prediction tasks and multi-modal learning to validate the generalizability of the proposed hierarchy. Additionally, investigating modern architectures such as Vision Transformers and ResNeXt with high-resolution inputs is essential to understand if these findings hold for models with different characteristics.

8.3. Synthetic Data Integration

Generative models, such as Denoising Diffusion Probabilistic Models (DDPMs) and Generative Adversarial Networks (GANs), offer promising avenues for addressing data scarcity (Vyver et al., 2025; Li et al., 2023; Ding et al., 2024). Given that our results identify Data Augmentation as the most significant contributor to performance improvement, extending the training distribution through synthetic data synthesis represents a natural progression. While this study focused exclusively on discriminative strategies, future work should rigorously evaluate generative augmentation, specifically characterizing the trade-offs between synthetic diversity and the risk of hallucinating non-existent features.

9. Conclusion

Our empirical study identifies Data Augmentation as the primary performance driver across all scarcity regimes, with Hard Adversarial Mining providing complementary gains by refining decision boundaries. Balancing and Reweighting strategies serve as specialized tools for enhancing rare-class sensitivity at the cost of overall precision, while Robustness-oriented training (EMA) offers low-overhead stabilization.

We propose a practical hierarchy for data-scarce settings: prioritize domain-specific augmentation, integrate adversarial mining to target hard samples, and apply balancing only when minority-class recall is critical.

Acknowledgments

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and the Zenith career development program at Linköping University.

References

Trivialaugment: Tuning-free yet state-of-the-art data augmentation. 2021. URL https://arxiv.org/pdf/2103.10158.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357, 2002. URL https://arxiv.org/pdf/1106.1813.

Manuel Cossio. Augmenting medical imaging: a comprehensive catalogue of 65 techniques for enhanced data analysis. *arXiv preprint arXiv:2303.01178*, 2023. URL https://arxiv.org/pdf/2303.01178.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. URL https://arxiv.org/pdf/1805.09501.

Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. 2019. URL https://arxiv.org/pdf/1909.13719.

Hongwei Ding, Nana Huang, and Xiaohui Cui. Leveraging gans data augmentation for imbalanced medical image classification. *Applied Soft Computing*, 165:112050, 2024. URL https://www.sciencedirect.com/science/article/pii/S156849462400824X.

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. *arXiv preprint arXiv:1912.02781*, 2019. URL https://arxiv.org/pdf/1912.02781.

Weizhe Hua, Yichi Zhang, Chuan Guo, Zhiru Zhang, and G Edward Suh. Bullettrain: Accelerating robust neural network training via boundary example mining. *Advances in Neural Information Processing Systems*, 34:18527–18538, 2021. URL https://arxiv.org/pdf/2109.14707.

Pratyush Kumar and Muktabh Mayank Srivastava. Example mining for incremental learning in medical imaging. pages 48–51, 2018. doi: 10.1109/SSCI.2018.8628895. URL https://arxiv.org/pdf/1807.08942.

Ming Li, Jiping Wang, Yang Chen, Yufei Tang, Zhongyi Wu, Yujin Qi, Haochuan Jiang, Jian Zheng, and Benjamin MW Tsui. Low-dose ct image synthesis for domain adaptation imaging using a generative adversarial network with noise encoding transfer learning. *IEEE transactions on medical imaging*, 42(9):2616–2630, 2023. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10081080.

Chang-Hui Liang, Wan-Lei Zhao, and Run-Qing Chen. Dynamic sampling for deep metric learning. *Pattern Recognition Letters*, 150:49–56, 2021. URL https://arxiv.org/pdf/2004.11624.

Chenhao Lin, Xiang Ji, Yulong Yang, Qian Li, Zhengyu Zhao, Zhe Peng, Run Wang, Liming Fang, and Chao Shen. Hard adversarial example mining for improving robust fairness. *IEEE Transactions on Information Forensics and Security*, 20:350–363, 2025. doi: 10.1109/TIFS.2024.3516554. URL https://arxiv.org/pdf/2308.01823.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. pages 2980–2988, 2017. URL https://arxiv.org/pdf/1708.02002.

Agnieszka Mikołajczyk and Michał Grochowski. Data augmentation for improving deep learning in image classification problem. pages 117–122, 2018. doi: 10.1109/IIPHDW.2018.8388338. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8388338.

Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. Exponential moving average of weights in deep learning: Dynamics and benefits. *Transactions on Machine Learning Research (TMLR)*, 2024. URL https://arxiv.org/pdf/2411.18704.

Marius Schmidt-Mengin, Théodore Soulier, Mariem Hamzaoui, Arya Yazdan-Panah, Benedetta Bodini, Nicholas Ayache, Bruno Stankoff, and Olivier Colliot. Online hard example mining vs. fixed oversampling strategy for segmentation of new multiple sclerosis lesions from longitudinal flair mri. *Frontiers in Neuroscience*, 16:1004050, 2022. URL https://pubmed.ncbi.nlm.nih.gov/36408404/.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019. URL https://link.springer.com/article/10.1186/s40537-019-0197-0.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. pages 761–769, 2016. doi: 10.1109/CVPR.2016.89. URL https://arxiv.org/pdf/1604.03540.

Elena Sizikova, Andreu Badal, Jana G. Delfino, Miguel Lago, Brandon Nelson, Niloufar Saharkhiz, Berkman Sahiner, Ghada Zamzmi, and Aldo Badano. Synthetic data in radiological imaging: Current state and future outlook. 2024. URL https://arxiv.org/pdf/2407.01561.

Wenhao Tang, Sheng Huang, Xiaoxian Zhang, Fengtao Zhou, Yi Zhang, and Bo Liu. Multiple instance learning framework with masked hard instance mining for whole slide image classification. pages 4078–4087, 2023. URL https://arxiv.org/pdf/2307.15254.

Valpola Harri Tarvainen Antti. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. pages 645–657, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf.

Gilles Van De Vyver, Aksel Try Lenz, Erik Smistad, Sindre Hellum Olaisen, Bjørnar Grenne, Espen Holte, Håavard Dalen, and Lasse Løvstakken. Generative augmentations for improved cardiac ultrasound segmentation using diffusion models. 2025. URL https://arxiv.org/pdf/2502.20100.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3): 107–115, 2 2021. ISSN 0001-0782. doi: 10.1145/3446776. URL https://dl.acm.org/doi/pdf/10.1145/3446776.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. 2018. URL https://arxiv.org/pdf/1710.09412.

Yingyi Zhong, Wen'an Zhou, and Zhixian Wang. A survey of data augmentation in domain generalization. *Neural Processing Letters*, 57(34), 2025. doi: 10.1007/s11063-025-11747-9. URL https://link.springer.com/content/pdf/10.1007/s11063-025-11747-9.pdf.

Table 4: Class-frequency for each dataset across the full, long-tail, and extreme long-tail profiles.

CIFAR10				
Class	Group	Full dataset train samples	Long-tail train samples	Extreme long-tail train samples
0	Rarest Classes	1000	100	10
1	Rarest Classes	1000	200	120
2	Moderate Classes	1000	300	230
3	Moderate Classes	1000	400	340
4	Moderate Classes	1000	500	450
5	Moderate Classes	1000	600	560
6	Moderate Classes	1000	700	670
7	Moderate Classes	1000	800	780
8	Frequent Classes	1000	900	890
9	Frequent Classes	1000	1000	1000
RetinaMNIST				
Class	Group	Full dataset train samples	Long-tail train samples	Extreme long-tail train samples
0	Frequent Classes	486	400	400
1	Moderate Classes	128	100	10
2	Moderate Classes	206	200	200
3	Moderate Classes	194	150	15
4	Rarest Classes	66	50	5
OrganCMNIST				
Class	Group	Full dataset train samples	Long-tail train samples	Extreme long-tail train samples
0	Moderate Classes	1148	700	700
1	Moderate Classes	619	300	30
2	Rarest Classes	595	100	10
3	Rarest Classes	600	200	20
4	Moderate Classes	1088	600	60
5	Moderate Classes	1170	800	800
6	Frequent Classes	2986	1500	1500
7	Moderate Classes	1002	400	40
8	Moderate Classes	1022	500	50
9	Moderate Classes	1173	900	900
10	Frequent Classes	1572	1000	1000
PathMNIST				
Class	Group	Full dataset train samples	Long-tail train samples	Extreme long-tail train samples
0	Moderate Classes	9366	3000	300
1	Moderate Classes	9509	5000	5000
2	Moderate Classes	10360	6000	6000
3	Moderate Classes	10401	7000	7000
4	Moderate Classes	8006	2000	200
5	Moderate Classes	12182	8000	8000
6	Rarest Classes	7886	1000	100
7	Moderate Classes	9401	4000	400
8	Frequent Classes	12885	10000	10000

Table 5: Data collection statistics for all experiments, including the number of configurations evaluated per dataset, scarcity profile, and training strategy

Statistic	Count
Total experiment configurations	1,344
Total experiment runs	3,051
By Dataset	
CIFAR10	863
OrganCMNIST	744
PathMNIST	719
RetinaMNIST	725
By Data Scarcity	
Full dataset	1,031
Long-tail	1,033
Extreme long-tail	987
By Strategy Component	
HEM	1,559
EMA	1,557
BatchBalance	1,494
LossWeight	1,463
HAM	1,358
BaseAug	924
RandAug	825
TrivAug	800
NoAug	502

Table 6: F1 scores for all datasets under the extreme long-tail scarcity profile. This table provides a comprehensive overview of how each strategy combination performs in terms of overall classification accuracy, highlighting the effectiveness of different approaches in handling severe class imbalance.

Dataset	Full dataset			Long-tail			Extreme long-tail		
	Base	Best	Δ	Base	Best	Δ	Base	Best	Δ
CIFAR10	0.804	0.947	+0.144	0.522	0.783	+0.261	0.444	0.714	+0.270
OrganCMNIST	0.895	0.949	+0.053	0.876	0.936	+0.059	0.776	0.876	+0.100
PathMNIST	0.851	0.938	+0.087	0.826	0.930	+0.103	0.694	0.891	+0.197
RetinaMNIST	0.351	0.432	+0.082	0.354	0.438	+0.083	0.241	0.413	+0.172

Class	CIFAR10 Full dataset			CIFAR10 Long-tail			CIFAR10 Extreme long-tail		
	Base	Best	Δ	Base	Best	Δ	Base	Best	Δ
Class 0	0.824	0.963	+0.139	0.371	0.754	+0.383	0.031	0.448	+0.418
Class 1	0.908	0.978	+0.070	0.628	0.915	+0.287	0.493	0.872	+0.379
Class 2	0.721	0.942	+0.221	0.374	0.717	+0.343	0.337	0.675	+0.338
Class 3	0.644	0.890	+0.246	0.323	0.652	+0.329	0.300	0.634	+0.334
Class 4	0.767	0.952	+0.185	0.457	0.784	+0.326	0.430	0.769	+0.339
Class 5	0.725	0.903	+0.178	0.476	0.754	+0.278	0.454	0.729	+0.275
Class 6	0.848	0.973	+0.125	0.649	0.878	+0.229	0.605	0.875	+0.270
Class 7	0.833	0.975	+0.142	0.620	0.863	+0.243	0.594	0.849	+0.255
Class 8	0.886	0.973	+0.087	0.659	0.871	+0.212	0.585	0.793	+0.207
Class 9	0.882	0.969	+0.087	0.660	0.886	+0.226	0.614	0.848	+0.234

Table 7: Per-class F1 scores for CIFAR-10 under the extreme long-tail scarcity profile. This detailed breakdown allows for an in-depth analysis of how each strategy combination affects individual class performance, particularly for rare versus common classes.

Class	OrganCMNIST Full dataset			OrganCMNIST Long-tail			OrganCMNIST Extreme long-tail		
	Base	Best	Δ	Base	Best	Δ	Base	Best	Δ
Class 0	0.879	0.960	+0.082	0.851	0.955	+0.105	0.760	0.906	+0.146
Class 1	0.826	0.948	+0.122	0.802	0.904	+0.102	0.623	0.781	+0.158
Class 2	0.893	0.950	+0.057	0.832	0.923	+0.091	0.585	0.790	+0.205
Class 3	0.933	0.982	+0.049	0.929	0.982	+0.053	0.796	0.952	+0.155
Class 4	0.788	0.875	+0.087	0.759	0.863	+0.104	0.599	0.755	+0.156
Class 5	0.797	0.900	+0.103	0.769	0.881	+0.113	0.690	0.838	+0.148
Class 6	0.984	0.997	+0.013	0.980	0.995	+0.014	0.972	0.992	+0.019
Class 7	0.970	0.996	+0.026	0.967	0.994	+0.026	0.905	0.986	+0.081
Class 8	0.983	0.994	+0.010	0.981	0.995	+0.014	0.948	0.990	+0.042
Class 9	0.874	0.961	+0.087	0.862	0.951	+0.089	0.801	0.916	+0.115
Class 10	0.923	0.965	+0.042	0.908	0.963	+0.054	0.853	0.931	+0.078

Table 8: Per-class F1 scores for OrganCMNIST under the extreme long-tail scarcity profile.

Class	PathMNIST Full dataset			Long-tail			Extreme long-tail		
	Base	Best	Δ	Base	Best	Δ	Base	Best	Δ
Class 0	0.974	0.993	+0.019	0.953	0.994	+0.041	0.934	0.992	+0.059
Class 1	0.925	0.999	+0.074	0.897	0.998	+0.101	0.920	0.991	+0.071
Class 2	0.707	0.957	+0.250	0.695	0.942	+0.246	0.533	0.895	+0.363
Class 3	0.964	0.994	+0.030	0.951	0.991	+0.040	0.872	0.991	+0.119
Class 4	0.925	0.982	+0.057	0.818	0.982	+0.164	0.847	0.970	+0.123
Class 5	0.768	0.911	+0.143	0.802	0.905	+0.103	0.612	0.901	+0.289
Class 6	0.924	0.977	+0.052	0.859	0.975	+0.116	0.535	0.940	+0.405
Class 7	0.545	0.796	+0.251	0.565	0.812	+0.247	0.195	0.715	+0.520
Class 8	0.926	0.971	+0.045	0.898	0.965	+0.068	0.799	0.946	+0.147

Table 9: Per-class F1 scores for PathMNIST under the extreme long-tail scarcity profile.

Class	RetinaMNIST Full dataset			Long-tail			Extreme long-tail		
	Base	Best	Δ	Base	Best	Δ	Base	Best	Δ
Class 0	0.768	0.788	+0.020	0.738	0.788	+0.050	0.679	0.790	+0.111
Class 1	0.215	0.362	+0.147	0.205	0.355	+0.150	0.013	0.367	+0.354
Class 2	0.382	0.494	+0.112	0.436	0.478	+0.042	0.427	0.511	+0.084
Class 3	0.389	0.537	+0.148	0.264	0.516	+0.252	0.053	0.443	+0.390
Class 4	0.000	0.417	+0.417	0.130	0.378	+0.248	0.030	0.359	+0.329

Table 10: Per-class F1 scores for RetinaMNIST under the extreme long-tail scarcity profile.