Generalization under Sub-population Shift: Equitable Models for Imbalanced, Long-tailed, and Fair Representation Learning

A thesis submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
in Computer Science

by

Faizanuddin Ansari

under the supervision of

Prof. Swagatam Das

Electronics and Communication Sciences Unit
Indian Statistical Institute

July, 2025

To Al-Hadi,
the One who guides to knowledge,
and to my parents,
the source of endless love and dua.

Acknowledgements

As I reach the final pages of this thesis, I find myself standing on a mountain of support, built over years by the kindness, brilliance, and companionship of many people. It's a humbling reminder that no PhD is ever a solo journey.

First and foremost, I would like to express my deepest gratitude to my supervisors, Professor Swagatam Das. Thank you for your guidance, insightful feedback, and the freedom to explore my own ideas (and sometimes, the patience to watch them fail!).

To my collaborators beyond the walls of this institute especially the brilliant students Abhrant, Agnish, and Biswajit, and Professor Dr. Tapabrata Chakraborty, with whom I had the privilege to work thank you for bringing fresh perspectives, challenging questions, and the occasional nudge that pushed my ideas toward maturity. Working across disciplines, institutions, and time zones taught me more than I could have ever anticipated.

A special shout-out goes to my hostel mates Gaurav, Suman, Gourab, Saeed, Rakesh, Kaushik, Meghna as well as our juniors Anup and Subho, and seniors Avisek Da, Bibhuti Da and Sankar Da my extended family away from home. I also want to thank my friends Anal and Priyobrata, who, though not hostelmates, have been an equally integral part of this journey. You've been my sounding boards, midnight food partners, stress therapists, and impromptu philosophers. Thank you for the chaos, the calm, and everything in between. Life in the hostel was an emotional rollercoaster, and I wouldn't trade it for anything.

To my family, for your unwavering belief in me, for enduring long silences and missed occasions, and for being my constant source of strength, thank you.

Finally, to everyone who cheered me on from the sidelines, knowingly or unknowingly - thank you.

Here's to the journey, and the people who made it worth it.

Abstract

Machine learning systems often experience performance degradation in real-world scenarios due to *subpopulation shift* defined as mismatches in the distribution of classes or attributes within datasets. This thesis investigates generalization failures arising from class imbalance, long-tailed distributions, and attribute-level biases (specifically, attribute-level biases that originate from demographic imbalances in sensitive domains, such as medical imaging). It proposes principled strategies to mitigate these effects in both classical and deep learning frameworks. *Class imbalance* and *long-tailed distributions* pose significant challenges, especially in real-world applications where minority classes are underrepresented yet critically important. To address these challenges, this work develops novel algorithms and frameworks that enhance model generalization on imbalanced and long-tailed datasets. The contributions encompass data-level, model-level, and loss-level innovations, each designed to mitigate bias and improve performance in minority classes while maintaining accuracy in majority classes.

First, we propose a data-level solution for classical class imbalance in tabular data through a novel oversampling technique that estimates minority class statistics using *neighborhood-based distributional calibration*. Unlike existing methods that rely on synthetic interpolation without accounting for class-specific geometry, the proposed approach preserves the fidelity of minority class distributions, leading to significant gains in both binary and multi-label imbalanced settings. Next, we introduce *STTP-Net*, a two-pronged framework for long-tailed learning in vision tasks. It integrates hybrid augmentation and sampling strategies with a newly proposed *Effective Balanced Softmax (EBS)* loss to correct label distribution shifts, enabling robust feature learning and improved accuracy across head, medium, and tail classes. Extensive evaluations on benchmark datasets such as CIFAR-LT, ImageNet-LT, and NIH-CXR-LT confirm its superiority over state-of-the-art methods. We address decision boundary distortion under class imbalance by introducing the *Goldilocks principle* to achieve "just-right" boundary fidelity. Our approach leverages this concept to design a training pipeline that produces smoother, more adaptive decision boundaries for tail classes. Specifically, we propose a *Dual-Branch Sampler-Guided Mixup (DBSGM)* strategy combined with an *Adaptive Class-Aware Feature Regularization (ACFR)* mechanism. These components jointly enhance intra-class compactness and inter-class separability, improving generalization, especially under extreme imbalance. By dynamically adjusting boundaries and applying adaptive regularization, our method achieves optimal fidelity for minority classes without compromising the performance of majority classes. Extensive experiments validate its effectiveness across a range of imbalance ratios. Furthermore, we extend these ideas to medical imaging, addressing both class imbalance and demographic fairness. This includes the *Mixture of Two Experts (Mo2E)* framework and fairness-aware lesion classification strategies that ensure equitable performance across subgroups. Mo2E combines asymmetric sampling with adaptive mixup to improve the detection of rare disease classes and is validated across tasks such as Gastrointestinal (GI) Tract Classification of Endoscopic Images and Diabetic Retinopathy (DR) grading.

Additionally, we introduce a *bias-aware training method* to mitigate both *class imbalance and skin tone bias*, achieving fair performance across *demographic subgroups*, as demonstrated on the ASAN and ISIC-2018 datasets. These results lay the groundwork for demographically fair model design in high-stakes medical applications.

Collectively, these contributions advance the field of imbalanced learning by offering scalable, practical solutions grounded in theoretical insight and empirical validation. This thesis provides a comprehensive toolkit for researchers and practitioners confronting the challenges of subpopulation shift, integrating principled data synthesis, loss rebalancing, and fairness constraints. It pushes the frontiers of robust, fair, and generalizable deep learning, particularly in domains where class rarity and demographic underrepresentation have tangible real-world consequences.

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

ACFR	Adaptive Class-Aware Feature Regularize
ADASYN	Adaptive Synthetic Sampling
ADL	Activities of Daily Living
BACC	Balanced Accuracy
BBN	Bilateral-Branch Network
BCL	Balanced Contrastive Learni
BLT	Balancing Long-Tailed datasets
BM	Balanced Mixup
BS	Balanced Softmax Loss
CBRW	Class Balanced Reweighing.
CBS	Class Balanced Sampling.
CDBNF	Cumulative Dual-Branch Network Framework
CDSS	Clinical Decision Support Systems
CE	Cross entropy
CIL	Class Incremental Learning.
CLS	Cumulative Learning Strategy
CMO	Context Rich Minority Oversampling
CNN	Convolutional Neural Network.
CSA	Context Shift Augmentation
DBSGM	d Dual-Branch Sampler-Guided Mixup
DODA	Dynamic Optional Data Augmentation
DOS	Deep Over-Sampling
DPM	Dual-phase Model
DR	Diabetic Retinopathy.
DRW	Deferred Re-Weighting.
EBS	Effective Balanced Softmax
FL	Focal Loss
FNR	False Negative Rate
FPR	False Positive Rate
GAMO	Generative Adversarial Minority Oversampling
GI	Gastrointestinal
GM	Geometric Mean

HAM	Human Against Machine
HH	Head of Head
HM	Head of Medium
HT	Head of Tail
IR	Imbalance Ratio
ISDA	Implicit Semantic Data Augmentation
LADE	LAbel distribution DisEntangling
LR	Learning Rate
LT	a Long-tailed
LTL	Long-tailed Learning
LTR	Long-tailed Recognition
LWS	Learnable Weight Scaling
MCC	Matthews Correlation Coefficient
ML	Machine Learning
MLP	Multi-layer Perceptron
MOO	Multi Objective Optimization.
MOOSF	Multi-Objective Optimization-based Strategy Fusion
MWMOTE	Majority Weighted Minority Oversampling Technique
OLTR	Open Long-Tailed Recognition
PS	Progressive Sampling
RIDE	Routing Diverse Distribution-Aware Experts
RS	Reverse Sampling
RUS	Random Undersampling.
RW	Reweighing
SCL	Supervised Contrastive Learning
SGD	Stochastic Gradient Descent
SMOTE	Synthetic Minority Oversampling Technique
SOTA	State of the ART
SSP	Self-Supervised Pre-training
TASCL	Temperature Adaptive Supervised Contrastive Loss
TH	Tail of Head
TM	Tail of Medium
TN	True Negative
TP	True Positive
TT	Tail of Tail

List of Publications by the Author Related to the Thesis

Journal Publications

1. F. Ansari, A. Panigrahi, and S. Das. "The Goldilocks Principle: Achieving Just Right Boundary Fidelity for Long-Tailed Classification." *IEEE Transactions on Emerging Topics in Computational Intelligence*, pp. 1–15, 2025. doi: https://doi.org/10.1109/TETCI.2025.3551950.
2. F. Ansari, A. Panigrahi, and S. Das. "Sampling-Tailored Two-Pronged Network for Long-Tailed Class Imbalance Learning." *Engineering Applications of Artificial Intelligence*, 2025. doi: https://doi.org/10.1016/j.engappai.2025.111466.

Conference Publications

1. F. Ansari, S. Das, and P. Shamsolmoali. "Handling Class Imbalance by Estimating Minority Class Statistics." In *2023 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2023. doi: https://doi.org/10.1109/IJCNN54540.2023.10191975.
2. F. Ansari, A. Bhattacharya, B. Saha, and S. Das. "Mo2E: Mixture of Two Experts for Class-Imbalanced Learning from Medical Images." In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2024. doi: https://doi.org/10.1109/ISBI56570.2024.10635212.
3. F. Ansari, T. Chakraborti, and S. Das. "Algorithmic Fairness in Lesion Classification by Mitigating Class Imbalance and Skin Tone Bias." In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, eds. M. G. Linguraru et al., pp. 373–382, Springer, Cham, 2024. doi: https://doi.org/10.1007/978-3-031-72378-0_35.

Chapter 1

Introduction

Synopsis

This introductory chapter frames the central challenge of subpopulation shift in machine learning (ML), a structured form of distributional shift that leads to uneven model performance across data subgroups, especially in high-stakes domains like healthcare. The chapter begins by defining subpopulation shift and illustrating its manifestations semantic, attribute, and class prior shifts followed by a taxonomy of four main types: spurious correlation, class imbalance, attribute imbalance, and attribute generalization. It then outlines the specific problem settings addressed in this thesis, focusing on class imbalance (in both tabular and long-tailed domains) and attribute generalization, particularly in fairness-critical settings. The subsequent sections discuss the foundations and motivation behind class imbalance, including its unique challenges in tabular data and deep learning, and elaborate on long-tailed learning as a severe, structured variant of class imbalance. Strategies ranging from sampling and reweighting to augmentation and representation learning are reviewed, along with the limitations that motivate the thesis contributions. The chapter further explores attribute generalization and fairness, with a focus on skin tone disparities in dermatological imaging. Finally, the chapter summarizes the core contributions of each thesis chapter.

1.1 Subpopulation Shift: The Central Challenge

The rapid integration of machine learning (ML) systems into high-stakes decision-making pipelines such as healthcare diagnostics (Ahsan et al., 2022), judicial risk assessments (Berk and Elzarka, 2020), algorithmic trading (Bhuiyan et al., 2025), loan approval systems (Nwafor et al., 2024), and autonomous vehicle control (Emory et al., 2022) has raised pressing concerns about their reliability, equity, and generalizability. While these models often demonstrate impressive overall performance metrics during validation and benchmarking, such metrics can mask a deeper systemic issue: a marked **performance disparity across subpopulations**. This disparity arises from a subtle yet pervasive phenomenon known as *subpopulation shift* a structured form of distributional shift

where the joint or marginal distributions of data subgroups encountered during deployment differ significantly from those seen during training (Figure 1.1, 1.5). In particular, some subpopulations defined by combinations of class labels, attributes, or latent factors such as demographic identifiers, background cues, or context-specific features are either sparsely represented or entirely absent in the training data. Consequently, ML models optimized for global accuracy may inadvertently overfit to majority patterns, while failing to generalize to underrepresented or rare subgroups. The result is a dangerous illusion of competence: models may exhibit high average accuracy yet simultaneously perform poorly sometimes catastrophically on critical minority cases. This failure mode not only degrades predictive performance but also poses severe ethical and societal risks by exacerbating inequities, undermining user trust, and violating principles of fairness and robustness that are fundamental to responsible AI deployment.

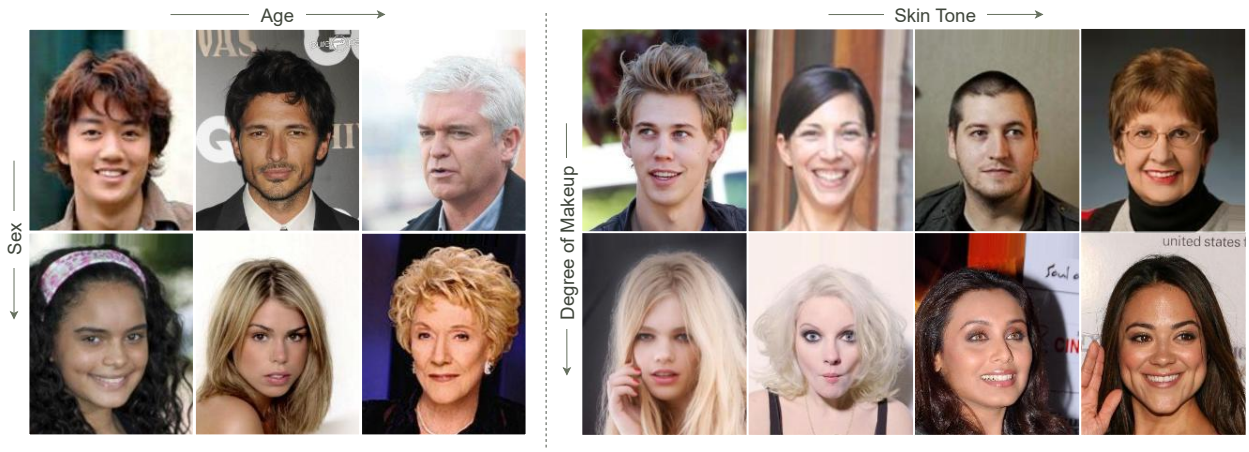


Figure 1.1: Illustration of subpopulation shift where underrepresented attribute combinations (e.g., young males, makeup on pale skin) pose significant generalization challenges for learning algorithms. The examples are sampled from the CelebA (Liu et al., 2015) dataset and showcase how specific intersections of class labels and attributes such as age, gender, makeup usage, and skin tone proxies may be underrepresented during training, leading to biased or inconsistent performance in real-world settings.

1.1.1 What is Subpopulation Shift?

Subpopulation shift refers to a structured form of distribution shift where the proportions or characteristics of distinct data subgroups differ between the training and deployment environments. Unlike traditional notions of distribution shift where global covariate or label shifts are considered subpopulation shift focuses on changes at the *granularity of subgroups*, which are defined by combinations of labels and attributes (e.g., class–gender pairs, class–background pairs, or class–demographic pairs). Despite achieving high average performance, machine learning (ML) models frequently fail to maintain consistent behavior across these subpopulations, often underperforming on the underrepresented or previously unseen ones (Yang et al., 2023).

Formally, let $x \in \mathcal{X}$ denote input features, $y \in \mathcal{Y}$ the class labels, and $a \in \mathcal{A}$ the auxiliary attributes (such as gender, race, background, or context). These elements collectively define a subgroup $g \in \mathcal{G}$ through a mapping $h : \mathcal{A} \times \mathcal{Y} \to \mathcal{G}$. The training distribution is expressed as a mixture over group-specific distributions:

$$P_{\text{src}} = \sum_{g \in \mathcal{G}} \alpha_g P_g,$$

where α_g denotes the mixing proportion of group g in the training data. At test time, the model encounters a potentially shifted distribution:

$$P_{\text{tar}} = \sum_{g \in \mathcal{G}} \beta_g P_g,$$

where the proportions β_g deviate from those seen in training. Importantly, while the conditional distributions P_g may remain fixed, the shift in mixing proportions ($\alpha_g \neq \beta_g$) can dramatically alter subgroup performance, especially when models overly rely on features that are correlated with specific groups.

To analyze this shift, Yang et al. (Yang et al., 2023) propose a structured factorization of the classification model via Bayes' theorem. Let x_{core} be the invariant, label-defining core features of x (e.g., shape of a bird or structure of a lesion), and a represent the contextual or spurious attributes. The posterior $P(y|x)$ can be decomposed as:

$$P(y|x) = \underbrace{\frac{P(x_{\text{core}}|y)}{P(x_{\text{core}})}}_{\text{Semantic Shift}} \cdot \underbrace{\frac{P(a|y, x_{\text{core}})}{P(a|x_{\text{core}})}}_{\text{Attribute Shift}} \cdot \underbrace{P(y)}_{\text{Class Prior Shift}}. \tag{1.1}$$

This decomposition reveals the three critical axes along which subpopulation shift can manifest:

1. **Semantic Shift (Pointwise Mutual Information)**: The core distribution $P(x_{\text{core}}|y)$ may differ across environments if the semantic content is non-invariant. However, in many scenarios, this component is assumed stable, allowing us to focus on the remaining two components.
2. **Attribute Shift**: Even if x_{core} remains fixed, the distribution of attributes a given the core and the label, $P(a|y, x_{\text{core}})$, may vary. When $P(a|y, x_{\text{core}}) \gg P(a|x_{\text{core}})$, the model may mistakenly associate a with y, resulting in spurious correlations that fail under shift.
3. **Class Prior Shift**: The marginal class distribution $P(y)$ may be skewed, particularly in long-tailed datasets. This imbalance biases the model towards majority classes, worsening performance on the tail.

The attribute term can further be decomposed under a conditional independence assumption

as:

$$\frac{P(a|y, x_{\text{core}})}{P(a|x_{\text{core}})} = \prod_i \frac{P(a_i|y, x_{\text{core}})}{P(a_i|x_{\text{core}})},$$

enabling fine-grained analysis of which attributes contribute most to shift.

This framework explains a variety of real-world phenomena:

- In medical imaging, attributes such as skin tone or lighting may dominate predictions despite being irrelevant to the diagnosis.
- In natural image classification, background features (e.g., "water" for waterbirds) may correlate with class labels in training but not deployment.
- In textual classification, dialectal variations (e.g., African-American English) can lead to model bias due to attribute imbalance.

Subpopulation shift thus encapsulates not a single failure mode, but a spectrum of nuanced and overlapping biases. The strength of this decomposition lies in its ability to attribute prediction error to the origin of distributional misalignment, guiding the development of robust and equitable ML systems. Figure 1.2 illustrates the contrast between the classic i.i.d. assumption where subpopulations are evenly represented in both training and test data and real-world subpopulation shift scenarios. These shifts introduce distributional mismatches in attribute-label combinations, including spurious correlations, attribute or class imbalance, and missing subgroups, all of which are discussed in detail in the next section. As a result, models trained under traditional assumptions may experience significant generalization failures on underrepresented or unseen subgroups, leading to uneven performance across populations.

1.1.2 Main Types of Subpopulation Shifts

Subpopulation shift is not a monolithic phenomenon. Instead, it manifests through a spectrum of interacting factors that distort the joint distribution of features, labels, and attributes between training and deployment settings. Based on recent literature (Yang et al., 2023), as well as the empirical patterns observed in real-world datasets, four principal types of subpopulation shifts exists: spurious correlation, class imbalance, attribute imbalance, and attribute generalization. These are visually represented in Figure 1.3a.

1. **Spurious Correlation**: Spurious correlations refer to non-causal associations between features and target labels that emerge in the training data due to confounding factors or environmental artifacts. These correlations may offer strong predictive signals during training but do not hold under distribution shift. For example, a model trained to classify birds might learn to associate "water background" with "waterbirds" simply because most training images of waterbirds appear on water. If a waterbird appears on land at test time, the model fails. This form of shortcut learning undermines the model's ability to generalize and can lead to brittle predictions when environmental conditions change.

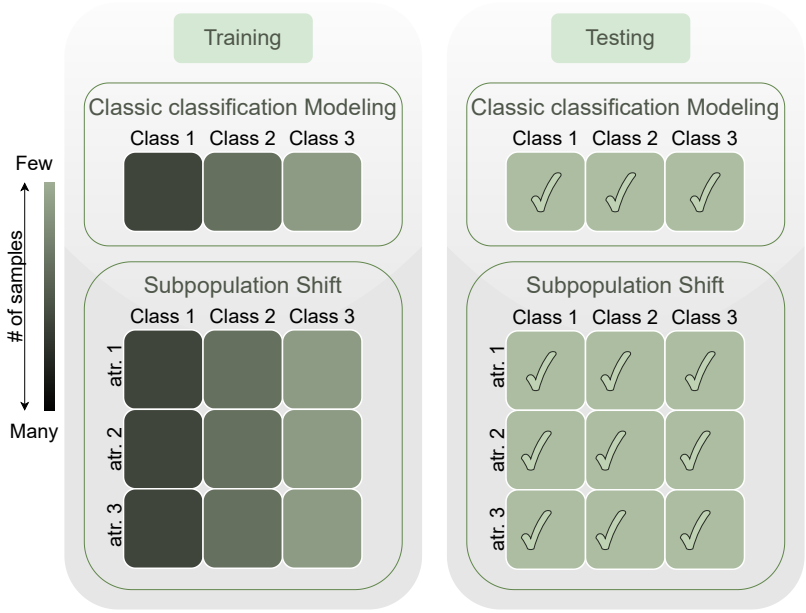
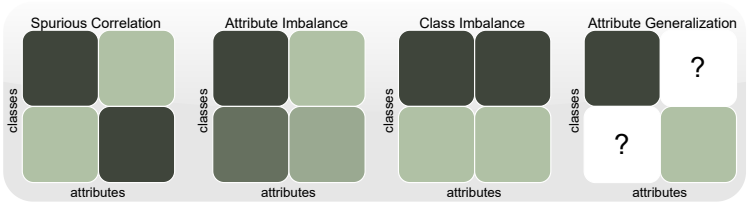
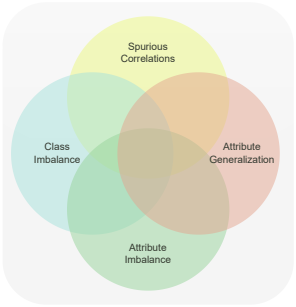


Figure 1.2: Illustrating the contrast between classic classification settings and subpopulation shift. While traditional modeling assumes uniform subgroup-class representation across training and test data, subpopulation shift introduces mismatches in attribute and label distributions, leading to unseen or underrepresented combinations at test time and performance disparities across subgroups.

2. **Class Imbalance Learning**: Class imbalance occurs when the frequency distribution of labels is skewed some classes (often referred to as head classes) have many instances, while others (tail classes) are underrepresented. This imbalance biases learning algorithms toward the majority classes, as the optimization objective (e.g., cross-entropy loss) is disproportionately influenced by frequent classes. Consequently, the model achieves high overall accuracy but performs poorly on minority classes. This problem is particularly exacerbated in long-tailed learning settings, where the number of tail classes is large, and each has very few samples.
3. **Attribute Imbalance**: Beyond label distributions, imbalance can also arise from features or attributes correlated with the labels. Attribute imbalance refers to cases where certain attribute values or combinations (e.g., gender, age, skin tone) are disproportionately represented in the dataset. For example, if most images of a specific disease show male patients with lighter skin tones, the model may implicitly learn to associate those attributes with the disease. When presented with female patients or darker skin tones, the model's performance may degrade due to lack of representation during training. This kind of imbalance often leads to biased and unfair decision-making across demographic groups.
4. **Attribute Generalization**: Attribute generalization captures the model's ability to apply learned representations to novel combinations of attributes not seen during training. This is particularly relevant in tasks where attribute-label pairs are sparse or combinatorially large.



(a) Taxonomy of subpopulation shifts. Each quadrant represents a distinct but often overlapping mode of data distribution misalignment that can impair model performance across subgroups.



(b) Cases arising from intersections of subpopulation shift types. Each component corresponds to a distinct failure mode requiring targeted intervention.

Figure 1.3: Taxonomy of subpopulation shifts and cases arising from their intersections.

A model with poor attribute generalization may overfit to specific training patterns and fail to correctly classify inputs from unseen attribute configurations. For example, in a facial recognition task, if the training data only includes images of young males wearing glasses, the model might not generalize to older females with glasses. Ensuring robust generalization requires not only feature-level disentanglement but also diverse and balanced data composition.

These four shift categories are not mutually exclusive. In fact, they frequently co-occur and interact in real-world settings (as shown in Figure 1.3b), amplifying model vulnerabilities. For instance, a class imbalance setting may also involve spurious background features that vary across demographic groups, or unseen attribute combinations during testing. Hence, a holistic understanding of these subpopulation shift types is crucial for designing robust and fair ML models.

1.1.3 Shifts Addressed in This Thesis

Within the broader taxonomy of subpopulation shift as discussed above, our research is focused on addressing a subset of challenges that are particularly prevalent and consequential in practical machine learning systems. These include class imbalance learning (both in tabular and long-tailed settings), and the challenge of attribute generalization.

The specific challenges tackled in this thesis are as follows:

1. **Class Imbalance Learning**: A persistent issue across many real-world datasets is the unequal representation of classes, leading to biased learning dynamics that favor the majority classes. We address this from two complementary perspectives:
   - **Tabular Data Imbalance**: Many practical applications such as credit scoring, medical diagnosis, and fraud detection rely on structured tabular data, which frequently suffers from severe class imbalance. We propose novel sampling and distribution modeling

techniques to enhance minority class representation.

- **Long-Tailed Imbalance**: In vision tasks, data often follows a long-tailed distribution where a few head classes dominate and many tail classes have scarce data. We design learning strategies that balance feature learning and classifier bias to improve generalization across the entire class spectrum, with a special emphasis on maintaining performance on rare classes.

2. **Attribute Generalization**: Subpopulations can differ along dimensions of demographic attributes, environmental conditions, or feature contexts. A key challenge is enabling the model to generalize beyond the attribute configurations seen during training. This is critical in applications where new attribute combinations appear at inference time (e.g., presence of skin tones in the test which is not present in the training set in medical skin diagnosis dataset). We address this by learning disentangled and transferable representations that are robust across attribute shifts, allowing for better generalization across unseen or rare subgroups. Importantly, this capability is closely tied to fairness models that fail to generalize to underrepresented attribute combinations often exhibit biased performance, disproportionately affecting marginalized or rare subpopulations. By improving attribute generalization, we also promote equitable treatment across demographic variations.

In summary, the thesis systematically addresses the data, model, and representation-level factors contributing to performance degradation under subpopulation shift. By tackling class imbalance, and enhancing attribute generalization, our goal is to build deep learning systems that are not only accurate but also resilient, fair, and deployable in real-world heterogeneous environments. In the following sections, we will discuss class imbalance and attribute generalization subpopulation shifts in existing literature, and also give the overview of the algorithms proposed in this thesis.

1.2 Class Imbalance: Background and Motivation

Class imbalance is a pervasive challenge in machine learning that significantly affects model performance, generalization, and fairness. This issue arises when certain classes are overrepresented while others appear infrequently, resulting in skewed learning dynamics. In traditional machine learning applications particularly with tabular data such as in credit scoring, fraud detection, or medical diagnostics imbalanced class distributions are common and have prompted the development of techniques like resampling, cost-sensitive learning, and ensemble strategies to rebalance the data. With the rise of deep learning, the problem of class imbalance has taken on new dimensions, particularly in high-dimensional data such as images. Unlike tabular models that rely on explicit feature engineering, deep networks learn representations directly from data, and hence are more vulnerable to skewed distributions. In vision-based tasks, this often results in biased feature extractors and decision boundaries that favor majority classes, especially when the imbalance is severe.

This vulnerability becomes especially pronounced in **long-tailed learning** a setting where class frequencies follow a power-law distribution, with a small number of "head" classes dominating the dataset and a large number of "tail" classes having very few samples. Unlike standard imbalance settings that involve a few minority classes, long-tailed scenarios present deeper representational and optimization challenges, including poor generalization to rare classes, overfitting to head-class patterns, and degraded performance in medium-frequency regimes.

To systematically address these challenges, we begin with the study of class imbalance in structured tabular datasets and progressively transition to long-tailed recognition in deep learning. In the subsections that follow, we provide a detailed discussion of these different paradigms namely, *Class Imbalance in Tabular Data*, *Class Imbalance in Deep Learning*, *Understanding Long-Tailed Distributions*, and *Long-Tailed Learning in Deep Networks* which together lay the foundation for our proposed solutions.

1.2.1 Class Imbalance in Tabular Data

Building on the broader motivation established in the previous section, we now turn our attention to the specific context of class imbalance in tabular datasets. Unlike high-dimensional data such as images or text, tabular datasets characterized by structured, columnar representations pose distinct challenges and opportunities for imbalance handling. In such settings, conventional learning algorithms are particularly prone to biasing toward the majority class due to uniform loss weighting and instance-driven optimization.

This subsection surveys key algorithmic strategies designed to mitigate class imbalance in tabular data, with a primary focus on oversampling techniques that synthetically augment the minority class. We begin with foundational methods such as SMOTE and its direct extensions, followed by more recent clustering-based and adaptive sampling approaches that aim to better capture the underlying data distribution. These methods vary in their assumptions, sampling strategies, and robustness to noise, making their comparative analysis critical for practical deployment. We conclude the subsection by motivating our proposed distributionally calibrated oversampling method, which builds on these foundations to address their limitations in complex tabular settings.

Among the earliest and most impactful solutions for this problem are oversampling methods, particularly those based on synthetic data generation. One of the most influential of these is the Synthetic Minority Over-sampling Technique (SMOTE), introduced by (Chawla et al., 2002a). SMOTE generates new synthetic samples by interpolating between existing minority class examples and their nearest neighbors. While this reduces overfitting compared to simple replication, it can still lead to noisy or less-informative samples, particularly when synthetic points are generated far from the true decision boundary.

To address these limitations, several enhancements to SMOTE have been proposed. Borderline-SMOTE (Han et al., 2005a) focuses on samples near the class boundary, aiming to generate more

decision-relevant examples. Distance SMOTE (de la Calleja and Fuentes, 2007) modifies the interpolation strategy by leveraging the mean vector of the nearest neighbors, thereby improving representational fidelity. ADASYN (He et al., 2008) further adapts the sampling strategy by generating more synthetic points in harder-to-learn regions. Another variant, SVM-SMOTE (Nguyen et al., 2011), integrates Support Vector Machines to guide the interpolation process toward the classifier's margin.

More recently, clustering-based oversampling approaches have gained prominence. These methods leverage the underlying structure of the minority class by organizing instances into clusters and generating synthetic data accordingly. Two general strategies exist: (i) clustering applied to only the minority class, and (ii) clustering applied to the entire dataset.

In the first category, ClusterSMOTE (Cieslak et al., 2006) applies clustering to the minority class and oversamples within each cluster. However, it requires prior knowledge of the number of clusters and the number of samples to generate. DBSMOTE (Bunkhumpornpat et al., 2012), based on the DBSCAN algorithm (Ester et al., 1996), constructs clusters and uses graph-based path computation (Dijkstra's algorithm) to guide sample generation. CURE-SMOTE (Ma and Fan, 2017) applies the CURE hierarchical clustering algorithm to identify representative minority samples but risks discarding valuable boundary instances. SOICJ (Sánchez et al., 2013) estimates local standard deviations to generate jittered synthetic samples, improving fidelity but still neglecting global data distribution.

In the second category, oversampling is performed after clustering the entire data space. Kmeans SMOTE (Xu et al., 2021) uses K-means clustering to exclude noisy minority-dominated clusters and applies SMOTE in relevant clusters. MSMOTE (Ghorab et al., 2022) builds on MeanShift clustering (Comaniciu and Meer, 2002) to locate high-density regions and generates samples accordingly. CE-SMOTE (Chen et al., 2010) utilizes cluster ensembles to identify boundary regions for targeted oversampling. MWMOTE (Barua et al., 2012), on the other hand, assigns weights to minority instances based on proximity to majority-class neighbors before generating samples via SMOTE; however, it may over-filter noisy instances, leading to information loss.

Despite the progress made by these techniques, several limitations remain. Many of the aforementioned methods either overfit to sparse regions, generate samples that do not reflect the true minority distribution, or fail to consider the interaction with majority-class data points. Additionally, clustering-based approaches often rely on heuristics or hyperparameters (e.g., number of clusters), which can reduce robustness across datasets.

A summary of the notable methods discussed above is provided in Table 1.1.

To overcome these limitations, our proposed method (Ansari et al., 2023) introduces a distributionally-aware oversampling technique that calibrates synthetic generation based on neighborhood statistics and class-conditional densities. This allows for more faithful sample generation

Table 1.1: Summary of some Notable Oversampling Methods for Tabular Class Imbalance

Method	**Core Idea**
SMOTE (Chawla et al., 2002a)	Interpolates between minority instances and nearest neighbors
Borderline-SMOTE (Han et al., 2005a)	Focuses on borderline instances to create samples near the decision boundary
Distance SMOTE (de la Calleja and Fuentes, 2007)	Uses mean of neighbors for smoother sample generation
ADASYN (He et al., 2008)	Adaptive generation of samples in difficult regions
SVM-SMOTE (Nguyen et al., 2011)	Uses SVM margins to guide sample generation
ClusterSMOTE (Cieslak et al., 2006)	Clusters minority class, oversamples within clusters
DBSMOTE (Bunkhumpornpat et al., 2012), (Ester et al., 1996)	DBSCAN-based clustering, shortest-path interpolation
CURE-SMOTE (Ma and Fan, 2017)	Uses CURE clustering to select core samples
SOICJ (Sánchez et al., 2013)	Uses local and global variance to jitter synthetic points
Kmeans SMOTE (Xu et al., 2021)	Clusters entire data using K-means, filters noisy clusters
MSMOTE (Ghorab et al., 2022), (Comaniciu and Meer, 2002)	MeanShift clustering to identify dense regions
CE-SMOTE (Chen et al., 2010)	Uses cluster ensembles to detect boundaries
MWMOTE (Barua et al., 2012)	Weighs minority points by majority proximity

in tabular settings, improving both classification performance and minority class representation. A detailed formulation and evaluation of this method will be presented in the subsequent chapter dedicated to tabular class imbalance learning.

1.2.2 Class Imbalance in Deep Learning

Having established the broader challenge of class imbalance across tabular dataset, we now focus on methods developed specifically for deep learning, where the complexity and high dimensionality of the data demand novel solutions. Unlike classical machine learning on tabular data, deep neural networks learn internal representations from raw data and are particularly sensitive to skewed class distributions. Consequently, imbalance in deep learning introduces challenges not only at the label level but also in the learned feature space, often resulting in biased representations and poor generalization to minority classes.

Among the prominent strategies to mitigate class imbalance in deep learning is the use of **deep generative models for oversampling**. These models aim to synthesize realistic data points in either the original input space or a learned latent space. Popular architectures such as the Variational Autoencoder (VAE) (Kingma and Welling, 2013) and the Generative Adversarial Network (GAN) (Goodfellow et al., 2014) have been widely adopted in this context. However, VAEs often struggle with generating detailed, high-fidelity samples, due to the limitations of the evidence lower bound (ELBO) objective, which often leads to blurred outputs in sparse data regions

Bredell et al. (2023). Similarly, GANs suffer from mode collapse and training instability when the discriminator easily overpowers the generator on underrepresented classes Bredell et al. (2023).

To improve upon this, Guo et al. (Guo et al., 2019) proposed modeling the latent representation using two Gaussian distributions with opposing means, tailored for binary classification. However, this approach does not generalize to multiclass imbalance scenarios. Conditional GANs (cGANs) (Gauthier, 2014) extend GANs to generate class-specific minority samples, yet they can suffer from feature entanglement and disruption of orientation-sensitive information due to the randomness in noise vectors. To mitigate this, BAGAN (Mariani et al., 2018) introduced a two-stage framework that first uses an Autoencoder (AE) to capture global structure, followed by a cGAN to generate samples in the latent space.

These models typically follow a two-step procedure: one module generates synthetic data, and another separately trains a classifier. This separation may result in a distributional mismatch between synthetic and real data. Addressing this, the Generative Adversarial Minority Oversampling (GAMO) method (Mullick et al., 2019) proposed a three-player adversarial game involving a convex generator, a classifier, and a discriminator, ensuring that new samples lie within the convex hull of minority class samples. Similarly, Arnab et al. (Mondal et al., 2023) proposed training a regularized autoencoder followed by synthetic data generation via convex combinations of latent vectors from minority samples. However, both methods rely on aggregating all feature vectors to generate even a single sample, reducing flexibility and scalability.

To address these limitations, DeepSMOTE (Dablain et al., 2021) was introduced. It employs an autoencoder trained with reconstruction and permutation losses to preserve class information in the latent space. Minority samples are then oversampled using the classical SMOTE algorithm in the encoded space. While more efficient, this method still does not explicitly consider the underlying data distribution in the sample space.

To explicitly address the distributional characteristics of class imbalance, Wang et al. proposed Deep Generative Classification (DGC) (Wang et al., 2020), which combines Bayesian inference with a mixture-model-based generator to synthesize class-conditional samples. This framework captures the complexity of both majority and minority class distributions and incorporates a discriminator to estimate class priors. This approach was later extended in DCGMM (Wang et al., 2022b), which integrates Gaussian mixture modeling in the latent space for improved class separation and sample fidelity.

A summary of the notable methods discussed above is provided in Table 1.2.

While these methods offer promising solutions to class imbalance in deep learning, they suffer from several key drawbacks. Many of them rely on decoupled generation and classification stages, which can lead to distributional mismatches between synthetic and real samples. Others require access to the entire class-wise sample set during generation, limiting their scalability and

Table 1.2: Summary of Notable Deep Learning-Based Imbalance Methods

Method	**Core Idea**
VAE (Kingma and Welling, 2013)	Latent generative modeling for sample synthesis
GAN (Goodfellow et al., 2014)	Adversarial generation of realistic samples
Discriminative VAE (Guo et al., 2019)	Latent modeling with opposite Gaussian priors (binary)
cGAN (Gauthier, 2014)	Class-conditional sample generation using GANs
BAGAN (Mariani et al., 2018)	Combines AE and cGAN for minority class generation
GAMO (Mullick et al., 2019)	Three-player game using convex generator for minority oversampling
Latent Mixing	Convex combination of autoencoder latent codes
DeepSMOTE (Dablain et al., 2021)	SMOTE in latent space of autoencoder with permutation loss
DGC (Wang et al., 2020)	Bayesian inference and generative mixture modeling
DCGMM (Wang et al., 2022b)	Extension of DGC using Gaussian mixture models

adaptability in more complex settings. Moreover, most of these approaches lack explicit modeling of the underlying sample-space density, resulting in less representative or redundant synthetic data.

Crucially, these methods are typically designed for binary or few-class imbalance scenarios and do not generalize well to *long-tailed distributions*. In such settings, assumptions like representative minority clusters or sufficient latent coverage often break down, leading to poor generalization and collapsed decision boundaries for rare classes. These limitations underscore the need for specialized strategies that can model inter-class relationships, account for severe sample sparsity, and ensure fair performance across the tail of the distribution.

1.2.3 Understanding Long-Tailed Distributions

The concept of *long-tailed distributions* represents a particularly severe and practically pervasive instance of class imbalance, where the class frequency follows a heavy-tailed or power-law distribution (as shown in Figure 1.4, the HyperKvasir dataset exhibits a long-tailed class distribution with severe imbalance across multiple gastrointestinal conditions, while the EyePACS dataset represents a normal imbalance scenario with fewer diabetic retinopathy classes and relatively milder skew.). This section explores the formal characteristics that distinguish long-tailed learning from generic class imbalance and contextualizes it within the broader framework of subpopulation shift.

Definition and Motivation. In standard multiclass classification, class imbalance refers to scenarios where some classes (majority or head classes) are significantly overrepresented compared to others (minority or tail classes). Long-tailed learning (LTL) extends this imbalance to a more extreme and structured regime, where the class frequency n_c for each class $c \in \mathcal{C} = \{1, 2, ..., C\}$

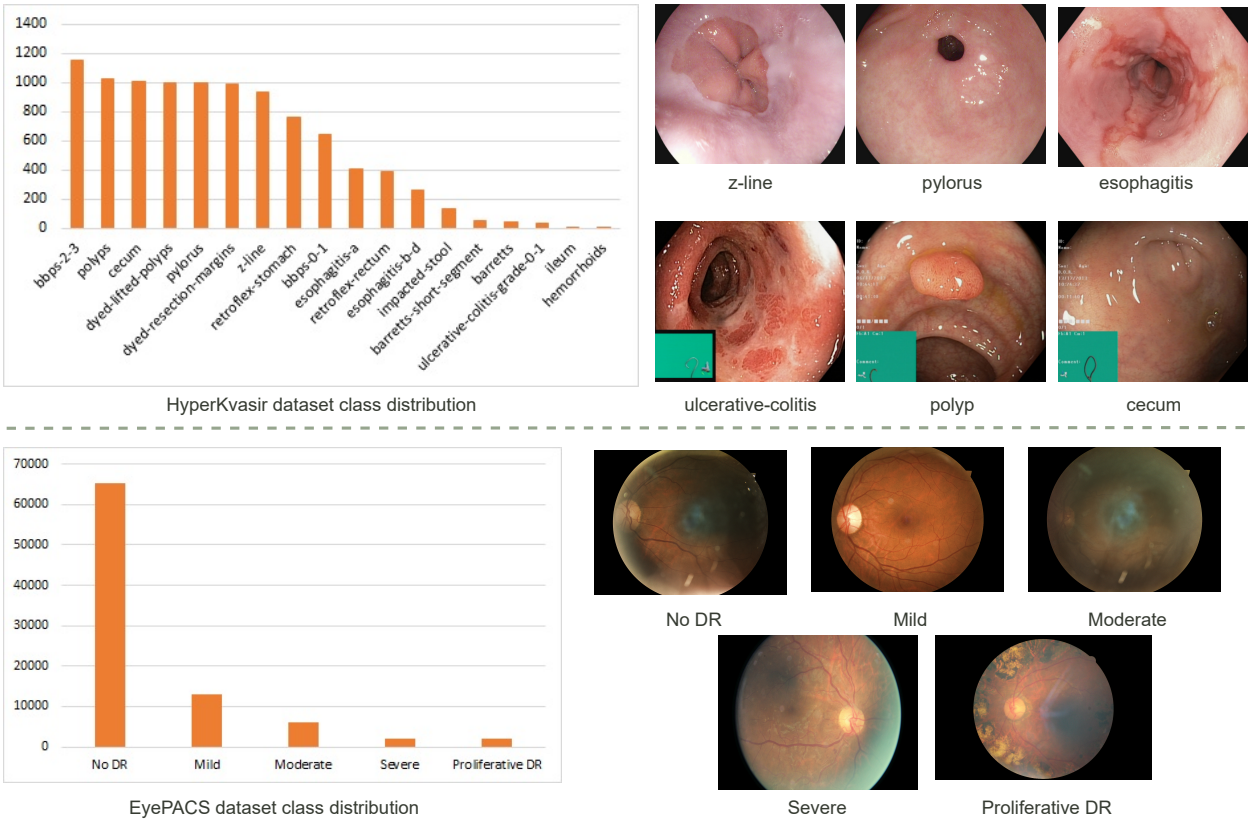


Figure 1.4: Illustration of long-tailed distribution in real-world medical datasets. The EyePACS dataset (Dugas et al., 2015) (bottom) shows an extreme imbalance across diabetic retinopathy severity levels, with a predominance of "No DR" cases. Similarly, the HyperKvasir (Borgli et al., 2020) dataset (top) demonstrates a skewed distribution across gastrointestinal conditions, reflecting the long-tailed nature of diagnostic categories in clinical imaging.

satisfies a power-law decay (Zhang et al., 2025):

$$n_c \propto c^{-\alpha}, \quad \alpha > 0, \tag{1.2}$$

such that a few classes dominate the sample space while a large number of classes (the "tail") have very limited examples. A long-tailed distribution is characterized not merely by a difference in instance counts between two classes, but by a continuous, skewed decay across the entire label spectrum where the 'head' (few classes with high frequency) contains the majority of samples, while the 'tail' (majority of classes) contains a small fraction of the total data. This distribution is common in real-world settings such as ecological recognition, rare disease classification, or product recommendation, where head classes provide strong gradient signals while tail classes are crucial for generalization and fairness (Zhang et al., 2025).

Mathematical Formulation. Let the input-label pair (x, y) be sampled from a distribution

$P(x, y)$, and denote $P(y = c) = p_c$ as the class prior for class c. In a long-tailed distribution, the priors $\{p_c\}$ are highly skewed such that:

$$\exists c_i, c_j \in \mathcal{C}, \text{ such that } \frac{p_{c_i}}{p_{c_j}} \gg 1. \tag{1.3}$$

Unlike balanced settings where $P(y = c) \approx \frac{1}{C}$, long-tailed scenarios present priors where

$$\sum_{c \in \text{head}} p_c \gg \sum_{c \in \text{tail}} p_c, \tag{1.4}$$

and yet the tail may comprise the majority of the class set: $|\text{tail}| \gg |\text{head}|$.

Contrast with Generic Class Imbalance. While both generic imbalance and long-tailed learning involve skewed class distributions, their *statistical and algorithmic challenges differ*:

- **Class Imbalance** typically involves a small number of minority classes (often binary or few-class problems), where resampling or reweighting methods (e.g., SMOTE, Focal Loss) can be effective.
- **Long-Tailed Learning** involves a large number of tail classes, each with extremely few examples. This leads to:
  1. **Sparse tail-class representations**, limiting feature learning.
  2. **Inconsistent decision boundaries**, due to insufficient intra-class variance modeling.
  3. **Overfitting to head classes**, driven by biased optimization trajectories under empirical risk minimization.

Subpopulation Shift Perspective. Following the formulation in (Yang et al., 2023), let each data point be defined by a tuple (x, y, a), where x is the input, y is the label, and a is an auxiliary attribute (e.g., demographic group, domain context). Then the classification model can be decomposed via Bayes' theorem (as already defined in 1.1):

$$P(y|x) \propto \underbrace{P(x_{\text{core}}|y)}_{\text{semantic shift}} \cdot \underbrace{\frac{P(a|y, x_{\text{core}})}{P(a|x_{\text{core}})}}_{\text{attribute shift}} \cdot \underbrace{P(y)}_{\text{class prior shift}}. \tag{1.5}$$

In long-tailed settings, the final term $P(y)$ is highly imbalanced and forms the *class prior shift*, a dominant component of the subpopulation shift. Importantly, when combined with *attribute imbalance* (e.g., some class–attribute pairs are never seen), long-tailed learning becomes even more challenging due to sparse coverage in the joint (a, y) space.

Implications for Model Learning. Long-tailed learning induces both statistical and optimization challenges. The head classes dominate gradient updates, causing feature extractors and classifiers to align with majority patterns. Meanwhile, tail classes may either be ignored or clustered

ambiguously, leading to:

$$\mathbb{E}_{(x,y)\sim\text{head}}[\ell(f(x),y)] \ll \mathbb{E}_{(x,y)\sim\text{tail}}[\ell(f(x),y)], \tag{1.6}$$

where ℓ is the classification loss. This leads to high overall accuracy but low fairness or robustness across subgroups.

We build on this understanding to first discuss the existing LTL methods in the literature and further also summarize our developed proposed methods, which aims to learn distribution-aware and transferable representations for rare classes. In the next subsection, we present the formulation and challenges of deep learning under such long-tailed regimes.

1.2.4 Long-Tailed Imbalance in Deep Learning

Having introduced the general concept of long-tailed distributions and their critical role in subpopulation shift scenarios, this subsection delves into the specific techniques developed to address long-tailed recognition (LTR) in deep learning. The severe skew in class frequencies where a small number of head classes dominate and a large number of tail classes are sparsely represented presents both algorithmic and representational challenges that are not adequately handled by standard training procedures. This has motivated a broad spectrum of methods aiming to restore balance through resampling, reweighting, augmentation, architecture modification, and optimization strategies. In what follows, we outline and categorize these efforts based on their methodological axes.

Data Resampling. Classical approaches such as undersampling (He and Garcia, 2009; Japkowicz and Stephen, 2002a; Van Hulse et al., 2007a) and oversampling (Van Hulse et al., 2007a; Mondal et al., 2024a) remain foundational. Undersampling reduces majority class samples to match the minority distribution, often at the cost of discarding useful data. Oversampling, conversely, involves sample duplication.

Loss Reweighting and Logit Adjustment. Another category modifies the loss landscape by emphasizing minority classes. Techniques include direct instance-level or class-aware reweighting (Huang et al., 2016a; Wang et al., 2017), balanced softmax (Ren et al., 2020b), and class-balanced loss (Cui et al., 2019b). Li et al. (Li et al., 2022c) further proposed perturbing logits via Gaussian noise to alleviate representation collapse in tail classes.

Data Augmentation. Augmentation-based solutions generate diversity for underrepresented classes through transformations or blending. These include traditional geometric augmentations, Implicit Semantic Data Augmentation (ISDA) (Wang et al., 2019a) that uses semantic covariance-based augmentation, and Mixup strategies (Zhang et al., 2017b; Chou et al., 2020; Yun et al., 2019a). More recent works such as MetaSAug (Li et al., 2021) and the use of context-guided augmentations in CSA (Shi et al., 2023) attempt to overcome limitations of naive sample mixing. Park et al. (Park et al., 2022) introduce majority class backgrounds to augment minority foregrounds,

while DODA (Wang et al., 2024) dynamically learns class-wise augmentation preferences to balance both data-wise and augmentation-wise disparities.

Contrastive and Representation Learning. Contrastive learning methods focus on improving representation uniformity. Wang et al. (Wang et al., 2021a) combine supervised contrastive objectives with classification loss, whereas Li et al. (Li et al., 2022e) regularize feature space uniformity to better represent tail classes. BBN (Zhou et al., 2020) and ResLT (Cui et al., 2022) employ multi-branch architectures trained under varying distributions (e.g., reverse samplers) to mitigate feature bias. A related stream includes CDBNF (Fan et al., 2022) and Bi-F3R (Chen et al., 2023), which simultaneously emphasize tail class learning while pruning redundant head features.

Two-Stage and Decoupled Learning. Inspired by the observation that joint optimization of feature extractors and classifiers can be suboptimal under imbalance, Kang et al. (Kang et al., 2020) propose a decoupled two-stage framework. The backbone is trained on uniformly sampled data, and the classifier is fine-tuned on balanced samples. Later extensions include the integration of mixup into this paradigm (Zhou et al., 2020).

Ensemble and Multi-Expert Methods. Approaches like RIDE (Wang et al., 2021b) and BalPoE (Aimar et al., 2023b) frame LTR as a multi-expert or mixture-of-experts problem. These models either route samples through expert branches or calibrate expert logits to different class regions. While effective, they introduce significant computational costs and require access to accurate class priors.

Calibration and Misconfidence Handling. In long-tailed distributions, classifier confidence is often misaligned with actual prediction accuracy. Zhong et al. (hong et al., 2021) address this by combining shifted batch normalization with label-aware smoothing and Mixup. These adjustments aim to reduce head-class overconfidence and provide better-calibrated predictions. **Method Combinations and Strategy Fusion.** Zhao et al. (Zhao et al., 2024) present a holistic approach through Multi-Objective Optimization-based Strategy Fusion (MOOSF), combining various techniques into an optimized fusion policy. While unifying, such systems often inherit the complexity of their components and may become hard to scale.

Cross-Domain Extensions. The scope of LTR has recently extended beyond vision tasks. ImbGNN (Xu et al., 2024a) tackles structural imbalance in graph learning, while Kandpal et al. (Kandpal et al., 2023) examine the limitations of large language models in handling infrequent patterns, showing that retrieval augmentation may help mitigate long-tail failure modes in NLP.

A summary of the notable methods discussed above is provided in Table 1.3. Moreover, *more detailed literature reviews of these methods along with a discussion of their limitations are provided in Chapters 3 and 4.*

Limitations and Motivation for Our Approach. Despite the broad landscape of methods discussed above, several persistent limitations remain. Many approaches struggle to jointly address

Table 1.3: Summary of Deep Learning-Based Long-Tailed Imbalance Methods

Method / Class	**Core Idea**
Resampling	Undersampling (He and Garcia, 2009; Japkowicz and Stephen, 2002a), Oversampling (Van Hulse et al., 2007a)
Reweighting / Logit Adjust.	Balanced Softmax (Ren et al., 2020b), Class Balanced Loss (Cui et al., 2019b), Gaussian Logit Perturbation (Li et al., 2022c)
Augmentation	ISDA (Wang et al., 2019a), MetaSAug (Li et al., 2021), Mixup / Remix / CutMix (Zhang et al., 2017b; Chou et al., 2020; Yun et al., 2019a), CSA (Shi et al., 2023), DODA (Wang et al., 2024)
Representation Learning and Dual Branch Networks	Contrastive Learning (Wang et al., 2021a), Uniformity-aware Loss (Li et al., 2022e), BBN (Zhou et al., 2020), ResLT (Cui et al., 2022), CDBNF (Fan et al., 2022), Bi-F3R (Chen et al., 2023)
Decoupled Learning	Two-Stage Training (Kang et al., 2020), Decoupled Mixup (Zhou et al., 2020)
Ensemble / Expert Models	RIDE (Wang et al., 2021b), BalPoE (Aimar et al., 2023b)
Calibration	Mixup with Label Smoothing and BatchNorm (hong et al., 2021)
Strategy Fusion	MOOSF (Zhao et al., 2024)
Cross-Domain Extensions	ImbGNN (Xu et al., 2024a), Retrieval-Augmented LLMs (Kandpal et al., 2023)

class imbalance and representational collapse, particularly under extreme data sparsity. Resampling methods may distort data distributions, reweighting often induces instability, augmentation strategies can create new forms of imbalance, and ensemble-based or multi-expert frameworks typically incur high computational overhead. Moreover, contrastive learning techniques face difficulties in maintaining uniformity and discriminability in the feature space for underrepresented classes.

To overcome these challenges, we proposes a unified, efficient, and theoretically grounded framework spread across two contributions presented in (**?**) and (Ansari et al., 2025). (**?**) introduces **STTP-Net**, a dual-expert architecture trained using *Hybrid-Mixup* and class-frequency-driven sampling to simultaneously address head, medium, and tail class imbalance. Additionally, the *Effective Balanced Softmax (EBS)* loss dynamically reweights logits based on class priors to reduce classifier bias.(Ansari et al., 2025) advances this further by addressing the challenge of boundary fidelity through the *Goldilocks Principle*. It proposes a single-stage pipeline integrating (i) **Dual-Branch Sampler-Guided Mixup (DBSGM)** to enforce class-aware interpolations across the class spectrum and (ii) **Adaptive Class-Aware Feature Regularizer (ACFR)** driven by *Temperature-Adaptive Supervised Contrastive Loss (TASCL)* that balances intra-class cohesion and inter-class separation. This holistic design not only mitigates overfitting to head classes but also

fosters discriminative yet generalizable representations for tail classes.

More detailed discussions, design rationales, and empirical evaluations of the proposed methods are provided in Chapters 3 and 4.

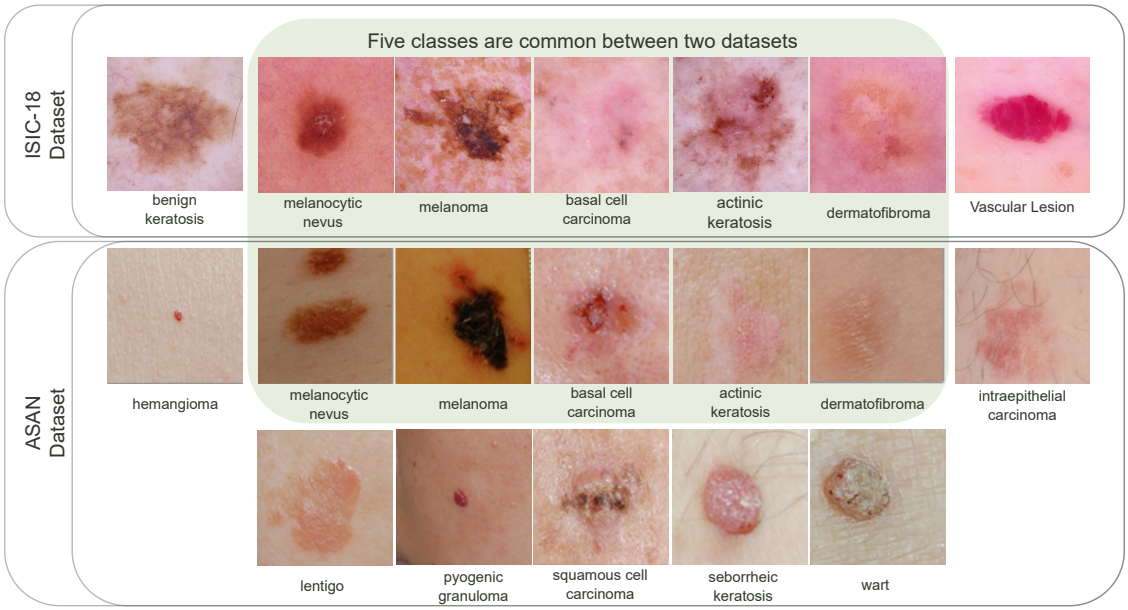


Figure 1.5: Visualization of subpopulation shift across skin tone domains using dermoscopic images from the ISIC-18 (Tschandl et al., 2018) and ASAN (Han et al., 2018) datasets. While both datasets share a set of common lesion classes (e.g., melanocytic nevus, melanoma, actinic keratosis), they differ significantly in the distribution of skin tones ISIC-18 predominantly features patients with lighter (Fitzpatrick I–II) skin, whereas ASAN includes images primarily from individuals with darker skin tones. This domain gap introduces attribute-based distribution shifts, which challenge model robustness and fairness, especially in cross-dataset generalization scenarios.

1.3 Attribute Generalization and Fairness

Machine learning models must often handle attribute shifts – changes in demographic traits, environmental conditions, or context features between training and deployment. Domain-generalization theory shows that learning invariant feature representations can improve out-of-distribution performance. In fact, if the model's features remain invariant across domains, they are provably "general and transferable" to new domains (Wang et al., 2022a). In practice, this means disentangling intrinsic task signals from attribute-specific variations (e.g., learning to ignore lighting or skin tone) so that the same model works well when attributes take novel values (Wang et al., 2022a). This is vividly illustrated in Figure 1.5, which shows a domain gap between the ISIC-18 (Tschandl et al., 2018) and ASAN (Han et al., 2018) datasets, where differences in the distribution of Fitzpatrick skin tones lead to attribute-based subpopulation shifts that impair generalization. Techniques like

adversarial invariant learning, augmentation, or meta-learning aim to enforce such disentanglement. By aligning the feature distributions of different attribute configurations during training, the model can better handle unseen combinations (for example, a disease image under a skin tone or illumination not seen in the training set).

Crucially, robust attribute generalization is tightly linked to fairness. When a model fails to generalize across an attribute (say, if one group is under-represented in training), it tends to produce biased predictions under distribution shift. For instance, covariate or demographic shifts (changes in the input distribution of a sensitive attribute) have been shown to induce biased errors: a model trained mostly on one group can systematically misclassify members of an underrepresented group (Shao et al., 2024). In other words, a classifier may appear fair on the training data but lose that fairness when the population mix changes. Recent work on fair domain generalization explicitly highlights this pitfall: standard fairness mechanisms typically assume fixed distributions and may not hold under new attribute distributions (Dong et al., 2025). Thus, improving attribute generalization – learning features that are invariant to those attributes – inherently promotes fairer outcomes. When feature representations no longer carry spurious signals (e.g., skin tone or other demographics), the model's accuracy tends to become more uniform across subgroups (Wang et al., 2022a).

In dermatological imaging, skin tone is exactly the kind of attribute that can shift and break model fairness. Most public skin-lesion datasets are heavily skewed toward lighter Fitzpatrick types: for example, Kinyanjui et al. found that benchmark dermatology datasets contain very few dark-skin images (Kinyanjui et al., 2019). This lack of diversity means that a model trained on these data will face a new "domain" at test time whenever it encounters darker skin. Empirically, such shifts degrade performance: Daneshjou et al. report that both AI algorithms and dermatologists perform worse on images of darker skin tones (Fitzpatrick V–VI) and uncommon diseases, and that balancing the training data (fine-tuning on the diverse DDI dataset) closes the gap (Daneshjou et al., 2022). In fairness terms, this is exactly the issue of a majority-dominated training set: a model biased toward the common (light) group underperforms on the rare (dark) group. Therefore, applying attribute-generalization ideas (e.g. enforcing skin-tone invariance or using synthetic augmentation across skin tones) addresses both robustness and equity. By learning disentangled, transferable features across skin-tone domains, dermatology models can maintain accuracy and fairness even on under-represented skin types (Alipour et al., 2024). Theoretically, disentangled representation learning aims to decompose the underlying factors of variation into independent components. In the context of subpopulation shift, this involves ensuring that the latent space $\mathcal{Z}$ can be partitioned such that $\mathcal{Z} = [\mathcal{Z}_{\text{core}}, \mathcal{Z}_{\text{attr}}]$, where $\mathcal{Z}_{\text{core}}$ is statistically independent of the auxiliary attributes a while remaining maximally informative of the label y.

Discussion on Fairness and Skin Tone Generalization in Dermatological Imaging. Recent studies have documented pronounced performance gaps in skin lesion classifiers across Fitz-

patrick skin types. Most dermatology datasets have "inadequate skin type diversity," with darker Fitzpatrick types (IV–VI) heavily under-represented. For example, the Fitzpatrick-17k dataset contains only ∼4% type IV–VI images (Groh et al., 2021). Similarly, the International Skin Imaging Collaboration (ISIC) archive lacks explicit skin-tone labels, forcing researchers to estimate skin color via proxies like the Individual Typology Angle (ITA). But ITA-based estimates disagree widely between methods, undermining confidence in fairness analyses on ISIC-18 (Tschandl et al., 2018). This dearth of diverse skin tones means that a model trained on predominantly light-skinned samples is subject to a classic subpopulation shift: it may not generalize to images from darker-skinned patients. In sum, inadequate dataset diversity leads to an *attribute generalization* problem for skin tone, where diagnostic performance degrades on under-represented subgroups.

Fairness evaluation in this setting has therefore attracted growing attention. Researchers adopt group-fairness criteria to quantify disparities. For instance, Xu *et al.* (2025) define multi-class versions of equalized opportunity (Eopp) and equalized odds (Eodd) to measure balance of true-positive and false-positive rates across skin-tone groups (Xu et al., 2025). Equalized opportunity requires that the true-positive rate be equal for, say, light- versus dark-skin classes, while equalized odds also matches false-positive rates. Xu *et al.* additionally introduce a combined fairness–accuracy tradeoff score (FATE) to capture how well a model balances overall performance with group fairness. Yang et al. (2024) likewise emphasize equal opportunity in dermatology: they examine "underdiagnosis" by comparing false-negative rates between age or race subgroups in ISIC, effectively measuring Eopp for the "No Finding" vs disease tasks. These metrics (Eopp, Eodd, subgroup FNR/FPR gaps, etc.) are now widely used to flag bias.

To address skin-tone disparities, several mitigation strategies have been proposed in the literature. At the data level, pre-processing techniques aim to rebalance under-represented groups. For example, stratified sampling or re-weighting during training can ensure each skin-tone group is seen equally. Xu et al. (2024b) showed that simple stratified batch resampling significantly improved fairness in dermatology models. Model-level interventions have also been widely explored. Bevan and Atapour-Abarghouei (2022) developed an automatic skin-tone labeling algorithm to annotate the ISIC archive, and then applied "unlearning" techniques to remove skin-tone signals during training. Their results show that explicitly stripping skin-color information from the network can improve generalization and reduce the accuracy gap between light- and dark-skin images. Du et al. (2022) propose FairDisCo, a disentanglement-based method: a dual-branch network where one branch learns to predict skin type (to be discarded) and the other learns invariant features. This contrastive disentanglement yields more equalized performance on Fitzpatrick17k and the Diverse Dermatology Images (DDI) dataset, as shown by improved fairness metrics and overall F1 scores compared to simple resampling or reweighting baselines. Ghadiri et al. (2024) introduce XTranPrune for vision transformers: they use explainability (Layer-wise Relevance Propagation) to identify "discriminatory" model modules associated with skin color, and prune them away. On

two skin lesion benchmarks, this pruning method attained better fairness scores than unpruned models. Xu et al. (2025) also explore model design: their FairMoE method employs a mixture-of-experts with group-specific pathways. By encouraging experts to specialize on particular skin-tone subgroups (while still sharing some information), FairMoE achieved higher accuracy and lower Eopp/Eodd than prior baselines on Fitzpatrick-17k and ISIC-2019. Another promising direction is domain and representation adaptation. Aayushman et al. (2024) propose PatchAlign, which aligns visual and textual (clinical label) representations via an optimal transport loss. This cross-domain alignment produces image features that are more invariant to skin color; empirically, PatchAlign improved in-domain accuracy by 2–6% and markedly reduced differences in true-positive rates across skin tones on Fitzpatrick17k and DDI. In related work, Xu et al. (2023) have experimented with group-aware batch normalization (FairAdaBN) and adversarial training to remove skin-type cues from features, though these are less explored in dermoscopy.

A critical gap in the existing literature concerns scenarios where certain skin tones are entirely missing or severely underrepresented in the training dataset. In such cases, even fairness-aware training methods fail to generalize due to the absence of representative data. When only a small number of samples from the missing tone are available and cannot be used directly for training most prior approaches offer limited solutions. To address this, we in (Ansari et al., 2024b) propose a specialized augmentation strategy that synthesizes minority skin-tone lesions by mixup of features from prevalent-tone images, guided by an adaptive sampler focused on hard-to-classify instances by checking the performance on the meta-set of unknown tone images. This method enables model recalibration without direct access to missing-tone samples during training, improving accuracy and fairness across demographic groups in ISIC and ASAN datasets.

These challenges manifest not only in aggregate performance metrics but also in systematic failure patterns observed in real-world classification scenarios. To contextualize the practical implications of class imbalance and fairness, we briefly discuss representative failure modes encountered in image classification tasks studied in this thesis.

1.4 Real-World Classification Failure Patterns Under Class Imbalance and Bias

Despite recent advances in deep learning–based medical image classification, real-world deployment continues to reveal systematic failure modes that are particularly pronounced under class imbalance and demographic bias. In this thesis, we focus on classification-centric failures observed in the medical datasets under study, rather than segmentation-related errors, as these directly impact diagnostic decision-making in imbalanced and fairness-sensitive settings. For example, in the HAM10000 dermatological dataset, minority disease categories such as melanoma, dermatofibroma, and vascular lesions are frequently misclassified as benign nevi due to strong visual overlap in color,

texture, and lesion boundaries. These errors are further amplified when lesions appear in darker skin tones (higher Fitzpatrick types), where reduced contrast and illumination variability obscure discriminative features. Such failures pose critical clinical risks, as they may delay diagnosis for already under-represented patient populations. Similarly, in endoscopic image classification tasks, rare pathological findings are often confused with normal mucosa or inflammatory patterns due to limited training samples and significant inter-patient variability. Motion blur, specular highlights, and variations in imaging devices further contribute to these misclassifications, disproportionately affecting classes with fewer annotated examples. Comparable failure patterns are also observed in chest X-ray classification tasks, where conditions such as pneumothorax, cardiomegaly, or early-stage pneumonia are under-represented relative to normal or common findings. In such cases, subtle radiographic cues may be overshadowed by dominant visual patterns learned from majority classes. These errors are often exacerbated by confounding factors such as patient positioning, image acquisition protocols, age-related anatomical differences, and comorbidities, which can introduce implicit bias into the learned representations. From a fairness perspective, these failure modes are not uniformly distributed across disease categories or patient subgroups. Minority disease classes, patients with atypical presentations, and under-represented demographic groups tend to experience consistently higher misclassification rates. These observations highlight the necessity of learning strategies that explicitly account for class imbalance and bias during training. The methods proposed in this thesis are motivated by such real-world failure patterns.

1.5 Organization and Chapter-wise Contributions of the Thesis

This thesis advances the field of subpopulation shift by proposing a sequence of principled, modular techniques that address **class imbalance**, **long-tailed distributions**, and **demographic bias**, spanning from data-level interventions to architectural and optimization-level innovations. The work begins with the challenge of class imbalance in structured tabular data, where we introduce a *calibrated oversampling method* that better estimates minority class statistics for improved representation and performance (Chapter 2). Building on this foundation, we extend our investigation to high-dimensional medical images, developing a *Mixture-of-Two-Experts (Mo2E)* framework (Chapter 5.2) that effectively handles long-tailed disease categories in clinical datasets. Insights gained from these two initial works revealed the limitations of static sampling and augmentation strategies, which motivated the design of **STTP-Net** (Chapter 3) a dual-branch architecture tailored for long-tailed visual recognition using hybrid mixup augmentation and logit reweighting. To further enhance this framework, Chapter 4 introduces the **Goldilocks Principle**, which leverages contrastive regularization and class-aware boundary smoothing to improve generalization across the class spectrum. The final part of the thesis (Chapter 5.3) addresses fairness under **attribute generalization**, particularly in skin tone disparities, by proposing a *bias-aware training strategy* that ensures equitable performance across demographic subgroups. Together, these contributions

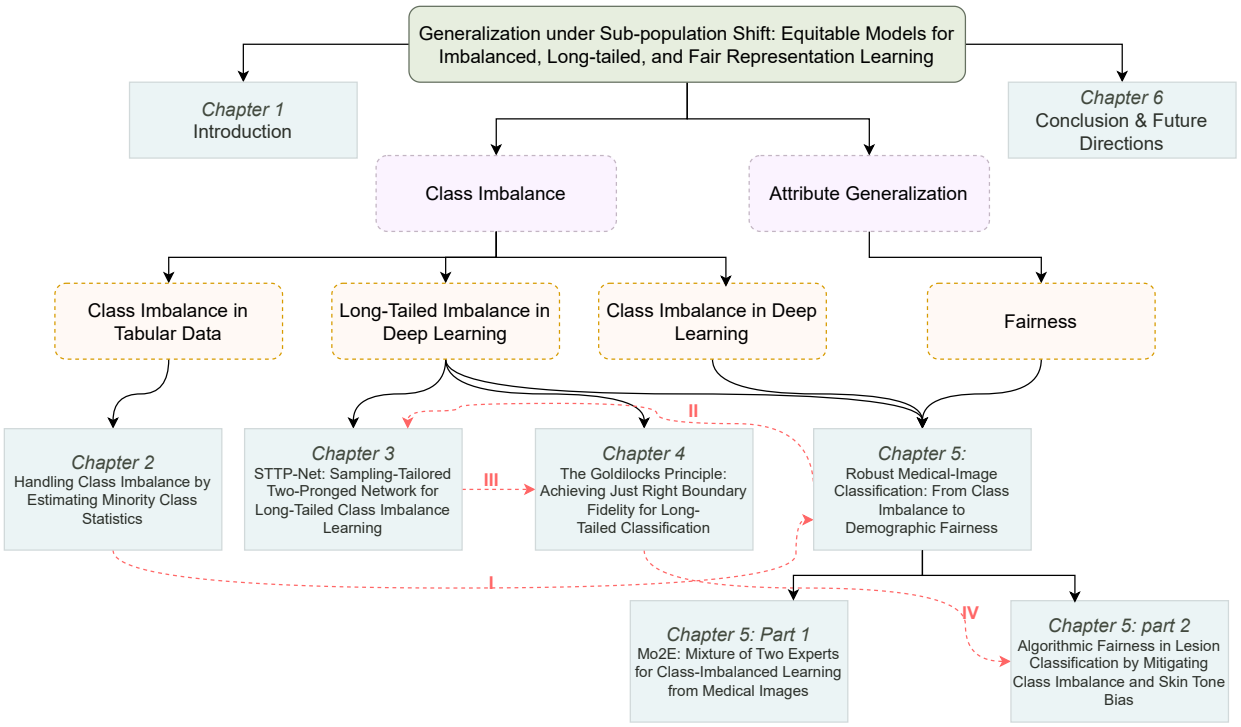


Figure 1.6: Layout of the Thesis

form a cohesive pipeline that evolves from calibrated data synthesis to dual-branch architectures and fairness-aware training. They collectively offer a robust solution for generalization under sub-population shift across both tabular and image-based domains.

Chronologically, the research began with Chapter 2 on tabular data imbalance, followed by the development of the Mo2E framework (Chapter 5.1). This led to further insights that shaped the design of STTP-Net (Chapter 3) and the Goldilocks boundary refinement strategy (Chapter 4). Since both Mo2E and the fairness-aware lesion classification (Chapter 5.2) are situated in the medical domain, they are unified under Chapter 5, titled: Robust Medical-Image Classification. Figure 1.6 gives the detailed chapterwise layout of the thesis.

1.5.1 Contributions of Chapter 2

In Chapter 2 presents a novel oversampling methodology for handling class imbalance by explicitly estimating the statistics of the minority class using information from nearby classes. Unlike standard oversampling methods, which often ignore the underlying distribution of the minority class, the proposed approach leverages weighted feature statistics and distributional calibration to approximate the original minority class distribution. This framework introduces a hyperparameter-driven mechanism to adaptively balance the spread and concentration of generated synthetic samples, mitigating the risk of overgeneralization while preserving distributional fidelity. Extensive exper-

iments across binary and multi-class tabular datasets demonstrate that our method outperforms established oversampling techniques, achieving superior classification performance and improved minority class representation. However, since the method relies on fractional-norm based distance calculations, its scalability to high-dimensional feature vectors (such as those extracted from images) remains challenging. Future research may explore dimensionality reduction or alternative neighborhood selection strategies to extend its applicability to high-dimensional feature spaces.

1.5.2 Contributions of Chapter 3

This chapter introduces **STTP-Net**, a novel two-pronged network framework for tackling long-tailed class imbalance in deep visual recognition. It presents a systematic investigation of mixed-sample data augmentation in conjunction with tailored sampling strategies to uncover the synergies that improve tail-class learning. Based on this, the chapter proposes a new **Hybrid-Mixup** augmentation method and a class-frequency-driven sampling strategy to train specialized experts for head/medium and tail classes within a unified architecture. Additionally, to mitigate classifier bias and label distribution shift, the chapter introduces the **Effective Balanced Softmax (EBS)** loss, which dynamically reweights logits based on class frequency priors. Extensive empirical evaluations on benchmark datasets like CIFAR-LT, ImageNet-LT, and NIH-CXR-LT demonstrate that STTP-Net surpasses several state-of-the-art long-tailed learning methods in both accuracy and robustness across many-shot, medium-shot, and few-shot classes. The contributions of this chapter offer a scalable and effective alternative to computationally expensive multi-expert and two-stage training paradigms, thus broadening the practical applicability of long-tailed learning in critical domains such as medical diagnostics.

1.5.3 Contributions of Chapter 4

Building upon the dual-branch paradigm introduced in Chapter 3, this chapter proposes an advanced, theoretically - grounded framework that addresses a deeper nuance of long-tailed recognition: **boundary fidelity**. Recognizing the limitations of overly hard or soft decision boundaries, this chapter introduces a principled strategy based on the **Goldilocks Principle** aiming for "just right" boundary sharpness to balance generalization and class discrimination. To that end, the chapter presents a novel single-stage framework that integrates two key modules: (1) **Dual-Branch Sampler-Guided Mixup (DBSGM)**, which orchestrates a class-aware mixup using instance, median, and reverse sampling strategies to balance learning across head, medium, and tail classes, and (2) **Adaptive Class-Aware Feature Regularizer (ACFR)**, driven by a newly formulated **Temperature Adaptive Supervised Contrastive Loss (TASCL)** that dynamically adjusts attraction-repulsion forces in the feature space depending on class frequencies and prediction confidence. The proposed approach offers significant improvements across multiple datasets, including CIFAR-LT, ImageNet-LT, and iNaturalist, particularly enhancing medium and tail class accuracy.

In contrast to Chapter 3, which primarily addressed class imbalance via tailored augmentation and logit-level adjustments (EBS), this chapter enhances intra-class cohesion and inter-class separation through dynamic feature-space regularization. Together, these chapters propose a holistic pipeline: from better sampling and augmentation (Chapter 3) to boundary refinement and class-aware contrastive learning (Chapter 4), all within an end-to-end efficient architecture.

1.5.4 Contributions of Chapter 5

Chapter 5 extends the progression from tailored augmentation and boundary-regularization techniques (Chapters 3 and 4) into real-world, high-stakes applications, focusing on the dual challenges of class imbalance and demographic fairness in medical image classification. The chapter makes two synergistic contributions. First, it proposes a multi-expert architecture, **Mo2E** (Mixture of Two Experts), which leverages class-specialized MixUp augmentation and domain-specific sampling strategies to learn robust decision boundaries across the long-tailed spectrum. Second, it introduces a novel **adaptive MixUp sampling framework** guided by a meta-learned heuristic to mitigate bias due to skin tone disparities in dermatological datasets. Mo2E improves performance across both head and tail classes in complex clinical datasets such as Hyper-Kvasir and Eyepacs, while the fairness-aware augmentation strategy generalizes effectively across patient subgroups in ISIC-2018 and the Asan dataset. Unlike previous chapters, which primarily tackled frequency imbalance, this chapter directly addresses algorithmic fairness, marking a significant evolution in the thesis' narrative from distribution-aware learning (Chapter 3), through boundary calibration (Chapter 4), to demographic generalization. This cohesive pipeline reflects a deepening resilience of deep learners to multiple sub-domain shifts that typify medical AI, moving the thesis closer to deployable, trustworthy clinical decision support systems.

1.5.5 Contributions of Chapter 6

Chapter 6 summarizes the contributions presented throughout the thesis and distills their collective impact on the broader problem of sub-domain shift in deep learning spanning class imbalance and demographic bias. It presents a critical evaluation of the methods proposed in Chapters 2 through 5, identifying common principles, strengths, and limitations across approaches. In doing so, the chapter constructs a unified perspective on resilient long-tailed learning, highlighting how targeted sampling, adaptive augmentation, and feature-space regularization converge to form a robust design paradigm. Furthermore, the chapter delineates open challenges such as dynamic expert assignment, label noise robustness, and demographic fairness under weak supervision and outlines promising future directions, including joint optimization of fairness and accuracy, cross-modal learning, and real-world deployment in low-resource clinical environments. This chapter not only concludes the thesis but also suggests a clear path for future research.

Chapter 2

Handling Class Imbalance by Estimating Minority Class Statistics

Synopsis

Class imbalance is a persistent challenge in tabular data classification, where the disproportionate representation of classes typically dominated by a majority class can severely degrade model performance on minority classes. A prevalent strategy to mitigate this issue involves oversampling, where synthetic samples are generated for underrepresented classes via convex combinations of existing minority instances. However, existing approaches often disregard the underlying distributional characteristics of the minority class, resulting in synthetic data that poorly approximates the true data manifold. In this work (Ansari et al., 2023) [1], we introduce a novel parametrization-based oversampling framework that estimates the minority class distribution by leveraging statistical cues from adjacent or semantically similar classes. Through tunable hyperparameters, our method allows precise control over the sampling behavior, enabling closer alignment with the true minority distribution. Extensive evaluations on both synthetic and real-world tabular datasets confirm the efficacy of our approach, demonstrating consistent gains across a range of standard performance metrics.

2.1 Introduction

Class-imbalanced datasets occur in several real-world applications, such as medical diagnosis, credit risk assessment, software defect detection, and fraud detection, where the (benign) majority class examples are much more frequent than the (target) minority class examples (Das et al., 2018). This very often leads to learning models that focus mainly on the majority class and overlook the minority class. Consequently, the model's accuracy decreases due to the class imbalance, as models

[1] F. Ansari, S. Das, and P. Shamsolmoali. "Handling Class Imbalance by Estimating Minority Class Statistics." In *2023 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2023. doi: https://doi.org/10.1109/IJCNN54540.2023.10191975.

may not accurately detect instances of the minority class or give them equal emphasis while making predictions.

Supervised learning algorithms create a model that is usually biased toward the majority class because such algorithms concentrate on reducing loss across the entire training set without paying close attention to the percentage of samples from each category. Due to the underrepresentation of the minority samples, minority class predictions may not reach the desired level of accuracy. To tackle this issue, various data rebalancing techniques like under- and oversampling were proposed (Branco et al., 2016)(Das et al., 2022). Among these, oversampling methods have become increasingly popular as, they are more effective in boosting the accuracy of classifiers compared to undersampling methods. While undersampling may result in loss of informative training points from the majority class, oversampling synthesizes artificial samples in the minority class to balance the class sizes in the training set, making the training set more representative, This, in turn, enhances the performance of the machine learning algorithms when applied to the augmented datasets. Typically, oversampling techniques create new samples from a minority class sample and its close neighbors. The distribution of minority class samples, however, is hardly considered by them. As a result, there may be distributional inconsistencies between real and artificial samples made using oversampling techniques.

This chapter focuses exclusively on class imbalance within the domain of structured tabular data. Unlike the long-tailed distributions discussed in Chapters 3 and 4, which involve severe power-law decay in high-dimensional image spaces, this work addresses the unique representational challenges of low-dimensional numeric features. In this work, we will estimate the minority class statistics so that the distribution of the generated samples draw closer to the original distribution of the minority class. Recent work on distribution calibration (Yang et al., 2021) shows how to use the statistics of the classes with a lot of samples to learn the distribution of the class with very few samples. Motivated by this, we propose an oversampling method that uses the statistics of its closest classes to estimate the statistics of the samples of the minority class and generates new samples that correspond to the samples in the minority class. We have parameterized our method so as to provide additional degrees of freedom to control the generated samples with a view to closely matching them with the original data points. Figure 2.1c depicts the synthetic samples ('+' green) generated by our proposed model, which can approximate the original distribution in Figure 2.1a, as opposed to Figure 2.1d, 2.1e, and 2.1f, which show the samples generated by the SMOTE variants and cannot approximate the original distribution in Figure 2.1a. To demonstrate the effectiveness of our work, various metrics for imbalanced learning are adopted.

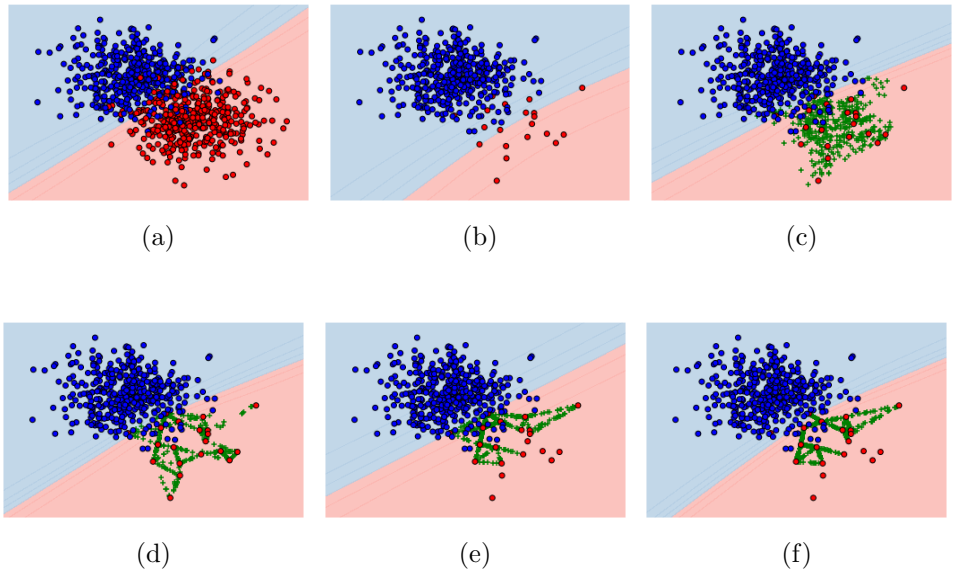


Figure 2.1: (a) A balanced dataset with equal number of samples in both classes (blue dot and red dot). (b) making the data class-imbalanced by removing samples from the red class. (c) Using our proposal, we generated the minority points (green "+"), which attempt to approximate the distribution of the red class as it was previously. (d) Points generated (green "+") using SMOTE. (e) Points generated (green "+") using ADASYN. (f) Points generated (green '+') using BoderlineSMOTE. As seen from (d), (e), and (f), new samples are generated based on a minority class sample and its nearest neighbors. As a result, the distribution of the real and synthetic samples differ significantly.

2.2 Related Works

Existing solutions to tackle the class imbalance problem have often relied on data resampling techniques such as oversampling and undersampling or weighting the minority classes more heavily during model training. Thus, these approaches can be classified into cost-sensitive, undersampling, and oversampling approaches. Cost-sensitive approaches use different penalties for misclassification of the minority and majority classes (Domingos, 1999) (Elkan, 2001). The penalty for misclassifying a minority point is higher than that for a majority sample to encourage the model to classify more of the minority instances correctly. The limitation of the cost-sensitive learning approach is that it is difficult to set the cost parameter for it to be appropriate for the given data. Moreover, they also do not take into account how samples are distributed within the minority and majority classes.

The random undersampling (RUS) technique (Kubat et al., 1997) helps to address this issue by randomly removing instances from the majority class until their size is balanced with that of the minority class. In the process, undersampling techniques can result in a loss of valuable data, as it is not guaranteed that the instances removed are not necessary for modelling.

The oversampling technique helps to counter this issue by randomly generating additional

instances of the minority class such that minority class size matches that of the majority class. The data points are synthesized by interpolating among several instances of minority classes that are located within a particular neighborhood (Fernández et al., 2018). SMOTE (Synthetic Minority Oversampling Technique) (Chawla et al., 2002b) is one of the most widely used oversampling techniques, where a target instance from the minority class is linearly interpolated with some of its randomly chosen nearest neighbors. BorderlineSMOTE (Han et al., 2005b) enhances SMOTE by generating artificial instances from only the wrongly classified data points at the boundaries of the minority classes. SVMSMOTE (Nguyen et al., 2011) is another enhanced version of SMOTE that uses support vector machines (SVMs) to identify instances of minority classes that are misclassified and then generates synthetic samples around them. ADASYN (Adaptive Synthetic Sampling Approach for Imbalanced Learning) (He et al., 2008) is a similar technique that dynamically adjusts the number of synthetic examples generated based on the local density of minority-class instances. MWMOTE (Barua et al., 2012) assigns a weight to each sample from the minority class based on how close it is to the sample from the majority class. Then, the weighted minority instances are used to synthesize artificial samples.

Even though SMOTE is favoured in the literature because of its simplicity and ease of implementation, it can suffer from the problem of overgeneralization (Fernández et al., 2018). Overgeneralization occurs when due to SMOTE like oversampling the minority class area blindly expands without caring for the majority class distribution. If the class distribution is highly skewed, i.e. the minority class is significantly sparse compared to the majority class, overgeneralization results in more overlap among the classes, thereby leading to deterioration of test accuracy. Moreover, (Elreedy et al., 2023) demonstrates theoretically that the patterns generated by SMOTE do not necessarily reflect the minority class distribution. Some methods, like (Yang et al., 2022), (Li et al., 2022a) have been proposed that generate samples based on how minority class samples are spread out in space. However, these algorithms first try to split minority classes into different clusters and then try to generate new points in these clusters using SMOTE. Because of the use of SMOTE, these methods still do not give a guarantee that the generated points will not overgeneralize.

To tackle the issue of overgeneralization and to generate distributions similar to minority class distributions, we propose a method that estimates the statistics of the minority class using the statistics of the closest classes. This method is based on the idea that two classes that are close to each other in data space should have similar distributions. Then, we'll make new points based on the estimated statistics of the minority class samples.

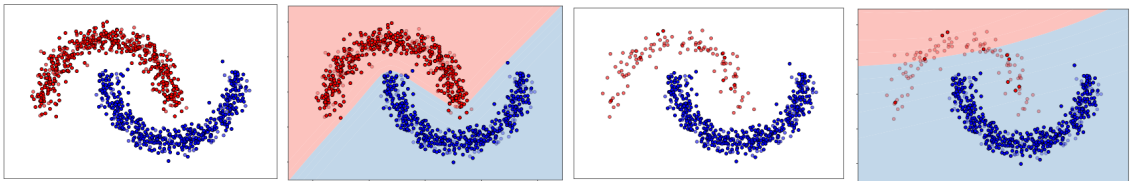
2.3 Proposed Methodology

2.3.1 Intuition and overview of the proposed method

In (Yang et al., 2021), the authors observed that the estimated mean and variance of each class are correlated to the semantic similarity of the class, if we assume the feature distribution is Gaussian. Knowing how similar the two classes are will help us transfer the statistics from the base classes to the novel classes. Thus, we can use this observation to tackle the imbalance problem by using the statistics of the majority classes having sufficient data points, and transferring those statistics to the minority classes. Now the question that naturally arises is how to transfer the statistics? As noted from (Yang et al., 2021), the transfer is determined by first finding the Euclidean distance between the mean of the features from the base classes and the feature space of the novel classes. The statistics of the closest base classes are then used to calibrate the distribution's mean and covariance. But in the case of the class imbalance problem, it is not possible to determine which classes can be considered the base classes for a particular minority class. Thus, the transfer of statistics between classes for the imbalanced problem is determined by first finding the distance between the mean of the features from all other classes and the feature space of the minority class. To determine the distance, we use the fractional norm to calculate the closeness between the minority class and the mean of the features of the other classes. The fractional norm to consider while finding the distance is a data-dependent hyperparameter that needs tuning. Now the statistics of the closest classes are used to calibrate the distribution of the minority class, but rather than using the statistics directly, we use the statistics from the weighted features of the closest classes (the use of weighted features is based on the assumption that the class that is closer has statistics more similar to that class, and therefore that class's attributes will be weighted higher compared to other classes), weighted by the degree of closeness depending upon the calculated distance. Moreover, we hyper-parameterize the weights (whether to increase or keep their value the same) as a function of distance, which helps us control the relative importance of the classes. Hyperparameters are used to play a kind of adversarial game for estimating the distribution of the minority class as a close approximation of the actual distribution.

2.3.2 Problem Definition

Let us consider C-class supervised classification problem with a training dataset $X = \{(x_i, y_i)\}_{i=1}^{N}$, where N is the number of samples, x_i is the i^{th} datapoint and y_i is the corresponding category label, such that $x_i \in \mathbb{R}^d$ and $y_i \in L$ and $L = \{1, 2, 3, \ldots, C\}$. X_c denotes the set of samples belonging to class c, and $N_c = |X_c|$ denotes the cardinality of X_c. And we consider the classes to be in order $N_1 \geq N_2 \geq N_3 \geq \cdots \geq N_C$ such that $\sum_{c \in C} N_c = N$. We intend to synthesize new points z to balance the imbalanced dataset.

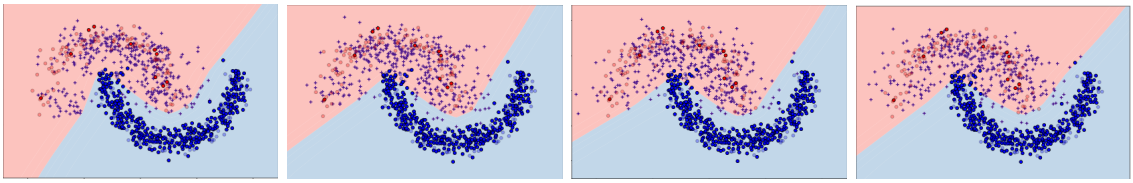


(a) Balanced dataset having two classes, each class having 1000 samples.

(b) Decisison boundary after classification on the imbalanced dataset.

(c) Imbalanced dataset created from the balanced dataset with IR = 100

(d) Decisison boundary after classification on the balanced dataset.

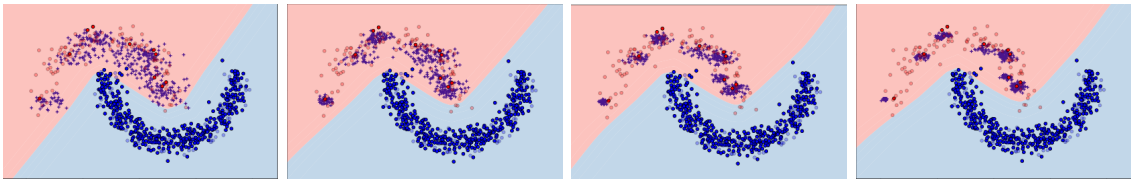


(e) f = 0.4

(f) f = 0.6

(g) f = 0.8

(h) f = 1



(i) $\beta = 0.4$

(j) $\beta = 0.6$

(k) $\beta = 0.8$

(l) $\beta = 1$

Figure 2.2: Rows 1 (a–d) depict the dataset used and the decision boundary after classifying on the balanced and imbalanced datasets. In the above plots, the dots with the light red and light blue color are the test dots, which are not used while training, and along with them, the purple '+' sign shows the generated data points. Row 2 from (e) to (h) shows the distribution when the f is changed and the new point is synthesized while the *beta* remains constant. Row 3 from (i) to (l) shows how the synthesized samples change when we change the β while keeping the f constant.

2.3.3 Proposed Method

Our proposed approach consists of 3 steps:

Step 1: Find out the class statistic of each given class, i.e., find out the mean for each class. Here we assume that each class's feature distribution is Gaussian.

$$\mu_c = \frac{1}{N_c} \sum_{x_i \in X_c} x_i, \tag{2.1}$$

where $\mu_c \in \mathbb{R}^d$ represents the mean of the class c.

Step 2: Taking data and transforming it using the Yeo-Johnson transformation (Weisberg, 2001). Transformation inflates low-variance data and deflates high-variance data to create a more uniform dataset. This helps to transform the skewed distribution and make the data more similar

to the normal distribution. It is specifically applied to each feature's dimension in the manner described below:

$$\tilde{x}_i = \begin{cases} ((x_i+1)^{\lambda}-1)/\lambda, & \text{if } \lambda \neq 0, x_i \geq 0 \ , \\ log(x_i+1), & \text{if } \lambda = 0,\ x_i \geq 0 \ , \\ -((-x_i+1)^{2-\lambda}-1)/(2-\lambda), & \text{if } \lambda \neq 2,\ x_i < 0, \\ -log(-x_i+1), & \text{if } \lambda = 2,\ x_i < 0 \ , \end{cases} \tag{2.2}$$

where $\lambda > 0$ is hyper-parameter.

Step 3: To compensate for the effect of imbalance, we generate $N_1 - N_c$ artificial points for the $c-th$ class. Each $\tilde{x}_{ic}$, $(i-th$ point in class $c)$, required to produces $\lceil \frac{N_1-N_c}{N_c} \rceil$ new points .

Now, calculate the distance between the selected data point $\tilde{x}_i$ from the minority class and the class means (calculated in step 1) using the fractional norm f as:

$$\forall c, d_{ic} = \|\tilde{x}_i - \mu_c\|_f, \tag{2.3}$$

$$\mathbb{S}_d = \{-d_{ic} | c \in L\}. \tag{2.4}$$

Now choose the top K closest classes of point $\tilde{x}_i$ using the distance calculated.

$$\mathbb{S} = \{c| - d_{ic} \in topK(\mathbb{S}_d)\}, \tag{2.5}$$

where $\mathbb{S}$ is the set containing the labels of the top K closest classes calculated using the distance between the centers of the classes and the data point $\tilde{x}_i$ from the minority class. $topK(.)$ is an operator to select top K elements from the distance set $\mathbb{S}_d$.
Given the K closest classes set $\mathbb{S}$ to the given minority class point $\tilde{x}_i$, we construct a new normal distribution whose mean and covariance are calculated using the K closest class's data points as follows, first calculate the weights as:

$$w_{ic} = \frac{1}{(1+d_{ic}^{\beta})}, \tag{2.6}$$

where w_{ic} is the weight of class c corresponding to the point $\tilde{x}_i$ where $c \in \mathbb{S}$, and β is the hyperparameter. Get the new mean corresponding to the point $\tilde{x}_i$ as:

$$\mu_{\tilde{x}_i} = \frac{\sum_{c \in \mathbb{S}} w_{ic}\mu_c + \tilde{x}_i}{1 + \sum_{c \in \mathbb{S}} w_{ic}}. \tag{2.7}$$

.

Finally, to calculate the covariance, we know that $\tilde{x}_i$ is a point from the minority class and $S \in \mathbb{R}^d$ is a random variable denoting a non-transformed points of class c such that $c \in \mathbb{S}$, we can compose random variable $S^{'}$, representing the estimate of minority class corresponding to the point

Algorithm 1 Our Proposed Method

Require: $X = \{(x_i, y_i)\}_{i=1}^{N}$, where N, number of samples, $x_i \in \mathbb{R}^d$, is the $i-th$ data point, and $y_i \in L$, such that L is the set of labels, $L = \{1, 2, 3, \ldots, C\}$. And X_c denotes set of samples belonging to class c.

Require: Hyperparametrs: β, f, K, and λ.

1: Determine the means for all classes, $\{\mu_i\}_{i=1}^{N}$, using Equation A.2.
2: Transform $(x_i)_{i=1}^{N}$ with Yeo-Jhonson Transformation, using Equation A.3. The transformed dataset is represented by $\tilde{X}$. And $\tilde{X}_c$ denotes transformed set of samples belonging to class c.
3: **for** $c \in L$ **do**
4: **if** $(N_1 - N_c)! = 0$ **do**
5: **for** $\tilde{x}_{ic} \in \tilde{X}_c$ **do** { // $\tilde{x}_{ic}$, i-th datapoint in class c.}
6: Calculate $\mu_{\tilde{x}_{ic}}$ using Equation A.8.
7: Calculate $\Sigma_{\tilde{x}_{ic}}$ using Equation A.1.1.
8: Sample $\lceil \frac{N_1 - N_c}{N_c} \rceil$, datapoints from $N(\mu_{\tilde{x}_i}, \Sigma_{\tilde{x}_i})$.
9: Add Sampled datapoints with corresponding labels to $\tilde{X}$.
10: **end for**
11: **end for**
12: Use the final balanced dataset $\tilde{X}$ to train a classifier.

$\tilde{x}_i$ as: $S^{'} = \frac{\tilde{x}_i + \sum_{c \in \mathbb{S}} w_{ic} S}{1 + \sum_{c \in \mathbb{S}} w_{ic}}$, thus the covariance is

$$\Sigma_{\tilde{x}_i} = Cov(S^{'}), \tag{2.8}$$

since $\tilde{x}_i$ is constant it does not contribute to the covariance. In practice, you can calculate covariance by using the weighted features of the classes in $\mathbb{S}$, weighted by the weights calculated using equation A.7. Furthermore, you can see that the equation A.8 can be obtained by taking the expectation of $S^{'}$ i.e. $\mathbb{E}(S^{'})$.

We found out the calibrated distribution associated with the minority class point $\tilde{x}_i$, which has the mean and covariance given as $\mu_{\tilde{x}_i}$ and $\Sigma_{\tilde{x}_i}$. We can sample points from this distribution and include them in the given minority class set as:

$$\tilde{X}_c \cup \{(z, c) | z \sim N(\mu_{\tilde{x}_i}, \Sigma_{\tilde{x}_i})\}, \tag{2.9}$$

where c represents the minority class to which the point $\tilde{x}_i$ belongs and $\tilde{X}_c$ represents the set of Yeo-Jhonson transformed points of class c. And z is the generated point from normal distribution $N(\mu_{\tilde{x}_i}, \Sigma_{\tilde{x}_i})$.

Algorithm 1 presents the proposed method pseudocode.

2.3.4 Hyper-parameter Explanation

For binary classification tasks, the two hyperparameters used are f and β. f is between $(0, 1]$ and β is between $[0, 1]$. From the toy dataset example in Figure 2.2 the effect of varying f and β can

be seen. As shown in Figure 2.2 [e-h], as the value of f increases while the value of β remains constant, the generated points ('+' in purple) try to spread from the minority points from which they are generated. Figure 2.2[i-l] shows, on the other hand, that as the value of f is held constant and the value of β is increased, it attempts to bring the generated points closer to the minority class point from which they are generated. Thus the value of f and β plays an adversarial game where increasing the value of f increases the spread of generated points while increasing the value of β decreases the spread. Thus, by using a grid search, we try to find the best combination of values for f and β, such that the final distribution we get can approximate the actual distribution and help classify the unseen points with high accuracy.

2.4 Experiments

2.4.1 Dataset

We conducted our experiments on both the imbalanced binary class classification and the imbalanced multi-class classification. Binary classification experiments are performed on 17 commonly used datasets. Among them, five open datasets for software defect detection, including pc1, mw1, kc3, pc4, pc3 are from OpenML (Vanschoren et al., 2014). The 5 datasets (poker-8_vs_6, poker-9_vs_7, poker-8-9_vs_5, poker-8-9_vs_6, shuttle-6_vs_2-3) are from the KEEL repository (Derrac et al., 2015). Among them are the poker dataset used for poker hand prediction and shuttle datasets used for predicting NASA space shuttle part failure. The other 7 datasets (optical_digits, wine_quality, letter_img, ozone_level, mammography, abalone_19, protein_homo) are taken from imbalanced-learn [2].

We provide a detailed description of characteristics for these 17 datasets in Table 2.1.

We also used ten multi-class imbalance classification datasets (wine, thyroid, hayes-roth, penbased, new-thyroid, balance, yeast, pageblocks, shuttle, and contraceptive) from the KEEL repository (Derrac et al., 2015) for benchmarking. Table 2.2 summarises the detailed description of the characteristics of these ten datasets. We randomly split the dataset into two parts for all experiments: the training set (80%) and the testing set (20%).

2.4.2 Baselines

The following state-of-the-art methods are used for comparison to evaluate the performance of the proposed method.

- **SMOTE** Synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002b). Use a minority class sample and its nearest neighbor to generate a new sample using linear interpolations, and the synthesized samples are used in training.

[2] https://imbalanced-learn.org/stable/datasets/index.html

Table 2.1: Dataset summary used for the binary labeled imbalanced classification

Dataset	#instance	#features	IR	#classes
pc1	1109	21	13.4:1	2
mw1	403	37	12:1	2
kc3	458	39	9.651:1	2
pc4	1458	37	7.191:1	2
pc3	1563	37	8.769:1	2
poker-8_vs_6	1477	10	85.88:1	2
poker-9_vs_7	244	10	29.5:1	2
poker-8-9_vs_5	2075	10	82:1	2
poker-8-9_vs_6	1485	10	58.4:1	2
shuttle-6_vs_2-3	230	9	22:1	2
optical_digits	5,620	64	9.1:1	2
wine_quality	4,898	11	26:1	2
letter_img	20,000	16	26:1	2
ozone_level	2,536	72	34:1	2
mammography	11,183	6	42:1	2
abalone_19	4,177	10	130:1	2
protein_homo	145,751	74	11:1	2

- **BorderlineSMOTE** (Han et al., 2005b) It first finds out the border point of the minority class. It uses those points to generate new points using the linear interpolation of the border point and its nearest neighbors.
- **SVMSMOTE** It makes use of the SVM to find boundary samples that are further used for synthesizing new samples as proposed in (Nguyen et al., 2011).
- **ADASYN** Adaptive synthesis (ADASYN) (He et al., 2008) is an extension of SMOTE. It produces more synthetic data for minority class samples that are challenging to learn than for minority class samples that are simple to learn. This is based on the idea that minority data samples should be created adaptively based on how they are distributed.
- **MWMOTE** Majority weighted minority oversampling technique (MWMOTE)(Barua et al., 2012). In this technique, each sample from the minority class is given a weight based on how close it is to the sample from the majority class. The weighted minority class is then used to create synthetic samples.

2.4.3 Evaluation Metrics Used

We have employed a number of evaluation metrics to test the effectiveness of our suggested methodology. The first metric used is Matthews correlation coefficient (MCC) (Boughorbel et al., 2017), also known as the phi coefficient, an evaluation metric primarily used for imbalanced datasets to generate a balanced measure using true and false positives and negatives. MCC only gives a high

Table 2.2: Dataset summary used for the multi-labeled imbalanced classification

Dataset	#instance	#features	IR	#classes
wine	178	13	1.5:1	3
thyroid	720	21	36.94:1	3
hayes-roth	132	4	1.7:1	3
penbased	1100	16	1.95:1	10
new-thyroid	215	5	4.84:1	3
balance	625	4	5.88:1	3
yeast	1484	8	23.15:1	10
pageblocks	548	10	164:1	5
shuttle	2175	9	853:1	7
contraceptive	1473	9	1.89:1	3

score if the classifier was able to correctly predict most of the positive data and most of the negative data. Another metric used is the geometric mean (GM), which is the root of the product of class-wise sensitivity. The value of GM is between $[1, 0]$. If the classifier fails to recognize at least one class, the G-mean resolves to zero. Also used is the "balanced accuracy" (BACC), which is the average recall of all the classes.

2.4.4 Parameter settings

All SMOTE-based algorithms utilized in this study employ K-nearest neighbors, where K is treated as a tunable parameter selected based on the highest achieved performance for each baseline. For our proposed method, hyperparameters were optimized through a systematic grid search to ensure reproducibility and distributional fidelity. For binary classification tasks, we conducted a grid search over the calibration scaling factor $f \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ and the mixing proportion $\beta \in \{0, 0.25, 0.5, 0.75, 1.0\}$. In multi-class scenarios, this search space was extended to include the neighborhood size K, which varied from 1 to $C - 1$, where C represents the maximum number of classes in the dataset. The Yeo-Johnson transformation parameter λ was fixed at 0.8 across all experiments. This value was selected based on preliminary variance-stabilization tests, where $\lambda = 0.8$ demonstrated the greatest consistency in normalizing feature distributions across diverse tabular datasets. Finally, to ensure fair comparisons among different methods, a Multi-Layer Perceptron (MLP) with consistent architectural parameters was used as the downstream classifier for all evaluated datasets.

2.4.5 Results Discussion and comparison with the SOTA methods

To show the effectiveness of our model, we use it to oversample the dataset in both the binary imbalanced classification problem and the multi-class imbalanced classification problem. We report the performance of all the methods in terms of MCC, geometric mean (GM), and balanced

Table 2.3: MCC, G-mean, and Balanced Accuracy for SMOTE, BorderlineSMOTE, SVMSMOTE, ADASYN, MWMOTE and Our method on the 17 public datasets for binary labeled imbalanced classification

Methods →	SMOTE			BorderlineSMOTE			SVMSMOTE			ADASYN			MWMOTE			Ours		
Dataset ↓	MCC	GM	BACC	MCC	GM	BACC	MCC	GM	BACC	MCC	GM	BACC	MCC	GM	BACC	MCC	GM	BACC
pc1	0.349	0.781	0.782	0.285	0.751	0.751	0.358	0.732	0.747	0.266	0.704	0.713	0.344	0.796	0.796	**0.381**	**0.829**	**0.829**
mw1	0.369	0.76	0.767	**0.494**	**0.856**	**0.857**	0.452	0.843	0.843	0.416	0.83	0.83	0.39	0.766	0.773	0.472	0.85	0.85
kc3	0.144	0.612	0.615	0.23	0.683	0.683	0.148	0.585	0.608	0.17	0.628	0.633	0.178	0.646	0.647	**0.27**	**0.718**	**0.72**
pc4	0.588	0.84	0.842	0.52	0.83	0.831	0.513	0.79	0.796	0.592	0.85	0.85	0.539	0.817	0.82	**0.621**	**0.85**	**0.85**
pc3	0.31	0.676	0.696	0.305	0.675	0.694	0.354	0.702	0.719	0.362	0.759	0.76	**0.436**	0.795	0.797	0.418	**0.805**	**0.805**
poker-8 _vs_6	0.772	0.997	0.997	0.865	0.998	0.998	1.000	1.000	1.000	0.65	0.993	0.993	0.65	0.993	0.993	**1.000**	**1.000**	**1.000**
poker-9_vs_7	0.271	0.676	0.707	0.315	0.684	0.718	0.467	0.934	0.936	0.211	0.66	0.686	0.211	0.66	0.686	**0.612**	**0.968**	**0.968**
poker-8-9_vs_5	0.198	0.742	0.759	0.181	0.618	0.678	0.193	0.62	0.68	0.26	0.853	0.855	0.181	0.618	0.678	**0.37**	**0.964**	**0.965**
poker-8-9_vs_6	0.625	0.888	0.893	0.625	0.888	0.893	0.797	0.893	0.898	0.725	0.891	0.897	0.797	0.893	0.898	**0.911**	**0.998**	**0.998**
shuttle-6_vs_2-3	0.611	0.965	0.966	0.807	0.989	0.989	0.807	0.989	0.989	0.611	0.965	0.966	0.691	0.977	0.977	**1.000**	**1.000**	**1.000**
optical_digits	0.97	0.989	0.989	0.96	0.984	0.984	0.966	0.993	0.993	0.966	0.989	0.989	0.965	0.984	0.985	**0.98**	**0.994**	**0.994**
wine_quality	0.239	0.632	0.679	0.3	0.641	0.692	0.304	0.659	0.703	0.291	0.674	0.71	0.267	0.67	0.704	**0.392**	**0.75**	**0.77**
letter_img	0.964	0.976	0.976	0.968	0.976	0.976	0.975	0.986	0.986	0.964	0.976	0.976	0.957	0.969	0.969	**0.982**	**0.989**	**0.99**
ozone_level	0.045	0.255	0.521	0.045	0.255	0.521	0.045	0.255	0.521	0.042	0.255	0.52	0.054	0.256	0.523	**0.237**	**0.673**	**0.673**
mammography	0.616	0.937	0.938	**0.693**	0.951	0.952	0.682	0.942	0.942	0.544	0.957	0.957	0.647	0.958	0.958	0.611	**0.964**	**0.964**
abalone_19	0.186	0.692	0.729	0.202	0.573	0.659	0.202	0.573	0.659	0.178	0.691	0.727	0.195	0.572	0.658	**0.246**	**0.738**	**0.738**
protein_homo	0.851	0.919	0.922	0.839	0.913	0.916	0.826	0.933	0.935	0.836	0.919	0.922	0.838	0.925	0.928	**0.858**	**0.925**	**0.928**

Table 2.4: MCC, G-mean, and Balanced Accuracy for SMOTE, BorderlineSMOTE, SVMSMOTE, ADASYN, MWMOTE and Our method on the 10 public datasets for multi-labeled imbalanced classification

Methods →	SMOTE			BorderlineSMOTE			SVMSMOTE			ADASYN			MWMOTE			Ours		
Dataset ↓	MCC	GM	BACC	MCC	GM	BACC	MCC	GM	BACC	MCC	GM	BACC	MCC	GM	BACC	MCC	GM	BACC
new-thyroid	0.783	0.912	0.871	0.623	0.855	0.827	0.819	0.922	0.883	0.558	0.842	0.819	0.739	0.834	0.783	**0.905**	**0.943**	**0.905**
wine	0.792	0.891	0.854	0.724	0.838	0.783	0.92	0.951	0.933	0.874	0.933	0.91	0.724	0.838	0.783	**1.000**	**1.000**	**1.000**
thyroid	0.862	0.959	0.953	0.769	0.921	0.909	0.734	0.817	0.761	0.806	0.887	0.842	0.297	0.557	0.444	**0.862**	**0.959**	**0.953**
shuttle	0.956	0.861	0.746	0.937	0.853	0.738	0.933	0.858	0.743	0.956	0.858	0.743	0.766	0.756	0.591	**1.000**	**1.000**	**1.000**
yeast	0.437	**0.783**	**0.649**	0.459	0.768	0.624	0.458	0.781	0.645	0.449	0.675	0.483	0.398	0.565	0.34	**0.496**	0.772	0.627
balance	0.647	0.818	0.745	0.714	**0.891**	**0.862**	0.787	0.858	0.789	0.674	0.859	0.814	0.69	0.836	0.762	**0.807**	0.88	0.826
penbased	0.82	0.907	0.837	0.87	0.933	0.883	0.845	0.92	0.86	0.835	0.915	0.851	0.82	0.906	0.836	**0.914**	**0.957**	**0.924**
hayes-roth	0.423	0.64	0.535	0.419	0.656	0.561	0.329	0.597	0.48	0.196	0.527	0.394	0.258	0.621	0.521	**0.56**	**0.748**	**0.677**
pageblocks	0.681	0.973	0.972	0.72	**0.978**	0.977	0.643	0.9	0.85	0.663	0.903	0.852	0.587	0.757	0.602	**0.789**	0.915	**0.865**
contraceptive	0.304	0.647	0.545	0.262	0.625	0.517	0.23	0.611	0.504	0.245	0.615	0.506	0.253	0.618	0.51	**0.33**	**0.66**	**0.56**

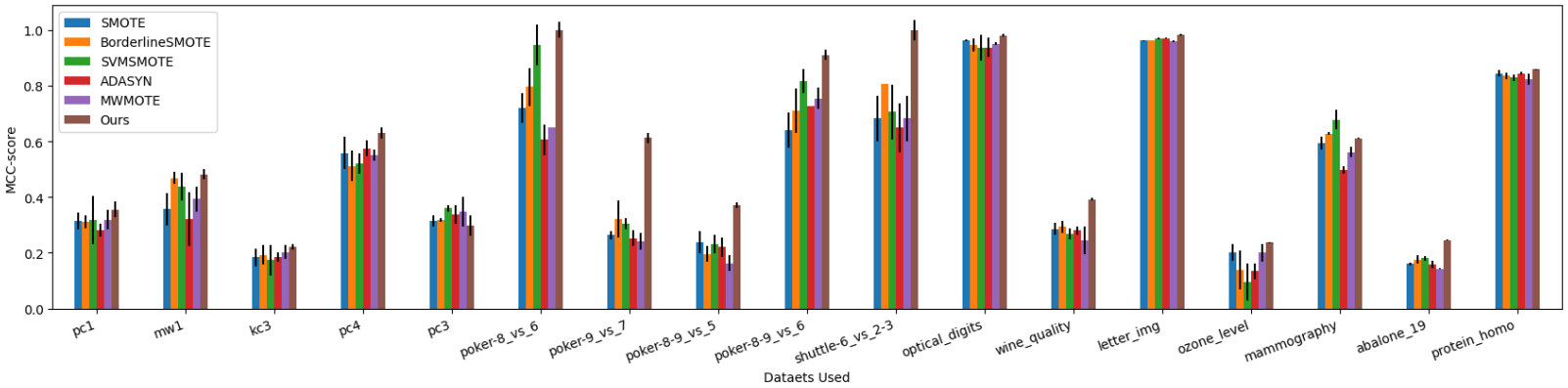


Figure 2.3: MCC average (height of the bars) and standard deviation (height of the error bars in black) over 5 random seeds for SMOTE, BorderlineSMOTE, SVMSMOTE, ADASYN, MWMOTE, and Our method on the 17 public datasets for binary labeled imbalanced classification.

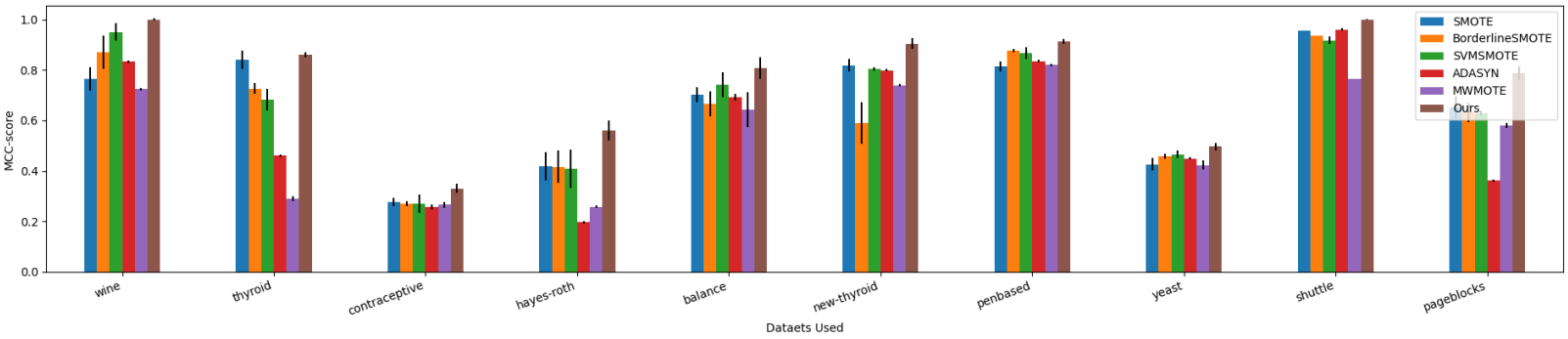


Figure 2.4: MCC average (height of the bars) and standard deviation (height of the error bars in black) over 5 random seeds for SMOTE, BorderlineSMOTE, SVMSMOTE, ADASYN, MWMOTE, and Our method on the 10 public datasets for multi-labeled imbalanced classification.

Table 2.5: Parameter used corresponding to different datasets for the binary labeled imbalanced classification and multi labeled imbalanced classification

Dataset with binary class	Parameters (f, β)	Dataset with multi class	Parameters (f, β, K)
pc1	(0.8, 0.5)	wine	(1, 0.9, 2)
mw1	(0.1, 0.5)	thyroid	(0.8, 1, 1)
kc3	(0.2, 0.8)	hayes-roth	(0.8, 0.2, 1)
pc4	(0.8, 0.4)	penbased	(0.3, 0.5, 2)
pc3	(0.7, 0.7)	new-thyroid	(0.2, 0.6, 1)
poker-8_vs_6	(0.2, 0.5)	balance	(0.9, 1, 2)
poker-9_vs_7	(0.7, 0.9)	yeast	(0.1, 0.8, 1)
poker-8-9_vs_5	(0.6, 1)	pageblocks	(0.1, 1, 4)
poker-8-9_vs_6	(0.9, 0.9)	shuttle	(0.7, 1, 2)
shuttle-6_vs_2-3	(0.2, 0.8)	contraceptive	(0.9, 1, 2)
optical_digits	(0.4, 1)		
wine_quality	(0.9, 0.2)		
letter_img	(1, 0.1)		
ozone_level	(1, 0.9)		
mammography	(0.2, 0.1)		
abalone_19	(0.3, 1)		
protein_homo	(0.7, 0.5)		

accuracy (BACC). The best results out of the random runs over 5 seeds on the binary imbalanced classification datasets are shown in Table 2.3 along with the parameters used for our method, given in Table 2.5. Moreover, average MCC over 5 random seeds are plotted in figure 2.3 along with the standard deviation. Our proposed method generally outperforms the state-of-the-art (SOTA) method in most of the 17 datasets. Especially in half of the datasets (pc1, kc3, poker-9_vs_7, poker-8-9_vs_5, poker-8-9_vs_6, shuttle-6_vs_2-3, wine, ozone_level, abalone), our approach consistently outperforms the competitor methods. On average, improvements in MCC, GM, and BACC are up to 11%, 8.94%, and 6.2%, respectively. The worst case of our model is shown on the mw1 dataset, which offers a 2% decrement in MCC score and similar GM and BACC scores as of BorderlineSMOTE, which implies that most of the samples are near the boundaries of the classes and our method might produce samples that overlap with the majority class samples. The same case goes with the mammography dataset, where the MCC of our method is less than the MCC of BorderlineSMOTE but there is improvement in the GM and BACC metrics. Nevertheless, our method outperforms all the other baselines, excluding BorderlineSMOTE for the mw1 dataset and all the other baselines in terms of GM and BACC for the Mammography dataset. Furthermore, it is worth noting that our method outperforms its competitors for highly skewed datasets such as ozone_level; when comparing the MCC of competitive methods, our method outperforms them

Table 2.6: BACC comparison with our methods and method when considering the Euclidean distance, i.e. $f = 2$ and considering $\beta = 1$

Binary classification			Multi class classification		
dataset	BACC Ours	BACC$^+$	dataset	BACC Ours	BACC$^+$
pc1	0.829	0.753	new-thyroid	0.905	0.883
mw1	0.85	0.823	wine	1	0.933
kc3	0.72	0.627	thyroid	0.953	0.837
pc4	0.85	0.811	shuttle	1	0.749
pc3	0.805	0.758	yeast	0.627	0.591
poker-8_vs_6	1	0.995	balance	0.826	0.791
poker-9_vs_7	0.968	0.691	penbased	0.924	0.901
poker-8-9_vs_5	0.965	0.854	hayes-roth	0.677	0.424
poker-8-9_vs_6	0.998	0.997	pageblocks	0.865	0.819
shuttle-6_vs_2-3	1	0.966	contraceptive	0.56	0.507
optical_digits	0.994	0.994			
wine_quality	0.77	0.694			
letter_img	0.99	0.989			
ozone_level	0.673	0.547			
webpage	0.898	0.9			
mammography	0.964	0.905			
abalone_19	0.738	0.794			
protein_homo	0.928	0.922			

$^+$ BACC when $f = 2$ and $\beta = 1$.

significantly.

To demonstrate that our method works not only for the binary imbalanced classification problem but also for the multi-class imbalanced classification problem, we report the performance on the multi-class datasets in Table 2.4 along with the parameters used for our method, given in Table 2.5. Moreover, average MCC over 5 random seeds are plotted in figure 2.4 along with the standard deviation, Table 2.4 contains the best results from the runs on the 5 random seeds. As seen from the Table 2.4 it consistently outperforms competitive methods in datasets (new-thyroid, wine, shuttle, penbased, hayes-roth, contraceptive). It improves the MCC, and GM by up to 15%, and 14%, respectively. And it also shows comparable performance in datasets like yeast and balance, where MCC is higher than the rest of the method, which means it classifies both majority and minority classes better than other algorithms. It also shows comparable results on the GM and BACC for the yeast and balanced datasets.

As discussed and shown in Tables 2.3 and 2.4, our proposed method outperforms the SOTA methods on all three metrics in most datasets. And it works fairly well for both the binary and the multi-class imbalanced classification problems.

2.5 Ablation Study

To evaluate the specific impact of the fractional norm and the parameterized mixing proportion β, we conducted a sensitivity analysis by comparing our optimized model against a baseline configuration where $f = 2$ (representing standard Euclidean distance) and $\beta = 1$ (neglecting the proposed parameterized weighting). This analysis demonstrates how model performance, measured via Balanced Accuracy (BACC), fluctuates when these hyperparameters deviate from their optimal settings.As detailed in Table 2.6, we observed that the inclusion of the fractional norm and the controlled spread through β consistently outperformed the standard configurations across both binary and multi-class classification tasks. Specifically, utilizing a fractional norm allows for a more accurate representation of the closeness between minority class points and their corresponding neighbors in high-dimensional tabular spaces. These results indicate a significant improvement in BACC for the majority of the datasets tested, confirming that the systematic tuning of f and β is essential for preserving distributional fidelity and controlling the spread of generated synthetic data.

2.6 Conclusion and future works

This work proposes an oversampling method for imbalanced data using distribution calibration. We use the statistics of the closest classes to estimate the statistics of the minority class sample and determine the gaussian distribution to which that particular sample belongs, then attempt to synthesize more samples using the mean and covariance corresponding to that specific minority sample's Gaussian distribution. We use the different parameters to control the generation of new samples, and thus, they approximate the original distribution of the minority class. The results of the experiments on the benchmark datasets for both binary and multi-class imbalance problems demonstrate the effectiveness of our proposed model. In the future, we plan to learn about hyperparameters rather than using grid search to find the best parameters.

Chapter 3

STTP-Net: Sampling-Tailored Two-Pronged Network for Long-Tailed Class Imbalance Learning

Synopsis

A long-tail class imbalanced learning problem is a scenario where the rare or minority classes, representing infrequent events or categories, make up the long tail of the class distribution and have disproportionately few examples compared to the dominant classes. The resulting imbalance makes it challenging to train models effectively for these underrepresented classes. We introduce a comprehensive solution - ***STTP-Net****:* ***S****ampling-****T****ailored* ***T****wo-****P****ronged* ***Net****work (****?****)* [1] *for long-tail class-imbalanced learning, which aims to address this issue holistically. The study thoroughly examines mixed sample data augmentation techniques in conjunction with various sampling strategies to identify the most effective approaches for handling long-tail imbalance. Based on this analysis, a hybrid mixup strategy tailored explicitly for data augmentation in long-tail imbalanced settings is proposed. The core of the proposed approach comprises a two-pronged network consisting of two classification heads designed to handle long-tail imbalanced datasets. One head specializes in learning the head and median classes in this design. In contrast, the other head becomes an expert in tail classes, striking a balance between accurate prediction of tail classes without compromising accuracy for the head classes. Additionally, we address the label distribution shifts in long-tail imbalance by introducing an Effective Balanced Softmax (EBS) function. The presented method achieves state-of-the-art performance in several benchmark categories for long-tail visual recognition datasets, surpassing the most prominent and pertinent end-to-end and dual-branch approaches.*

[1] F. Ansari, A. Panigrahi, and S. Das. "Sampling-Tailored Two-Pronged Network for Long-Tailed Class Imbalance Learning." *Engineering Applications of Artificial Intelligence*, 2025. doi: https://doi.org/10.1016/j.engappai.2025.111466.

3.1 Introduction

Deep Neural Networks have emerged as cutting-edge tools for diverse visual tasks, including image classification, object detection, semantic, and instance segmentation. The performance of Convolutional Neural Networks (CNNs) in visual tasks has experienced a significant boost, owing to the rise in computational capabilities and the abundance of training data. However, this progress heavily relies on the availability of meticulously curated, high-quality, and well-balanced datasets like ImageNet (Liu et al., 2019b) characterized by a uniform distribution of samples across all classes. Despite excelling on these datasets, these networks encounter significant challenges when dealing with highly imbalanced data. Real-world data often exhibit a long-tailed distribution, where a few classes possess high cardinality, representing the head classes. In contrast, many classes have very low cardinality, constituting the tail classes. In applied engineering domains, this long-tailed imbalance poses critical operational challenges. For instance:(1)In **medical imaging** (e.g., chest X-ray diagnostics), rare pathologies (tail classes) like pneumothorax are underrepresented yet demand high detection accuracy to avoid life-threatening oversights. (2)In **industrial automation**, defect detection systems must identify rare failure modes (e.g., micro-cracks in semiconductor wafers) to prevent costly production downtimes.(3)In **autonomous driving**, recognizing uncommon road scenarios (e.g., overturned vehicles) is critical for safety but often lacks sufficient training data. In such practical scenarios, deep networks struggle to yield satisfactory results for the tail classes due to an overwhelming disparity of information available for the head and the tail classes. Consequently, the models tend to overfit the head classes while underfitting the tail classes. Furthermore, dissimilarities in data distribution between the training and testing sets can lead to reduced performance on the test set. In response to these challenges, in this work, we propose a novel approach to handle the complexities posed by long-tailed data effectively.

As data unavailability scales up, the challenge of learning from long-tailed data intensifies, with a notable array of available techniques still insufficiently addressing this issue. However, the popular techniques used to handle class imbalance with fewer classes like SMOTE (Chawla et al., 2002a), DeepSMOTE (Dablain et al., 2022), and GAMO (Mullick et al., 2020) require considerable computational resources, suffer from mode collapse, and thus, are not suitable for a large number of classes as dealt in long-tailed datasets. To solve this issue, resampling strategies (Van Hulse et al., 2007b; He and Garcia, 2009; Buda et al., 2018a; Japkowicz and Stephen, 2002a) and re-weighting methods like (Huang et al., 2016b; Wang et al., 2017; Cao et al., 2019; Cui et al., 2019b) were introduced that work by balancing the head and tail classes. Nevertheless, the implementation of re-weighting strategies could pose challenges in optimizing models when trained on large-scale and real-world datasets, and these strategies sometimes deteriorate the performance of models in the majority classes to a large extent. Instead of reweighing or oversampling approaches, a separate line of work focuses on spatial-level augmentation by using mixed data augmentation techniques like

Mixup(Zhang et al., 2017a), CutMix (Yun et al., 2019a), Remix(Chou et al., 2020), MetasAug (Li et al., 2021), and CMO (Park et al., 2022). Multi-branch networks have been introduced to improve performance. These networks can be further classified into ensemble models (consisting of more than 2 branches, with the option to increase the number of branches to enhance performance) (Xiang et al., 2020; Cai et al., 2021; Zhu et al., 2022; Li et al., 2022b; Aimar et al., 2023a; Jin et al., 2023) and dual-branch models (Zhou et al., 2020; Wang et al., 2021b; Cui et al., 2022; Fan et al., 2022; Chen et al., 2023). Ensemble models, while improving accuracy, can suffer from drawbacks such as high network complexity, resource intensiveness, interpretability challenges, potential overfitting, and scalability issues. Dual-branch methods also demonstrate improved accuracy for medium and few classes but may compromise accuracy for many classes. Thus, to further overcome this challenge of improving the accuracy of all classes without hampering the performance of Many classes, two-stage methods (hong et al., 2021; Kang et al., 2020; Wang et al., 2021b; Zhang et al., 2023a) are proposed. However, these methods require multi-level tuning of multiple networks and a costlier two-stage training process, which might not be feasible for large datasets.

To address challenges in improving the accuracy of the tail classes without significantly deteriorating the accuracy of the head and median classes, we have come up with an STTP-Net, a sampling-tailored two-pronged network for this particular task. This network incorporates two specialized "experts," each focusing on distinct data segments. One expert concentrates on learning from the prevalent head and median classes, while another expert hones in on learning from the tail classes. Utilizing these distinct experts allows us to discern the distribution of tail classes without significantly compromising the accuracy of the head and median classes. The model categorizes the data into three groups based on frequency: the "head" classes, the "median" classes, and the "tail" classes. Each group is equipped with a dedicated sampler. The sampler for the tail classes employs a class-aware reverse data loader strategy, selectively sampling a more significant number of points from the tail classes compared to the median and head classes. Our method utilizes a proposed data augmentation technique called "Hybrid-Mixup", which was developed after analyzing sampling strategies and mixed-sample data augmentation techniques. During training, "Hybrid-Mixup" generates two augmented inputs for each expert in our two-prong network based on data samples from different samplers. Both experts learn independently from these inputs, maintaining a shared feature extraction network. Specifically, the feature extractor network shares its weights across both inputs, with the divergence lying in the linear classifier layer, where weights are not shared. To address the challenge of label distribution shift and mitigate classifier bias, we also proposed a modification of the softmax function for long-tailed data distribution termed the Effective Balanced Softmax (EBS) function. This function plays a crucial role during the training process. It dynamically adjusts the loss function, placing greater emphasis on under-represented classes within the training data. Consequently, the classifier is encouraged to learn more effectively from these classes, resulting in a reduction in overall bias.

The contribution of our research can be summarized as follows:

1. *Analysis of Mix-data Augmentation with Sampling Strategies:* The research delves into the effectiveness of mix-data augmentation techniques for imbalanced and long-tailed class scenarios. This analysis explores a spectrum of sampling approaches to identify the most efficient strategies.
2. *Hybrid Mixup Technique:* Based on the analysis, this work proposes a "hybrid mixup" technique. This technique aims to improve representation learning across all classes (head, medium, and tail) by strategically combining samples during augmentation.
3. *Two-Pronged Network Architecture:* The proposed framework, STTP-Net, utilizes a two-pronged network architecture. This architecture shares a feature extraction network but employs separate classifier heads for each branch.
4. Effective Balanced Softmax (EBS) Function: To address label distribution shifts, this work introduces the "Effective Balanced Softmax" (EBS) function.
5. *Engineering-Driven Validation:* We conducted a comprehensive assessment of the effectiveness of the proposed methodology, utilizing a rigorous evaluation process on prominent long-tailed visual recognition datasets, namely: CIFAR-LT-10, CIFAR-LT-100, ImageNet-LT, and NIH-CXR-LT (a medical imaging benchmark for chest X-ray diagnosis). Our method surpasses the performance of conventional end-to-end training techniques tailored for long-tailed visual recognition, attaining a remarkable state-of-the-art performance in specific benchmark categories. The improved accuracy on NIH-CXR-LT underscores STTP-Net's applicability to critical engineering systems like healthcare diagnostics, where detecting rare pathologies is vital.

The rest of the chapter is organized in the following way. Section 2 reviews related works and provides context. Section 3 analyzes the existing data augmentation methods and sampling techniques. Section 4 outlines the proposed methods. Section 5 discusses the quantitative and qualitative findings. Section 6 does an ablation study of the proposed method, and Section 7 concludes with some intriguing future research avenues.

3.2 Related Works

Researchers deploy data resampling methods to address challenges stemming from imbalanced training data (Johnson and Khoshgoftaar, 2019). These techniques fall into two categories: Undersampling and Oversampling. Undersampling approaches (He and Garcia, 2009; Buda et al., 2018a; Japkowicz and Stephen, 2002a; Van Hulse et al., 2007b) tackle data imbalance by removing samples from the majority class. While effective for slightly imbalanced larger datasets, they risk information loss and data variability issues, especially with smaller datasets or significantly imbalanced class distributions. In contrast, oversampling methods aim to balance the dataset by increasing

the number of minority class samples. Random oversampling (Buda et al., 2018a; Van Hulse et al., 2007b), a simple form of oversampling, duplicates all minority class samples for balance. However, this computationally efficient approach amplifies the dataset without enhancing data variability, often leading to overfitting. A more advanced oversampling technique, Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002a), addresses this by generating synthetic samples through interpolation between pairs of existing minority samples. Multiple SMOTE variants, like Borderline-SMOTE (Han et al., 2005a) and Safe-level SMOTE (Bunkhumpornpat et al., 2009), have emerged following SMOTE's success. Recently, Mondal et al. (2024) (Mondal et al., 2024b) introduced the CCO technique, which designates "Cluster Cores" as dense central regions in the feature space and employs a clustering-driven oversampling strategy to generate synthetic samples within the convex hull of minority class points. However, these traditional methods, while effective in machine learning, struggle with scalability for image datasets due to the computational demands of nearest-neighbor calculations and remain inadequate for modeling intricate, high-dimensional data distributions due to their reliance on Euclidean distance metrics and linear interpolation.

Beyond resampling techniques which used in traditional machine learning, there is an active research strand involving the generation of synthetic data points for the deep learning techniques. Deep oversampling techniques such as Deep Over-Sampling (DOS) (Ando and Huang, 2017), Generative Adversarial Minority Oversampling (GAMO) (Mullick et al., 2020), and DeepSMOTE (Dablain et al., 2022) aim to mitigate class imbalance by generating synthetic samples within a learned feature space rather than performing naive resampling. DOS (Ando and Huang, 2017) extends traditional over-sampling to the deep feature space of a convolutional neural network (CNN) by interpolating synthetic embeddings from a linear subspace of in-class neighbors, reducing intra-class variance and improving feature discrimination. However, its assumption of local linearity does not always hold, particularly for highly complex or non-linearly separable data, and its iterative target updating adds computational overhead. To address this, GAMO (Mullick et al., 2020) adopts a generative adversarial approach, where a convex generator creates synthetic minority samples that fool both a classifier and a discriminator, thereby refining decision boundaries. While this adversarial game effectively positions synthetic instances near minority class peripheries, GAMO suffers from mode collapse, leading to poor sample diversity, and its need for a separately trained generator further increases computational costs. DeepSMOTE (Dablain et al., 2022) attempts to balance these trade-offs by combining an encoder-decoder framework with SMOTE-based interpolation, generating synthetic samples in feature space before decoding them back into the original domain. Unlike GAN-based methods, it avoids discriminator training, reducing complexity, but its reliance on nearest-neighbor selection remains computationally expensive and ineffective in sparse minority distributions. While these methods have shown promise in handling standard class imbalance, they remain inadequate for long-tailed settings where extreme imbalance introduces further challenges. The presence of a large number of classes in long-tailed distributions makes these meth-

ods increasingly complex, as architectures must scale with class diversity, rendering them infeasible for real-world long-tailed imbalances. To address the long-tailed setting, approach known as Balancing Long-Tailed Datasets with Adversarially-Perturbed Images (BLT) (Kozerawski et al., 2021) has been introduced. BLT (Kozerawski et al., 2021) employs a gradient-ascent-based adversarial image generation technique to perturb tail-class images into hard examples, which are then used for training. Here, 'adversarial perturbations' refer to gradient-level pixel shifts that can create visual artifacts ('unnatural distortions') not present in real data. Unlike generative models such as GANs or VAEs, BLT does not require additional neural networks, reducing training complexity. While BLT improves tail-class accuracy without compromising head-class performance, its reliance on adversarial perturbations may lead to unnatural distortions that do not always align with real-world data distributions. Additionally, its effectiveness depends on the quality of the confusion matrix, which may not generalize well across different datasets.

Rather than generating additional samples, an alternative paradigm focuses on adjusting the learning dynamics by modifying the contribution of different samples or classes during training, which seems more feasible in case of long-tailed distribution. This shift in strategy has led to the emergence of re-weighting techniques, which offer a more direct and computationally efficient way to address long-tailed class imbalance. Unlike resampling or synthetic data point generation, re-weighting strategies work by assigning different weights to different classes and samples. The most traditional approach assigns class weights inversely proportional to the number of samples (Huang et al., 2016b; Wang et al., 2017). However, this naive strategy can lead to overcompensation, causing minority class overfitting. More advanced formulations have been proposed to mitigate such limitations. A widely used method for re-weighting long-tailed datasets is LDAM-DRW (Label-Distribution Aware Margin with Deferred Re-Weighting) (Cao et al., 2019), which enhances generalization by combining class-specific decision margins with an adaptive re-weighting schedule. The LDAM loss modifies the cross-entropy loss by introducing a margin that is inversely proportional to the number of samples in each class. This ensures that minority classes receive larger margins, preventing the classifier from developing overly tight decision boundaries around head classes, which typically dominate training. LDAM is often used in conjunction with Deferred Re-Weighting (DRW), a two-phase training strategy where uniform class weighting is initially used for better feature learning and class-balanced re-weighting is applied in later stages to refine decision boundaries. While LDAM-DRW improves tail-class performance, it may still be sensitive to outliers that can distort decision margins. Another loss known as Class Balanced Loss (Cui et al., 2019b) introduces the concept of *effective number of samples* to determine class weights dynamically. Instead of relying on simple inverse frequency, the effective number of samples for a class is defined. This formulation prevents extreme weighting while maintaining balanced weights across classes. However, its reliance on hyperparameters requires careful tuning, and in cases of severe imbalance, the loss may still suffer from performance degradation in the tail classes. Another refined approach

is Balanced Softmax (Ren et al., 2020a), which directly modifies the softmax normalization to integrate class frequency priors. Instead of computing standard probabilities, Balanced Softmax introduces frequency-based corrections into the logit computation. This adjustment ensures that majority classes are not disproportionately favored by the classifier, thereby mitigating bias. While Balanced Softmax improves class-level balance, it assumes that the dataset distribution remains stable throughout training, which may not be valid in dynamically evolving datasets. Beyond class-wise re-weighting, instance-level re-weighting techniques provide an alternative mechanism by dynamically adjusting sample importance during training. One widely used approach is Focal Loss (Lin et al., 2018), which down-weights well-classified instances to emphasize hard examples. This loss ensures that easily classified instances contribute minimally to gradient updates. While this improves performance on minority classes, Focal Loss assumes that all hard-to-classify instances are useful, which may not hold in scenarios with label noise or highly overlapping class distributions.

A separate line of work focuses on various image augmentation strategies to increase the frequency of tail classes. Image augmentation strategies have shown favorable results in dealing with imbalanced datasets. Traditional augmentation strategies like rotation, cropping, flipping, etc. are widely used to mitigate data imbalance. They are applied on input images, and hence, they inherently limit the data variability of minority classes in long-tailed datasets. To counter this problem, *Implicit Semantic Data Augmentation (ISDA)* (Wang et al., 2019b) performs semantic augmentations on images by statistically estimating the class-wise covariance matrices of the data points. This calculation, however, depends on the calculation of covariance matrices, which is difficult to compute for minority classes in long-tailed datasets due to the extremely limited number of samples. *MetaSAug* (Li et al., 2021), introduced by Li et al. (2021), provides a framework to optimize the augmentation strategy of minority classes dynamically during training. Instead of relying on precomputed covariance matrices, MetaSAug (Li et al., 2021) estimates the class-wise covariance matrices "Σ_i" in an online meta-learning fashion, applying semantic augmentations that enhance the performance of ISDA in long-tailed datasets.

Mixup (Zhang et al., 2017a) and Remix (Chou et al., 2020) are augmentation techniques where images are created by interpolating between two images. Remix favors the minority classes while doing this interpolation, pushing the decision boundary toward the majority class. However, Mixup-based methods assume that interpolated feature representations retain semantic meaning, which may not always hold in highly imbalanced distributions where minority class characteristics are not well represented. CutMix (Yun et al., 2019a) cuts out random patches from images and fills the empty patch with other training images. Although CutMix improves generalization, it may suffer in long-tailed settings where tail-class features are underrepresented, making the mixed images overly influenced by majority classes. Park et al. (Park et al., 2022) describes a method that utilizes the context-rich majority class images as background images and takes cutout minority class images as foreground images. By explicitly placing the cutouts of minority class instances in

majority class instances as a background, it diversifies the minority samples.

While augmentation techniques provide a valuable alternative to oversampling or reweighting methods, their effectiveness is highly dependent on the characteristics of the data set. Simple transformations like cropping and flipping offer limited benefits in cases of extreme imbalance, while Mixup-based approaches require sufficient feature diversity to produce meaningful interpolations. Advanced techniques like ISDA and MetaSAug improve semantic augmentation but introduce computational overhead due to covariance estimation. Similarly, methods like CutMix and contextual augmentation refine class distributions but may still struggle with severe underrepresentation of tail classes, leading to suboptimal decision boundaries. As a result, augmentation-based techniques often benefit from integration with other strategies, such as meta-learning frameworks or adaptive loss functions, to enhance robustness in long-tailed recognition.

Another area of research explores dual-branch networks, which aim to balance feature learning and classification by leveraging multiple training streams. Zhou et al. introduced a dual-branched network called BBN (Zhou et al., 2020), where one branch is trained on the original data distribution while the other is trained using a reverse sampler that prioritizes tail classes by increasing their sampling probability and reducing that of head classes. Both branches share a common ResNet backbone for feature extraction, and their outputs are combined using a cumulative learning approach to maintain overall class balance. Building upon this, ResLT (Cui et al., 2022) extends BBN (Zhou et al., 2020) into a triple-branch architecture by categorizing the dataset into three distinct subsets: head, median, and tail classes. Each branch specializes in different class combinations—(head, median, tail), (median, tail), and (tail)—which enhances parameter specialization and improves tail-class recognition. Residual connections aggregate the outputs, ensuring smoother feature adaptation across imbalanced class distributions. Another state-of-the-art approach, Routing Diverse Distribution-Aware Experts (RIDE) (Wang et al., 2021b), introduces a sophisticated multi-expert network where each branch employs a specific loss function and incorporates a diversity loss based on divergence measures. This enforces collaboration among multiple experts, leading to mutually reinforcing predictions. While RIDE enhances generalization and robustness in long-tailed classification, its reliance on multiple experts significantly increases computational overhead and memory consumption, making it less scalable for large-scale datasets. The Cumulative Dual-Branch Network Framework (CDBNF) (Fan et al., 2022) extends the dual-branch paradigm by integrating class imbalance learning and feature representation learning simultaneously. One branch focuses on learning from imbalanced data to improve tail-class classification, while the other branch employs few-shot learning to enhance feature representation. Additionally, the framework implements a Cumulative Learning Strategy (CLS), progressively emphasizing tail classes throughout training. Despite its effectiveness, CLS requires careful tuning of curriculum schedules, and the reliance on two separate optimization objectives may lead to instability if not well-calibrated. To further refine long-tailed learning, the Bilateral expert-based Feature Redundancy Reduction

and Rebalancing (Bi-F3R) (Chen et al., 2023) approach was introduced, drawing inspiration from information theory to address the trade-off between tail-class and head-class performance. Bi-F3R actively removes redundant information from well-represented classes while preserving critical tail-class features. This is achieved through a feature redundancy extraction module, along with two complementary loss functions: one encouraging the model to forget redundant information, and the other guiding it to learn more discriminative representations for rare classes. While Bi-F3R improves tail-class differentiation, its dependency on explicitly modeling feature redundancy can introduce sensitivity to dataset shifts, requiring additional regularization strategies for robustness.

Although dual-branch architectures and multi-expert frameworks have demonstrated strong performance in long-tailed learning, they introduce significant architectural complexity. These models often require additional computational resources for multiple branches or expert modules, making them challenging to scale efficiently. Furthermore, designing an optimal fusion strategy for multi-branch outputs remains an open research problem, as naive aggregation may fail to preserve class-discriminative properties.

Beyond architectural modifications and augmentation strategies, recent advancements in representation learning have introduced alternative approaches for tackling long-tailed distributions. One such direction is contrastive learning, which focuses on learning discriminative feature representations by pulling similar instances together and pushing dissimilar ones apart in the embedding space. Wang et al. (Wang et al., 2021a) propose an innovative hybrid network design that combines a supervised contrastive loss for image representation learning and a cross-entropy loss for classifier learning. The learning process transitions gradually from acquiring features to learning classifiers. However, contrastive learning suffers from poor uniformity in feature space, which is an inherent nature of long-tailed data distributions. Hence, Li et al. (Li et al., 2022e) propose a novel framework that enhances the uniformity of feature distribution on the hypersphere, thus improving the performance of supervised contrastive learning. Despite these improvements, contrastive learning-based approaches require large batch sizes and effective negative sampling strategies, which can be computationally expensive and challenging to optimize for highly imbalanced datasets.

Another prominent research direction focuses on two-stage learning strategies, which aim to mitigate class imbalance by decoupling feature representation learning from classifier training. Two-stage methods train a single network for both representation and classifier learning. However, this can hamper the learning process as there is overwhelmingly more information about the head classes than in the tail classes. Hence, decoupling the Representation Learning and Classifier training stages to create a two-stage approach has been gaining popularity. In their work, Kang et al. (Kang et al., 2020) demonstrated decoupled representation and classifier learning, where they trained the representation learning network on uniformly sampled data points and the classifier on class-balanced samples. Zhou et al. (Zhou et al., 2020) later integrated Mixup into this learning paradigm. While these methods improve generalization across imbalanced distributions, they rely

on the assumption that a well-trained feature extractor can generalize to tail classes, which may not hold when tail-class features are severely underrepresented. Moreover, the effectiveness of two-stage training is dependent on the choice of balancing strategies during classifier training, requiring additional hyperparameter tuning.

Another crucial aspect of long-tailed learning is calibration, as neural networks tend to be overconfident in their predictions, particularly for head-class samples. Miscalibration is a vital issue in long-tailed classification, where the predicted probability distribution of classes is heavily influenced by the number of training samples per class. Moreover, classifiers tend to be highly overconfident in the head classes. To solve these issues, Zhong et al. (hong et al., 2021) proposed Shifted Batch Normalization and Label-Aware Smoothing, which, when combined with Mixup (Zhang et al., 2017b), effectively counteract varying degrees of over-confidence in classifiers. Although these techniques improve model calibration, they require additional hyperparameter tuning to determine the extent of smoothing and shifting required for different imbalance levels. Furthermore, these methods assume that label smoothing and batch normalization shifts will be sufficient to correct calibration errors, which may not generalize well to extremely skewed datasets.

Finally, evaluation practices in long-tailed learning pose a fundamental challenge, as conventional validation sets often fail to reflect real-world class distributions. Long-tailed classifiers are generally validated on a test set with a uniform class distribution. Hong et al., in their work LADE (Hong et al., 2021), argue that this practice is suboptimal, as real-world datasets may follow arbitrary class distributions. Hence, they propose to disentangle the source label distribution from the model prediction so that it can adapt to any random probability distribution during inference. While LADE provides a more flexible evaluation framework, it requires an accurate estimation of prior class distributions at inference time, which is often unavailable in real-world scenarios. Additionally, disentangling label distributions may lead to suboptimal performance when the underlying data shifts significantly.

Critical Limitations of Existing Approaches and Our Resolution. Despite significant advancements in long-tailed learning, existing methods exhibit fundamental limitations that hinder their effectiveness in real-world scenarios. Resampling techniques and synthetic sample generation methods (e.g., GAMO) often suffer from computational inefficiency, mode collapse, and limited sample diversity, restricting their generalizability. Re-weighting strategies (e.g., LDAM, Focal Loss) attempt to balance class contributions but frequently degrade head-class accuracy, leading to biased decision boundaries. Augmentation-based approaches (e.g., Mixup, CutMix) lack tailored mechanisms for handling long-tailed data, failing to ensure sufficient feature diversity across minority classes. Dual-branch architectures (e.g., BBN, ResLT) enhance tail-class recognition but often come at the expense of head- and median-class performance, while more complex frameworks (e.g., RIDE, CDBNF) introduce significant architectural and computational overhead. Two-stage methods, which decouple representation learning from classifier training, mitigate class imbalance

but require additional training phases and careful tuning, making them less practical for large-scale applications.

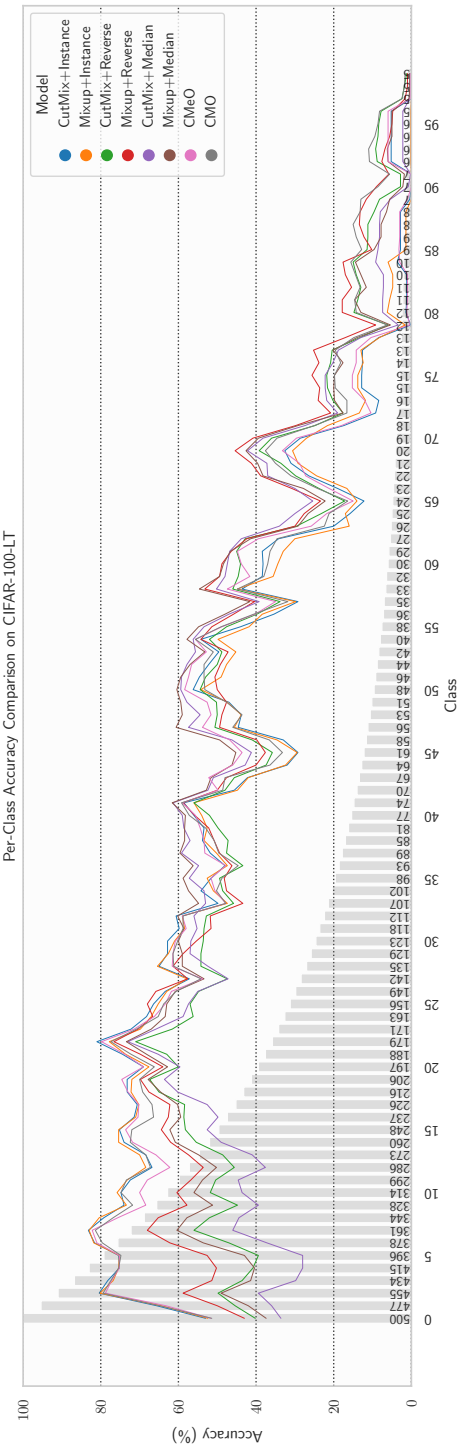
To overcome these limitations, we propose STTP-Net, a novel framework integrating three key innovations: (1) a Hybrid Mixup strategy, which combines Context-Rich Median Oversampling (CMeO) with reverse-sampled Mixup to promote diverse feature representations across all class distributions; (2) a Two-Pronged Network Architecture, featuring a shared feature extractor and dual classifiers specialized for head/median and tail classes, thereby eliminating performance trade-offs; and (3) an Effective Balanced Softmax (EBS) loss, which dynamically adjusts label distributions to mitigate bias and prevent overfitting. By holistically addressing computational inefficiency, label shift, and class-specific performance degradation, STTP-Net achieves state-of-the-art results without sacrificing head-class accuracy, offering a robust and scalable solution for long-tailed recognition.

In the above discussion, our method is closely related to the dual-branch network. We attempt to address the drawbacks of these methods, as most of them aim to improve the performance of the tail classes. However, in the process, they hinder the performance of the head and median classes. We propose a method that also emphasizes the performance of both the head and median classes, in addition to enhancing the performance of the tail classes.

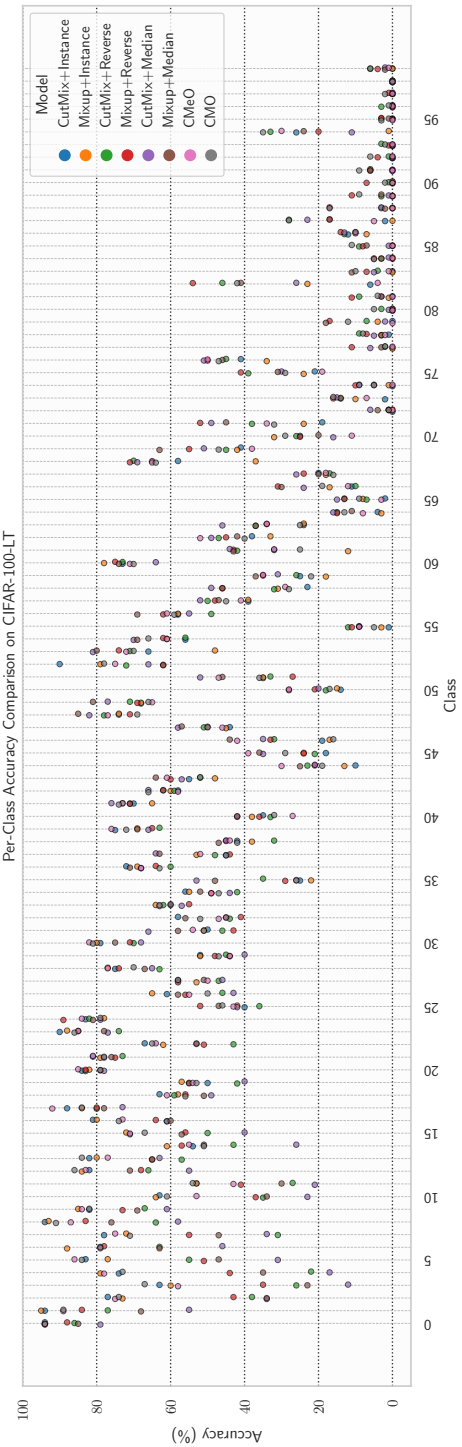
3.3 Analysis of Augmentation Method along with different sampling strategies

Augmentation Strategies. Data augmentation plays a crucial role in improving the generalization performance of deep neural networks by augmenting the training data to create diverse samples without collecting additional labeled data. In this section, we introduce state-of-the-art data augmentation techniques, CutMix(Yun et al., 2019a) and Mixup(Zhang et al., 2017b), along with these also discuss CMO (Context rich minority oversampling technique) (Park et al., 2022) and define new augmentation technique as the Context rich median oversampling (CMeO). Let us explore how these techniques can improve representation learning and their impact on different data sampling strategies in a long-tail imbalanced scenario.

Mixup(Zhang et al., 2017b) is a powerful data augmentation technique proposed by Zhang et al. (2017) that encourages the model to learn more generalized and smoother decision boundaries. Mixup operates at the pixel level by linearly interpolating pairs of images and their corresponding labels. Specifically, the algorithm creates a new training sample for two randomly chosen images by combining their pixel values and labels in the same ratio. Blending different images prevents overfitting and reduces the likelihood of the model memorizing noise in the training data. Moreover, Mixup encourages the network to exhibit greater sensitivity to variations in the input space, enhancing its ability to handle out-of-distribution samples more effectively. Mixup can represented



(a) Bar plot shows class-wise sample counts, with overlaid smoothed accuracy curves for method comparison.



(b) Strip plot of class-wise accuracy, with color-coded dots for different methods.

Figure 3.1: Comparison of the per-class accuracy on CIFAR-100-LT across various combinations of augmentations and sampling techniques. (a) The light-colored bar plot in the background represents the sample count for each class, with the corresponding values displayed at the bottom of each bar. Overlaid on this plot are smoothed line plots depicting accuracy across different classes, providing a clearer comparison of how each method performs on various classes. (b) Strip plot displaying accuracy values for different classes, with color-coded dots representing various methods for each class. (Based View in 300% zoom and in color)

as mixing two data points using a parameter λ as:

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j, \tag{3.1}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda)y_j, \tag{3.2}$$

where (x_i, x_j) are the inputs from classes i and j, and (y_i, y_j) are the one hot encoded labels of the corresponding inputs. The parameter λ is drawn from a beta distribution i.e $\lambda \sim \beta(\alpha, \alpha)$. The choice of a symmetric Beta distribution follows established Mixup protocols. Mixup works by regularizing the network to favor simple linear behavior between training samples, which effectively smooths the decision boundaries. The parameter α controls the intensity of this interpolation; smaller values ensure the model remains focused on the primary characteristics of the rare tail classes while still benefiting from the broadened feature space. $(\tilde{x}, \tilde{y})$, is the newly generated data point, with data as $\tilde{x}$, and corresponding label as $\tilde{y}$.

CutMix(Yun et al., 2019a) introduced by Yun et al. (2019) is a data augmentation technique introduced as an extension of the traditional Cutout and Mixup methods. The key idea behind CutMix is to combine information from different images by cutting and pasting random patches during training. This process encourages the model to learn robust features and enhances its generalization ability to unseen data. Blending two images through CutMix results in a more diverse dataset, effectively preventing the model from relying too heavily on specific image regions during training. By introducing this novel form of regularization, CutMix mitigates overfitting and improves the network's performance on challenging test samples. CutMix generates a new training sample $(\tilde{x}, \tilde{y})$ by combining two training samples (x_1, y_1) and (x_2, y_2). combining operation defined as:

$$\tilde{x} = \mathbf{M} \odot x_1 + (\mathbf{1} - \mathbf{M})x_2, \tag{3.3}$$

$$\tilde{y} = \lambda y_1 + (1 - \lambda)y_2, \tag{3.4}$$

where $\mathbf{M} \in \{0, 1\}^{W \times H}$ is a binary mask that identifies areas for dropout and replacement from two images. The symbol $\mathbf{1}$ represents a binary mask filled entirely with ones, and $\odot$ denotes element-wise multiplication.

CMO (context-rich minority oversampling technique) (Park et al., 2022) is the modification of the CutMix for the long-tail imbalance dataset, which mixes the sample similarly to CutMix but considers the background as the image taken from the majority classes and pastes the masked image taken from the minority classes.

Based on this, we introduce the *Context rich median oversampling (CMeO)* technique, which is similar to the CMO. However, the difference is that we focus on the Median classes. That is, while applying CutMix, the background image is chosen from the head classes, but the masked

image pasted on it will be taken from the Median classes.

Sampling strategies. Moreover, Sampling techniques play a crucial role in training machine learning models, especially when dealing with imbalanced datasets where the distribution of classes is uneven. Let us define the sampling using the common formula to represent the different sampling strategies:

$$p_j = \frac{(|n_j - \psi| + \epsilon)^q}{\sum_{i=1}^{C} (|n_i - \psi| + \epsilon)^q}, \tag{3.5}$$

where C total classes, $q \in [-1, 1]$, ψ represent some statistics corresponding to the number of samples in the dataset, and ϵ is the small value added to avoid dividing by zero error. Different sampling strategies arise based on the different values of the q and ψ.

Instance-balanced sampling. For instance sampling, the probability p_j^{inst} is given by Eqn 3.5 by using $q = 1$ and $\psi = 0$, i.e., a data point from class j will be sampled proportionally to the cardinality n_j of the class in the training set.

Reverse-balanced sampling. For Reversed sampling, the probability p_j^{rev} as per Eqn 3.5 is calculated using $q = 1$ and $\psi = N$, where $N = \sum_i^C n_i$, i.e. a data point from the minority class will be sampled with the high probability corresponding to the sample from the majority class.

Median-balanced sampling. For Median sampling, the probability p_j^{med} the Eqn 3.5 by using $q = -1$ and $\psi = M$, where $M = Med\{n_1, n_2, n_3, \cdots, n_i, \cdots, n_C\}$, and Med is the function to calculate the median from the given numbers i.e., a data point from the median (or medium) classes will be sampled with the high probability corresponding to the sample from the majority or minority classes.

Now, we need to analyze how combining different augmentation methods with different sampling strategies affects the network's performance. Moreover, to show the performance of different classes, we have divided the classes into six groups for better performance analysis: Head of Head (HH), Tail of Head (TH), Head of Medium (HM), Tail of Medium (TM), Head of Tail (HT), and Tail of Tail (TT). This approach provides a more comprehensive understanding of performance instead of dividing the classes into only three groups (Many, Medium, and Few).

We conducted experiments on the CIFAR-100-LT dataset with an imbalance ratio of 100, trained on Resnet32 using cross-entropy loss. We tested 8 combinations of sampling strategies and augmentation techniques, as depicted in Figure 3.1. The figure displays the accuracy of various classes for different sampling and augmentation strategy combinations. The bar plot on the same figure indicates the number of samples in each category. From Figure 3.1, we can see that the CutMix augmentation technique with the instance sampler gives higher head class accuracy than the tail and median classes. In contrast, in the case of the tail classes, the mixup augmentation, along with the combination of the reverse sampler, gives much better accuracy compared to the head

Table 3.1: Comparison of performance metric accuracy across different groups for various augmentation methods and samplers.

(a) Head of Head and Tail of Head (Table 1-a)

Augmentation Method	**Sampler Used**	**Head of Head** ($n > 250$)	**Tail of Head** ($250 \geq n > 100$)
Mixup	Instance	77.2	66.2
	Reverse	60.13	62.05
	Median	53.2	61.8
CutMix	Instance	76.8	67.35
	Reverse	50.47	57.4
	Median	41.00	57.20
CMeO	Instance + Median	74.4	65.70
CMO	Instance + Reverse	76.00	64.30

(b) Head of Median and Tail of Median (Table 1-b)

Augmentation Method	**Sampler Used**	**Head of Median** ($100 \geq n > 50$)	**Tail of Median** ($50 \geq n > 25$)
Mixup	Instance	46.87	38.14
	Reverse	49.6	46.07
	Median	57.67	49.21
CutMix	Instance	47.33	41
	Reverse	48.73	44.14
	Median	56.33	49.14
CMeO	Instance + Median	52.27	46.21
CMO	Instance + Reverse	48.07	41.86

(c) Head of Tail and Tail of Tail (Table 1-c)

Augmentation Method	**Sampler Used**	**Head of Tail** ($25 \geq n \geq 15$)	**Tail of Tail** ($n < 15$)
Mixup	Instance	18.00	3.04
	Reverse	29.50	11.50
	Median	26.75	8.33
CutMix	Instance	16.92	3.67
	Reverse	24.25	10.00
	Median	27.17	6.42
CMeO	Instance + Median	17.92	4.29
CMO	Instance + Reverse	23.33	11.50

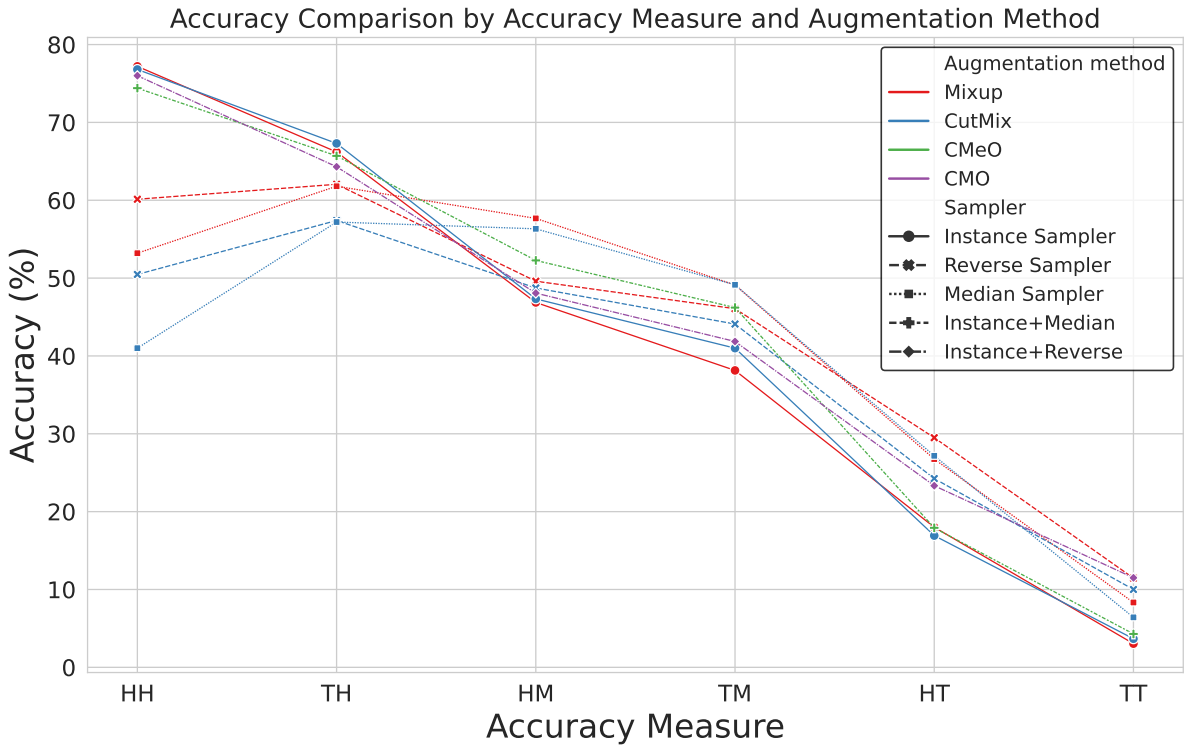


Figure 3.2: Plot showing Comparison based on accuracy for different groups for different augmentation methods and samplers

and median classes. Let us look at the Table 3.1 for a more detailed analysis. The table compares performance metrics, specifically accuracy, for different augmentation methods and samplers across various class subsets. Mixup with Instance sampler excels in the "HH" and "TH," achieving high accuracy (77.2% and 66.2%, respectively). In contrast, Mixup with Reverse sampler performs well in the "HT" and "TT," with notable accuracy (29.5% and 11.5%, respectively). CutMix with Instance sampler is effective for "HH" and "TH," achieving high accuracy (76.8% and 67.35%, respectively), while CutMix with Reverse sampler stands out in "TT" with a significant accuracy of 10%. CMeO demonstrates balanced performance across subsets, particularly excelling in "TH" (65.7%) and "TM" (52.27%). On the other hand, CMO is well in "HT" (23.33%) and "TT" (11.5%). To summarize, the selection of techniques relies on the attributes of the data subset. CMeO displays potential for balanced results in the head and medium classes, while mixup with reverse sampler performs well in the tail classes, with over 6% improvement compared to CMO in the "HT" classes. Well, as seen from Figure 3.2, CMeO shows balanced results till "TM". After that, the mixup with the reverse sampler started showing better performance. This analysis highlights the significance of customizing augmentation strategies for specific subsets to achieve optimal outcomes.

3.4 Proposed Methodology

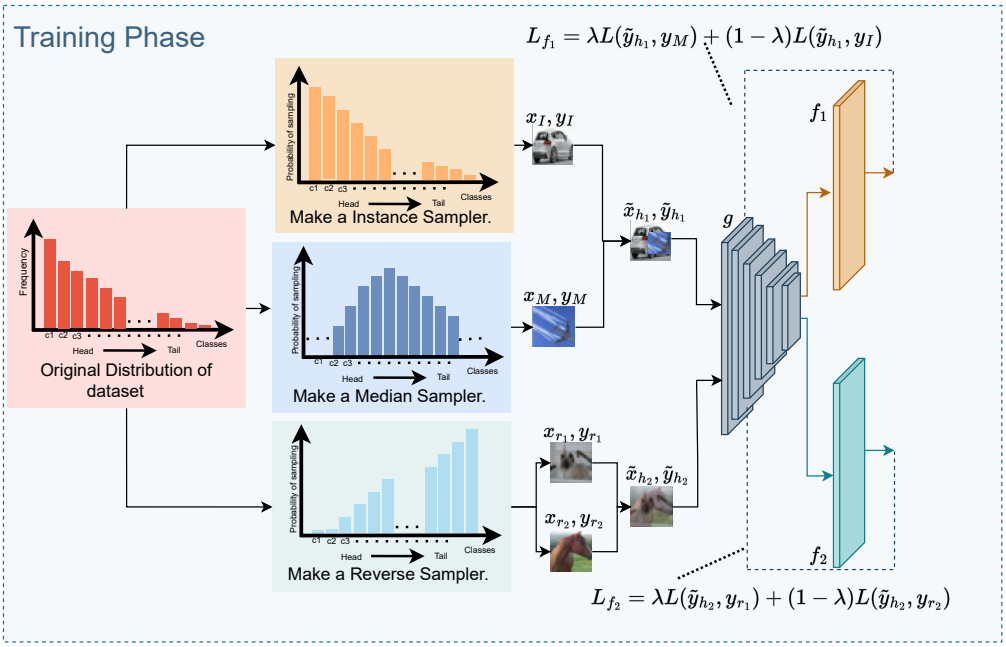
Based on our analysis of the previous section, we have devised a new data augmentation technique that is ideal for long-tail imbalanced datasets. In this section, we introduce **STTP-Net**, a novel Sampling-Tailored Two-Pronged Network for long-tailed classification. Our approach leverages a *hybrid augmentation strategy (HybridMix)* and a reweighted loss function i.e. *Effective Balanced Softmax (EBS) Loss* to mitigate class imbalance by redistributing decision boundaries. The proposed method ensures robust feature learning across all class distributions by integrating adaptive data sampling, augmentation-driven feature expansion, and classifier bias correction. The overall framework is illustrated in Figure 3.3. We begin by formalizing the long-tailed classification problem and defining class groups (Subsection 3.4.1), followed by an overview of the STTP-Net architecture (Subsection 3.4.2). Next, we detail HybridMix (Subsection 3.4.3) and than derive the Effective Balanced Softmax (EBS) function 3.4.4 followed by the derivation of EBS Loss (Subsection 3.4.4.1) to address label distribution shifts. The two-phase training strategy (Subsection 3.4.5) is then explained, balancing hybrid augmentation with boundary refinement, and finally, the inference mechanism (Subsection 3.4.6) is described, where weighted logit fusion ensures balanced predictions. This structured progression ensures systematic alignment between theoretical formulation, implementation, and evaluation.

3.4.1 Problem Definition and Notation

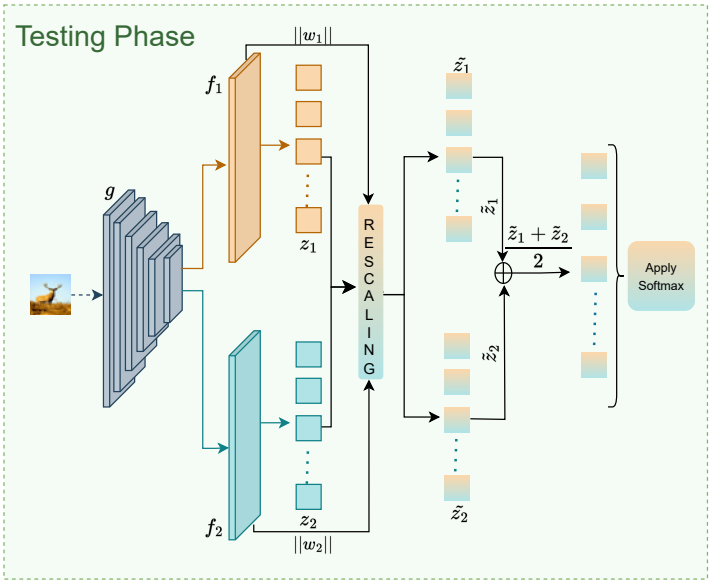
We define the long-tailed classification problem as a multi-class learning task where the class distribution is highly imbalanced, leading to biased model predictions. Given a dataset $D = (X, Y)$, where $X = \{x_1, x_2, \ldots, x_N\}$ represents input images and $Y = \{y_1, y_2, \ldots, y_N\}$ are their corresponding labels, the goal is to train a classifier that maintains performance across all classes despite significant differences in sample sizes. Each image $x_i \in \mathbb{R}^{W \times H \times C}$ belongs to one of k classes, where the number of samples per class n_i varies significantly, forming a long-tailed distribution. The dataset size is defined as $N = \sum_{i=1}^{k} n_i$, where n_i represents the number of instances in class i. The classes are labeled such that $n_1 > n_2 > ... > n_k$, meaning the head classes have the highest sample counts while the tail classes have the least. To handle the imbalance, we define three distinct class groups:(1) *Head classes:* Classes with a large number of samples ($n_i > 100$). (2) *Medium classes:* Classes with a moderate number of samples ($20 \leq n_i \leq 100$).(3) *Tail classes:* Classes with very few samples ($n_i < 20$).

3.4.2 An Overview of our Framework

As shown in the figure 3.3, our proposed method consists of a shared backbone network g to extract the image features or to learn the representation of the image, followed by the multi-head, multi-layer perceptron network f_1, and f_2 for classification. In the first stage, three different data



(a) Training Phase of STTP-Net



(b) Testing Phase of STTP-Net

Figure 3.3: Above, the training phase is shown, where the backbone network g, with two heads f_1 and f_2, are trained using our hybrid mixup technique using effective balanced softmax loss. Below, the testing phase is shown, which shows how the network is used during inference time.

samplers are constructed. The data are sampled from these samplers and are augmented according to our proposed *HybridMix* approach and training is done using the proposed *Effective Balanced Softmax* (EBS) Cross-Entropy loss function. Now, we define each part of our algorithm one by one, starting with the augmentation strategies followed by the proposed effective balanced softmax function to tackle the problem of label distribution shift.

3.4.3 HybridMix: Combined augmentation with diverse sampling

Modified CutMix strategy between head classes and medium classes: We will be going to apply the *Context rich median oversampling (CMeO)* technique discussed above. It selects one sample from the instance sampler, and another from the median sampler. Mathematically, let (x_I, y_I) be the sample selected from the instance sampler and (x_M, y_M) be the sample selected from the median sampler. Now perform the CutMix between these two samples taken from the different samplers. Such that:

$$\tilde{x}_{h_1} = \mathbf{M} \odot x_M + (\mathbf{1} - \mathbf{M}) \odot x_I, \tag{3.6}$$

$$\tilde{y}_{h_1} = \lambda y_M + (1 - \lambda) y_I, \tag{3.7}$$

where $\mathbf{M} \in \{0,1\}^{W \times H}$ is a binary mask that identifies areas for dropout and replacement from two images. The symbol $\mathbf{1}$ represents a binary mask filled entirely with ones, and $\odot$ denotes element-wise multiplication. To sample $\mathbf{M}$, we first get the bounding box coordinates $\square = (cx, cy, wd, ht)$ for the cropping regions. We remove the region $\square$ from x_I and fill it with the corresponding patch from the x_M. $(\tilde{x}_{h1}, \tilde{y}_{h_1})$, is the newly generated data point, with data as $\tilde{x}_{h1}$, and corresponding label as $\tilde{y}_{h1}$. To get the $\square$ coordinates, $cx \sim \mathcal{U}(0, W)$, $cy \sim \mathcal{U}(0, H)$, $wd = W\sqrt{1-\lambda}$, and $ht = H\sqrt{1-\lambda}$. The mask $\mathbf{M} \in \{0,1\}^{W \times H}$ is decided by filling with 1 within the $\square$ in $\mathbf{M}$, and rest of the values with 0.

Mixup on the samples taken from the reverse sampler: Let (x_{r_1}, y_{r_1}) and (x_{r_2}, y_{r_2}) be the samples selected from the reverse sampler. Now perform the mixup between these two samples as:

$$\tilde{x}_{h_2} = \lambda x_{r_1} + (1 - \lambda) x_{r_2}, \tag{3.8}$$

$$\tilde{y}_{h_2} = \lambda y_{r_1} + (1 - \lambda) y_{r_2}, \tag{3.9}$$

The parameter λ is drawn from a beta distribution $beta(\alpha, \alpha)$. $(\tilde{x}_{h_2}, \tilde{y}_{h_2})$, is the newly generated data point, with data as $\tilde{x}_{h_2}$, and corresponding label as $\tilde{y}_{h_2}$. In the pursuit of optimizing our network training methodology, we incorporate the hybrid mixup technique. Notably, we strategically allocate the final epochs (τ) for distinctively training our network. During this phase, we deliberately abstain from applying the hybrid mixup, opting instead to leverage samples directly from

the instance sampler for the Head 1 classifier (f_1) and samples from the reverse sampler for the Head 2 classifier (f_2), as elucidated in Algorithm 2. This nuanced approach harnesses the benefits of a hybrid mixup for enhanced representation learning. Furthermore, the deliberate omission of a hybrid mixup in the concluding epochs serves a specific purpose—it facilitates more insightful segregation of decision boundaries, thereby contributing to the overall efficacy of our model.

3.4.4 Effective Balanced Softmax (EBS)

Under the long-tailed scenario, the Softmax function exhibits intrinsic bias. Consequently, there arises a necessity to adapt the Softmax function to capture the shift in label distribution during test time explicitly. We will see from the probabilistic perspective how it can be modified. Let's consider our classification task, where the goal is to train a model for accurately predicting the class y when presented with an input image x. The distributions of the training and test sets are denoted as $\hat{p}(x, y)$ and $\tilde{p}(x, y)$, respectively. In real-world scenarios, a model trained on a dataset with a long-tailed class distribution often faces challenges due to a shift in label distribution. This implies that the distribution of labels in the training set, $\hat{p}(y)$, differs from that in the test set, $\tilde{p}(y)$, expressed as $\hat{p}(y) \neq \tilde{p}(y)$. However, it is important to note that the conditional distributions remain consistent, meaning that $\hat{p}(x|y) = \tilde{p}(x|y)$. To solve this problem of label distribution shift, let's revisit the multiclass softmax regression problem, conceptualized as a multinomial distribution denoted by Φ:

$$\Phi = \Phi_1^{\mathbf{1}\{y=1\}} \Phi_2^{\mathbf{1}\{y=2\}} \Phi_3^{\mathbf{1}\{y=3\}} \dots \Phi_k^{\mathbf{1}\{y=k\}}, \tag{3.10}$$

where $\mathbf{1}(\cdot)$ is the indicator function, $\Phi_j = \frac{e^{\varphi_j}}{\sum_{i=1}^{k} e^{\varphi_i}}$, where $\varphi_j = f_\theta(g(x))[j]$ is the logit of class-j of the model, where x is the input image and $\sum_{j=1}^{k} \Phi_j = 1$. As per Bayes' rule, the interpretation of Φ_j can be expressed as:

$$\Phi_j = p(y = j|x) = \frac{p(x|y = j)p(y = j)}{p(x)}. \tag{3.11}$$

From the above equation, to address the class imbalance, our primary focus lies on the variable $p(y = j)$. While we assume that both the training and test datasets originate from an identical process $p(x|y = j)$, disparities may arise during testing due to variations in label distributions $p(y = j)$, as well as the available evidence $p(x)$. That is when we examine the label shift problem, wherein the target label distribution differs from the source (train) label distribution , it becomes apparent that the model's predictive estimates of $\hat{p}(y|x)$ cannot effectively anticipate the shifted distribution. This limitation arises from the significant interdependence between $\hat{p}(y|x)$ and $\hat{p}(y)$, as substantiated by Bayes' rule.

To address the difference in posterior distributions between the training and testing phases, we will use the concept of effective number (Cui et al., 2019b) as proposed by Cui et al. They consider a class and denote the set encompassing all potential data points within the feature space of this class as 'S'. They assume that the volume of 'S' is represented by '$\mathcal{N}$,' with '$\mathcal{N}$' being equal

to or greater than 1. Each data point is referred to as a subset of 'S', characterized by a unit volume of 1, allowing for potential overlaps between these subsets. This methodology aligns with the principles of a random covering problem. Within this framework, conduct a random sampling of subsets from 'S' to comprehensively cover the entire set 'S'. Notably, as the quantity of data samples drawn increases, the coverage of 'S' also improves; this, in turn, leads to an expected increase in the total volume of the sampled data, and such growth is bounded by '$\mathcal{N}$.' Thus, they define the effective number of samples as the anticipated volume of these individual samples. Mathematically, *effective number* is given by $E_n = (1 - \beta^n)/(1 - \beta))$, where $n \in \mathbb{Z}^+$ is the number of samples and $\beta = (\mathcal{N} - 1)/\mathcal{N}$.

In the context of our problem, to address the issue of differences in the posterior distributions of our training and testing phases, we can apply the *effective number* concept to find the probability of samples in the given class when there is an imbalance in the classes. We redefine the softmax function for imbalanced data and named it as *Effective Balanced Softmax*. Consider conditional probabilities $\hat{\Phi}$ for the imbalanced training set and $\tilde{\Phi}$ for the balanced testing set. And $\tilde{p}(x|y) = \hat{p}(x|y) = p(x|y)$, by the assumption that instances in both the test and training datasets originate from the identical process.

Theorem 1. *Let's consider $\tilde{\Phi}$ as the target conditional probability for a balanced dataset, defined as $\tilde{\Phi}_j = \tilde{p}(y = j|x) = \frac{\tilde{p}(x|y=j)}{\tilde{p}(x)} \frac{1}{k}$. Similarly, let's define $\hat{\Phi}$ as the target conditional probability for an imbalanced training set, given by $\hat{\Phi}_j = \hat{p}(y = j|x) = \frac{\hat{p}(x|y=j)}{\hat{p}(x)} \frac{1-\beta^{n_j}}{\sum_{i=1}^{k}(1-\beta^{n_i})}$. When expressing $\tilde{\Phi}$ using the standard Softmax function of the model's output φ, we can then represent $\hat{\Phi}$ as:*

$$\hat{\Phi}_j = \frac{\exp(\varphi_j)(1 - \beta^{n_j})}{\sum_{i=1}^{k} \exp(\varphi_i)(1 - \beta^{n_i})} \tag{3.12}$$

Outline of the Proof. Theorem 1 establishes that the posterior class probabilities $\hat{\Phi}_j$ under a label-imbalanced training distribution can be expressed as a reweighted softmax function over model outputs, where the weights are driven by a class-specific smoothing factor $(1 - \beta^{n_j})$. The proof begins by modeling both the balanced and imbalanced posteriors $\tilde{\Phi}_j$ and $\hat{\Phi}_j$ using the exponential family structure, exploiting the canonical link between the logit space and conditional probabilities. Starting from this formulation, we use the canonical transformation to express the logits in terms of the log-ratio between $\tilde{\Phi}_j$ and $\hat{\Phi}_j$. Through algebraic manipulation, we isolate the expression for $\hat{\Phi}_j$ in terms of φ_j and a ratio of the balanced-to-imbalanced probabilities. Substituting the definitions of $\tilde{\Phi}_j$ and $\hat{\Phi}_j$ derived from Bayes' rule under their respective distributions, we simplify the ratio into a form involving the known label priors and density estimates. Under the assumption that the conditional likelihood $p(x|y = j)$ is invariant across the balanced and imbalanced settings, we express the ratio of posteriors as a function of class frequencies and derive the final softmax-like form with class-dependent weights. The final equation confirms that $\hat{\Phi}_j$ is

normalized and sums to one over all classes, which is then rigorously verified through a separate induction proof. The detailed proof is provided in the supplementary material (see Section A.1.1)

Theoretical Analysis. The result offers an elegant probabilistic interpretation of the modified softmax form (termed *Effective Balanced Softmax*) as a likelihood-adjusted posterior under a distributionally shifted prior. Specifically, the reweighting term $(1-\beta^{n_j})$ reflects the marginal label frequency, correcting the bias induced by imbalanced training distributions. This formulation not only preserves the normalization property of softmax but also implicitly aligns the model's predictions with the class distribution observed at test time (typically assumed uniform or balanced). As such, it provides a theoretically sound basis for mitigating classifier bias under long-tailed label distributions. Importantly, this result also bridges the gap between heuristic logit-adjustment methods and principled Bayesian calibration, thereby justifying the use of such reweighting schemes in modern class-imbalanced learning frameworks. The induction proof further reinforces the robustness of the formulation by confirming its validity across any number of classes k. Complete derivation steps and mathematical justifications are provided in the supplementary material.

Moreover, Utilizing effective numbers to compute $\hat{p}(y)$ aligns with our data augmentation strategy. In this context, effective numbers consider two scenarios for newly sampled data points: they either fall entirely within the set of previously sampled data with a probability of ρ or entirely outside with a probability of $1-\rho$. This mirrors the principles of combining data through techniques like CutMix or Mixup. The generated data point either resides within the sampled points of the class, indicating that the mixed data points belong to the sampled set, or it exists outside the sampled points, implying that one data point is not from the existing set.

3.4.4.1 EBS Cross Entropy Loss

If $\mathscr{Y}$ represents the one-hot encoded vector corresponding to the true class y and $\varphi_y = f(g(x_i))[y]$, then we characterize the Effective Balanced Softmax (EBS) Cross Entropy Loss function as follows:

$$CE_{\text{EBS}}(\varphi_y, \mathscr{Y}) = -\sum_{j=1}^{k} \mathscr{Y}_j log(\hat{\Phi}_j) \tag{3.13}$$

$$= -log\left(\frac{\exp\{(\varphi_y)\}(1-\beta^{n_y})}{\sum_{i=1}^{k}\exp\{(\varphi_i)(1-\beta^{n_i})\}}\right) \tag{3.14}$$

$$= \underbrace{-(\varphi_y + \log(1-\beta^{n_y}))}_{\text{Data Term}} + \underbrace{\log\left(\sum_{i=1}^{k}\exp\{(\varphi_i + \log(1-\beta^{n_i}))\}\right)}_{Normalization Term} \tag{3.15}$$

The Equation 3.15 above combines two terms: the *Data Term* and the *Normalization Term*. The Data Term, $-(\varphi_y + \log(1-\beta^{n_y}))$, focuses on the correct class (y). Here, φ_y penalizes high logit values for the correct class, pushing it towards stronger activation. The term $\log(1-\beta^{n_y})$ in-

troduces a class-balancing factor. For a minority class with low n_y, β^{n_y} becomes smaller, increasing the overall loss and forcing the model to pay more attention to it. The Normalization Term, $\log\left(\sum_{i=1}^{k}\exp\{(\varphi_i + \log(1-\beta^{n_i}))\}\right)$, normalizes the loss across all classes, similar to standard Cross-Entropy. However, the β factor within the summation again gives more weight to the minority classes due to their lower n_i values.

We will utilize the EBS Cross Entropy Loss (CE_{EBS}) to compute the mixup losses L_1 and L_2 at head 1 and head 2, respectively.

3.4.5 Training Phase

The training procedure (Algorithm 2) operates in two distinct phases to optimize representation learning and decision boundary calibration. In the *HybridMix phase* (epochs 1 to $T-\tau$), we simultaneously sample from three distributions: instance sampler I (prioritizing head classes), median sampler M (medium-frequency classes), and reverse sampler R (tail classes). For each batch, three parallel streams are processed: 1) instance data $(x^I, y^I) \sim I$, 2) median-class samples $(x^M, y^M) \sim M$, and 3) reverse-sampled pairs $(x^{r_1}, y^{r_1}), (x^{r_2}, y^{r_2}) \sim R$. These are transformed into hybrid samples through two augmentation pathways: *Head-Medium CutMix* combines (x^I, y^I) and (x^M, y^M) via Eqs. 3.6–3.7 using a mask $\mathbf{M}$ with $\lambda \sim beta(\alpha, \alpha)$, while *Reverse Mixup* blends (x^{r_1}, y^{r_1}) and (x^{r_2}, y^{r_2}) through Eqs. 3.8–3.9 with $\lambda \sim beta(\alpha, \alpha)$. The backbone g and classifier heads f_1, f_2 are jointly optimized using the hybrid samples $(\tilde{x}_{h_1}, \tilde{y}_{h_1})$ and $(\tilde{x}_{h_2}, \tilde{y}_{h_2})$, with gradients computed via the EBS loss (Eq. 3.15):

$$(\theta, \phi_1) \leftarrow (\theta, \phi_1) - \zeta \nabla L_1(\tilde{x}_{h_1}, \tilde{y}_{h_1}; \theta, \phi_1),$$

$$(\theta, \phi_2) \leftarrow (\theta, \phi_2) - \zeta \nabla L_2(\tilde{x}_{h_2}, \tilde{y}_{h_2}; \theta, \phi_2).$$

In the *Boundary Refinement phase* (epochs $T-\tau$ to T), HybridMix is disabled to eliminate synthetic sample interference. The instance sampler I trains head f_1 on raw head-class data (x^I, y^I), while the reverse sampler R trains f_2 on unaltered tail-class instances (x^R, y^R). The transition parameter τ (typically $10\% - 20\%$ of T) balances hybrid diversity early in training with boundary precision late in training. As shown in Fig. 3.3a, this phased approach ensures the backbone g learns robust cross-class features while f_1 and f_2 specialize in head- and tail-class discrimination, respectively.

3.4.6 Inference Phase

During the testing phase, the test image is passed through the backbone g and then through both the heads f_1, and f_2, and output logits $z_1 \in \mathbb{R}^{1\times k}$ and $z_2 \in \mathbb{R}^{1\times k}$ (before the softmax layer) of f_1, and f_2 are collected. They are further adjusted as $\tilde{z_1}$ and $\tilde{z_2}$ to have the comparable scales weighted using the norm of the weights of the fully connected layer, given by $\tilde{z_1} = \frac{||w_1||_2}{||w_1||_2+||w_2||_2} \cdot z_1$, and $\tilde{z_2} = \frac{||w_2||_2}{||w_1||_2+||w_2||_2} \cdot z_2$, where $||w_1||_2$, $||w_2||_2$ are the norms of the weights of the fully connected

Algorithm 2 Proposed Training Algorithm

Input: Dataset $D_{i=1}^{N}$, model parameters θ for backbone g and parameters ϕ_1 and ϕ_2 for classifier head 1 and head 2 respectively, Instance sampler I, Median Sampler M, Reverse Sampler R, loss functions $L_1(.)$ for first head and $L_2(.)$ for second head, parameter of epoch τ, learning rate ζ.
Output: The trained network backbone, g, along with the trained classifier heads 1 and 2.

Computation:

Randomly initialize θ.
Dataset sampled using I: $D_{i=1}^{N} \sim I$
Dataset sampled using M: $\hat{D}_{i=1}^{N} \sim M$
Dataset sampled using R: $\tilde{D}_{i=1}^{N} \sim R$
for epoch $= 1, ..., T$ **do**
  if epoch $\geq T - \tau$ **then**
    for batch $i = 1, ..., B$ **do**
      Draw a mini-batch (x_i^I, y_i^I) from $D_{i=1}^{N}$
      Draw a mini-batch (x_i^M, y_i^M) from $\hat{D}_{i=1}^{N}$
      Draw a mini-batch $(x_i^{r_1}, y_i^{r_1})$ and $(x_i^{r_2}, y_i^{r_2})$ from $\tilde{D}_{i=1}^{N}$
      Generate Sample $(\tilde{x}_{h_1}, \tilde{y}_{h_1})$ using samples (x_i^I, y_i^I) and (x_i^M, y_i^M) by using Eqn's 3.6 and 3.7
      Generate Sample $(\tilde{x}_{h_2}, \tilde{y}_{h_2})$ using samples $(x_i^{r_1}, y_i^{r_1})$ and $(x_i^{r_2}, y_i^{r_2})$ by using Eqn's 3.8 and 3.9
      $(\theta, \phi_1) \leftarrow (\theta, \phi_1) - \zeta \nabla L_1((\tilde{x}_{h_1}, \tilde{y}_{h_1}); (\theta, \phi_1))$
      $(\theta, \phi_2) \leftarrow (\theta, \phi_2) - \zeta \nabla L_2((\tilde{x}_{h_2}, \tilde{y}_{h_2}); (\theta, \phi_2))$
    end for
  else
    for batch $i = 1, ..., B$ **do**
      Draw a mini-batch (x_i^I, y_i^I) from $D_{i=1}^{N}$
      Draw a mini-batch $(x_i^{r_1}, y_i^{r_1})$ from $\tilde{D}_{i=1}^{N}$
      $(\theta, \phi_1) \leftarrow (\theta, \phi_1) - \zeta \nabla L_1((x_i^I, y_i^I); (\theta, \phi_1))$
      $(\theta, \phi_2) \leftarrow (\theta, \phi_2) - \zeta \nabla L_2((x_i^{r_1}, y_i^{r_1}); (\theta, \phi_2))$
    end for
  end if
end for

layer of f_1 and f_2 respectively. The final logits are calculated as : $\tilde{z} = \frac{\tilde{z_1} + \tilde{z_2}}{2}$. Now, the softmax function is applied to $\tilde{z}$ to obtain the confidence for the classification. The complete process for the testing phase is shown in Fig. 3.3b.

3.5 Experiments and Results Discussion

3.5.1 Datasets and Experimental Setup

3.5.1.1 Dataset Overview

We evaluate the proposed method by employing four datasets designed for long-tailed visual recognition tasks: CIFAR-LT-10, CIFAR-LT-100, ImageNet-LT, and NIH-CXR-LT. *Long-tailed CIFAR-10* and *Long-tailed CIFAR-100* datasets are the long-tailed version of the CIFAR10 and CIFAR100 (Krizhevsky, 2009) datasets. To ensure a fair comparison, we adopt the same settings as (Cao et al., 2019) for the long-tailed versions of CIFAR datasets. Our experiments cover both CIFAR-LT-10 and CIFAR-LT-100, exploring imbalance ratios of 10, 50, 100, and 200.

The *Long-tailed ImageNet* dataset represents a skewed distribution of the original ImageNet (Deng et al., 2009) dataset. This long-tailed variant was generated by introducing an exponential distribution to the class distribution. We utilize the *Imagenet-LT* dataset proposed in (Liu et al., 2019b) by Liu et al., comprising 115.8K images across 1000 categories, with class cardinality ranging from 5 to 1,280.

We also employed the long-tailed chest X-ray benchmarks, specifically the *NIH-CXR-LT* from (Holste et al., 2022), which comprises 88,637 images. The NIH-CXR-LT dataset encompasses 20 classes, categorized into 7 many-shot classes (cardinality>1000), 10 medium-shot classes (cardinality 100-1000), and 3 few-shot classes (or tail classes) (cardinality<100). The cardinality ranges from 45,439 in the head class to 7 in the tail class.

3.5.1.2 Metrics Used for Evaluation

Performances are mainly reported as the overall top-1 accuracy. Consistent with Liu et al.'s approach (Liu et al., 2019b) in the context of CIFAR-100-LT and Imagenet-LT datasets, we present accuracy for three distinct subsets: many-shot classes (>100 training samples), medium-shot classes ($20 \leq$ training samples ≤ 100), and few-shot classes (<20 training samples).

3.5.1.3 Training Details

[2] To ensure alignment with the established conventions for long-tail datasets, we maintain consistency with the parameters defined in (Cao et al., 2019) for CIFAR-10-LT and CIFAR-100-LT. The chosen backbone is ResNet-32. Training involves a batch size of 128 over 200 epochs, utilizing the SGD optimizer with a momentum of 0.9. The initial learning rate is set at 0.1, with a linear warm-up learning rate scheduler applied for the first 5 epochs. Subsequently, the rate undergoes a decay of 0.01 at the 160^{th} epoch and again at the 180^{th} epoch. For ImageNet-LT, our backbone networks are ResNet-50 and ResNet-10, following the experimental settings outlined in (Kang et al., 2020). In the case of ResNet-10 and ResNet-50, a cosine learning rate schedule is adopted, gradually decreasing from 0.1 to 0. The image resolution is set at 224×224, with a batch size of 256. Throughout all experiments of Resnet-50/10, the SGD optimizer with a momentum of 0.9 is consistently applied. Moreover, for NIH-CXR-LT, we adopt the experimental settings outlined in (Holste et al., 2022). Specifically, we utilize a ResNet50 pre-trained on ImageNet, employing the Adam optimizer with a learning rate of $1e-4$. All models undergo training for a maximum of 60 epochs, with early stopping based on overall validation accuracy.

3.5.1.4 Methods used for comparison

The methods we considered for comparison with our method are broadly classified as follows: (1.) Reweighing-based Methods: LDAM/LDAM-DRW (Cao et al., 2019), CE-CBRW/DRW (Cui et al.,

[2] Code is available at: https://github.com/fa-submit/STTP-Net

2019b), IB-Focal/CB (Park et al., 2021), BALMS (Ren et al., 2020a). (2.) Augmentation-based Methods: Mixup (Zhang et al., 2017b), CutMix (Yun et al., 2019a), CMO+CE/DRW/BS (Park et al., 2022), MetasAug (Li et al., 2021). (3.) Dual-Branch Networks: BBN (Zhou et al., 2020), ResLT (Cui et al., 2022), RIDE (Wang et al., 2021b), CBD (Iscen et al., 2021) (Class-Balanced Distillation), CDBNF (Fan et al., 2022), Bi-F3R (Chen et al., 2023). (4.) Two-stage Methods: τ-Norm (Kang et al., 2020), cRT (Kang et al., 2020), LWS (Kang et al., 2020), DPM (Zhang et al., 2023a), MisLAS (hong et al., 2021). (5.) Miscellaneous Methods (including Knowledge Distillation-based methods, logit adjustment methods, self-supervised learning, and contrastive learning-based methods): CBS+RRS (Zhang et al., 2019), DisAlign (Zhang et al., 2021), TDE (Tang et al., 2020a), SEQL (Yang and Xu, 2020), OLTR (Liu et al., 2019b), SSP (Yang and Xu, 2020), Hybrid-SC (Wang et al., 2021a), TSC (Li et al., 2022e), LADE (Hong et al., 2021), Causal (Tang et al., 2020a), GCL (Li et al., 2022d). We did not consider ensemble models involving multiple experts.

3.5.2 Results & discussions: Comparison With Previous Methods

In order to ensure a fair comparison with our method for CIFAR10-LT and CIFAR-100-LT datasets, we replicated the results of competing methodologies utilizing their provided source code, while maintaining consistency in seed values, optimizer configurations, and learning rates across all methodologies. In this section, we will compare our method with the previous works that help in handling long-tail imbalance.

3.5.2.1 Comparison on CIFAR-10-LT Dataset & CIFAR-100-LT

In our comprehensive evaluation of the CIFAR-10-LT dataset in Table 4.1, we present a detailed analysis of top-1 accuracy across various imbalance ratios (IR) for different state-of-the-art methods. Notably, our proposed method consistently outperforms other approaches across all imbalance ratios. At IR=10, our method achieves an impressive 90.47% accuracy, showcasing its robustness even in scenarios with minimal class imbalance. As the imbalance ratio increases, our method continues to demonstrate superior performance, achieving 86.70%, 84.15%, and 81.07% accuracy at IR=50, 100, and 200, respectively. The closest competitors, CMO+BS (Park et al., 2022) and MisLAS (hong et al., 2021), lag behind by 1.29%, 1.26%, 1.45%, and 3.26% for CMO+BS (Park et al., 2022), and 0.14%, 1.14%, 1.72%, and 4.01% for MisLAS (hong et al., 2021), respectively. Highlights the substantial improvements our approach brings, especially in comparison to MisLAS (hong et al., 2021), showcasing its efficacy in handling class imbalance.

In our experimental evaluation of the CIFAR-100-LT dataset in Table 4.2, we compared the performance of our proposed method with several SOTA methods, focusing on top-1 accuracy across various imbalance ratios (IR). Our method consistently demonstrated superior results across all imbalance ratios, showcasing its effectiveness in handling class imbalance for the ResNet-32 architecture. Specifically, at IR = 10, our method achieved an accuracy of 63.85%, outperforming

Table 3.2: Comparison on CIFAR-10-LT dataset with different State-of-the-art methods in terms of top-1 accuracy (%) for ResNet-32 architecture with different imbalance ratios.

Methods	IR=10	IR=50	IR=100	IR=200
CE	86.81	77.00	71.21	66.05
LDAM (Cao et al., 2019)	86.79	79.05	74.64	69.55
CE-CBRW (Cui et al., 2019b)	86.88	78.53	73.16	65.65
CB-Focal (Cui et al., 2019b)	83.98	70.13	64.58	35.61
CE-DRW (Cao et al., 2019)	88.28	80.44	76.36	70.38
LDAM-DRW (Cao et al., 2019)	87.78	82.36	78.33	73.38
IB-Focal (Park et al., 2021)	87.27	77.42	71.71	65.96
IB+CB (Park et al., 2021)	88.32	81.74	78.31	74.79
BALMS (Ren et al., 2020a)	88.24	81.66	78.4	72.87
Mixup (Zhang et al., 2017b)	88.24	79.16	72.75	66.13
CutMix (Yun et al., 2019a)	88.22	79.59	75.61	69.75
CMO (Park et al., 2022)	88.72	80.51	73.84	71.07
CMO+DRW (Park et al., 2022)	89.37	84.60	81.76	77.41
CMO+BS (Park et al., 2022)	89.18	85.44	82.70	78.33
MetasAug (Li et al., 2021)	89.23	83.9	82.07	76.27
Causal (Tang et al., 2020a)	88.20	83.10	79.90	76.60
BBN (Zhou et al., 2020)	87.95	80.67	77.40	73.76
LADE (Hong et al., 2021)	87.60	82.80	79.70	75.20
MisLAS (hong et al., 2021)	90.33	85.56	82.43	76.06
Hybrid-SC (Wang et al., 2021a)	91.12	85.36	81.40	-
GCL(Li et al., 2022d)	89.83	84.99	81.98	78.24
TSC (Li et al., 2022e)	88.7	82.9	79.7	-
ResLT (Cui et al., 2022)	88.98	82.96	78.87	75.07
STTP-Net ($\beta = 0.999$)	90.47	86.70	84.15	81.07
STTP-Net ($\beta = 0.9999$)	91	86.4	83.65	80.88

the closest competitor by a notable margin. As the imbalance ratio increased, the performance gap widened, with our method consistently yielding higher accuracy than alternative approaches. At IR = 50, 100, and 200, our method exhibited top-1 accuracies of 53.65%, 49.19%, and 44.67%, respectively, showcasing its robustness in mitigating the impact of class imbalance. Notably, the nearest best-performing method fell significantly behind our proposed approach. Against CMO+DRW (Park et al., 2022), our method displayed a remarkable superiority of 1.94%, 2.36%, and 2.78% at IR=50, 100, and 200, respectively. Furthermore, when compared to ResLT§ (Cui et al., 2022), which is a ResLT (Cui et al., 2022) network trained using strong data augmentation (mixup and auto augment), our method shows performance improvement by ∼1% in all cases. However, under normal training conditions of ResLT (Cui et al., 2022) without strong augmentation, our method exhibited substantial improvement, showcasing 3.71%, 4.59%, 4.36%, and 4.05% performance enhancement

Table 3.3: Comparison of CIFAR-100-LT dataset with different State-of-the-art methods in terms of overall top-1 accuracy (%) for IR= 10, 50, 100, and 200 along with its top-1 accuracy (%) for Many-shot, Medium-shot, and Few-shot classes for ResNet-32 architecture with IR=100 and 200. (Entries with * denote results directly taken from the corresponding research work.)

Methods	Venue	IR = 10	IR = 50	IR = 100				IR = 200			
		Top1-acc	Top1-acc	Top1-acc	Many	Medium	Few	Top1-acc	Many	Medium	Few
CE		56.81	43.99	37.58	64.8	35.97	7.7	34.8	66.33	37.39	8.49
LDAM (Cao et al., 2019)	[NeurIPS'19]	56.01	42.97	39.7	67.54	39	8.03	36.02	67.1	41.61	7.67
CE-CBRW (Cui et al., 2019b)	[CVPR'19]	56.58	43.53	38.31	65.37	35.54	9.97	34.56	66.9	36.45	8.18
CB-Focal (Cui et al., 2019b)	[CVPR'19]	52.49	40.22	34.88	59.94	32.03	8.97	31.92	62.17	33.52	7.38
CE-DRW (Cao et al., 2019)	[NeurIPS'19]	58.01	46.6	40.98	62.71	40.74	15.9	37.81	63.5	43.23	13.74
LDAM-DRW(Cao et al., 2019)	[NeurIPS'19]	57.45	46.61	43.44	62.11	44.57	20.33	37.86	62.27	41.16	16.46
IB-Focal (Park et al., 2021)	[ICCV'21]	57.82	43.74	42.67	56.71	44.6	24.03	37.72	56	41.29	20.82
IB+CB (Park et al., 2021)	[ICCV'21]	54.78	42.21	37.27	68.09	33.43	5.8	33.02	68.37	33.13	5.74
BALMS(Ren et al., 2020a)	[NeurIPS'20]	59.54	47.19	42.54	58.66	42.31	24	37.93	60.57	41	18.08
Mixup (Zhang et al., 2017b)	[ICLR'18]	58.06	44.72	40.08	70.91	38.91	5.47	36.15	67.22	32.26	2.28
CutMix (Yun et al., 2019a)	[ICCV'19]	59.21	46.72	40.74	71.40	40.57	5.17	36.33	70.5	40.84	6.46
CMO (Park et al., 2022)	[CVPR'22]	60.22	47.46	41.95	69.51	40.6	11.37	38.7	70.2	43.74	10.46
CMO+DRW (Park et al., 2022)	[CVPR'22]	62.08	51.71	46.83	63.26	47.77	26.57	41.89	61.93	48.16	21.49
CMO+BS (Park et al., 2022)	[CVPR'22]	61.52	51.2	45.98	57.37	45.4	33.37	40.52	56.07	42.16	27.26
MetasAug (Li et al., 2021)	[CVPR'21]	61.83	51.09	47.24	63.86	47.74	27.52	42.91	62.86	43.6	19.41
Causal (Tang et al., 2020a)	[NeurIPS'20]	59.4	48	45	64.8	47.6	18.8	39.7	62.3	47.9	15.9
BBN (Zhou et al., 2020)	[CVPR'20]	58.33	46.62	41.96	53.37	51.71	17.27	37.14	53.63	50.1	14.15
CBD (Iscen et al., 2021)	[BMVC'21]			44.4	64.5	45.2	20.6	40.4	65.4	45.0	18.9
LADE (Hong et al., 2021)	[CVPR'21]	60.3	50.4	45.3	61.3	43.3	29.1	39.6	60.9	43.2	20.3
MisLAS (hong et al., 2021)	[CVPR'21]	62.05	51.4	47.25	62.83	48.14	26.83	42.92	61.64	44.11	18.24
Hybrid-SC (Wang et al., 2021a)	[CVPR'21]	63.05	51.87	46.72	-	-	-	-	-	-	-
TSC (Li et al., 2022e)	[CVPR'22]	59.0*	47.4*	43.8*	-	-	-	-	-	-	-
GCL(Li et al., 2022d)	[CVPR'22]	61.66	53.55	48.71	-	-	-	-	-	-	-
CDBNF (Fan et al., 2022)	[EAAI'22]			45.1	61.1	45.1	25.4	40.9	62.5	46.4	20.9
Bi-F3R (Chen et al., 2023)	[EAAI'23]	-	53.15	50.54	-	-	-	46.89	-	-	-
DPM (Zhang et al., 2023a)	[TNNLS'23]	59.47	49.32	44.8*	60.7*	46.4*	24.4*	-	-	-	-
RIDE (Wang et al., 2021b)	[ICLR'21]	-	-	47.62	66.14	48.31	25.2	-	-	-	-
ResLT (Cui et al., 2022)	[TPAMI'23]	60.14	49.06	44.56	60.25	47.14	21.97	40.62	58.25	44.86	13.62
ResLT§ (Cui et al., 2022)	[TPAMI'23]	62.01*	52.71*	48.21*	-	-	-	-	-	-	-
STTP-Net ($\beta = 0.99$)		63.85	53.65	48.46	61.8	49.94	31.17	44.67	62.27	48.26	28.28
STTP-Net ($\beta = 0.999$)		63.58	53.02	49.19	63.94	50.83	30.07	43.02	51.37	49.06	31.79

at IR=10, 50, 100, and 200, respectively. Additionally, our method showcased improvement over the RIDE (Wang et al., 2021b) method, a multi-expert two-step process (we have trained RIDE (Wang et al., 2021b) using 3 experts) with an improvement of around 1.57% at IR=100.

For CIFAR-100-LT in Table 4.2, analysis is also done focusing on accuracy for Many-shot, Medium-shot, and Few-shot classes under IR of 100 and 200. Let us consider the case of IR=100 for Few-shot classes in our method. The few-shot accuracy is 31.17 (for $\beta = 0.99$) or 30.07 (for $\beta = 0.999$), which is better than all the other methods with huge margin except for the case of CMO+BS (Park et al., 2022) whose few-shot accuracy is 33.37 which comes at the expense of deteriorating performance of both Many-shot and Medium-shot classes whose performances are less than ours by almost $\sim 5\%$. For the case of medium-shot classes for IR=100, our method shows accuracy of 49.94 (for $\beta = 0.99$) or 50.83 (for $\beta = 0.999$), which less than Medium shot accuracy of BBN (Zhou et al., 2020) by 1% but as seen for BBN (Zhou et al., 2020) the accuracy of both Many Shot classes and Few Shot classes is less than ours by almost 10%, and 13% respectively. Furthermore, if we see the case of a higher imbalance ratio of 200, our few-shot accuracy is better in all the cases. In the case of medium shot classes, our accuracy is better than all the other methods except for the BBN (Zhou et al., 2020), which is better than ours by 1%. However, its accuracy deteriorate heavily for many-shot and few-shots classes by almost 10% and 14%, compared to our method.

3.5.2.2 Comparison on ImageNet-LT

Table 4.3 reports the results on ImageNet-LT. Mostly used those methods for comparison which uses ResNet-10/50 as network architecture. Decoupled methods like τ-norm (Kang et al., 2020), TDE (Tang et al., 2020a), and DisAlign (Zhang et al., 2021) tend to under-fit the head and over-fit the tail classes. Our method has performed better than all the decoupled methods with an improvement of approximately 1% for ResNet-10 and 1.4% for ResNet-50 in top-1 accuracy. The superiority of our method can be seen from Table4.4, for ResNet-10 backbone, our method surpasses cRT (Kang et al., 2020), τ-norm (Kang et al., 2020), and LWS (Kang et al., 2020) by almost 2% for Medium-shot accuracy and cRT (Kang et al., 2020) by 4.5%, τ-norm (Kang et al., 2020) by 3.2%, and LWS (Kang et al., 2020) by 2.3% in terms of few-shot accuracy. Similarly, for ResNet-50 backbone, our method surpasses cRT (Kang et al., 2020), τ-norm (Kang et al., 2020), and LWS (Kang et al., 2020) by 6.9%, 6.7%, and 5.7% for Medium-shot accuracy and by 11%, 9.7%, and 7.8% respectively in terms of few-shot accuracy. MetasAug (Li et al., 2021) and OLTR (Liu et al., 2019b) are two methods that generate data using complex modules. Our method outperforms MetasAug (Li et al., 2021) and OLTR (Liu et al., 2019b) with 4.9% and 5.03% improvement in top-1 accuracy for ResNet-50 and ResNet-10 backbone, respectively, while using less overhead. It also surpasses the augmentation method CMO (Park et al., 2022) trained with CE loss by 3.2% for the ResNet-50 backbone. Self-supervised learning with SSP yields good feature initialization. Our approach

Table 3.4: Comparison on Imagenet dataset with different State-of-the-art methods in terms of top-1 accuracy (%) for ResNet-10 and ResNet-50 architecture. (Results for the peer algorithms are taken from the respective papers and "†" denotes the results from (Cui et al., 2022).)

Epoch	Method	Resnet-10	Resnet-50
90	Focal(Lin et al., 2018)	30.5	-
	BALMS(Ren et al., 2020a)	41.8	-
	τ-norm(Kang et al., 2020)	40.6	46.7
	cRT(Kang et al., 2020)	41.8	47.3
	CBS+RRS(Zhang et al., 2019)	41.9	47.3
	LWS(Kang et al., 2020)	41.4	47.7
	TDE(Tang et al., 2020a)	-	51.1
	DisAlign(Zhang et al., 2021)	-	51.3
	CBD(Iscen et al., 2021)	37.9	51.6
	MetaSAug(Li et al., 2021)	-	47.4
	SEQL(Yang and Xu, 2020)	36.4	-
	OLTR(Liu et al., 2019b)	37.3	-
	CDBNF(Fan et al., 2022)	38.5	-
	CMO(Park et al., 2022)	-	49.1
	STTP-Net(β =0.999)	42.33	52.3
>180	τ-norm(Kang et al., 2020)	42.7†	46.7
	cRT(Kang et al., 2020)	43.2†	47.3
	LWS(Kang et al., 2020)	43†	47.7
	SSP(Yang and Xu, 2020)	43.2	51.3
	ResLT(Cui et al., 2022)	43.8	48.5
	DPM(Zhang et al., 2023a)	42.0	51.0
	STTP-Net(β =0.999)	44.0	53

outperforms SSP. With ResNet10, it achieves approximately 1% higher top-1 accuracy. With ResNet50, it achieves 1.7% higher top-1 accuracy. Along with this, our model surpasses the Dual-phase model (DPM) (Zhang et al., 2023a) by 2% for both ResNet-10 and ResNet-50 backbone. Our method and ResLT (Cui et al., 2022) both use a multi-head framework for long-tailed recognition. Our method outperforms ResLT (Cui et al., 2022), with ResNet50 as its backbone by almost 4.5% in terms of top-1 accuracy, with increment in both Many and Medium class accuracies. However, there is a slight decrement in a few classes' accuracy by 5.3%. However, the increment of a few classes' accuracy of ResLT (Cui et al., 2022) comes at the cost of decrement of Many class accuracy by almost 9.2% compared with our method.

3.5.2.3 Comparison on NIH-CXR-LT

We employed the benchmark methods outlined in (Holste et al., 2022) to assess the performance of our approaches on the NIH-CXR-LT dataset. The results, presented in Table 3.6, demonstrate the superior performance of our method. Specifically, our method achieved a top-1 accuracy of 40.17%,

Table 3.5: Top-1 Accuracy(%) of Many-shot, Medium-shot and Few-shot on ImageNet-LT with ResNet-10 & ResNet-50 backbone("*","†", and "‡" denotes the results are from the original papers, (Cui et al., 2022) and (Kang et al., 2020), respectively.)

Backbone Model	Methods	Top1-acc	Many	Medium	few
ResNet-10	CE†	37.3	59.7	29.4	5.7
	CBD	37.9	49.2	36.9	21.5
	CDBNF	38.5	46.0	36.1	27.0
	cRT†	43.2	53.8	41.3	25.4
	τ-norm†	42.7	50.4	42.1	26.7
	LWS†	43	51.8	41.6	27.6
	ResLT†	43.8	52.3	41.6	29.5
	STTP-Net(β =0.999)	44	50.1	43.3	29.9
ResNet-50	CBD	51.6	65.2	48.0	25.9
	NCM‡	44.3	53.1	42.3	26.5
	cRT‡	47.3	58.8	44	26.1
	τ-norm‡	46.7	56.6	44.2	27.4
	LWS‡	47.7	57.1	45.2	29.3
	CE+CMO‡	49.1	67	42.3	20.5
	DPM*	51	64.6	48.3	22.1
	ResLT	48.5	52.4	47.4	42.4
	STTP-Net(β =0.999)	53	61.6	50.9	37.1

surpassing all other methods. Notably, the Many-shot top-1 accuracy of 50% outperformed all alternative methods. However, our method exhibited a slightly lower performance in Medium-shot top-1 accuracy, recording a value of 35.56%. This result is inferior to RW-LDAM (Cao et al., 2019), RW-LDAM-DRW (Cao et al., 2019), and cRT (Kang et al., 2020). It is important to highlight that the enhanced performance of these methods in Medium-shot classes comes at the expense of significantly deteriorating performance in Many-shot classes compared to our method. Specifically, our method outperformed RW-LDAM (Cao et al., 2019) by 19.5%, RW-LDAM-DRW (Cao et al., 2019) by 9%, and cRT (Kang et al., 2020) by 6.7% in terms of Many-shot classes. Moreover, our method excelled in Few-shot classes top-1 accuracy, achieving a notable accuracy of 33.33%, surpassing all other methods in this category.

3.5.3 Summarizing Comparative Insights with State-of-the-Art Methods

Our method consistently demonstrates superior performance across multiple long-tailed datasets, including CIFAR-10-LT, CIFAR-100-LT, ImageNet-LT, and NIH-CXR-LT, surpassing various state-of-the-art approaches designed for handling class imbalances. Reweighting-based methods such as LDAM/LDAM-DRW (Cao et al., 2019), CE-CBRW/DRW (Cui et al., 2019b), and IB-Focal/CB (Park et al., 2021) aim to mitigate imbalance through class-aware loss adjustments but struggle with overfitting to minority classes, leading to suboptimal generalization. Augmentation-based

Table 3.6: Results on NIH-CXR-LT. Accuracy is reported for the balanced test set in terms of overall top-1 accuracy (%) and top-1 accuracy (%) for Many-shot, Medium-shot, and Few-shot for ResNet-50 Architecture.

Methods	Top1-acc	Many	Medium	Few
Softmax	17.5	41.9	5.6	1.7
CB-Softmax	33.3	29.5	41.5	21.7
RW-Softmax	30	24.8	35.9	25.8
Focal-Loss	16	36.2	5.6	4.2
CB-Focal-Loss	30.3	37.1	33.3	11.7
RW-Focal-Loss	25.5	28.6	29.3	11.7
LDAM	23.2	41	13.3	14.2
CB-LDAM	29.5	35.7	28.5	20.8
CB-LDAM-DRW	37.7	47.6	35.6	25
RW-LDAM	35.3	30.5	41.9	29.2
RW-LDAM-DRW	37	41	36.7	30.8
MixUp	17	41.9	4.4	1.7
Balanced-MixUp	21.3	44.3	8.1	10.8
cRT	38	43.3	37.4	30
τ-norm	28	45.7	23	8.3
STTP-Net	40.17	50	35.56	33.33

techniques like Mixup (Zhang et al., 2017b), CutMix (Yun et al., 2019a), and CMO+CE/ DRW/ BS (Park et al., 2022) enhance representation learning but often introduce label noise that affects decision boundaries. Dual-branch networks, including BBN (Zhou et al., 2020), ResLT (Cui et al., 2022), and RIDE (Wang et al., 2021b), leverage specialized feature extraction for different class groups but suffer from increased model complexity and computational overhead. Similarly, two-stage methods like τ-Norm (Kang et al., 2020), cRT (Kang et al., 2020), LWS (Kang et al., 2020), DPM (Zhang et al., 2023a), and MisLAS (hong et al., 2021) address imbalance through feature decoupling yet tend to underperform in highly imbalanced settings due to reliance on separate training phases. Knowledge-distillation-based and contrastive learning approaches, such as CBS+RRS (Zhang et al., 2019), DisAlign (Zhang et al., 2021), TDE (Tang et al., 2020a), SSP (Yang and Xu, 2020), and GCL (Li et al., 2022d), improve feature space discrimination but often require extensive pretraining or auxiliary networks, making them less practical. Unlike these methods, our approach achieves a favorable balance between model efficiency and accuracy, as evidenced by its consistent outperformance across datasets. Specifically, our method achieves state-of-the-art accuracy on CIFAR-10-LT and CIFAR-100-LT, particularly excelling in high imbalance ratios where prior techniques exhibit significant performance drops. On ImageNet-LT, our model surpasses decoupled and feature-reweighting strategies by a notable margin, demonstrating robustness in large-scale, real-world scenarios. In medical imaging, our approach outperforms baseline methods on NIH-CXR-LT by effectively maintaining strong performance across Many-shot, Medium-shot,

and Few-shot classes, unlike reweighting-based methods that compromise head-class accuracy. The widespread effectiveness of our method across diverse domains validates its adaptability and efficiency in mitigating long-tailed imbalances, making it a more generalized and computationally feasible alternative to existing solutions. The results validate two core design principles: (1) unified feature-classifier alignment mitigates the overfitting risks inherent in decoupled frameworks, and (2) dynamic regularization ensures robustness across varying imbalance ratios. For instance, on ImageNet-LT (Table 4.4), our method surpasses τ-Norm (Kang et al., 2020) and LWS (Kang et al., 2020) by 6.7–9.7% in Few-shot accuracy while maintaining Many-shot performance, demonstrating that holistic optimization outperforms fragmented two-stage training. Practically, the method's computational efficiency—achieving 53% top-1 accuracy on ImageNet-LT with ResNet-50 versus ResLT's (Cui et al., 2022) 48.5%—positions it as a scalable solution for real-world applications, from autonomous systems to healthcare.

3.6 Further Analysis, Limitations and Future Directions

In this section, we will discuss the three main ideas that have contributed to overall performance: 1) Hybrid Mixup augmentation, 2) Effective balanced softmax loss, and 3) the Effect of Hyperparameter β used.

3.6.1 Why Hybrid Mixup?

Based on the results illustrated in Figure 3.2, we evaluate three distinct network configurations: "Net-1," utilizing Instance Sampling with CutMix on the first head and Reverse Sampling with Mixup on the second head; "Net-2," a three-head network incorporating Instance Sampling with CutMix on the first head, Median Sampling with Cutmix on the second head, and Reverse Sampling with Mixup on the third head; and our proposed approach. As indicated in Table 3.7, our method surpassed both Net-1 and Net-2, achieving a top-1 accuracy of 49.19%. While Net-1 exhibits superior performance for Medium and Few classes by approximately 2.5%, this improvement comes at the expense of a significant 7.5% decline in accuracy for Many classes, compared to our proposed approach. Similarly, Net-2 demonstrates enhanced accuracy for Few classes by almost 3% and comparable Medium classes accuracy, but at the cost of a 6.14% decrease in Many classes accuracy compared to our method. Thus, the integration of CMeO (Context-rich Median Oversampling in the first branch) and Mixup with reverse sampling in other branches strives to maintain a balance between class accuracies without substantial deterioration in the performance of different classes.

3.6.2 Why is Effective Balanced Softmax Loss Required?

To check the advantage of Effective Balanced Softmax loss, we will compare it with the Cross entropy loss (CE), Balanced Softmax loss (BS), and Effective balanced softmax loss (EBS) and

Table 3.7: Results on different network configurations on CIFAR-100-LT in terms of top-1 accuracy (%) for many-shot, medium-shot and few-shot for ResNet-32 architecture with IR=100

Networks	top1-acc	Many	Medium	Few
Net-1	48.34	56.43	53.43	32.97
Net-2	47.77	57.8	50.03	33.43
STTP-Net	49.19	63.94	50.83	30.07

Table 3.8: Comparison on different loss types for CIFAR-100-LT dataset with IR=100 for our (STTP-Net) proposed method

Loss-Type	top1-acc	Many	Medium	Few
CE	47.36	67.46	49.86	20.33
BS	47.72	53.63	50.8	35.8
EBS	49.19	63.94	50.83	30.07

see how our loss is more beneficial compared to other losses. Table 3.8, shows the comparison of different losses for CIFAR-100-LT with IR=100. As seen from the table, CE performs better on Many and Medium classes at the expense of the accuracy of Few-shot classes, less than our method by almost 10%. In contrast, BS gives better accuracy in the Few shot classes but comes at the huge expense of Many-shot classes, which is less than our method by almost 10% in many-shot classes accuracy. Thus, EBS acts as a method that balances the different groups' accuracies. Moreover, the softmax function used in CE uses $\tilde{p}_i = \frac{1}{k}$, BS uses $\tilde{p}_i = \frac{n_i}{\sum_{j=1}^{k} n_j}$, and EBS uses $\tilde{p}_i = \frac{1-\beta^{n_i}}{\sum_{j=1}^{k} 1-\beta^{n_j}}$, which can be interpreted as CE is not penalizing the logits obtained, whereas the BS is penalizing the logits corresponding to tail classes very heavily, while EBS tries to penalize the logits not too heavily but in a balancing manner.

3.6.3 Effect of Hyperparameter β

Table 3.9 presents accuracy values for different values of beta (β) across two datasets, CIFAR-100-LT and Imagenet-LT, and two corresponding backbones, ResNet-32 and ResNet-50. In the case of CIFAR-100-LT with ResNet-32, as beta decreases from 0.9999 to 0.99, there is a general trend of improvement in top-1 accuracy, reaching 63.85%. This improvement is particularly notable in the "Few" category, where accuracy increases from 42.63% to 44.67%. Similarly, for Imagenet-LT with ResNet-50, a decrease in beta leads to an increase in top-1 accuracy, reaching 53% at beta 0.999. Interestingly, this improvement is in all three categories, i.e., rising from 59% to 61.6% in "Many," rising from 49.3% to 50.9% in "Medium," and rising from 36.4% to 37.1% in "Few." However, if we further decrease the beta value, it leads to deterioration in the "Few" category by 3%, but no significant change in "Many" and "Medium." The results suggest that adjusting beta values impacts model performance, with lower beta values generally associated with improved accuracy, and the effect varies across different backbone-dataset combinations.

Table 3.9: Accuracy values based on different values of β for CIFAR-100-LT with IR=100 and Imagenet-LT datasets on different backbones corresponding to our method

Dataset \ Backbone	β	top1-acc	Many	Medium	Few
CIFAR-100-LT \ ResNet-32	0.9999	62.98	53.2	47.06	42.63
	0.999	63.58	53.02	49.19	43.02
	0.99	63.85	53.65	48.46	44.67
Imagenet-LT \ ResNet-50	0.9999	51.1	59	49.3	36.4
	0.999	53	61.6	50.9	37.1
	0.99	52.3	61.8	50.1	34.3

3.6.4 Limitations and Future Directions

While STTP-Net demonstrates state-of-the-art performance across multiple benchmarks, several limitations merit discussion. First, the dual-head architecture and hybrid augmentation strategy introduce moderate computational overhead compared to single-branch models, potentially limiting scalability in resource-constrained environments. Future work could explore lightweight architectural variants or knowledge distillation to mitigate this cost. Second, the method's performance is sensitive to hyperparameters such as β (in the EBS loss) and τ (epochs for hybrid mixup), requiring careful calibration for optimal results. To address this, future work will integrate a meta-learned hyperparameter controller into the training loop, leveraging reinforcement learning to adaptively optimize β and τ based on validation performance. Third, the hybrid mixup technique is tailored for visual data, and its applicability to non-image modalities (e.g., text or tabular data) remains an open question. We plan to extend the algorithm by replacing hybrid mixup with modality-agnostic augmentation, such as semantic mixup in text via synonym replacement or feature-space interpolation for tabular data. Finally, STTP-Net assumes static class distributions, limiting its adaptability to dynamic environments where new classes emerge incrementally, as in long-tailed class-incremental learning (CIL). To address this, we will augment the framework with replay buffers or elastic weight consolidation to support lifelong learning scenarios, enhancing its real-world applicability. By addressing these limitations through algorithmic extensions and cross-domain validation, future research could further advance the robustness and generality of long-tailed learning systems.

3.7 Discussion

We introduce STTP-Net a novel framework to effectively tackle the complexities of learning in imbalanced, long-tailed class scenarios. Our initial investigation delves deeply into the effectiveness of mixed sample data augmentation strategies, considering a spectrum of sampling approaches. Thorough analysis culminates in the conceptualization of a *hybrid mixup* technique, strategically

designed to enhance representation learning across the spectrum of classes, encompassing head, medium, and tail. This involves constructing a network with separate branches, maintaining an identical feature extraction network, yet modifying the classifier heads. Furthermore, we address the vital aspect of countering label distribution shifts to enhance classifier learning. For this purpose, we propose the *Effective Balanced Softmax* loss function, a specialized tool to rectify such shifts in the labels. Our proposed methodology is rigorously validated through extensive ablation studies and experimental evaluations across diverse benchmarks, convincingly demonstrating its effectiveness. Looking ahead, future work will focus on reducing computational overhead via lightweight architectures, automating hyperparameter tuning with meta-learned controllers, extending hybrid mixup to non-visual domains through modality-agnostic augmentations, and enabling lifelong learning via replay buffers for dynamic class distributions. Furthermore, STTP-Net's capability to handle long-tailed data makes it particularly valuable for applied engineering systems. In healthcare, it can be integrated into diagnostic pipelines to enhance the detection of rare diseases, while in industrial IoT, it facilitates robust defect classification in imbalanced manufacturing datasets. Future efforts will also include deploying STTP-Net on edge devices for real-time quality inspection, leveraging its efficient architecture. Additionally, we acknowledge the imperative of conducting further theoretical analyses on mixed sample data augmentation techniques, specifically tailored for handling imbalance, a direction we reserve for future investigations.

Chapter 4

The Goldilocks Principle: Achieving Just Right Boundary Fidelity for Long-Tailed Classification

Synopsis

In this chapter, we present a extension to the STTP-Net framework proposed in Chapter 3 by introducing a novel approach titled The Goldilocks Principle: Achieving Just Right Boundary Fidelity for Long-Tailed Classification *(Ansari et al., 2025)* [1]*. The study addresses the persistent challenges posed by long-tailed class imbalances in deep neural networks, particularly in image recognition tasks, where a few head classes dominate the dataset while numerous tail classes remain underrepresented. Traditional classification models often overfit to frequent classes, neglecting rare ones, and struggle to learn consistent decision boundaries that are both sharp enough to differentiate classes and smooth enough to generalize well. Hard decision boundaries typically arising from tail class overfitting amplify intra-class variation and degrade generalization, while overly soft boundaries blur inter-class distinctions, harming classification accuracy. While Chapter 3 tackled these issues using a hybrid augmentation strategy called HybridMix combining CutMix with static sampling and a dual-head classifier trained with Effective Balanced Softmax (EBS) loss this approach lacks the flexibility to dynamically control boundary regularization across class strata. In contrast, this chapter enhances the prior framework with two key innovations:* Dual-Branch Sampler-Guided Mixup (DBSGM) *and* Adaptive Class-Aware Feature Regularization (ACFR)*. DBSGM utilizes Mixup-based augmentation across two parallel branches, one focusing on head and medium classes using instance and median samplers, and the other emphasizing tail class learning using a reverse sampler. This structured, sampler-guided Mixup not only improves representation diversity but also encourages the model to learn smoother, semantically meaningful decision boundaries. Complementing DBSGM, ACFR intro-*

[1] F. Ansari, A. Panigrahi, and S. Das. "The Goldilocks Principle: Achieving Just Right Boundary Fidelity for Long-Tailed Classification." *IEEE Transactions on Emerging Topics in Computational Intelligence*, pp. 1–15, 2025. doi: https://doi.org/10.1109/TETCI.2025.3551950.

duces a feature-level regularizer that dynamically aligns feature representations using a temperature-adaptive supervised contrastive loss (TASCL). This regularization promotes intra-class cohesion and inter-class separability, reducing the model's sensitivity to irrelevant intra-class variations. The integration of DBSGM and ACFR within a single-stage, end-to-end dual-branch architecture achieves the optimal balance between decision boundary sharpness and smoothness termed "just right boundary fidelity." Together, these components significantly improve class-balanced performance, particularly for tail classes, offering a principled and empirically robust solution to long-tailed classification beyond the capabilities of the hybrid strategy proposed in Chapter 3.

4.1 Introduction

The term "Long Tail" comes from Chris Anderson's Long Tail theory, which he first introduced in a Wired magazine article in 2004 and later expanded upon in his book "The Long Tail: Why the Future of Business is Selling Less of More" in 2006, illustrates a shift away from the traditional retail model that focuses on popular items to a decentralized marketplace empowered by the internet. The idea behind the "long tail" is that businesses can now cater to specialized audiences rather than just mainstream hits, which has significant implications for industries ranging from entertainment to retail. In AI, this theory aligns with challenges posed by tasks involving long-tailed data distributions, such as pedestrian re-identification (An and Wang, 2023) and real-time recognition of Activities of Daily Living (ADL) (Chaudhary et al., 2022). Building effective models for tasks with long-tailed distribution is termed Long-tailed learning (LTL), and the subdomain that deals with recognition/classification in image domain is termed Long-tailed image recognition (LTR).

Various methods are used to address the issue of long-tailed class imbalance in deep learning, particularly in the context of image recognition. These include re-sampling (Van Hulse et al., 2007a; Chawla et al., 2002a; He and Garcia, 2009; Japkowicz and Stephen, 2002a), balancing losses or gradients (Huang et al., 2016a; Wang et al., 2017; Ren et al., 2020b; Cui et al., 2019b), two-stage methods (hong et al., 2021; Kang et al., 2020; Li et al., 2022c), adjustments to logits, margin learning approaches (Li et al., 2022c), post hoc techniques, augmentation techniques (Zhang et al., 2017b; Yun et al., 2019a; Park et al., 2022; Li et al., 2021), and multi-branch networks (Zhou et al., 2020; Wang et al., 2022c; Cui et al., 2022) . However, these methods have drawbacks, such as overfitting, degradation of generalization ability, and deterioration of performance in head classes. Furthermore, multi-expert systems like RIDE achieve state-of-the-art (SOTA) results by incorporating a combination of advanced techniques, such as data augmentation, knowledge distillation, and self-supervised pretraining, which may complicate their implementation and make them less straightforward to adopt in simpler models.

Long-tailed class imbalance presents a significant challenge in achieving consistent and accurate decision boundaries across head, medium, and tail classes. Existing methods, such as re-

sampling, balancing losses, and multi-branch networks, have attempted to address this issue but come with notable trade-offs, particularly in shaping class decision boundaries. **Re-sampling techniques** often rely on aggressively over-sampling tail classes to balance class distributions. While effective in increasing the representation of minority classes, this approach tends to create excessively *hard boundaries*, which are overly sharp and restrictive (Kim and Jung, 2023). Hard boundaries overfit to the limited data available for tail classes, amplifying intra-class variability and reducing generalization to unseen instances. This results in poor performance for tail classes and worsens the imbalance in learning. Conversely, **balancing loss functions** aim to downscale the contribution of dominant head classes by applying lower weights to their gradients during training. While this prevents the over-dominance of head classes, it frequently leads to excessively *soft boundaries*, where decision boundaries become too smooth and indistinct (Fernando and Tsokos, 2022; Lee et al., 2021). Soft boundaries blur the separation between classes, diminishing precision for head classes and leading to underfitting, particularly in scenarios where class distinctions are critical.

Multi-branch networks attempt to strike a balance by learning separate decision boundaries for different class groups. However, these methods often require complex architectures and extensive fine-tuning to achieve meaningful boundary adjustments. Moreover, they struggle to maintain consistent *boundary fidelity* across the spectrum of head, medium, and tail classes, leading to imbalances in how decision boundaries are formed and applied. The crux of the problem lies in achieving "just-right" boundary fidelity decision boundaries sharp enough to separate classes while being smooth enough to generalize well to unseen data (Kim and Kim, 2020). In the context of long-tailed imbalances, this balance is challenging due to (1) over-dominance of head classes, which occupy a large portion of feature space, skewing boundaries in their favor; (2) under-representation of tail classes, leading to noisy or overly restrictive boundaries that hinder generalization; and (3) medium class inconsistency, where medium classes receive insufficient attention, resulting in suboptimal performance.

Given the outlined challenges, building on the architecture of the STTP-Net of chapter 3 we propose a dual-branch network featuring a shared feature extractor network and distinct classifier heads. One branch amalgamates data from both the instance sampler, where the sampling probability depends on the provided distribution, and the median sampler, prioritizing medium classes with a higher sampling probability. This fusion is facilitated through mixup, highlighting the importance of both head and medium classes. Concurrently, the other branch integrates data from the reverse sampler, elevating the probability of data from tail classes, thereby prioritizing the tail classes. Our findings reveal that using softer probabilities during sampling significantly enhances results. The introduction of the median sampler proves instrumental in augmenting the accuracy of medium classes without substantially affecting head class accuracy and contributing to tail class accuracies. We term this augmentation method a Dual-Branch Sampler-Guided Mixup (DBSGM).

DBSGM encourages the model to acquire a broader understanding and develop smoother decision boundaries.

Additionally, to enhance model robustness and reduce sensitivity to irrelevant intra-class variations, we introduce Adaptive Class-Aware Feature Regularization (ACFR). This method directly applies a specialized loss function called Temperature Adaptive Supervised Contrastive Loss (TASCL) to extracted features, encouraging similar representations for samples from the same class and distinct representations for samples from different classes. By combining DBSGM and ACFR, our model strikes a balance between learning smooth decision boundaries (DBSGM) and preserving class-specific information (ACFR). This delicate equilibrium is analogous to the *Goldilocks principle: the model's decision boundaries are neither too sharp nor too smooth but just right, aligning with the underlying class structure.* We refer to this optimal balance as achieving "just right boundary fidelity." This single-stage, end-to-end framework yields promising results in addressing the challenges of long-tailed recognition.

In summary, our contributions can be distilled into three key aspects: (1) A novel **Dual-Branch Sampler-Guided Mixup (DBSGM)** that ensures balanced learning across head, medium, and tail classes by leveraging specialized samplers and mixup strategies; (2) the introduction of **Adaptive Class-Aware Feature Regularization (ACFR)**, which employs a temperature-adjusted supervised contrastive loss to improve class-specific robustness and representation learning; and (3) the integration of these components into an **end-to-end framework** that achieves state-of-the-art performance on long-tailed recognition tasks by effectively balancing decision boundary smoothness and class specificity.

We thoroughly evaluated the effectiveness of the proposed methodology by employing a rigorous assessment process on well-known datasets for long-tailed visual recognition, specifically CIFAR-LT-10, CIFAR-LT-100, ImageNet-LT and iNaturalist 2018. Our approach surpasses conventional end-to-end training methods, 2-stage methods, contrastive learning methods, and dual branch methods designed for long-tailed visual recognition, thus demonstrating the usefulness of our proposed method.

Advancing Beyond Chapter 3. This chapter builds directly upon the dual-branch framework introduced in Chapter 3 (STTP-Net), which tackled long-tailed class imbalance by combining sampling strategies with a two-pronged network and a hybrid mixup augmentation method. While STTP-Net effectively improved representation learning across head, medium, and tail classes by using distinct samplers and classification heads, it did not explicitly address the underlying issue of *decision boundary fidelity*, which plays a critical role in generalization. In contrast, the current chapter presents a more refined framework motivated by the *Goldilocks Principle*, aiming to achieve *just right* decision boundaries neither too sharp (which may overfit the tail classes) nor too soft (which may blur class distinctions). This is achieved by introducing two major innovations: (1) **Dual-Branch Sampler-Guided Mixup (DBSGM)**, which improves upon the earlier Hy-

brid Mixup by incorporating median-based sampling in addition to instance and reverse sampling, thereby strengthening medium-class representations without degrading head-class accuracy; and (2) **Adaptive Class-Aware Feature Regularization (ACFR)**, which applies a temperature-adaptive supervised contrastive loss (TASCL) directly on the feature space, encouraging better intra-class compactness and inter-class separability. Furthermore, this work formalizes the concept of *boundary thickness* and shows its impact on classifier overfitting and underfitting across the long-tailed spectrum. Unlike Chapter 3, which largely focused on architecture and sampling-augmentation integration, this chapter offers a deeper theoretical and empirical treatment of the decision boundary behavior in long-tailed settings. Thus, it not only generalizes the STTP-Net framework into a more unified and theoretically grounded system but also introduces a single-stage, end-to-end solution with better performance.

The chapter is structured as follows: Section 2 discusses the related work. Section 3 presents the proposed methodology, beginning with the preliminaries, including augmentation techniques and sampling strategies, followed by a detailed explanation of the proposed modules. Section 4 covers the experiments and results, followed by Section 5, which focuses on the ablation study. Finally, Section 6 concludes the chapter.

4.2 Related Works

We already discuss the literature related with Long-tailed classification in the chapter 3 but we further extend the discussion in the context of this chapter and the way we further trying to solve the problem. This section will discuss the different methods used to handle LTR, which vary from rebalancing in the data space, data augmentation methods, instance level and class-specific reweighing methods, calibration of logits, multi-stage methods, and ensemble-based methods.

Undersampling (He and Garcia, 2009; Japkowicz and Stephen, 2002a; Van Hulse et al., 2007a) and Oversampling (Van Hulse et al., 2007a; Mondal et al., 2024a) are the most commonly used data resampling techniques to address the problem of imbalanced data. While undersampling reduces the number of instances in the majority class, hence reducing the overall data points, oversampling does the exact opposite via repetition of existing samples (Van Hulse et al., 2007a) or generation of synthetic samples(Chawla et al., 2002a).

Loss function modification via reweighting strategies like (Huang et al., 2016a; Wang et al., 2017), Balanced Softmax (Ren et al., 2020b), Class balanced Loss (Cui et al., 2019b), focuses on assigning weights to samples according to their class cardinality to ensure that models pay more attention to minority classes. Normal Cross-Entropy loss squeezes the features of tail classes. Hence, Li et al. (Li et al., 2022c) suggest a Gaussian perturbation of the logits to enlarge the distribution of the tail classes.

An alternative avenue of research delves into various image augmentation techniques, such

as cropping and rotation, designed to amplify the representation of minority classes. These techniques typically act directly on input images, thereby inherently limiting the diversity of minority classes within datasets exhibiting a long-tailed distribution. Resampling techniques can learn spurious correlations between irrelevant contexts and labels. To address this, Shi et al. (Shi et al., 2023) introduce the Context Shift Augmentation (CSA) framework, which dynamically extracts and transfers well-separated contexts using Grad-CAM and a memory bank, improving tail-class representation. While CSA integrates with class-balanced resampling and other augmentation methods, it is not standalone and relies on existing frameworks like balanced sampling. Its drawbacks include the need for accurate context separation, increased computational complexity, and potential performance degradation in noisy or complex head-class contexts. ISDA (Wang et al., 2019a) employs semantic augmentations by estimating class-wise covariance to address these inherent limitations matrices. However, computing these matrices faces challenges with minority classes in long-tailed datasets due to limited samples. To tackle this issue, MetaSAug (Li et al., 2021) introduces a dynamic framework that optimizes augmentation strategies for minority classes using online meta-learning. Mixup (Zhang et al., 2017b), Remix (Chou et al., 2020), and CutMix (Yun et al., 2019a) are effective augmentation strategies in long-tailed classification. Mixup combines input data and labels through convex combinations, while CutMix replaces random image patches with patches from other training images. Park et al. (Park et al., 2022) enhance CutMix by incorporating context-rich majority class images as background for minority class images in the foreground. Applying augmentation to the minority classes, despite balancing the long-tailed data distribution, also introduces imbalance inherently as the augmentation strategies are only applied to minority classes (Wang et al., 2024). This creates an intertwined imbalance in the data: *inherent data-wise imbalance* that comes from the data distribution and *extrinsic augmentation-wise imbalance* that arises from the augmentation strategy. To address this issue, Wang et al. (Wang et al., 2024) proposed DODA (Dynamic Optional Data Augmentation), which allows each class to choose appropriate augmentation methods during training by maintaining a 'preference list' for each class. While DODA effectively reduces class sacrifices and improves tail class accuracy, it introduces computational overhead due to dynamic preference updates, which may limit scalability in extremely large datasets.

Learning proper image representations is a central challenge in long-tailed classification. BBN (Zhou et al., 2020) is a dual-branched network with a common resnet backbone in which one branch is trained on the given data distribution, and the other branch is trained on a reverse sampler, which has a higher probability of sampling tail classes. ResLT (Cui et al., 2022) furthers this approach to propose a network with three branches that each learn from different combinations of *head, medium & tail* classes. Separate representation and classifier training is another way to ensure proper representation learning of tail classes. Kang et al. (Kang et al., 2020) introduced a two-stage method where they trained the representation learning backbone on uniformly sampled

data and the classifier head on class-balanced samples.

Another state-of-the-art approach is Routing Diverse Distribution-Aware Experts (RIDE) (Wang et al., 2021b), a multi-expert network architecture employing branch-specific loss functions. It also uses a diversity loss based on divergence measures to ensure collaborative decision-making among multiple prediction models, with experiments utilizing two such models. Aimar et al. (Aimar et al., 2023b) introduced the *Balanced Product of Experts (BalPoE)*, which combines diverse expert models using logit adjustment to calibrate experts for specific class regions, ensuring unbiased predictions and minimizing balanced error. While BalPoE achieves state-of-the-art performance, it requires accurate class distribution priors, introduces computational overhead due to ensemble training, and may suffer from performance decline if training and test distributions differ. This trade-off between robustness and simplicity makes BalPoE less suitable for resource-constrained applications.

Fan et al.'s cumulative dual-branch network framework (CDBNF) (Fan et al., 2022) tackles class imbalance and feature representation simultaneously through a network with two branches. One of the branches emphasizes tail class performance, while the other performs few-shot learning to learn better representations. The network also uses a cumulative learning strategy to emphasize tail classes progressively. To address the head-tail trade-off in long-tailed data, the Bilateral expert-based Feature Redundancy Reduction and Rebalancing (Bi-F3R) approach (Chen et al., 2023), inspired by information theory actively removes redundant head class information through a dedicated module and complementary losses that guide forgetting redundant head class features and learning discriminative tail-class features. This makes the model forget excessive information about head classes, which counterbalances the imbalance in the dataset.

Representation learning has witnessed promising advancements through the application of contrastive learning techniques. Wang et al.(Wang et al., 2021a) introduce an innovative hybrid network architecture that seamlessly integrates a supervised contrastive loss for image representation learning with a cross-entropy loss for classifier training. This approach facilitates a gradual transition from feature acquisition to classifier learning. However, a fundamental challenge arises as contrastive learning inherently struggles with achieving uniform feature distributions, particularly in the context of long-tailed data distributions. To address this limitation, Li et al.(Li et al., 2022e) propose a novel framework that enhances the uniformity of feature distribution on the hypersphere, thereby improving the performance of supervised contrastive learning.

Learning the representations and classifier in a single network might hinder the learning process due to the extremely imbalanced nature of long-tailed datasets. Hence, a two-stage approach of decoupling the representation and classifier learning has been gaining popularity. Kang et al. (Kang et al., 2020) trained a representation learning network with a uniform sampler and the classifier network on a class-balanced sampler to learn better representations. Mixup was later brought into this network by Zhou et al. (Zhou et al., 2020).

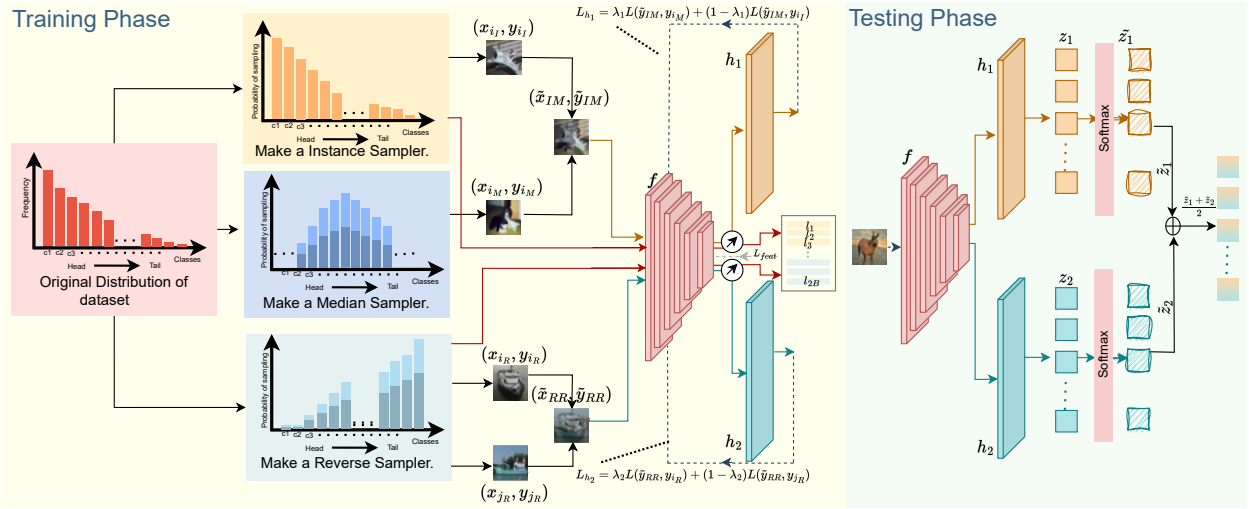


Figure 4.1: The figure depicts the training and testing phases of the proposed framework. The training phase is shown on the left, while the testing phase is illustrated on the right. The sampler used in the training phase is visually represented, demonstrating the probabilities of different classes. The loss function employed for updates in different branches is also mentioned, along with an explanation of how the loss on features is calculated.

Miscalibration poses a critical challenge in long-tailed calibration, where predicted class probabilities are heavily influenced by sample counts. Overconfidence, especially in head classes, is common in classifiers. Zhong et al. (hong et al., 2021) addressed these issues by introducing shifted batch normalization and label-aware smoothing, combined with Mixup (Zhang et al., 2017b), to handle varying degrees of classifier overconfidence.

Zhao et al. (Zhao et al., 2024) frame the trade-off between head and tail class performance as a *multi-objective optimization (MOO)* problem and propose consolidating existing techniques via Multi-Objective Optimization-based Strategy Fusion (MOOSF). While MOOSF outperforms more complex methods by fusing simpler strategies, it introduces increased computational complexity and a limited performance ceiling when combining numerous strategies.

Long-tailed imbalance impacts not only vision tasks but also graph classification and transformers. Xu et al. (Xu et al., 2024a) introduce ImbGNN to address long-tail and structural imbalances in graph classification by leveraging graph-specific features. Similarly, Kandpal et al. (Kandpal et al., 2023) show that large language models (LLMs) struggle with infrequent patterns due to pre-training data reliance, with retrieval augmentation offering a promising solution to enhance long-tail information capture.

4.3 Proposed Methodology

4.3.1 Preliminaries

4.3.1.1 Boundary Thickness and Soft vs. Hard Boundaries

The concept of boundary thickness as mentioned in (Yang et al., 2020) provides a formal mechanism to characterize the transition between classes in a classifier's decision space. Let $\mathbb{I}(\cdot)$ denote the indicator function and $\mathbb{E}[\cdot]$ represent the expectation over the data distribution. The boundary parameters α and β define the 'Goldilocks' region—the optimal thickness that prevents both overfitting (too sharp) and underfitting (too soft). For $\alpha, \beta \in (-1, 1)$ and a distribution p over pairs of points $(x_r, x_s) \sim p$, with predicted labels i and j for x_r and x_s respectively for defining the boundary thickness let's first define:

$$\Delta(x_r, x_s) = \|x_r - x_s\|,\ \Phi(t) = \mathbb{I}\{\alpha < g_{ij}(x(t)) < \beta\} \text{ and,}$$

$$\Psi = \int_0^1 \Phi(t)\, dt.$$

Then, the boundary thickness is defined as: :

$$\Theta(f, \alpha, \beta, p) = \mathbb{E}_{(x_r, x_s)\sim p}\left[\Delta(x_r, x_s) \cdot \Psi\right]. \tag{4.1}$$

where $g_{ij}(x) = f(x)_i - f(x)_j$ represents the difference in prediction probabilities, $x(t) = tx_r + (1 - t)x_s$, $t \in [0, 1]$, and $f(x) : X \to [0, 1]^C$ is the prediction function, with $f(x)_i$ denoting $\Pr(y = i \mid x)$. For binary linear classifiers, the boundary thickness reduces to:

$$\Theta(f, \alpha, \beta) = \frac{g^{-1}(\beta) - g^{-1}(\alpha)}{\|w\|},$$

where $g(\cdot) = 2\sigma(\cdot) - 1$, and $\sigma(\cdot)$ is the sigmoid function.

Hard boundaries are characterized by sharp transitions, minimal transition regions, and high-confidence predictions near the decision boundary. Mathematically, this corresponds to a small difference between α and β in $\Theta(f, \alpha, \beta)$, leading to behavior resembling a step function. While such boundaries can result in overfitting especially for tail classes in long-tailed distributions they provide strong separation between classes. In contrast, soft boundaries introduce gradual transitions with a wider decision region, yielding smoother changes in predictions across boundaries. This is represented by a larger difference between α and β in $\Theta(f, \alpha, \beta)$, which enhances robustness for tail classes by reducing overfitting but may cause poor class separation in some cases. These observations underscore the importance of balancing boundary thickness to optimize classification performance across the head and tail of the distribution.

4.3.1.2 Boundary Effects on Long-tailed Distributions

In long-tailed distributions with n classes, the boundary effects critically affect model performance. Hard boundaries, defined by decision functions such as:

$$g_{\text{tail},j}(x) = \text{sign}(w_{\text{tail}}^T x + b_{\text{tail}})$$

often overfit on tail classes due to their limited sample sizes, creating unreliable decision regions with high variance in the predictions:

$$\text{variance}(\|w_{\text{tail}}\|) > \text{variance}(\|w_{\text{head}}\|).$$

In contrast, soft boundaries with increased thickness, quantified by:

$$\Theta_{\text{tail}}(f, \alpha, \beta) > \Theta_{\text{head}}(f, \alpha, \beta),$$

provide better regularization for tail classes, leading to more robust decision regions, improved generalization for rare classes, and reduced sensitivity to sample sparsity. However, there are trade-offs: excessively soft boundaries can result in poor class separation, while overly hard boundaries exacerbate overfitting. The optimal boundary thickness, governed by the regularization term: $R_{\text{thick}}(f) = \lambda \sum_{i \in \text{tail}} \left(\Theta_i(f, \alpha, \beta) - \Theta_{\text{target}}\right)^2$, varies across head and tail classes and requires balancing classification accuracy and generalization. This highlights the need for adaptive strategies to dynamically adjust boundary thickness across the class distribution. Because of this reason, we present a method that will dynamically adjust the boundary thickness.

Sampling strategies. To compensate for the skewed distribution of classes in imbalanced datasets, machine learning models leverage specific sampling techniques that aim to balance the representation of different classes. The uneven presence of classes within a dataset requires the use of effective sampling techniques in machine learning. To capture the essence of various sampling approaches, we can define a unified formula for defining sampling techniques, which gives probabilities of sampling from individual classes as:

$$p_j = \frac{(|n_j - \zeta| + \varepsilon)^{-\gamma}}{\sum_{i=1}^{K} (|n_i - \zeta| + \varepsilon)^{-\gamma}}. \tag{4.2}$$

This approach defines various sampling strategies for imbalanced datasets with K classes. Two key parameters govern these strategies: γ, a "softness" factor in the range of 0 to 1, which adjusts the influence of class distribution, and ζ, some statistics for the whole dataset (like median, max, or min class counts). A small value, ε, is added to prevent division by zero errors.

The sampling formulation presented above is mathematically equivalent to the version introduced in Chapter 3, but offers a more generalized and expressive perspective. Specifically, by

substituting $\gamma = -q$, $\zeta = \psi$, and $\varepsilon = \epsilon$, we obtain:

$$p_j = \frac{(|n_j - \psi| + \epsilon)^q}{\sum_{i=1}^{C}(|n_i - \psi| + \epsilon)^q},$$

which is the original formulation used to define instance-balanced, reverse-balanced, and median-balanced sampling. The revised parameterization in this chapter introduces two intuitive control variables: γ, which acts as a "softness" factor controlling the influence of class imbalance (with higher values emphasizing balancing), and ζ, which flexibly anchors the sampling strategy through dataset-level statistics (e.g., 0 for instance-balanced, total count N for reverse, or median M for median-balanced sampling). This makes the current formulation a strict generalization of the earlier one, offering broader interpretability and facilitating integration into adaptive or learnable sampling mechanisms.

Mixup Augmentation. Mixup is a data augmentation technique for training machine learning models, particularly on image classification tasks. It leverages the linear interpolation of both features and labels from two randomly chosen data points:

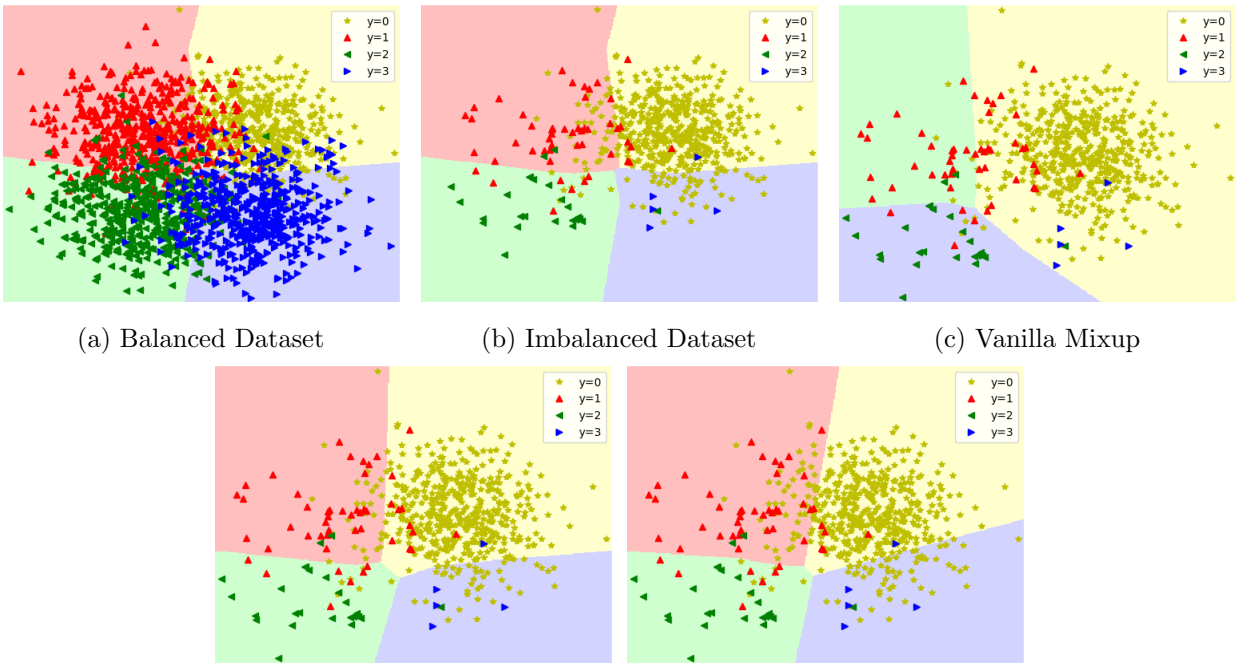
$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j, \tag{4.3}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda)y_j, \tag{4.4}$$

Where $\tilde{x}$ and $\tilde{y}$ are the interpolated features and labels, respectively. (x_i, y_i) and (x_j, y_j) are the features and labels of randomly chosen original data points. λ is a mixing parameter $\sim \text{Beta}(\alpha, \alpha)$. Mixup creates virtual training examples by blending features and labels from existing data points, encouraging the model to learn generalizable features. This technique helps prevent overfitting, reduces the chances that the model memorizes specific training points, and reduces the sensitivity of the model to noise.

4.3.2 Method Overview

Consider the K-class supervised classification problem with a training dataset $\mathcal{D} = (x_i, y_i)_{i=1}^{N}$, where $x_i \in \mathbb{R}^{W \times H \times \mathcal{C}}$ and $y_i \in C$, with $C = \{1, 2, 3, \ldots, K\}$. Let X_k represent the collection of samples associated with class k, and $N_k = |X_k|$ signify the size or cardinality of the set X_k. We assume the classes are arranged in the order $N_1 \geq N_2 \geq N_3 \geq \cdots \geq N_K$, where $\sum_{k \in K} N_k = N$. As illustrated in Figure 4.1, our proposed method consists of a feature extraction backbone network f (considered as a CNN-based network for extracting features from images), and two MLP classifiers h_1 and h_2 which take extracted features from f. Three samplers are defined as instance sampler ($\mathcal{I}$) (formed by putting $\gamma = 1$, and $\zeta = 0$ in equation 4.2), median sampler ($\mathcal{M}$) (formed by putting $\gamma \in (0, 1]$, and $\zeta = \textit{median}(\{N_1, \cdots, N_K\})$ in equation 4.2), and reverse sampler ($\mathcal{R}$) (formed by putting $\gamma \in (0, 1]$, and $\zeta = \textit{maximun}(\{N_1, \cdots, N_K\})$ in equation 4.2). Data is drawn from these



(a) Balanced Dataset (b) Imbalanced Dataset (c) Vanilla Mixup

(d) DBSGM (e) DBSGM+TASCL

Figure 4.2: Figure (a) shows a balanced dataset along with the decision region produced by the classifier if the data is balanced. Figure (b) shows an imbalanced dataset and the decision boundaries when the data is balanced. Figure (c) shows decision regions produced by vanilla mixup trained on an imbalanced dataset. Figure (d) shows decision regions by DBSGM. Figure (e) shows the decision region produced by DBSGM+TASCL. In Figure (c), the majority class (yellow) dominates the decision region, with only the majority class being classified correctly, while the other minority classes are misclassified. This highlights the limitation of vanilla mixup in handling imbalanced datasets, which is addressed by DBSGM and DBSGM+TASCL in Figures (d) and (e), respectively.

samplers and augmented based on our proposed Dual-Branch Sampler-Guided Mixup (DBSGM). Training is performed by simultaneously applying DBSGM and Adaptive Class-Aware Feature Regularization (ACFR). Detailed explanations of the augmentation method, feature regularization, training procedure, and testing phase are provided in the subsequent subsections.

4.3.3 Dual-Branch Sampler-Guided Mixup (DBSGM)

For our augmentation method DBSGM, the data is first sampled from instance sampler, $(x_{i_I}, y_{i_I}) \sim \mathcal{I}(\mathcal{D}, 1)$ and median sampler, $(x_{i_M}, y_{i_M}) \sim \mathcal{M}(\mathcal{D}, \gamma)$, and combined using mixup as follows:

$$\tilde{x}_{IM} = \lambda_1 x_{i_I} + (1 - \lambda_1) x_{i_M}, \tag{4.5}$$

$$\tilde{y}_{IM} = \lambda_1 y_{i_I} + (1 - \lambda_1) y_{i_M}, \tag{4.6}$$

Secondly, augmentation is applied to two samples taken from the reverse sampler as $(x_{i_R}, y_{j_R}) \sim \mathcal{R}(\mathcal{D}, \gamma)$ and $(x_{j_R}, y_{j_R}) \sim \mathcal{R}(\mathcal{D}, \gamma)$. Mixup is then applied between these two samples as:

$$\tilde{x}_{RR} = \lambda_2 x_{i_R} + (1 - \lambda_2) x_{j_R}, \tag{4.7}$$

$$\tilde{y}_{RR} = \lambda_2 y_{i_R} + (1 - \lambda_2) y_{j_R}, \tag{4.8}$$

These two augmented samples are then passed to the feature extractor network f. The feature corresponding to $\tilde{x}_{IM}$ is fed to h_1, and features corresponding to $\tilde{x}_{RR}$ are fed to h_2. Branch 1 tackles the bias towards head classes by carefully mixing them with medium classes. Here, an instance sampler captures the natural abundance of classes, while a dedicated median sampler prioritizes medium classes through increased selection probability. Mixup then augments these samples, ensuring both head and medium classes contribute meaningfully to the learning process. Meanwhile, branch 2 dedicates itself to elevating the often-neglected tail classes. It accomplishes this via a reverse sampler, which boosts their representation by increasing their selection probability. This targeted approach directly addresses the scarcity of tail class data, providing the network with crucial information to learn their distinguishing features. By strategically balancing head and medium class learning while also improving tail class performance, our dual-branch network with mixup achieves superior overall classification accuracy compared to standard methods, demonstrating its effectiveness in handling imbalanced datasets. In DBSGM, we observe that varying the probability of sampling samples using γ in equation 4.2 helps improve accuracy between different classes and achieves the best overall balanced accuracy. Losses corresponding to each mixup sample are calculated at the respective head using the balanced softmax loss as $\mathcal{L}_{h1}$ and $\mathcal{L}_{h2}$.

4.3.4 Justification for DBSGM Over HybridMix

In Chapter 3.4.3, we introduced *HybridMix*, a dual-augmentation strategy combining CutMix between samples drawn from instance and median samplers, and Mixup applied to samples from a reverse sampler. While HybridMix offers improved generalization under class imbalance by diversely augmenting different class strata (head, medium, and tail), it is fundamentally limited by the properties of CutMix as an augmentation strategy. CutMix, as a spatial composition technique, replaces a patch of one image with a region from another and interpolates the corresponding labels. While effective for standard classification settings, recent investigations have uncovered limitations of CutMix under adversarial or distributionally shifted settings. Notably, Rebuffi et al. (**?**) conducted a comprehensive study demonstrating that augmentation strategies such as CutMix and Mixup behave differently with respect to robustness and generalization. They argue that although CutMix achieves strong performance in clean accuracy and alleviates robust overfitting to some extent, it does not outperform Mixup in maintaining robustness across training epochs

when combined with model weight averaging. Mixup, by linearly interpolating both input images and labels, introduces smoother transitions in the data manifold and encourages the model to learn more globally linear decision boundaries. These smoother boundaries inherently reduce the model's sensitivity to small perturbations, making Mixup particularly beneficial in the context of underrepresented (tail) class generalization and adversarial robustness. The study by Rebuffi et al. (**?**) shows that:(1) Mixup effectively mitigates robust overfitting, a known issue in long-tailed and adversarially robust training. (2) When combined with model weight averaging (WA), Mixup consistently improves the model's robust accuracy throughout training, surpassing CutMix under ℓ_∞ perturbations. (3) Unlike CutMix, Mixup promotes consistency across model snapshots, making it more compatible with temporal ensembling strategies such as WA.

Empirically, Rebuffi et al. observed that while CutMix yields high clean accuracy and performs well on robust test sets when WA is applied, Mixup stands out in preserving robustness across epochs without suffering from abrupt degradation post-learning rate decay. This property is critical in our setting, where preserving tail class information and preventing overfitting on overrepresented head classes is crucial. Further, the smooth interpolation of feature and label space in Mixup complements the dual-branch architecture proposed in DBSGM. Branch 1 of DBSGM integrates Mixup between instance and median samplers, targeting the imbalance between head and medium classes. Branch 2 employs Mixup on reverse-sampled tail-class data, enhancing rare class representation. The use of Mixup across both branches in DBSGM ensures: (1) Controlled blending of underrepresented and overrepresented class samples without introducing abrupt spatial discontinuities (which can occur with CutMix). (2) Smooth label transitions that regularize class boundaries and avoid overfitting to noisy or sparse tail samples. (3) Compatibility with balanced loss functions and weight averaging, both of which synergize with DBSGM's dual-branch architecture to boost balanced accuracy. Therefore, to leverage the theoretical robustness guarantees and empirical superiority of Mixup over CutMix in imbalanced and robustness-sensitive settings, we replace the CutMix-based hybrid strategy with our proposed Dual-Branch Sampler-Guided Mixup (DBSGM). DBSGM builds upon the strengths of Mixup by combining it with carefully designed sampling strategies and also using the fractional value of γ to further get the smooth sampling that target different parts of the class frequency spectrum (head, medium, and tail), resulting in a more principled and robust augmentation pipeline for long-tailed classification.

This strategic shift from HybridMix to DBSGM is not only supported by emerging literature but also by our empirical findings which demonstrate superior balanced and tail-class accuracy when training under class-imbalanced regimes.

4.3.5 Proof of concept demonstrating why DBSGM provides better class boundaries and the necessity of contrastive loss.

To demonstrate that our Mixup-based algorithm provides better decisions compared to the standard Mixup, we used a 2D - toy dataset with four classes, each containing 500 (yellow), 50 (red), 25 (green), and 5 (blue) samples, resulting in a class imbalance ratio of 100. Figure 2 illustrates how decision regions change across different methods. Subfigure (a) shows the decision boundaries produced by the classifier when the dataset is balanced, while subfigure (b) depicts the imbalanced dataset and the corresponding decision regions when the data is balanced. Subfigure (c) demonstrates the decision regions obtained using Vanilla Mixup on the imbalanced dataset, highlighting its limitations in effectively separating classes, particularly minority classes. Subfigure (d) shows the improved decision regions achieved by our proposed DBSGM method, which provides better separation between the majority and minority classes and intermediate classes. Finally, subfigure (e) illustrates how adding TASCL to DBSGM further enhances the decision boundaries, resulting in improved separation between minority and median classes. Together, these visualizations demonstrate the efficacy of DBSGM and TASCL in addressing class imbalance and improving decision boundaries, particularly for minority and intermediate classes.

However there are some drawbacks associated with the Mixup, it creates new data points by interpolating between existing ones. This can lead to smoother decision boundaries, which might be desirable sometimes. However, excessive smoothing can cause the model to miss important features in the data, potentially reducing its ability to distinguish certain classes, especially when the data is highly separable (Oh and Yun, 2023; Yang et al., 2020). Also, as discussed in (Yang et al., 2020), mixup can lead to thicker, more gradual decision boundaries that are less sharp than standard training. The decision regions may also be more fragmented in some cases. Also, as seen from the figure mixup is unable to classify (or cptures) the features close to the boundaries of the different classes.

Thus, to resolve the above issues associated with Mixup augmentation, we use the concept of contrastive learning in the form of supervised contrastive learning (SCL). SCL loss offers a potential solution to these drawbacks. By incorporating class labels during training, SCL loss steers the model towards learning class-relevant distinctions, potentially leading to more interpretable decision boundaries. Additionally, SCL encourages similar classes to be closer in the embedding space while pushing dissimilar ones apart. This can help maintain sharper decision boundaries compared to Mixup's smoothing effect, which might be detrimental to well-separated data. However, it is important to remember that SCL's effectiveness can be data-dependent, achieving the optimal balance between smooth and sharp boundaries.

However, directly applying the supervised contrastive loss function (Khosla et al., 2020) to data with a long-tailed distribution is infeasible. This can be demonstrated by examining the lower

bound of the SCL loss, which is calculated under the assumption of balanced data in Lemma S1 in (Graf et al., 2021) as given below.

Lemma 2. *Given that normalized feature embeddings are used, let* $L = (l_1, l_2, \cdots, l_N) \in \mathfrak{L}^N$ *represent a configuration of* N *points, each associated with labels* $Y = (y_1, y_2, y_3, \cdots, y_N) \in C^N$. *Here,* $\mathfrak{L}$ *is defined as* $\{l \in \mathbb{R}^h : ||l|| = 1\}$, *indicating that each point* l *in* $\mathbb{R}^h$ *is normalized to have a unit norm. The class-specific batch-wise loss can then be expressed as follows:*

$$\mathcal{L}(L; Y, B, \tau, y) \geq \sum_{i \in B_y} \log \left((|B_y| - 1) + |\overline{B_y}| \exp \left(\underbrace{\frac{1}{|\overline{B_y}|} \sum_{k \in \overline{B_y}} \frac{l_i \cdot l_k}{\tau}}_{\textit{repulsion term}} - \underbrace{\frac{1}{|B_y| - 1} \sum_{j \in B_y \setminus i} \frac{l_i \cdot l_j}{\tau}}_{\textit{attraction term}} \right) \right) \tag{4.9}$$

where B is the batch and $|B|$ is its cardinality, $B_y \subseteq B$ contains all data points belonging to class y, and $|B_y|$ is its cardinality. In the given boundterm $\frac{1}{|\overline{B_y}|} \sum_{k \in \overline{B_y}} \frac{l_i . l_k}{\tau}$, can be though of a repulsion term and the term $\frac{1}{|B_y| - 1} \sum_{j \in B_y \setminus i} \frac{l_i . l_j}{\tau}$, can be thought of as a attraction term. The attraction term in SCL aims to pull similar data points closer together in the feature space. In a balanced setting, this helps refine class-specific features. However, with imbalanced data, this term can excessively collapse features within the minority classes, regardless of their actual variation. This essentially reduces the distinctiveness of minority class data points. The repulsion term in SCL pushes dissimilar data points apart. Ideally, this fosters a clear separation between classes. However, in imbalanced datasets, the majority classes heavily influence the repulsion term. Since there are many more samples from these classes in each mini-batch, the repulsion term prioritizes pushing everything away from the majority, potentially neglecting the specific features of minority classes. This can lead to uneven feature separation, with larger distances between the majority classes and smaller distances between minority classes themselves. Furthermore, the imbalanced nature amplifies the impact of gradients during training. Since there are more samples from the majority classes, the gradients calculated from pushing data points away from these classes will be much stronger. This can lead the training process to prioritize optimizing the representation of the majority classes at the expense of the minority classes.

Solution: Redistribute attraction and repulsion terms by adjusting them differently for the majority and minority classes. The loss defined in Equation A.15 is a class-specific, batch-wise loss calculated on a balanced dataset. We first convert this class-specific loss to a loss computed across all classes within each batch to address the imbalanced dataset. By summing these class-level losses over all classes in the batch, we obtain a lower bound for the total batch loss.

Theorem 3. *Consider a dataset comprising* N *normalized feature embeddings,* $L = \{l_1, l_2, ..., l_N\} \subset \mathfrak{L}^N$, *where* $\mathfrak{L} = \{l \in \mathbb{R}^h : ||l||_2 = 1\}$ *is the unit hypersphere in* $\mathbb{R}^h$. *Associated with each embedding*

l_i is a class label $y_i \in C$, where C is the set of all classes. Partition the label set into majority and minority subsets, denoted Y_M and Y_m respectively, such that $Y = Y_M \cup Y_m$. The aggregate batch-wise loss, $\mathcal{L}(L; Y, B, \tau)$, is defined as the summation of class-conditional batch-wise losses across the entire class space:

$$\mathcal{L}(L;Y,B,\tau) \geq \log\left(\prod_{y\in Y}|\overline{B_y}|^{|B_y|}\right)+ \sum_{y\in Y^M}\left(\sum_{i\in B_y}\left(\underbrace{\frac{\sum_{k\in B\setminus B_y}\frac{l_i\cdot l_k}{\tau_M}}{|\overline{B_y}|}}_{\text{Repulsion by Majority}} - \underbrace{\frac{\sum_{j\in B_y\setminus i}\frac{l_i\cdot l_j}{\tau_M}}{|B_y|-1}}_{\text{Attraction by Majority}}\right)\right) + \sum_{y\in Y^m}\left(\sum_{i\in B_y}\left(\underbrace{\frac{\sum_{k\in B\setminus B_y}\frac{l_i\cdot l_k}{\tau_m}}{|\overline{B_y}|}}_{\text{Repulsion by Minority}} - \underbrace{\frac{\sum_{j\in B_y\setminus i}\frac{l_i\cdot l_j}{\tau_m}}{|B_y|-1}}_{\text{Attraction by Minority}}\right)\right), \quad (4.10)$$

Proof see Supplementary A.2.1.

comment. The lower bound stated in 3 characterizes the minimum achievable value of the batch-wise loss under normalized embeddings and balanced attraction–repulsion forces. Similar to the analysis in (Graf et al., 2021), the bound is approached when intra-class similarities are maximized and inter-class similarities are uniformly minimized on the unit hypersphere. However, unlike the balanced-data assumption in prior work, class imbalance causes the repulsion term of minority classes to be disproportionately weakened due to normalization by larger complementary sets. As a result, the lower bound is generally not tight in long-tailed settings. This observation motivates the introduction of class-dependent temperature scaling, which redistributes attraction and repulsion forces and tightens the bound under imbalanced data. A formal derivation and detailed analysis are provided in Appendix A.2.

We divide the attraction and repulsion terms into four components: attraction by head, attraction by tail, repulsion by head, and repulsion by tail. For simplicity in explaining the tail part, we can consider its median classes. As seen from equation 4.10, the repulsion term by the head class repels the tail classes to a much greater extent than the tail classes repel the head classes. This imbalance may result in the collapse of the class boundaries associated with the tail classes. Therefore, adjusting the parameters to alter this phenomenon is more feasible. Specifically, we can change the temperatures associated with the head class repulsion and the tail class repulsion terms so that the power of repulsion by the majority class is less than that of the tail classes. Thus, different temperatures will be associated with these repulsion terms. Let τ_M and τ_m be the temperatures associated with the majority and minority classes, respectively. It is worth noting that τ_M must be greater than τ_m to meet the required repulsion conditions for instances of majority

and minority classes.

Moreover, if we talk about both attraction terms, learning the instance-related features for the tail classes to achieve much better classification accuracy will be more feasible. Thus, if we associate a lower temperature with this, it will only affect the instances nearer to it rather than those quite far away. However, in the case of majority classes, a higher temperature can also work, as it has many instances and can influence the faraway neighbors to make them similar to the particular term being applied. Therefore, for the attraction, τ_M must be greater than τ_m.

Based on this analysis, we will propose the temperature adaptive supervised contrastive loss (TASCL) next. The toy dataset in Figure 4.2 (e) demonstrates how supervised contrastive loss modifies the decision boundaries to learn instances of the minority class better. Specifically, it optimizes the discrimination between instances of the minority class, thereby improving overall classification accuracy.

Based on the above discussion, some queries must be tackled. We introduce the feature regularization module ACFR to achieve this:

Query1. *Mixup creates smoother decision boundaries, which can be undesirable for highly separable data. This smoothing can lead the model to miss crucial features and hinder class distinction, especially for minority classes.*

Solution. we explore contrastive learning approaches such as Supervised Contrastive Learning (SCL) to address this. SCL encourages sharper boundaries by pushing dissimilar classes apart and pulling similar ones closer, leveraging class labels during training. To achieve this, we introduce the feature regularization module ACFR, as discussed in the next section.

Query2. *Directly applying SCL to long-tailed datasets (where some classes have significantly fewer samples) presents challenges. The attraction term in SCL can excessively collapse features within minority classes, reducing their distinctiveness. Additionally, the repulsion term gets dominated by the majority classes, neglecting minority class features and leading to uneven separation.*

Solution. The principles outlined in (Wang and Isola, 2020) emphasize that a representation should extract the shared information between positive pairs while remaining invariant to noise. Therefore, the loss function must satisfy two properties: alignment and uniformity (Wang and Isola, 2020). Alignment ensures that samples from the same class map to nearby features, thus becoming invariant to irrelevant noise. Uniformity, however, suggests that feature vectors should be roughly uniformly distributed to preserve as much class-specific information as possible without compromising the model's discriminative capabilities. Supervised Contrastive Learning (SCL) aims to bring all class representations to a single central representation, avoiding the properties of uniformity and alignment and potentially reducing the model's ability to discriminate between instances within the same class. To address this, we propose a modification of the SCL loss, ensuring that representations remain distinguishable from their neighbors, are appropriately grouped with instances from the same classes, and retain instance-specific features. This is achieved using the

proposed loss function called Temperature Adaptive Supervised Contrastive Loss (TASCL).

Query3. *While sharper boundaries are often desirable, some level of smoothness might still be beneficial for generalization. Finding the optimal balance between these two aspects can be challenging.*

Solution. To address the challenge of balancing sharp boundaries with smoothness for optimal generalization, we devised a strategic training process for the network using a long-tailed dataset. This process alternates between training the network with our proposed modules, DBGSM and ACFR, effectively leveraging both mixup augmentation strategies and contrastive learning approaches. The integration and utilization of these modules are detailed in the overall training algorithm section.

Query 4 *How to mitigate the issue of imbalanced gradients during training with SCL for long-tailed data?*

Solution. We propose a batching strategy to address imbalanced gradients during training with supervised contrastive loss (SCL) on long-tailed data. This strategy involves creating a new batch that combines features extracted from two sampling methods: an instance and a reverse sampler. The instance sampler draws samples to reflect the natural class distribution, while the reverse sampler ensures an equal representation of classes by oversampling the minority classes. By merging these two batches, we achieve a balanced representation of different classes within each training batch. This approach mitigates the problem of gradient imbalance during training with contrastive loss, ensuring that the number of samples from each class is roughly equal. Consequently, the gradients computed during backpropagation are more balanced, leading to more stable and effective model training on long-tailed datasets.

The ACFR module, the overall training strategy, and the testing phase are discussed in the following section.

4.3.6 Adaptive Class-Aware Feature Regularization (ACFR)

While Mixup promotes smoother decision boundaries, it may not intrinsically address irrelevant intra-class variations that can hamper classification accuracy. This limitation arises from its focus on interpolating existing data points, which can inadvertently amplify such variations. We introduce a technique called Adaptive Class-Aware Feature Regularization (ACFR) to mitigate this issue. ACFR leverages a loss function that operates directly on the extracted features, effectively guiding the model's learning process toward a feature space that exhibits strong intra-class cohesion and pronounced inter-class separation. This complements the benefits of Mixup and contributes to developing a more robust and accurate classification model. Based on the previous section's supervised contrastive loss analysis, we propose a Temperature Adaptive Supervised Contrastive Loss (TASCL) that operates directly on the extracted features to achieve these goals.

The loss is defined as follows:

$$\mathcal{L}^{feat}(L,\tau) = \sum_{i=1}^{2B} \mathcal{L}_{x_i}^{feat}, \tag{4.11}$$

$$\mathcal{L}_{x_i}^{feat}(L,\tau) = \frac{-1}{N_{y_i}^{2B}-1} \sum_{j=1}^{2B} \mathbb{1}_{i\neq j}\mathbb{1}_{y_i=y_j} \log \frac{\exp\left(l_{x_i}\cdot l_{x_j}/\tau_{y_i}\right)}{\sum_{k=1}^{2B}\mathbb{1}_{i\neq k}\exp(l_{x_i}\cdot l_{x_k}/\tau_{y_i})}, \tag{4.12}$$

where $l_{x_i} = f(x_i)$ (extracted feature of image x_i with label y_i), y_i is the class of the sample x_i, and τ_{y_i} is the temperature corresponding to class y_i of the sample x_i. To calculate the loss, batches are taken from both the instance sampler and the reverse sampler, and these batches are combined to create a batch of size $2B$. This combined batch is then used to calculate the loss as defined above. Thus we can construct the set $L = \{l_{x_1}, \cdots, l_{x_i}, \cdots, l_{x_B}, \cdots, l_{x_{2B}}\}$ to calculate loss $\mathcal{L}^{feat}(L,\tau)$. $N_{y_i}^{2B}$ is the number of images in the combined batch $2B$ belonging to class y_i. Moreover, another important parameter in the loss is the temperature set, $\tau = \{\tau_{y_1}, \cdots, \tau_{y_i}, \cdots, \tau_{y_B}, \cdots, \tau_{y_{2B}}\}$. We calculate τ_{y_i} corresponding to each sample x_i, depending on the prediction probability of each sample corresponding to its label from their respective head in the batch $2B$. Samples from $\{1,\ldots,B\}$ obtain their sample probability from head h_1 as they belong to the instance sampler. On the other hand, samples from $\{B+1,\ldots,2B\}$ obtain their sample probability from head h_2 as they belong to the reverse sampler. We determine the temperature τ using these probabilities.

Since temperature is a crucial parameter in loss and plays a role in imposing penalties on the samples, as temperature decreases, it results in more significant penalties in the loss term. Conversely, a higher temperature value leads to a lower penalty. Additionally, considering the classifier's ability to provide probabilities corresponding to the sample (x_i, y_i), we can assert that a sample with a lower probability should incur a higher penalty than a sample with a higher probability during prediction. This aligns with our analysis in Theorem 3, which demonstrates the necessity of assigning different temperatures to majority and minority classes. Depending on this, we can say that the probability of predicting sample x_i given class y_i i.e. $p(x_i|y_i) \propto \tau_{y_i}$. Thus, we can propose the function for τ_{y_i} as:

$$\tau_{y_i} = \tau_{min} + (\tau_{max} - givenclass\tau_{min}) \cdot p(x_i|y_i), \tag{4.13}$$

where τ_{max} and τ_{min} are the minimum and maximum values the temperature can take. Using this, we can apply more penalties to the less confident sample and fewer penalties to the more confident sample.

By pushing similar samples closer together and driving dissimilar samples further apart, ACFR promotes the learning of semantically meaningful features. This means the features extracted by the model capture the essence of a class, going beyond superficial appearance and focusing

on the underlying concepts that define it. This focus on semantics translates to more robust and generalizable representations, ultimately improving classification accuracy across diverse data points.

4.3.7 Overall Training Algorithm

Algorithm 3 illustrates the overall proposed training algorithm. Initially, for the first few warm-up epochs, denoted as T_{warm}, only the $\mathcal{L}^{\text{feat}}$ loss is applied to the features with a temperature parameter $\tau_{\min}$, while keeping the classifier heads frozen. Subsequently, from epoch T_{warm} to T_{AugMax}, we engage in simultaneous training using the DBSGM method. This involves updating both the feature backbone f and the heads h_1 and h_2 using the balanced softmax losses $\mathcal{L}_{h_1}$ and $\mathcal{L}_{h_2}$. During this phase, we also update only the backbone f using the $\mathcal{L}^{\text{feat}}$ loss with temperature, calculated using equation 4.13. The transition between exclusively updating the backbone using $\mathcal{L}^{\text{feat}}$ and updating both the backbone and heads using $\mathcal{L}_{h_1}$ and $\mathcal{L}_{h_2}$ depends on a probability parameter ν and given its threshold p_{aug} (representing whether to perform augmentation or not), calculated before each epoch.

After reaching T_{AugMax}, the heads are frozen, and only the backbone is updated using the $\mathcal{L}^{\text{feat}}$ loss with temperature, calculated using equation 4.13. Additionally, please note that during the training in the ACFR phase, to enhance training efficiency, we utilize a single-view approach for contrastive learning. Unlike dual-view methods such as contrastive loss, such as BCL (Zhu et al., 2022), which employ multi-view approaches that double the batch computation, we rely on the diversity introduced by our DBSGM sampler to simulate contrastive pairs, thereby reducing computational overhead without sacrificing representational quality.

4.3.8 Testing Phase

In the testing phase, a test image passes through the feature extractor f and then through both the classifier heads h_1 and h_2. Softmax is applied to each head after collecting the logits from both heads. Finally, the mean of the softmax outputs is used to predict the class of the test image. Let's define the prediction function for the test image x as:

$$\mathcal{P}(x) = \frac{softmax(h_1(f(x))) + softmax(h_2(f(x)))}{2}, \tag{4.14}$$

where $\mathcal{P}(x) = [p_1, \cdots, p_K]$ that is the K-dimension array of predicted probability corresponding to each class. Predicted class c can be found out using $c = \{i \in \mathbb{Z} \mid 1 \leq i \leq K \text{ and } p_i = max(\mathcal{P}(x))\}$. The complete testing phase is also illustrated in Figure 4.1.

Algorithm 3 Proposed Training Algorithm

Input: For a dataset D, the model comprises parameters θ for the backbone f, and parameters ψ_1 and ψ_2 for classifier heads h_1 and h_2. The instance sampler is denoted as $\mathcal{I}(\mathcal{D}, 1)$, and the Median Sampler and Reverse Sampler are $\mathcal{M}(\mathcal{D}, \gamma)$ and $\mathcal{R}(\mathcal{D}, \gamma)$. Loss functions for the first and second heads are $\mathcal{L}_{h_1}(.)$ and $\mathcal{L}_{h_2}(.)$, and the feature loss is $\mathcal{L}^{feat}$. The parameter for the epoch probability of choosing between augmentation or direct feature update is ν with the threshold p_{aug}. The learning rate is denoted as η. Total epochs are T, maximum warm-up epochs are T_{warm}, and maximum augmentation epochs are T_{AugMax}.
Output: Trained network f, h_1 and h_2.

Computation:

Randomly initialize θ, ψ_1 and ψ_2.
Dataset sampled using $\mathcal{I}(\mathcal{D}, 1)$: $D_{\mathcal{I}} \sim \mathcal{I}(\mathcal{D}, 1)$
Dataset sampled using $\mathcal{M}(\mathcal{D}, \gamma)$: $D_{\mathcal{M}} \sim \mathcal{M}(\mathcal{D}, \gamma)$
Dataset sampled using $\mathcal{R}(\mathcal{D}, \gamma)$: $D_{\mathcal{R}} \sim \mathcal{R}(\mathcal{D}, \gamma)$
for epoch $= 1, ..., T$ **do**
  if epoch $\geq T - T_{\text{warm}}$ and $\nu < p_{aug}$ and $T \leq T_{\text{AugMax}}$ **then**
    // Augmentation phase: activate sampler-guided Hybrid-Mixup for tail-aware training
    for batch $i = 1, ..., B$ **do**
      Draw a mini-batch (x_{i_I}, y_{i_I}) from $D_{\mathcal{I}}$
      Draw a mini-batch (x_{i_M}, y_{i_M}) from $D_{\mathcal{M}}$
      Draw a mini-batch (x_{i_R}, y_{i_R}) and (x_{j_R}, y_{j_R}) from $D_{\mathcal{R}}$
      Generate Sample (x_{IM}, y_{IM}) using samples (x_{i_I}, y_{i_I}) and (x_{i_M}, y_{i_M}) by using Eqn's 5.5 and 5.6
      Generate Sample (x_{RR}, y_{RR}) using samples (x_{i_R}, y_{i_R}) and (x_{j_R}, y_{j_R}) by using Eqn's 4.7 and 4.8
      $(\theta, \psi_1) \leftarrow (\theta, \psi_1) - \eta \nabla \mathcal{L}_{h_1}((x_{IM}, y_{IM}); (\theta, \psi_1))$
      $(\theta, \psi_2) \leftarrow (\theta, \psi_2) - \eta \nabla \mathcal{L}_{h_2}((x_{RR}, y_{RR}); (\theta, \psi_2))$
    end for
  else
    // Feature-regularization phase: update backbone using adaptive feature loss without augmentation
    for batch $i = 1, ..., B$ **do**
      Draw a mini-batch (x_{i_I}, y_{i_I}) from D_I
      Draw a mini-batch (x_{i_R}, y_{i_R}) from D_R
      Extract the features using the backbone f as:
      $L_I = f(x_{i_I})$, and $L_R = f(x_{i_R})$,
        where, L_I is the set of all feature in batch (x_{i_I}, y_{i_I}), and
          L_R is the set of all feature in batch (x_{i_R}, y_{i_R})
      $p_I = softmax(h_1(L_I))$
      $p_R = softmax(h_1(L_T))$
      Calculate τ_I and τ_R using p_I and p_R using equation 4.13
      Concatenate temperature sets τ_I and τ_R to make temperature set $\tau = \tau_I \cup \tau_R$
      Concatenate features sets L_I and L_R to make feature set $L = L_I \cup L_R$
      Calculate $\mathcal{L}^{feat}$ using L and τ, using equation 4.11
      If $T \leq T_{warm}$ then $\tau = \tau_{min}$ else $\tau = \tau$
      Finally update the backbone f using $\mathcal{L}^{feat}$
      $(\theta) \leftarrow (\theta) - \eta \nabla \mathcal{L}^{feat}(L, \tau; (\theta))$
    end for
  end if
end for

4.4 Experiments and Discussion

4.4.1 Datasets and Experimental Setup

4.4.1.1 Dataset Overview

The assessment of our proposed method involves a rigorous examination utilizing three distinct datasets tailored for long-tailed visual recognition tasks: CIFAR-LT-10, CIFAR-LT-100, ImageNet-

LT and iNaturalist 2018. The datasets, denoted as *Long-tailed CIFAR-10* and *Long-tailed CIFAR-100*, are specifically crafted long-tailed versions of the CIFAR10 and CIFAR100 datasets, respectively (Krizhevsky, 2009). To maintain methodological consistency, we adhere to the experimental settings outlined in (Cao et al., 2019) for the long-tailed adaptations of CIFAR datasets. Our experimental investigations encompass CIFAR-LT-10 and CIFAR-LT-100, exploring various imbalance ratios, precisely 10, 50, 100, and 200.

The *Long-tailed ImageNet* dataset is a skewed distribution derived from the original ImageNet dataset (Deng et al., 2009). This skewed variant is produced by incorporating an exponential distribution into the class distribution. We utilize the *ImageNet-LT* dataset introduced by Liu et al. in (Liu et al., 2019a), which encompasses 115.8K images categorized into 1000 classes, exhibiting class cardinalities ranging from 5 to 1,280.

The *iNaturalist 2018* (Horn et al., 2018) dataset is a large-scale benchmark commonly used for evaluating methods addressing long-tailed class imbalance. It comprises 437.5K datapoints across 8,142 classes, with an extreme imbalance ratio of 500, making it particularly challenging for conventional learning algorithms.

4.4.1.2 Experimental Training Overview

This section provides an overview of training procedures for long-tail datasets, specifically CIFAR-LT-10, CIFAR-LT-100, and ImageNet-LT. We follow parameter configurations from (Cao et al., 2019) for CIFAR-LT-10 and CIFAR-LT-100, using ResNet-32 as the backbone. For both, a batch size of 128 is used over 200 epochs with SGD optimizer (momentum 0.9), an initial learning rate of 0.1, with a linear warm-up learning rate scheduler applied for the first 5 epochs, and decay at the 160^{th} and 180^{th} epochs. For ImageNet-LT, ResNet-50 and ResNet-10 are used as backbones, following settings in (Kang et al., 2020). A cosine learning rate schedule is applied, decreasing from 0.1 to 0, with an image resolution of 224×224 and a consistent batch size 256. In all ResNet-50/10 experiments, the SGD optimizer with a momentum of 0.9 is employed. The code is available at: https://anonymous.4open.science/r/Submission2-213D/.

4.4.1.3 Evaluation Metrics

We focus on top-1 accuracy as the key performance metric. Following the method outlined by (Liu et al., 2019a) on CIFAR-LT-100 and Imagenet-LT, accuracy is grouped into three subsets: many-shot classes (> 100 samples), medium-shot classes ($20 - 100$ samples), and few-shot classes (< 20 samples).

4.4.2 Results and Discussion

To compare our method, we evaluated several categories of existing approaches. The reweighing-based methods include LDAM/LDAM-DRW (Cao et al., 2019), CE-CBRW/DRW (Cui et al.,

2019b), IB-Focal/CB (Park et al., 2021), and BALMS (Ren et al., 2020b). Augmentation-based methods considered are Mixup (Zhang et al., 2017b), CutMix (Yun et al., 2019a), CMO+CE/DRW/BS (Park et al., 2022), and MetasAug (Li et al., 2021). We also examined dual-branch networks like BBN (Zhou et al., 2020), ResLT (Cui et al., 2022), RIDE (Wang et al., 2021b), CBD (Iscen et al., 2021), CDBNF (Fan et al., 2022), and Bi-F3R (Chen et al., 2023). Two-stage methods such as τ-Norm (Kang et al., 2020), cRT (Kang et al., 2020), LWS (Kang et al., 2020), DPM (Zhang et al., 2023a), and MisLAS (hong et al., 2021) are also included. Additionally, miscellaneous methods encompassing knowledge distillation, logit adjustment, self-supervised learning, and contrastive learning-based methods (except for Balanced Contrastive Learning (BCL) (Zhu et al., 2022), which utilizes multiple views of a single image during training, our approach uses only a single view.) are compared, including CBS+RRS (Zhang et al., 2019), DisAlign (Zhang et al., 2021), TDE (Tang et al., 2020a), SEQL (Yang and Xu, 2020), OLTR (Liu et al., 2019a), SSP (Yang and Xu, 2020), Hybrid-SC (Wang et al., 2021a), TSC (Li et al., 2022e), LADE (Hong et al., 2021), Causal (Tang et al., 2020a), and GCL (Li et al., 2022c).

Since our method is a single-stage dual-branch network, we did not consider ensemble models involving multiple experts, though we used models like RIDE and compared them when they used two branches.

4.4.2.1 Results comparison on CIFAR-LT-10 and CIFAR-LT-100 Dataset

To ensure an unbiased comparison of our approach with other methods on CIFAR-LT-10 and CIFAR-LT-100 datasets, we replicated the outcomes of contrasting methods by implementing their source code. We maintained uniformity across different techniques using the same seed values, optimizer setups, and learning rates. Table 4.1 presents top-1 accuracy on CIFAR-LT-10 for various imbalance ratios (IR) from 10 to 200. Our method consistently outperforms reweighing-based methods across different IRs, showing almost 2% for IR = 10 improvements, increasing to around 6% for IR = 200. Compared to Balanced Softmax and IB+CB, our method excels with the highest improvement observed at IR = 200. Additionally, against augmentation-based methods like CMO+BS, our method exhibits improvements ranging from 1% to 1.72%. In comparisons with multi-branch networks (BBN and ResLT) and a two-stage method (MisLAS), our method consistently outperforms them, with substantial improvements at various IRs. Furthermore, against contrastive learning-based methods (Hybrid-SC and TSC), our method shows significant improvements of 2.55% and 3.89%, respectively, for IR = 100.

Table 4.2 compares the CIFAR-LT-100 dataset, showcasing top-1 accuracy for many-shot, medium-shot, and few-shot classes. Our method outperforms reweighing-based approaches across various imbalance ratios. Compared to CMO combined with balanced softmax (CMO+BS) and DRW, our method demonstrates top-1 accuracy improvements of 3.37% and 2.52% for IR=100, and 4.61% and 3.24% for IR=200. Notably, our method exhibits improvements in medium and few-

Table 4.1: Comparison on CIFAR-LT-10 dataset with different State-of-the-art methods in terms of top-1 accuracy (%) for ResNet-32 architecture with different imbalance ratios.

Methods	IR=10	IR=50	IR=100	IR=200
CE	86.81	77.00	71.21	66.05
LDAM (Cao et al., 2019)	86.79	79.05	74.64	69.55
CE-CBRW (Cui et al., 2019b)	86.88	78.53	73.16	65.65
CB-Focal (Cui et al., 2019b)	83.98	70.13	64.58	35.61
CE-DRW (Cao et al., 2019)	88.28	80.44	76.36	70.38
LDAM-DRW (Cao et al., 2019)	87.78	82.36	78.33	73.38
IB-Focal (Park et al., 2021)	87.27	77.42	71.71	65.96
IB+CB (Park et al., 2021)	88.32	81.74	78.31	74.79
Balanced Softmax (Ren et al., 2020b)	88.24	81.66	78.4	72.87
Mixup (Zhang et al., 2017b)	88.24	79.16	72.75	66.13
CutMix (Yun et al., 2019a)	88.22	79.59	75.61	69.75
CMO (Park et al., 2022)	88.72	80.51	73.84	71.07
CMO+DRW (Park et al., 2022)	89.37	84.60	81.76	77.41
CMO+BS (Park et al., 2022)	89.18	85.44	82.70	78.33
MetasAug (Li et al., 2021)	89.23	83.9	82.07	76.27
Causal (Tang et al., 2020b)	88.20	83.10	79.90	76.60
BBN (Zhou et al., 2020)	87.95	80.67	77.40	73.76
LADE (Hong et al., 2021)	87.60	82.80	79.70	75.20
MisLAS (hong et al., 2021)	90.33	85.56	82.43	76.06
Hybrid-SC (Wang et al., 2021a)	91.12	85.36	81.40	-
GCL(Li et al., 2022c)	89.83	84.99	81.98	78.24
TSC (Li et al., 2022e)	88.70	82.90	79.70	-
ResLT (Cui et al., 2022)	88.98	82.96	78.87	75.07
STTP-Net (ours)	91	86.4	83.65	80.88
BalPOE (no. expert: 2)(Aimar et al., 2023b)	83.16	79.69	72.07	64.69
BalPOE (no. expert: 7)(Aimar et al., 2023b)	73.83	76.89	75.06	64.57
Ours	90.52	86.09	83.59	80.05

shot classes for IR=100, with gains of almost 6.46% and 4.09% for CMO+BS and CMO+DRW in medium classes, and a 4.1% improvement over CMO+DRW for few-classes. However, CMO+BS has a 2.7% higher accuracy for few-shot classes but at the cost of a significant drop in many-shot class performance. Similar trends are observed for IR=200. Additionally, compared to MetasAug, our method shows overall accuracy improvements of 2.11% and 2.22% for IR=100 and 200, with gains in medium and few-shot classes. Compared to multi-branch networks (BBN, ResLT, and RIDE), our method outperforms with improvements of 7.39%, 4.79%, and 1.73% for IR=100, and 7.99% and 4.51% for BBN and ResLT in IR=200. Compared to contrastive learning methods (Hybrid-SC and TSC), our method shows improvements of 2.63% and 5.55% for top-1 accuracy. Furthermore,

employing RandAugment during training enhances our method's accuracy by approximately 2.85% and 1.6% for medium and few-shot classes with IR=100, and 4.61% and 2.91% for IR=200.

4.4.2.2 Results comparison on Imagenet-LT Dataset

Table 4.3 displays ImageNet-LT results, comparing ResNet-10/50 backbones. Our approach outperforms τ-norm by around 2% for both 90 and >180 epochs and cRT by approximately 1% for ResNet-10. It surpasses OLTR with a top-1 accuracy gain of 4.73% for ResNet-10. SSP's self-supervised learning provides effective feature initialization, but our method is superior by about 1.6% (ResNet-10). Additionally, our model outshines DPM by 2.03% for ResNet-10 and beats ResLT and SEQL (using ResNet-10) by 1.09% and 5.63%, respectively. Table 4.4 demonstrates good performance across decoupled methods for top-1, many-shot, medium-shot, and few-shot accuracy. It also surpasses ResLT in top-1, many-shot, and medium-shot accuracy, with slightly lower few-shot accuracy. For ResNet-50, our method excels against most decoupled methods and is comparable to TDE and DisAlign, even as a single-stage end-to-end framework. It outperforms multi-branch networks like ResLT and self-supervised networks like SSP.

4.4.2.3 Results Comparisons on iNaturalist2018 Dataset

The table 4.5 demonstrates the performance comparison of our method with existing approaches on the iNaturalist2018 dataset. Our approach achieves competitive results, with a top-1 accuracy of 70.01%, surpassing several methods such as LDAM-DRW (68.0%), OLTR (63.9%), and Decouple-LWS (68.5%). Notably, our method outperforms many of these baselines in handling medium and few classes, as indicated by the accuracy scores of 70.70% and 70.028%, respectively. This highlights the robustness of our approach in addressing the long-tail distribution challenge. Compared to methods like Balanced Softmax (70.5%) and LADE (70.6%), our approach maintains comparable performance, while offering a distinct advantage in the medium and few categories. For instance, while LADE achieves 70.3% for the medium class, our method achieves a higher accuracy of 70.70%, demonstrating its efficacy in managing subpopulation shifts and class imbalances. Furthermore, our results highlight a balanced performance across different class categories, unlike approaches such as CMO, which achieves a high score for the many classes but struggles with the few classes (62.9%). In summary, the results validate the effectiveness of our method, particularly in scenarios with class imbalances, by achieving a better balance across the many, medium, and few categories without compromising overall classification accuracy.

4.4.2.4 Justification of the Comparative Study

While our method demonstrates robust performance across multiple benchmarks, its design philosophy and computational feasibility guided our choice of comparative baselines. Specifically, we focused on single-stage approaches, including reweighting-based methods (e.g., LDAM-DRW,

Table 4.2: Comparison of CIFAR-LT-100 dataset with different State-of-the-art methods in terms of overall top-1 accuracy (%) for IR= 10, 50, 100, and 200 along with its top-1 accuracy (%) for Many-shot, Medium-shot, and Few-shot classes for ResNet-32 architecture with IR=100 and 200. (Entries with * denote results directly taken from the corresponding research work.)

Methods	Venue	IR = 10	IR = 50	IR = 100				IR = 200			
				Top1-acc	Many	Medium	Few	Top1-acc	Many	Medium	Few
CE		56.81	43.99	37.58	64.8	35.97	7.7	34.8	66.33	37.39	8.49
LDAM (Cao et al., 2019)	[NeurIPS'19]	56.01	42.97	39.7	67.54	39	8.03	36.02	67.1	41.61	7.67
CE-CBRW (Cui et al., 2019b)	[CVPR'19]	56.58	43.53	38.31	65.37	35.54	9.97	34.56	66.9	36.45	8.18
CB-Focal (Cui et al., 2019b)	[CVPR'19]	52.49	40.22	34.88	59.94	32.03	8.97	31.92	62.17	33.52	7.38
CE-DRW (Cao et al., 2019)	[NeurIPS'19]	58.01	46.6	40.98	62.71	40.74	15.9	37.81	63.5	43.23	13.74
LDAM-DRW(Cao et al., 2019)	[NeurIPS'19]	57.45	46.61	43.44	62.11	44.57	20.33	37.86	62.27	41.16	16.46
IB-Focal (Park et al., 2021)	[ICCV'21]	57.82	43.74	42.67	56.71	44.6	24.03	37.72	56	41.29	20.82
IB+CB (Park et al., 2021)	[ICCV'21]	54.78	42.21	37.27	68.09	33.43	5.8	33.02	68.37	33.13	5.74
Balanced Softmax(Ren et al., 2020b)	[NeurIPS'20]	59.54	47.19	42.54	58.66	42.31	24	37.93	60.57	41	18.08
Mixup (Zhang et al., 2017b)	[ICLR'18]	58.06	44.72	40.08	70.91	38.91	5.47	36.15	67.22	32.26	2.28
CutMix (Yun et al., 2019a)	[ICCV'19]	59.21	46.72	40.74	71.40	40.57	5.17	36.33	70.5	40.84	6.46
CMO (Park et al., 2022)	[CVPR'22]	60.22	47.46	41.95	69.51	40.6	11.37	38.7	70.2	43.74	10.46
CMO+DRW (Park et al., 2022)	[CVPR'22]	62.08	51.71	46.83	63.26	47.77	26.57	41.89	61.93	48.16	21.49
CMO+BS (Park et al., 2022)	[CVPR'22]	61.52	51.2	45.98	57.37	45.4	33.37	40.52	56.07	42.16	27.26
MetasAug (Li et al., 2021)	[CVPR'21]	61.83	51.09	47.24	63.86	47.74	27.52	42.91	62.86	43.6	19.41
Causal (Tang et al., 2020b)	[NeurIPS'20]	59.4	48	45	64.8	47.6	18.8	39.7	62.3	47.9	15.9
BBN (Zhou et al., 2020)	[CVPR'20]	58.33	46.62	41.96	53.37	51.71	17.27	37.14	53.63	50.1	14.15
LADE (Hong et al., 2021)	[CVPR'21]	60.3	50.4	45.3	61.3	43.3	29.1	39.6	60.9	43.2	20.3
MisLAS (hong et al., 2021)	[CVPR'21]	62.05	51.4	47.25	62.83	48.14	26.83	42.92	61.64	44.11	18.24
Hybrid-SC (Wang et al., 2021a)	[CVPR'21]	63.05	51.87	46.72	-	-	-	-	-	-	-
TSC (Li et al., 2022e)	[CVPR'22]	59.0*	47.4*	43.8*	-	-	-	-	-	-	-
GCL(Li et al., 2022c)	[CVPR'22]	61.66	53.55	48.71	-	-	-	44.88	-	-	-
CDBNF (Fan et al., 2022)	[EAAI'22]			45.1	61.1	45.1	25.4	40.9	62.5	46.4	20.9
Bi-F3R (Chen et al., 2023)	[EAAI'23]	-	53.15	50.54	-	-	-	46.89	-	-	-
DPM (Zhang et al., 2023a)	[TNNLS'23]	59.47	49.32	44.8*	60.7*	46.4*	24.4*	-	-	-	-
RIDE (Wang et al., 2021b)	[ICLR'21]	-	-	47.62	66.14	48.31	25.2	-	-	-	-
ResLT (Cui et al., 2022)	[TPAMI'23]	60.14	49.06	44.56	60.25	47.14	21.97	40.62	58.25	44.86	13.62
ResLT§ (Cui et al., 2022)	[TPAMI'23]	62.01*	52.71*	48.21*	-	-	-	-	-	-	-
ResLT§ (Cui et al., 2022)	[TPAMI'23]	62.01*	52.71*	48.21*	-	-	-	-	-	-	-
BalPOE (No. Exp: 2)(Aimar et al., 2023b)	[CVPR'23]	44.07	42.17	40.72	52.88	44.05	22.63	33.94	47.51	40.37	10.59
BalPOE (No. Exp: 7)(Aimar et al., 2023b)	[CVPR'23]	52.81	44.78	43.66	59.62	44.40	24.16	34.86	49.60	38.28	13.66
DODA (Wang et al., 2024)	[ICLR'24]	58.68	50.82	46.62	61.19	49.11	26.7	43.77	63.26	49.45	24.25
MOOSF(BS+CE-DRW) (Zhao et al., 2024)	[ICML'24]	61.55	51.21	47.16	61.68	49.14	27.89	42.40	61.86	46.35	24.28
STTP-Net (Ours)	[EAAI]	63.58	53.02	49.19	63.94	50.83	30.07	43.02	51.37	49.06	31.79
Ours		62.67	54.36	49.350	62.86	51.86	30.67	45.13	65.90	48.32	26.62
Ours§		60.780	53.60	50.34	61.46	54.71	32.27	45.94	60.03	52.93	29.53

Table 4.3: Evaluating ResNet-10 and ResNet-50 architectures on the Imagenet dataset involves comparing their top-1 accuracy (%) with various state-of-the-art methods. The top-1 accuracy results for peer algorithms were sourced from their respective papers, and the results marked with "†" are obtained from (Cui et al., 2022).

Epoch	Method	Resnet-10	Resnet-50
90	Focal(Lin et al., 2020)	30.5	-
	BALMS(Ren et al., 2020b)	41.8	-
	τ-norm(Kang et al., 2020)	40.6	46.7
	cRT(Kang et al., 2020)	41.8	47.3
	CBS+RRS(Zhang et al., 2019)	41.9	47.3
	LWS(Kang et al., 2020)	41.4	47.7
	TDE(Tang et al., 2020a)	-	51.1
	DisAlign(Zhang et al., 2021)	-	51.3
	CBD(Iscen et al., 2021)	37.9	51.6
	MetaSAug(Li et al., 2021)	-	47.4
	SEQL(Yang and Xu, 2020)	36.4	-
	OLTR(Liu et al., 2019a)	37.3	-
	CDBNF(Fan et al., 2022)	38.5	-
	LADE(Hong et al., 2021)	-	51.9
	CMO(Park et al., 2022)	-	49.1
	STTP-Net(ours)	42.33	52.3
	Ours	42.03	52.8
>180	τ-norm(Kang et al., 2020)	42.7†	46.7
	cRT(Kang et al., 2020)	43.2†	47.3
	LWS(Kang et al., 2020)	43†	47.7
	SSP(Yang and Xu, 2020)	43.2	51.3
	ResLT(Cui et al., 2022)	43.8	48.5
	DPM(Zhang et al., 2023a)	42.0	51.0
	TSC(Li et al., 2022e)	-	52.4
	DODA (Wang et al., 2024)	-	48.1
	LADE(Hong et al., 2021)	-	53.0
	MiSLAS(hong et al., 2021)	-	52.7
	GCL(Li et al., 2022c)	-	54.8
	MOOSF(BS+CE-DRW)(Zhao et al., 2024)	-	53.2
	STTP-net (Ours)	44.00	53.00
	Ours	44.86	54.12

CE-DRW, IB-Focal, BALMS), augmentation techniques (e.g., Mixup, CutMix, CMO, MetasAug), dual-branch networks (e.g., BBN, Bi-F3R), and two-stage methods (e.g., τ-Norm, cRT, LWS, MisLAS). These methods align closely with our single-stage, dual-branch, end-to-end framework, making them natural points of comparison. In contrast, ensemble-based methods such as RIDE, ResLT, and BalPoE, which depend on multiple experts to achieve enhanced performance, diverge funda-

Table 4.4: Top-1 Accuracy(%) of Many-shot, Medium-shot and Few-shot on ImageNet-LT with ResNet-10 and 50 backbones (for networks trained for >180 epochs) ("*","†", and "‡" denotes the results are from the original papers, (Cui et al., 2022) and (Kang et al., 2020), respectively.)

Backbone Model	Methods	Top1-acc	Many	Medium	few
ResNet-10	CE†	37.3	59.7	29.4	5.7
	cRT†	43.2	53.8	41.3	25.4
	CBD	37.9	49.2	36.9	21.5
	CDBNF	38.5	46.0	36.1	27.0
	τ-norm†	42.7	50.4	42.1	26.7
	LWS†	43	51.8	41.6	27.6
	ResLT†	43.8	52.3	41.6	29.5
	Ours	44.86	52.83	43.45	28.35
ResNet-50	CBD	51.6	65.2	48.0	25.9
	NCM‡	44.3	53.1	42.3	26.5
	cRT‡	47.3	58.8	44	26.1
	τ-norm‡	46.7	56.6	44.2	27.4
	LWS‡	47.7	57.1	45.2	29.3
	CE+CMO‡	49.1	67	42.3	20.5
	DPM*	51	64.6	48.3	22.1
	ResLT	48.5	52.4	47.4	42.4
	TSC	52.4	63.5	49.7	30.4
	LADE	53.0	65.1	48.9	33.4
	DODA	48.1	67.4	47.5	13.9
	Ours	54.12	65.56	50.58	35.34

mentally from our design principles. Although we included ResLT with two heads on ImageNet for a broader perspective, such ensemble-based methods inherently conflict with our goal of achieving simplicity and computational efficiency. Similarly, multi-strategy fusion frameworks like MOOSF, DODA, and CSA were omitted due to their reliance on complex mechanisms. While these approaches excel in specific scenarios, their significant computational and methodological overhead renders them less relevant for direct comparison with our streamlined framework. Despite this, we compared these methods on smaller datasets like CIFAR-10/100, where resource constraints were manageable, to offer additional context. Our method consistently outperformed these alternatives in these benchmarks. For larger datasets like ImageNet, we included the base versions of MOOSF and DODA, demonstrating our method's superior performance under comparable conditions. Finally, while training on iNaturalist required approximately 5 days & 12 hrs on an NVIDIA RTX 6000 Ada GPU for a single run, resource limitations restricted exhaustive parameter tuning. Instead, we used parameters similar to those optimized for ImageNet, which served as a reasonable approximation. While further fine-tuning could potentially enhance performance on iNaturalist, the existing results already demonstrate the robustness and scalability of our approach. In summary,

Table 4.5: Comparison of state-of-the-art methods on the iNaturalist2018 dataset. The table reports classification accuracy (%) for ResNet-50 on iNaturalist2018. Symbols ∗ † and ‡ denote results from the original papers, (Zhou et al., 2020), and (Cui et al., 2022) respectively.

Method	**Top-1 acc**	**Many**	**Medium**	**Few**
Cross Entropy (CE)	61.0	73.9	63.5	55.5
IB Loss (Park et al., 2021)*	65.4	-	-	-
LDAM-DRW (Cao et al., 2019)†	68.0	72.1	67.2	66.2
OLTR (Liu et al., 2019a)*	63.9	59.0	64.1	64.9
Decouple-cRT (Kang et al., 2020)*	68.9	73.2	68.6	66.6
Decouple-LWS (Kang et al., 2020)*	68.5	73.2	68.4	66.1
BBN (Zhou et al., 2020)*	69.6	-	-	-
Balanced Softmax (Ren et al., 2020b)	70.5	-	-	-
LADE (Hong et al., 2021)*	70.6	73.8	70.3	69.2
DisAlign (Zhang et al., 2021)*	70.06	-	-	-
MiSLAS (hong et al., 2021)*	71.6	-	-	-
GCL (Li et al., 2022c)*	72.01	-	-	-
Hybrid-SC (Wang et al., 2021a)*	66.74	-	-	-
TSC (Li et al., 2022e)*	69.7	72.6	70.6	67.8
CMO (Park et al., 2022)*	68.9	76.9	69.4	62.9
DODA (Wang et al., 2024)*	63.6	74.9	66.0	58.4
Ours	70.01	66.00	70.70	70.028
Ensemble-based and Multitask Methods				
RIDE (3 experts) (Wang et al., 2021b) ‡	71.7	68.3	72.6	71.8
ResLT (3 experts) (Cui et al., 2022) ‡	72.9	73.0	72.6	73.1
BalPoE (Aimar et al., 2023b)*	75.0	-	-	-
MOOSF(CE+LDAM-DRW) (Zhao et al., 2024)*	72.6	75.1	72.3	71.9

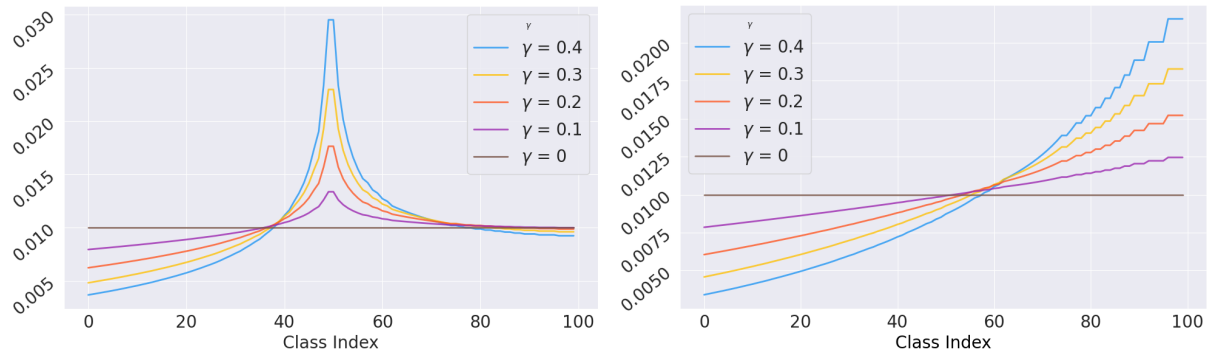
while we acknowledge the strengths of ensemble-based and multi-strategy fusion methods, their fundamental differences in design and computational demands make them less relevant than our lightweight, single-stage framework. Our focus remains on achieving a balance between simplicity, scalability, and competitive performance, which aligns with the goals of our proposed approach.

4.5 Ablation Study

How do key components influence the performance? This study explores the impact of individual and combined components (ACFR and DBSGM) on two imbalance ratios in our network framework. Table 4.6 shows that, at an imbalance ratio of 100, ACFR+DBSGM achieves a Top-1 accuracy of 49.35%, with notable improvements in many classes (62.86%), medium classes

Table 4.6: Analysis of ACFR and DBSGM components using CIFAR-LT-100 (IR= 100 & 200) dataset. Evaluation includes Top-1 accuracy and many, Medium, and Few-shot accuracies.

Module Used		Top 1-acc		Many		Medium		Few	
ACFR	DBSGM	IR = 100	IR = 200	IR = 100	IR = 200	IR = 100	IR = 200	IR = 100	IR = 200
✓	✓	49.35	45.13	62.86	65.9	51.86	48.32	30.67	26.60
✗	✓	48.38	43.6	63.08	63.03	50.65	48.35	28.56	24.87
✓	✗	44.63	40.72	62.77	64.33	45.42	44.16	22.53	19.82



(a) Median Sampler $\mathcal{M}(\mathcal{D}, \gamma)$ (b) Reverse Sampler $\mathcal{R}(\mathcal{D}, \gamma)$

Figure 4.3: The figure shows how the sampling probabilities change as we vary the value of γ in Equation 4.2. The plot depicts 'Sampling Probabilities' v/s 'Class Index' for (a) the Median sampler and (b) the Reverse Sampler, with varying values of γ for the CIFAR-LT-100 dataset.

Table 4.7: Comparison is made on top-1 accuracy, as well as on many-shot, medium-shot, and few-shot accuracies for various values of γ on the CIFAR-LT-100 dataset with imbalance ratios (IR) set to 100 and 200.

γ	Top 1-acc		Many		Medium		Few	
	IR=100	IR=200	IR=100	IR=200	IR=100	IR=200	IR=100	IR=200
0.4	48.89	43.91	60.71	58.7	51.85	50.67	31.63	27.15
0.3	48.39	43.70	63.17	62.86	50.00	49.06	29.26	24.69
0.2	48.48	45.13	62.11	65.90	49.85	48.32	30.96	26.6
0.1	49.35	44.16	62.86	63.90	51.86	48.06	30.67	25.87
0	47.89	44.20	64.91	65.86	48.71	49.06	27.06	23.6

(51.86%), and few-shot classes (30.67%). Removing ACFR decreases Top-1 accuracy to 48.38%, affecting few-shot class accuracy more (28.56%). Excluding DBSGM reduces Top-1 accuracy to 44.63%, particularly impacting many and medium classes and showing a significant decline in few-shot class accuracy (22.53%). Similar trends are observed at an imbalance ratio of 200, emphasizing the synergistic effects of ACFR and DBSGM in achieving robust performance across diverse class imbalances. ACFR handles imbalances in medium and few-shot classes, while DBSGM contributes significantly to overall model robustness. The combined architecture demonstrates improved performance, highlighting the importance of a comprehensive approach to address class imbalances.

How the parameter γ influence the performance? Understanding the γ parameter in Equation 4.2 is vital for addressing class imbalance. The sensitivity of γ (in the Figures 4.3a and 4.3b) demonstrates the trade-off between standard classification loss and contrastive regularization. A value that is too high restricts the classifier's ability to learn specific class nuances, while a value that is too low leads to a collapsed feature space for tail classes. Figures 4.3a and 4.3b illustrate how γ affects sampling probabilities in the CIFAR-LT-100 dataset. In the median sampler, $\gamma = 0$

equalizes class weights, with values between 0 and 1 favoring medium-frequency classes while maintaining some focus on lower-frequency ones. Increasing γ decreases head class probabilities faster than tail class probabilities, indicating preferential sampling of less frequent classes. The reverse sampler shows a similar trend, boosting tail class probabilities more than head class probabilities as γ increases. Examining γ impact on accuracy, experiments on CIFAR-LT-100 at IR=100 and IR=200 were conducted with varying γ values. In Table 4.7, optimal γ for max accuracy was 0.1 (IR=100) and 0.2 (IR=200). Notably, $\gamma = 0$ boosts Many-class accuracy for both IRs, while $\gamma \approx 0.4$ enhances Medium and Few-shot class accuracy. These findings underscore γ's crucial role in tailoring sampling behavior and its impact on classification performance. Achieving the right balance requires identifying the optimal γ for a specific imbalance scenario while maximizing overall accuracy and maintaining class-specific performance

4.6 Discussion

In conclusion, this study presents an innovative dual-branch network with a shared feature extractor designed to address the challenges of long-tailed class imbalances in deep neural networks, particularly for image recognition. By incorporating the DBSGM and ACFR modules, the framework demonstrates significant improvements in robustness and performance. It enhances accuracy for head and medium classes without sacrificing performance for tail classes. The model's feature regularization increases its resilience to intra-class variations, and the use of softer probabilities during sampling results in smoother decision boundaries that align with class semantics. This single-stage, end-to-end framework offers a promising solution to the problem of long-tailed class imbalance. Future research will focus on generalizing this approach to various domains and exploring its scalability and transfer learning potential.

As a potential future research avenue, we aim to explore combining our method with frameworks like MOOSF or DODA, which balance class-specific trade-offs and optimize augmentation strategies. Integrating our dual-branch sampling with these methods could enhance class separability and resolve conflicts between strategies. However, challenges arise from differing objectives, such as potential conflicts between our predefined sampling distributions and MOOSF's dynamic adjustments and the computational cost of merging DODA's adaptive components with our approach. Overcoming these challenges will require careful co-design to balance performance and computational feasibility across datasets.

Chapter 5

Robust Medical-Image Classification: From Class Imbalance to Demographic Fairness

Synopsis

This chapter presents two complementary contributions that address critical challenges in medical image classification arising from sub-domain shifts specifically, class imbalance and demographic bias. The first contribution proposes a dual-expert framework ***Mo2E*** *(Ansari et al., 2024a)* [1] *that improves classification under long-tailed distributions by learning decision boundaries between head and tail classes using expert-specific MixUp augmentation with targeted sampling strategies. We would like to highlight that this was the first idea we developed to address class imbalance in deep learning networks as part of our thesis. Building upon this dual-branch network concept, we subsequently proposed the methods presented in Chapter 3 and their improvements in Chapter 4. The second contribution in this chapter addresses algorithmic bias in skin lesion classification by introducing an* ***adaptive MixUp sampling method*** *(Ansari et al., 2024b)* [2]*, which uses a meta-learned heuristic to guide data augmentation, thereby mitigating the model's bias toward dominant skin tones while maintaining class-level balance. These two works are unified in this chapter because they both aim to enhance the* resilience of deep learning models to real-world sub-domain shifts *be it statistical (class imbalance) or demographic (skin tone variation). While Mo2E improves robustness across class frequency spectrum, the adaptive MixUp framework extends robustness to demographic fairness across skin tones. Together, they demonstrate how principled data-driven augmentation and expert-guided strategies can build equitable and generalizable medical AI systems, directly aligning with the central theme of this thesis: developing deep learners resilient to sub-domain shifts in medical imaging.*

[1] F. Ansari, A. Bhattacharya, B. Saha, and S. Das. "Mo2E: Mixture of Two Experts for Class-Imbalanced Learning from Medical Images." In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2024. doi: https://doi.org/10.1109/ISBI56570.2024.10635212.

[2] F. Ansari, T. Chakraborti, and S. Das. "Algorithmic Fairness in Lesion Classification by Mitigating Class Imbalance and Skin Tone Bias." In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, eds. M. G. Linguraru et al., pp. 373–382, Springer, Cham, 2024. doi: https://doi.org/10.1007/978-3-031-72378-0_35.

5.1 Introduction

The promise of deep learning in medical image analysis lies in its ability to automate diagnostic decision-making with high accuracy, consistency, and speed. However, this promise remains largely unrealized in real-world clinical deployments due to the persistent presence of *sub-domain shifts* distributional discrepancies between training and deployment environments. These shifts can emerge from several sources, including unequal class prevalence, demographic variability, device heterogeneity, or institutional bias. In this chapter, we focus on two such dominant and orthogonal challenges: **class imbalance** and **demographic bias**, both of which are widely encountered in medical image datasets but insufficiently addressed.

Class imbalance is an intrinsic property of most medical datasets due to the natural scarcity of certain pathological conditions and the difficulty of acquiring rare disease annotations. This imbalance is often long-tailed in nature, where a few common classes dominate the sample distribution (head classes), while rare but clinically critical categories (tail classes) remain severely underrepresented. Standard deep neural networks, when trained on such skewed data, tend to overfit to the majority classes and severely underperform on the minority ones, thereby undermining clinical reliability. The problem is exacerbated in fine-grained multi-class settings such as gastrointestinal (GI) lesion classification and diabetic retinopathy (DR) grading, where subtle visual differences and class overlap challenge even human experts. A variety of approaches have been proposed to mitigate class imbalance, broadly categorized into re-weighting (Japkowicz and Stephen, 2002b; Lin et al., 2017; Cui et al., 2019a; Cao et al., 2019), re-sampling (Kubat et al., 1997; Buda et al., 2018b; Nekooeimehr and Lai-Yuen, 2016; Koto, 2014), and augmentation-based techniques (Zhang et al., 2017a; Chou et al., 2020; Galdran et al., 2021). While re-weighting techniques attempt to compensate for imbalance via loss scaling, they are often sensitive to hyperparameter selection and can destabilize training. Re-sampling methods, though intuitively appealing, may result in information loss or overfitting due to redundant or discarded examples. Recently, augmentation methods such as MixUp have been shown to improve generalization through convex combinations of samples and labels. However, standard MixUp fails to consider the class semantics during interpolation and tends to enhance performance only for median-frequency classes, leaving tail class performance largely unaffected or even degraded. To address these limitations, we propose a structured and semantically guided framework, **Mo2E (Mixture of Two Experts)** (Ansari et al., 2024a), designed to improve class-specific decision boundaries in long-tailed distributions. Mo2E consists of two CNN-based experts that are trained with tailored MixUp strategies: one using a uniform-reverse sampling scheme to optimize head and head-tail boundaries, and another using reverse-reverse sampling to focus exclusively on tail-tail separability. During inference, class-specific confidence scores are weighted and fused based on the experts' domain of specialization, resulting in robust predictions across the distribution spectrum. We validate this framework using

two benchmark datasets:

- **Hyper-Kvasir**, a 23-class dataset for GI lesion classification, with imbalance ratios exceeding 180:1 across pathological categories.
- **Eyepacs**, a large-scale DR grading dataset with over 31,000 images, where later-stage DR conditions (e.g., proliferative DR) are significantly underrepresented.

These datasets represent two anatomically and clinically diverse tasks one endoscopic and one fundoscopic allowing us to rigorously demonstrate the effectiveness and generality of the Mo2E framework in real-world, long-tailed settings.

In parallel, we turn our attention to another critical source of sub-domain shift: **demographic bias**, particularly with respect to skin tone diversity in dermatological imaging. While skin cancer affects individuals across skin tones, the incidence is higher in patients with lighter skin, leading to benchmark datasets such as ISIC-2018 being disproportionately composed of fair-skinned patients. Although this may reflect epidemiological trends, it poses a severe risk for algorithmic bias: models trained on such datasets underperform on underrepresented skin tones, failing to generalize fairly across demographic lines (Groh et al., 2021). This imbalance in representation not only reduces diagnostic performance for marginalized populations but also raises critical ethical concerns in clinical AI deployment. To mitigate this, we introduce a novel **adaptive mixup-based framework with meta-learning-guided sampling** (Ansari et al., 2024b), which addresses both class and demographic imbalance in skin lesion classification. The method synthesizes training samples via MixUp, where one instance is drawn from an instance-based sampler and the other from a *meta-adaptive sampler*. The adaptive sampler dynamically adjusts class selection probabilities based on a heuristic computed over a separate meta-validation set that reflects target demographic diversity. This enables targeted exploration of underperforming or underrepresented classes during training, leading to fairer and more generalizable decision boundaries.

We validate our framework using two demographically diverse datasets:

- **ISIC-2018**, a global skin lesion classification benchmark with predominantly Caucasian skin tones.
- **Asan Dataset**, a dermatology dataset collected from East Asian populations, characterized by darker skin tones and different class distributions.

Through cross-domain evaluations and fairness metrics such as Equalized Odds and Equalized Opportunity, we show that our method achieves balanced accuracy across demographic groups and outperforms conventional augmentation and re-weighting strategies on both performance and fairness grounds.

Rationale for Combined Inclusion. These two works are deliberately presented in a single chapter because they address orthogonal but equally significant sub-domain shifts - *frequency - based* and *demographic - based* using a common underlying principle: **context - aware sample**

mixing and learning adaptation. Both frameworks utilize MixUp not as a generic regularizer but as a guided augmentation strategy tuned to real-world distributional challenges. Moreover, the validation datasets span a diverse spectrum of imaging modalities (endoscopy, fundus photography, dermoscopy), anatomical regions (GI tract, retina, skin), and populations (Western and Asian), collectively demonstrating the broad applicability and clinical relevance of our proposed techniques. This chapter thus aligns with the central thesis goal: developing deep learning frameworks that are not only accurate under ideal conditions but resilient to the kinds of distributional and population-level shifts that characterize real-world clinical deployment. By advancing the robustness and fairness of deep medical image classifiers under long-tailed and demographically imbalanced data regimes, our work contributes to the development of **trustworthy and equitable AI systems for healthcare**.

5.2 Mo2E: Mixture of Two Experts for Class-Imbalanced Learning from Medical Images

Class imbalance in the medical image dataset is almost inherent due to the limited availability of clinical data for certain diseases and patient populations. Under-represented classes in the training set affect the classification task because the classifier tends to learn more from the majority classes, which are more common in the dataset and ignore data from the minority classes. To mitigate this issue, we propose a method to learn using two different convolutional neural network-based experts; such experts try to learn boundaries within the head classes, between the head and tail classes, and within the tail classes. During expert training, we integrate the MixUp regularization method to augment imbalanced data, employing distinct data sampling strategies for more effective mixing compared to random selection in traditional MixUp. During the inference phase, we combine the logits of the different experts based on their expertise in the corresponding classes. This way, we can improve the accuracy of the head and tail classes. The effectiveness of the suggested framework is demonstrated by experiments using highly imbalanced and long-tailed datasets.

5.2.1 Background: Medical Image Classification and Class Imbalance

Medical image datasets are inherently class-imbalanced. Deep convolutional neural networks (CNN) perform well on balanced datasets, but training with imbalanced image data affects test accuracy. The imbalance worsens when the available classes are more concentrated at the head than at the tail. Typically, real-life clinical datasets contain this distribution, called a "long-tailed" (LT) dataset. Due to the data disparity, over-represented classes (i.e., the majority) tend to lead training, which lowers the performance of under-represented classes (i.e., the minority). Thus, medical image classification on data with skewed class-wise distributions is challenging. Various techniques address imbalances, including cost-sensitive approaches, re-sampling, and data augmentation. Re-

weighting methods assign different weights to each class, enhancing the network's sensitivity to minority groups. Evolving from cost-sensitive learning (Japkowicz and Stephen, 2002b), advanced algorithms like Focal Loss (Lin et al., 2017) and LDAM (Cao et al., 2019) prioritize higher margins for minority classes. Class-balanced loss (Cui et al., 2019a) considers effective image numbers, capturing true class volumes. However, re-sampling-based methods are preferred to deal with long-tailed problems, focusing on obtaining more evenly distributed data. These include either under-sampling the majority class instances (Kubat et al., 1997) or over-sampling those of the minority class (Buda et al., 2018b). The main problem is that undersampling may lose valuable data, while random oversampling may cause overfitting. As a result, modified sampling techniques were introduced, like the adaptive semi-unsupervised weighted oversampling (Nekooeimehr and Lai-Yuen, 2016) and the SMOTE-based methods (Koto, 2014). Even though these resampling-based methods change the weights of the data classes, they do not expand the feature ranges of the training data. This makes the data less diverse. In addition, Zhou et al. proposed a dual-branched network, BBN(Zhou et al., 2020), to address LT-class imbalance. One branch is trained on the original data, while the other is trained on a reverse sampler that samples tail classes more frequently.

Moreover, when combined with re-sampling, data augmentation techniques like MixUp (Zhang et al., 2017a) can improve results in long-tailed visual recognition. These techniques, such as Balanced MixUp (Galdran et al., 2021) and Rebalanced MixUp (Remix) (Chou et al., 2020), leverage regularization. Balanced MixUp achieves a balanced distribution through regular (instance-based) and balanced (class-based) sampling. Remix assigns greater weight to minority class labels during data mixing, shifting classifier boundaries towards majority instances and reducing generalization errors. These MixUp techniques may not effectively address imbalanced data as they primarily boost overall performance by improving median class accuracy while having limited impact on tail classes and potentially harming head classes' accuracy. Thus, to solve this problem, our proposed approach improves the overall accuracy and the accuracies of the tail classes by using two different CNN-based experts. Each expert is trained using modified training data sampling strategies and MixUp regularization. Finally, these experts are combined at the time of the inference phase. The effectiveness of the suggested framework is demonstrated by experiments on various medical image classification tasks using highly imbalanced and LT-imbalanced data.

5.2.2 Methodology Proposed

5.2.2.1 Problem Definition and Preliminaries

Let us consider the C-class supervised classification problem with a training dataset $D = \{(x_k, y_k)\}_{k=1}^{N}$, where N is the number of samples, x_k is the k^{th} image datapoint and y_k is the corresponding category label, such that $y_k \in L = \{1, 2, 3, \ldots, C\}$. We consider the two CNN networks f_1 and f_2, as expert-1 and expert-2. The training dataset coupled with the specific sampling strategies can be

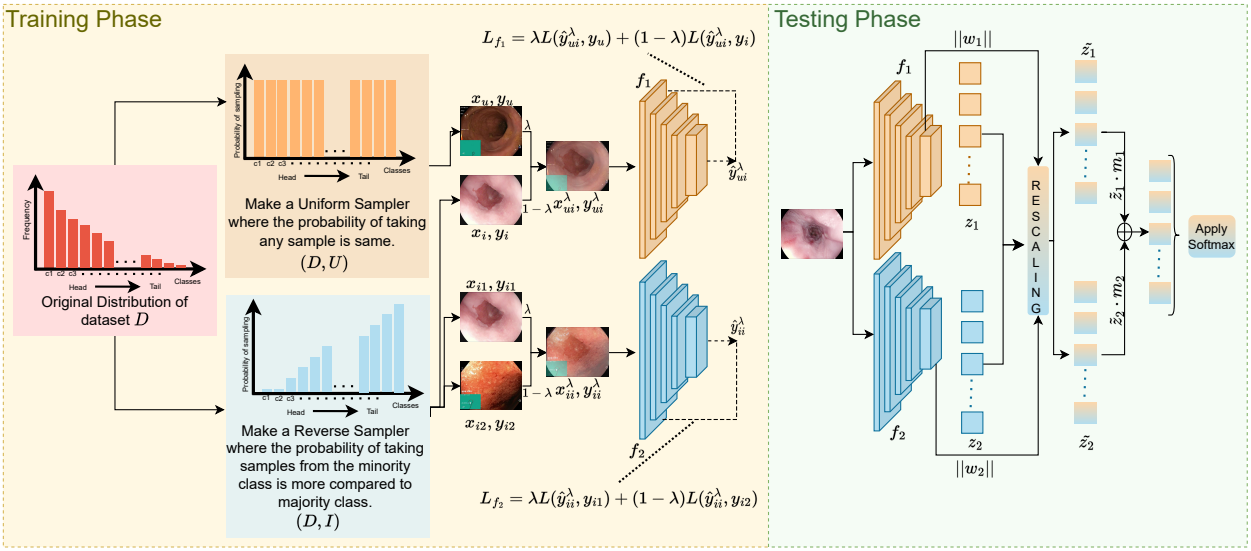


Figure 5.1: On the left, the training phase is shown, where the two experts are trained using specific sampling strategies and consequent MixUp. On the right, the testing phase is shown, which shows how experts are used during inference time.

represented as (D, U) (for the dataset with a uniform-balanced sampling strategy) and (D, I) (for the dataset with a reverse-balanced sampling strategy). In the uniform-balanced sampling strategy (D, U), samples are selected with equal probability. In the Reverse-balanced sampling strategy (D, I), samples are selected with a higher likelihood for the classes with fewer data samples than those with more data samples.

5.2.2.2 Proposed Method

We propose a multi-expert-based model that is independently optimized at the first step. The experts are trained so that they try to learn the boundaries between the head classes, head-tail classes, and tail classes. The experts were trained using the different MixUp techniques, where the MixUp is not applied normally but in a more specialized manner (see Figure 5.1) to learn boundaries in the dataset more effectively. Finally, during the inference phase, experts are combined by collecting scaled logits from the different experts corresponding to the classes for which they hold expertise in classification. This section will first define the experts used, followed by how these experts combine at inference time.

Uniform-Reverse Balanced MixUp Expert (f_1): This expert mainly focuses on learning the boundaries within the head classes and between the head and tail classes. While training this expert, two data points are sampled, one using (D, I) and the other using (D, U), rather than choosing them randomly without considering their category. The image-target pairs are generated

using the following expression for training using the MixUp method:

$$\left(x_{ui}^{\lambda}, y_{ui}^{\lambda}\right) = (\lambda x_u + (1-\lambda)x_i, \lambda y_u + (1-\lambda)y_i), \lambda \sim \beta(\alpha, \alpha), \tag{5.1}$$

where $(x_u, y_u) \sim (D, U)$ and $(x_i, y_i) \sim (D, I)$ and $\lambda \sim \beta(\alpha, \alpha)$, with $\alpha \in [0.1, 0.4]$. Now the neural network f_1 is trained using the generated datapoint $(x_{ui}^{\lambda}, y_{ui}^{\lambda})$.

Reverse-Reverse Balanced MixUp Expert (f_2): Contrary to the above expert, f_2 focuses on learning the boundaries between the tail classes. Here, both data points are sampled using (D, I). The image-target pairs are generated using the following expression for training using the MixUp method:

$$\left(x_{ii}^{\lambda}, y_{ii}^{\lambda}\right) = (\lambda x_{i1} + (1-\lambda)x_{i2}, \lambda y_{i1} + (1-\lambda)y_{i2}), \tag{5.2}$$

where $(x_{i1}, y_{i1}) \sim (D, I)$ and $(x_{i2}, y_{i2}) \sim (D, I)$, and $\lambda \sim \beta(\alpha, \alpha)$, with $\alpha \in [0.1, 0.4]$. Now the neural network f_2 is trained using the generated datapoint $(x_{ii}^{\lambda}, y_{ii}^{\lambda})$.

During the *testing phase*, the test image is passed through both the experts f_1, and f_2, and output logits $z_1 \in \mathbb{R}^{1\times C}$ and $z_2 \in \mathbb{R}^{1\times C}$ (before the softmax layer) of f_1, and f_2 are collected. They are further adjusted as $\tilde{z_1}$ and $\tilde{z_2}$ to have the comparable scales weighted using the norm of the weights of the fully connected layer, given by: $\tilde{z_1} = \frac{||w_1||^2}{||w_1||^2+||w_2||^2} \cdot z_1$, and $\tilde{z_2} = \frac{||w_2||^2}{||w_1||^2+||w_2||^2} \cdot z_2$, where $||w_1||$, $||w_2||$ are the norms of the weights of the fully conncted layer of f_1 and f_2 respectively. Now multiply $\tilde{z_1}$ and $\tilde{z_2}$ with masks m_1 and m_2, where $m_1, m_2 \in \{0,1\}^{1\times C}$. Here, m_1 signifies a vector containing ones at the places of head classes and m_2 containing ones at the places of tail classes, the rest entries being all zeros. The final logits are calculated as : $\tilde{z} = \tilde{z_1} \cdot m_1 + \tilde{z_2} \cdot m_2$. The softmax function is then applied to $\tilde{z}$ to obtain the confidence for the classification. The complete process for the training and testing phase is shown in Figure 5.1.

5.2.2.3 Proof of Concept

In examining the impact of diverse data sampling methods on neural network decision boundaries in MixUp, we employ a toy dataset (Figure 5.2(a)). The dataset comprises 4 classes with sample sizes of 500 (yellow), 50 (red), 25 (green), and 5 (blue), resulting in a class imbalance ratio of 100. In Figure 5.2(b), vanilla MixUp fails to correctly classify minority classes (red, blue, and green). Thus, the vanilla MixUp method is unable to generate better decision regions. The method struggles due to random sampling favoring the majority class, leading to decision regions predominantly influenced by the majority class and suppressing minority class representation. This results in the misclassification of minority data points. To tackle the imbalance in the data, the sampling strategy can be modified so that while drawing data points for mixing, one can be sampled from the uniform sampler, followed by the reverse-balanced sampler. As Figure 5.2(c) shows, this sampling strategy can generate better class boundaries between the majority and minority classes. However, it still fails to create accurate boundaries between the minority classes, which may lead to a minority

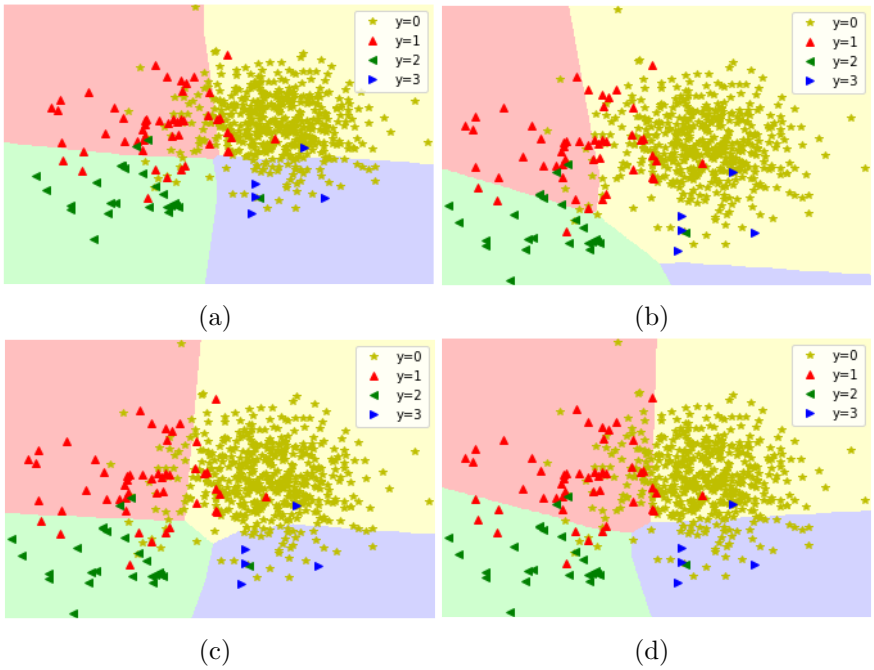
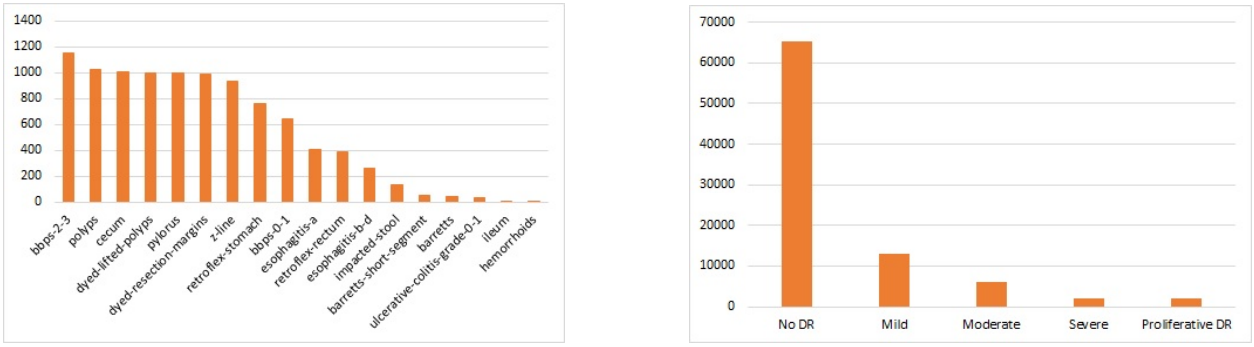


Figure 5.2: (a) Actual decision regions and dataset used. (b) Decision regions on the MixUp method. (c) Decision regions when we consider the MixUp method, where the data used for mixing are taken one from the uniform sampler and the other from the reverse sampler. (d) Decision regions when we consider the MixUp method, where both data points used for mixing are taken from the reverse sampler.

point being misclassified. On the other hand, Figure 5.2(c) indicates that the majority class decision region fills the region between the green and blue minority classes. Thus, to get better accuracy on minority classes, we use a reverse-balanced sampling strategy that enforces the increase in the probability of mixing from minority classes. This helps to create better class boundaries between the minority classes, as demonstrated in 5.2(d). Finally, we use the network trained using the uniform-reverse balanced MixUp and reverse-reverse balanced MixUp to get a better inference at test time.

5.2.2.4 Datasets, Training Details and Evaluation Metrics

We use the Hyper-Kvasir dataset (Borgli et al., 2020) for gastrointestinal (GI) image classification tasks. It presents a difficult classification problem because there are 23 classes with varying class frequencies and little representation of minority classes, as shown in Figure 5.3 (a). The last seven classes, starting with Esophagitis-B-D, are taken as tail classes; the remaining classes are considered head classes. The dataset comprises 10,662 images annotated for landmarks, pathological, and normal findings. Images are resized to 512 x 512. The dataset's official test split lacks some classes. To address this, we randomly split the training set into 4 (Stratified, 80:20 train-test ratio splits), reporting average performance across all splits. The next task we consider is the DR grading from



(a) Endotract dataset

(b) Eyepacs dataset

Figure 5.3: Class Distribution of Endotract and Eyepacs dataset

Table 5.1: Performance comparison of different methods on ResNext Model for GI image classification on the Hyper-Kvasir dataset for tail class, head class, and all classes.

Methods	Tail Class		Head Class		All Classes		
	BACC	F1	BACC	F1	BACC	F1	GM
CS*(Buda et al., 2018b)	4.57	5.86	87.09	86.67	61.97	62.07	78.58
RS*	9.86	11.82	86.13	84.89	62.92	62.65	79.15
SS*(Mahajan et al., 2018)	7.1	9.31	86.76	85.83	62.52	62.54	78.9
CB-RW(Cui et al., 2019a)	10.18	11.29	87.09	86.18	63.68	63.38	79.63
Focal(Lin et al., 2017)	10.65	10.52	86.16	85.01	63.18	62.34	79.31
BBN(Zhou et al., 2020)	8.7	10.36	87.27	86.13	63.35	63.07	79.43
BM*(α=0.3)	9.27	12.35	86.95	86.86	63.31	64.18	79.55
Mo2E(α=0.1)	**15.72**	**17.39**	85.3	85.43	**64.13**	**64.72**	**79.92**
Mo2E(α=0.2)	8.75	11.93	**87.75**	**87.16**	**64.11**	**64.56**	**79.93**
Mo2E(α=0.3)	**13.97**	**14.97**	86.02	86.45	**64.1**	**64.69**	**79.91**

*: CS: Class-balanced Sampling; RS:Reverse-balanced Sampling; SS:Square-root-Sampling; BM: Balanced Mixup

the retinal fundus images. For DR grading, we utilize Eyepacs dataset [3]. This dataset comprises 31,613 training examples and 55,000 official test set examples, with DR grades ranging from DR0 (no DR) to DR4 (proliferative DR). Additionally, 3,513 examples are allocated for validation and early halting during training.

In all the experiments, a CNN-based ResNeXt-50 (Kolesnikov et al., 2020) (initialized with weights pre-trained on the ImageNet dataset) is used for the classification. The loss function used is cross-entropy loss, and the optimizer used is stochastic gradient descent (SGD). During training, the batch size used is 8 with a learning rate of 0.01, and the learning rate decays over the period of training. In order to ensure a fair comparison, we train all comparative methods with the same setting (except for those methods whose codes are not given; their results have been directly

[3] https://www.kaggle.com/c/diabetic-retinopathy-detection

Table 5.2: Performance comparison of Mo2E and Balanced MixUp method for Eyepacs dataset

	$\alpha = 0.1$		$\alpha = 0.2$		$\alpha = 0.3$	
	k	MCC	k	MCC	k	MCC
Bal-Mix	73.95	51.35	77.00	56.95	80.49	63.11
Mo2E	**74.78**	**53.19**	**77.52**	**57.79**	**80.62**	**63.36**

Table 5.3: Comparison of quad- k performance with competing methods trained on the Eyepacs dataset and evaluated on the official Eyepacs test split

QWKL* PRL'18 (de La Torre et al., 2018)	**Cost-Sensitive*** MICCAI'20 (Galdran et al., 2020)	**DR-graduate*** MIA'20 (Araújo et al., 2020)	**Iterative Augmentation*** TMI'20 (González-Gonzalo et al., 2019)	**Mo2E** ($\alpha = 0.3$)
74.00	78.71	74.00	80	**80.62**

*Results of k are taken from the respective published papers.

taken from the published work, mainly for the Eyepacs dataset). The model's performance on the validation set is tracked using the given metric, and the best one is saved. The proposed work has been compared to other methods using F1-Score, Geometric Mean (GM), and Balanced Accuracy (BACC) for the GI classification task and using quadratic-weighted kappa score (k) along with MCC (Matthews Correlation Coefficient) for the DR-grading task.

5.2.2.5 Results and Discussion

The task of GI classification has an imbalance ratio of around 183 in the presence of 23 classes, which turns this into a very challenging dataset. Table 5.1 shows the average performance over four randomly created splits from the HyperKavasir training dataset. We compared our results with methods addressing imbalanced data, including sampling-based, cost-sensitive, mixup-based techniques, and a dual-branch network. The results have been found for the head, tail, and overall classes. As seen from Table 5.1, our method shows improvements in both F1-score and BACC compared to the other competing methods for the tail classes. In the head class, almost all the methods show comparable performance. Thus, our method tries to improve the performance of the tail class without much affecting the performance of the head class. In the performance comparison of all classes setup, our method outperforms almost all the other methods.

We used the Eyepacs dataset to compare the performance of the DR-Grading task. To test the performance, we use the already available official test split of the Eyepacs dataset. In this case, we consider the last two classes, as shown in Figure 5.3(b), severe and proliferative DR, as the tail classes and the rest as the head classes. In table 5.2, we compared our method with the balanced MixUp for different values of α, in terms of MCC and k, and for all α values, our method performed better. We also compared our method with other recently published techniques in terms of quad-k,

Table 5.4: Performance comparison of different MixUp-based methods on ResNext Model for GI image classification on the Hyper-Kvasir dataset for tail class, head class, and all classes. The best results are highlighted in bold.

Methods	α	Tail Class		Head Class		All Classes		
		BACC	F1	BACC	F1	BACC	F1	GM
MixUp	0.1	6.02	8.18	87.38	86.93	62.62	62.97	78.98
Bal-Mix	0.1	7.94	9.61	86.14	86.05	62.34	62.79	78.8
Mo2E	0.1	**15.72**	**17.39**	85.3	85.43	**64.13**	**64.72**	**79.92**
MixUp	0.2	5.06	6.51	87.19	86.82	62.2	62.38	78.72
Bal-Mix	0.2	5.35	6.96	87.4	86.98	62.42	62.63	78.87
Mo2E	0.2	**8.75**	**11.93**	**87.75**	**87.16**	**64.11**	**64.56**	**79.93**
MixUp	0.3	5.54	7.49	87.98	87.39	62.89	63.07	79.16
Bal-Mix	0.3	9.27	12.35	86.95	86.86	63.31	64.18	79.55
Mo2E	0.3	**13.97**	**14.97**	86.02	86.45	**64.1**	**64.69**	**79.91**

Table 5.5: Result of 4 Random splits (min, median, max) on the Hyper-Kvasir dataset. The best results are highlighted in bold.

Methods	All Classes								
	BACC			F1-Score			GM		
	Min	Median	Max	Min	Median	Max	Min	Median	Max
CS(Buda et al., 2018b)	60.34	63.13	65.06	59.46	63.07	64.99	77.51	79.29	80.5
RS	61.25	61.93	62.78	61.23	62.06	62.94	78.11	78.54	79.1
SS (Mahajan et al., 2018)	59.84	62.67	64.88	59.38	62.8	65.17	77.19	79.01	80.39
CB-RW (Cui et al., 2019a)	60.98	63.45	**66.82**	61.11	62.92	**66.57**	77.92	79.5	**81.59**
Focal (Lin et al., 2017)	61.42	63.17	64.95	60.52	62.11	64.6	78.2	79.3	80.42
BBN(Zhou et al., 2020)	61.8	63.19	65.21	61.11	63.19	64.78	78.45	79.34	80.59
BM(α=0.3)	60.97	63.59	65.08	61.03	64.69	66.31	78.42	79.7	80.38
Mo2E(α=0.3)	**62.31**	**64.48**	65.12	**62.66**	**65.22**	65.67	**78.78**	**80.15**	80.56

*: CS: Class-balanced Sampling; RS:Reverse-balanced Sampling; SS:Square-root-Sampling; BM: Balanced Mixup

as shown in table 5.3, demonstrating how our results were in line with recently published methods.

5.2.3 Conclusion and Discussions

Imbalanced data is pervasive in medical image classification tasks, necessitating the development of robust algorithms to address and alleviate its impact on accurate diagnosis. As a result, designing an effective and efficient algorithm for handling and mitigating the effects of class imbalance is crucial for accurate medical image diagnosis. This work presents a novel expert-based approach tailored for heavily imbalanced data distributions. Trained through a specialized adaptation of the MixUp regularization method, these experts are strategically combined during inference to enhance the performance of tail classes while minimally impacting the head classes.

Table 5.6: Comparative analysis of performance using k, MCC, F1-Score on Eyepacs dataset tested on official test split. The best results are highlighted in bold.

Methods	α	k	MCC	F1-Score
Reverse-sampling	-	72.19	48.28	55.87
Sqrt-sampling (Mahajan et al., 2018)	-	78.46	59.33	59.16
Class-sampling	-	80.42	62.65	59.88
Focal (Lin et al., 2017)	-	77.26	54.68	58.47
CB-RW (Cui et al., 2019a)	-	80.51	62.29	60.25
Bal-Mix	0.1	73.95	51.35	55.61
Mo2E	0.1	74.78	53.19	56.39
Bal-Mix	0.2	77.00	56.95	57.60
Mo2E	0.2	77.52	57.79	57.75
Bal-Mix	0.3	80.49	63.11	**60.53**
Mo2E	**0.3**	**80.62**	**63.36**	60.34

5.2.4 Extended GI image classification task and DR-Grading task results

In table 5.4, different alpha values are used to compare the GI image classification task results with those of the other MixUp-based algorithms. The small number of instances in the minority classes makes the results for the GI classification task noisy. In this work, we show the average over the 4 random splits. To show that not only the average results are the best, but also the minimum (min) and median are also best, except for maximum (max) results, which favor CB-RW as shown in table 5.5. Table 5.6 compares results for different sampling-based methods, cost-sensitive methods, and MixUp-based methods on the Eyepacs dataset, all using the ResNeXt-50 architecture.

5.3 Algorithmic Fairness in Lesion Classification by Mitigating Class Imbalance and Skin Tone Bias

Deep learning models have shown considerable promise in the classification of skin lesions. However, a notable challenge arises from their inherent bias towards dominant skin tones and the issue of imbalanced class representation. This study introduces a novel data augmentation technique designed to address these limitations. Our approach harnesses contextual information from the prevalent class to synthesize various samples representing minority classes. Using a mixup-based algorithm guided by an adaptive sampler, our method effectively tackles bias and class imbalance issues. The adaptive sampler dynamically adjusts sampling probabilities based on the network's meta-set performance, enhancing overall accuracy. Our research demonstrates the efficacy of this approach in mitigating skin tone bias and achieving robust lesion classification across a spectrum of diverse skin colors from two distinct benchmark datasets, offering promising implications for improving dermatological diagnostic systems.

5.3.1 Introduction

The field of skin lesion classification using deep learning models has shown great promise in achieving high performance in recent years. However, a significant challenge these models face is the presence of bias towards dominant skin tones and imbalanced class representation within datasets. This is because skin lesions and cancer occurs more on skin with less melanin, and thus is more prevalent among patients with pale skin, compared to darker skin. This has given rise to benchmark skin cancer image datasets that are biased towards samples from white patients, which is statistically proportionate, but causes training imbalance for machine learning algorithms. For a machine learning algorithm to be equitable and generalisable, it needs to perform robustly across demographics in a fair manner, because skin cancer can happen for darker skin, though lesser in number, but no patient can be treated as an outlier. Hence this is an open persisting problem within the area of machine learning fairness and health equity, which we address in this work.

Researchers have proposed various methods to handle imbalanced datasets. The methods mainly used for handling imbalances in the skin lesion dataset are widely classified into sampling-based, reweighing, and augmentation-based methods. The reweighing-based methods include RW (Reweighing) (Huang et al., 2016b), Focal-loss (Lin et al., 2017), CBRW (Class balanced reweighing) (Cui et al., 2019a), and Balanced Softmax (Ren et al., 2020a), are some of the state-of-the-art reweighing based methods used to handle imbalance. This genre of approach focuses on assigning different weights to each data point during training. While effective in some cases, reweighing techniques can be sensitive to the chosen weight function and might not always capture the true importance of each data point. Other methods include using modified sampling techniques, such as oversampling the under-represented categories. We can further divide this oversampling and undersampling technique into instance, class-balanced (Shen et al., 2016; Mahajan et al., 2018), reverse, and progressive-balanced sampling. Oversampling techniques, while seemingly intuitive, can introduce redundancy and lead to the model overfitting on the minority data. Undersampling, on the other hand, discards valuable information from the majority class, potentially affecting overall model performance. Another line of work focuses on augmentation-based techniques, such as Mixup (Zhang et al., 2017a), CutMix (Yun et al., 2019b), and Balanced Mixup (Galdran et al., 2021), aim to regularize training data by mixing instance and class-based sampling. Rebalanced Mixup (Remix) (Chou et al., 2020) gives greater weight to labels of minority classes, improving generalization. Another technique, CMO (Park et al., 2022), incorporates minority class images into majority class backgrounds. Methods like CutMix and CMO improve performance on tasks like CIFAR and ImageNet by replacing image parts with patches from others. However, in medical datasets, such as skin cancer detection, they often reduce accuracy by losing crucial diagnostic information. These methods also lack dynamic adaptation during mixing, limiting their effectiveness on unseen data.

Moreover, not many studies have been done on mitigating the racial skin-tone bias in lesion

classification. Groh et al. (Groh et al., 2021) study highlights a bias in skin tone classification using convolutional neural networks (CNNs). CNNs trained on datasets with limited skin tone diversity perform better on images with skin tones similar to those in the training data. This leads to lower accuracy for individuals with darker skin tones, which are often underrepresented in current datasets, thus giving rise a serious problem in machine learning fairness and health equity.

By critically evaluating strengths and limitations discussed above, our work develops a more robust approach for handling imbalanced datasets and inherent color-tone bias in skin lesion classification simultaneously, ultimately leading to more accurate and fair models for skin lesion diagnosis.

We propose a novel approach utilizing a mixup-based algorithm guided by an adaptive sampler. This method tackles these issues on two fronts:

1. *Mixup-based Algorithm*: We augment and create new training data points that blend features from existing samples by leveraging the mixup technique. This process encourages the model to learn more robust representations that are less susceptible to biases based on skin tone.
2. *Meta Adaptive Sampler*: We introduce an adaptive sampling strategy that dynamically adjusts the selection probabilities of training data during each iteration. This strategy prioritizes samples that pose more significant challenges to the network based on their performance on a dedicated meta-validation set. This targeted approach fosters more balanced learning and improves overall classification accuracy.

Our research demonstrates significant progress in mitigating skin tone bias within the model through this combined strategy. This translates to achieving robust lesion classification performance across a diverse range of skin colors on multiple datasets. These findings hold promising implications for developing more accurate dermatological diagnostic systems, that work across patient demographics with equity and fairness.

5.3.2 Methodology

In this section, we first present some concepts relevant to this work, and then introduce the proposed method.

5.3.2.1 Preliminaries

Sampling Strategy Sampling strategies of training data refer to techniques used to select and represent data instances in deep learning models, particularly when dealing with imbalanced datasets. These strategies aim to mitigate the underfitting of minority classes and prevent the overfitting of majority classes. Modified sampling strategies can include oversampling under-represented categories, leading to counter-productive outcomes like repeatedly showing the same training examples to the model. Given the provided training set notation, one can describe data sampling strategies

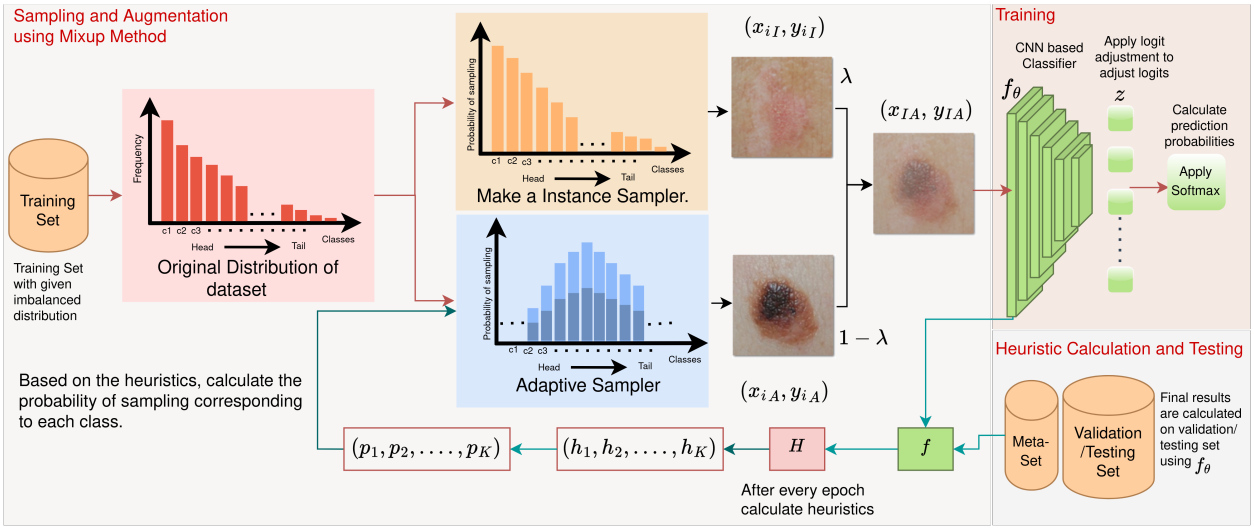


Figure 5.4: Our Proposed Framework

mathematically as: given a training set $D = \{(x_i, y_i), i = 1, ..., N\}$ for a multi-class problem with K classes where each class k contains n_k examples and $\sum_{k=1}^{K} n_k = N$, we can describe some common data sampling strategies mathematically with the probability p_k, associated with the particular class as $p_k = (\frac{(|n_k - \psi|)^\gamma}{\sum_{l=1}^{K}(|n_l - \psi|)^\gamma + \epsilon})$ (already explained in detail in chapter 3 and 4), where different values of γ and ψ, guides the different sampling strategies, where $\gamma \in [0, 1]$, and ψ, some statistic related with the examples in the dataset and ϵ added to avoid divide by zero error. $\gamma = 1$, and $\psi = 0$ forms the instance sampling strategy, followed by $\gamma = 1$, and $\psi = N$ forms reverse sampling strategy. However, such static strategies that fix the sampling probability at the start and use it throughout training are not feasible.

Mixup The mixup technique constitutes a data augmentation approach wherein synthetic training instances are created through the linear interpolation between pairs of authentic examples alongside their corresponding labels. The formula for mixup is:

$$\hat{x} = \lambda x_1 + (1 - \lambda) x_2 \tag{5.3}$$

$$\hat{y} = \lambda y_1 + (1 - \lambda) y_2, \tag{5.4}$$

where λ is a random variable sampled from a Beta distribution $B(\alpha, \alpha)$. One underlying concept behind mixup is that by employing linear interpolation among data points, we encourage the network to smoothly and seamlessly transition between data points, minimizing abrupt changes. Mixup improves deep neural network performance by enhancing robustness to adversarial attacks and promoting better generalization through data-adaptive regularization, ultimately leading to more accurate and reliable model predictions.

5.3.2.2 Proposed Method

Let us consider the K-class supervised classification problem with a training dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, where N is the total number of datapoints, x_i is the i^{th} image datapoint and y_i is the corresponding category label, such that $y_i \in L = \{1, 2, 3, \ldots, K\}$. Let f denote the classifier with parameter θ, which we need to train. The training set D is imbalanced with $n_1 > n_2 > \cdots > n_K$. Along with this we also have a meta set, $\mathcal{M} = \{(x_i, y_i)\}_{i=1}^{M}$, with total M datapoints. And also have heuristic function $\mathcal{H} : \mathcal{M} \rightarrow \{h_1, h_2, \cdots, h_K\}$, which gives heuristics corresponding to each class. The heuristic value increases in proportion to the degree of representation refinement achieved by the network or the level of accuracy attained by the class. The heuristic function dynamically evaluates model gradients.

This research proposes a novel imbalanced class learning approach that leverages a synergistic combination of instance sampling, adaptive sampling, and Mixup augmentation. Instance sampling ensures that all classes have a base representation during training. Adaptive sampling dynamically adjusts sampling probabilities based on a heuristic function utilizing a meta-set. This allows the model to focus on informative minority class examples. At the same time, the heuristic function, informed by the network's state and the meta-set, guides the sampling process to prevent overfitting the majority classes.

First, let us define the samplers to sample data points. The first sampler we will be using is the instance sampler where the probability of sampling from the Dataset D depends on the actual number of samples in the different classes, that is, the probability of choosing samples from the particular class k, $p_k = \frac{n_k}{N}$, where the Instance sampler denoted as $I(D, 1)$. This sampling approach will help to leverage the contexts present in the majority samples to enhance the limited context of the minority samples during the augmentation process.

The other sampler we will be using is the adaptive sampler $A(D, \mathcal{H})$, where the probability of sampling from the particular classes, depends on the heuristic calculate corresponding to each class using the meta-set $\mathcal{M}$ (the meta-set acts as a validation proxy for unseen attribute configurations, allowing the sampler to prioritize augmentations for subgroups where the model currently exhibits the highest error rates), using the heuristic function as $\mathcal{H}(\mathcal{M}, f_\theta) = \{h_1, h_2, \cdots, h_K\}$, the calculate heuristic depend on the state of the network f and the meta set $\mathcal{M}$. The probability corresponding to each class for sampling calculated as $p_k = 1 - \frac{h_k}{\sum_{i=1}^{K} h_i}$. This technique dynamically adjusts the sampling probability based on a 'heuristic' function. This function considers the current state of the learning model and a separate 'meta-set' of data. The adaptive sampler refines the sampling process based on the model's learning progress. It can potentially identify classes still challenging for the model to learn, even within the minority class, and prioritize those for further training. This can lead to more focused learning and faster improvement in difficult classes.

For our augmentation method, the data is first sampled from instance sampler, $(x_{i_I}, y_{i_I}) \sim$

$\mathcal{I}(\mathcal{D}, 1)$ and adaptive sampler, $(x_{i_A}, y_{i_A}) \sim \mathcal{A}(\mathcal{D}, f_\theta)$, and combined using mixup as follows:

$$\tilde{x}_{IA} = \lambda x_{i_I} + (1 - \lambda) x_{i_A}, \tag{5.5}$$

$$\tilde{y}_{IA} = \lambda y_{i_I} + (1 - \lambda) y_{i_A}. \tag{5.6}$$

This heuristic augmentation provides a better representation of the minority class by not neglecting classes that are not well-learned but instead striving to make the accuracy of such classes comparable to that of the majority classes using samples from the adaptive sampler. Furthermore, the data augmentation technique cleverly leverages the rich context found in abundant examples (majority samples) using an instance sampler to enhance the limited context surrounding the rarer examples (minority samples). By incorporating these additional details in the creation of new training data, we can significantly enhance our model's understanding of minority samples.

The Meta-set we are using does not necessarily need to belong to the same dataset as the training data. It might belong to a different dataset with varying skin tones, or it could be a combination of data similar to the training samples or samples with different skin tones. Thus, this data is not directly involved in the training samples used by the network for learning, but it will help to refine the heuristic. It will affect how the samples are chosen during augmentation, thus indirectly modifying the decision boundaries without exposing those samples to the network during training. The proposed framework is illustrated in Figure 5.4.

5.3.3 Experimental Analysis

5.3.3.1 Datasets, training protocol, comparison and evaluation metrics.

Dataset Used We have used the Asan Dataset (Han et al., 2018), mainly containing patients from the Asia with darker skin tone, and the ISIC-2018 dataset (Tschandl et al., 2018) containing mainly caucasian patients with pale skin tone. We divide the Asan Dataset training set into two parts: 10% of the images are used as a Meta-set, while the remaining 90% are used for training. The provided test-set images are used for testing. ISIC-2018 skin lesion classification challenge adopted the HAM10000 dataset (HAM) as a training dataset. The HAM dataset is one of the largest and most used skin image datasets publicly available in the ISIC archive. It consists of 10,015 skin lesion images in seven skin lesion types. The test set comprises 1512 skin lesion images without published labels. The only method for performance evaluation is to upload the predicted results to the ISIC website. So, we divided the training set into three splits in the ratio 70:10:20 (train:meta:test). For both datasets, we convert the images to a size of 100x100px. The Asan dataset has 12 classes, and the ISIC-2018 dataset has 7. The classes both data have in common are five, namely melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, and dermatofibroma.

Training Details We utilize the CNN ResNeXt-50 (as used in (Kolesnikov et al., 2020)) for image classification. We trained a network using stochastic gradient descent (SGD) with a batch size

of 128 for 100 epochs, starting with a learning rate (LR) of 0.01 that decayed if metrics didn't improve (via Reduce LR on plateau). We used early stopping if the LR reached 0, and set α to 0.2 for mixup ($B(\alpha, \alpha)$). We evaluated the performance of each model using task-specific metrics on a separate validation set, and we retained the best-performing model for further analysis. The code is available at: https://github.com/fa-submit/Submission_M.

Comparison Methods Not much work has addressed both the imbalanced problem and skin-tone bias together. To compare, we selected cost-sensitive methods: Reweighing-based methods like simple reweighing (RW) (weighting by the inverse of class frequencies), focal loss (FL), and class balanced reweighing (CBRW). We also considered resampling-based methods such as class balanced sampling (CBS), reverse sampling (RS), and progressive sampling (PS) - a combination of CBS and RS. Additionally, we examined Class balanced retraining, which adjusts sampling probabilities using the inverse frequency of each class raised to the power of $(1/8)^{\text{th}}$. We also evaluated balanced mixup (BalMixup) and mixup-based methods.

Evaluation Metrics We measure balanced accuracy (Bacc) (i.e. average recall of all classes). The macro-F1 score reflects improvements in smaller categories. We also include the geometric mean (GM) score. The heuristic used to determine sampling probabilities is calculated using per-class accuracy (i.e., for example, h_1 represents the accuracy of class 1). We assess model fairness using multi-class equalized odds (Eodd) and equalized opportunity (Eopp). Eopp0 evaluates the disparity in the True Negative Rate, Eopp1 evaluates the disparity in the True Positive Rate, and Eodd sums the disparities in the True Positive and False Positive Rates.

5.3.3.2 Results and Discussions

To demonstrate the effectiveness of our proposed model, we conducted several experiments. In the initial experiment, we aimed to address the existing imbalance in the dataset while enhancing performance. We accomplished this by training the ResNeXt-50 model using our proposed framework with the Asan dataset, comprising 12 classes. Subsequently, we compared its performance with other methods (as shown in Table 5.7) using the test set of the Asan dataset. Our findings indicate that our model exhibits superior performance compared to other methods. Additionally, when tested on the ISIC-2018 dataset, it demonstrates better performance compared to alternative methods. However, it is important to note that the ISIC-2018 dataset only shares 5 classes with the Asan dataset. To address this limitation, we utilized another subset of the Asan dataset containing the 5 common classes with the ISIC-2018 dataset. Furthermore, to showcase the generalizability of our trained model across various skin tones, we evaluated its performance on the entire ISIC dataset. Our method outperforms other methods, underscoring its superior generalizability. The lower value of fairness metrics in Table 5.8 highlight our method's predictive fairness compared to others, demonstrating its effectiveness. We demonstrate the performance of our main proposition regarding the use of the meta-set to regulate the sampling of samples in Table 5.9. We present

Table 5.7: The ResNeXt-50 network was trained on the Asan dataset and evaluated on two sets: one with all 12 classes and another with 5 classes common to the ISIC-2018 dataset. The evaluation metrics include F-1 score (in %), GM (in %), and Bacc (in %).

Method	12 Classes						5 Classes					
	Tested on the Asan Datasetset			Tested on the ISIC-2018 Dataset			Tested on the Asan Dataset Test Set			Tested on the ISIC-2018 Dataset		
	F1-score	GM	Bacc	F1-score	GM	Bacc	F1-score	GM	Bacc	F1-score	GM	Bacc
RW	59.18	74.78	57.89	8.92	25.06	26.71	79.75	86.3	78.48	28.71	58.86	40.12
FL	57.04	73.4	55.88	7.47	22.22	21.19	77.83	85.43	77.15	27.8	59.82	41.67
CBRW	58.62	74.72	57.88	8.04	23.63	23.75	79.84	86.45	78.72	33.37	59.35	40.95
CBS	58.13	74.11	56.94	7.96	23.95	24.52	**82.38**	**88.12**	**81.46**	22.67	55.93	36.79
RS	59.56	74.95	58.14	7.54	22.28	21.2	81.26	87.07	79.59	31.95	62.4	44.32
Cbrt	58.51	74.64	57.68	8.15	23.44	23.49	79.67	86.68	79.01	29.96	60.68	42.21
PS	56.91	73.7	56.33	7.81	25.33	27.36	81.1	87.61	80.54	26.67	56.81	37.62
BalMixup	60.04	75.91	59.62	9.77	26.64	30.18	81.10	87.25	79.58	34.25	60.91	42.63
Mixup	56.09	72.39	54.36	8.49	25.03	26.67	82.34	88.04	80.92	25.61	55.98	36.24
Ours	**61.05**	**77.54**	**62.16**	**10.36**	**28.97**	**35.5**	79.84	86.95	79.53	**40.57**	**65.33**	**47.81**

Table 5.8: Fairness results of different methods trained on the ASAN dataset (5 classes), tested on a combined ASAN test set and ISIC-2018 dataset.

Method(→) Metric(↓)	RW	FL	CBRW	CBS	RS	Cbrt	PS	BalMixup	Mixup	**Ours** (Meta-Set 10% Asan)	**Ours** (Meta-Set 10% Asan + 10% ISIC)
EOpp0	0.115	0.087	0.108	0.079	0.115	0.083	0.111	0.100	0.105	**0.075**	**0.055**
EOpp1	0.489	0.355	0.378	0.353	0.489	0.386	0.429	0.37	0.447	**0.317**	**0.336**
EOdds	0.298	0.185	0.19	0.186	0.298	0.200	0.241	0.164	0.185	**0.153**	**0.146**

Table 5.9: The table presents the ResNeXt-50 Network trained on the Asan Dataset, showcasing how results (in %) change with varying compositions of the Meta-Set.

Composition of Meta-Set	Tested on the Asan Data Test Set			Tested on the ISIC-2018 Dataset		
	F1-score	GM	Bacc	F1-score	GM	Bacc
10% of Asan Dataset	79.84	86.95	79.53	40.57	65.33	47.81
10% of ISIC Dataset	74.54	83.01	73.62	34.25	59.55	39.98
10% of ISIC Dataset+10% of Asan Dataset	**82.34**	**88.34**	**81.62**	**40.60**	**66.00**	**49.11**
20% of ISIC Dataset	73.31	82.69	72.81	33.37	59.91	41.37
30% of ISIC Dataset	74.81	83.87	75.17	34.95	61.98	43.58

Table 5.10: The results of the ResNeXt-50 network trained on the ISIC-2018 dataset are evaluated using F-1 score (in %), GM (in %), and Bacc (in %).

Method	Tested on the ISIC-2018 Dataset Test Set			Tested on the Asan Dataset		
	F1-score	GM	Bacc	F1-score	GM	Bacc
RW	**69.08**	81.61	71.53	**38.9**	55.97	35.98
FL	67.43	78.94	66.96	21.24	48.97	29.07
CBRW	69.27	81.94	71.6	31.17	52.18	32.53
CBS	62.48	74.13	59.79	22.15	50.48	30.67
RS	62.48	74.13	59.79	30.61	53.19	33.84
Cbrt	60.73	73.28	58.6	17.85	44.32	24.23
PS	62.61	74.51	60.1	25.88	51.17	31.63
BalMixup	65.89	75.90	62.77	14.95	41.97	21.85
Mixup	70.32	78.92	66.45	22.71	47.32	27.10
Ours	65.86	**83.6**	**74.79**	38.77	**57.4**	**38.77**

various compositions of the meta-set and the corresponding results on the Asan test set and the ISIC-2018 dataset. Since the images in the meta-set are not directly used for training, the entire ISIC dataset can be used for testing. From the table, it is evident that the optimal composition involves utilizing both the Asan and ISIC-2018 datasets in the meta-set, resulting in an improvement in performance of nearly 2% in terms of Bacc for both test cases. Additionally, there is an enhancement of approximately 3% in the F-1 score for the Asan dataset when compared with the meta-set containing only Asan dataset images, indicating an improvement in the performance of the minority class. These results show that performance can be enhanced by using a small number of images in the meta-set, even without directly including images of different skin tones in the training set. Regulating the sampling process aids in better augmentation, resulting in improved and more generalizable decision boundaries. We also computed results on the ISIC-2018 dataset using a subset of 5 classes common with the Asan dataset. We used the trained network to evaluate results on the Asan dataset and the ISIC-2018 test set (Table 5.10). Our method performs well in both test cases, with nearly a 3% improvement in both Bacc and GM, albeit with a slight dip in the F-1 score compared to RW and Mixup. Overall, our method aims for balanced performance across all classes, enhancing generalization to unseen data, thus showing better performance in unseen skin-tone images than other methods. Fairness metrics are excluded in this case because the ISIC test set's minority class has far fewer samples than the ASAN dataset's minority, making estimates for the underrepresented group potentially unreliable.

5.3.4 Conclusion and future work

In conclusion, this study presents an innovative data augmentation technique that addresses the challenges of skin tone bias and imbalanced class distribution in deep learning models for skin lesion classification simultaneously through a novel adaptive mixup sampling strategy that uses cross sampling between the diverse skin tones in a judicious manner. To demonstrate the efficacy of our method across different skin tones and class imbalance (skin cancer is more common among caucasian patients and hence usually data available is biased towards that patient demographic), we choose two benchmark datasets: ISIC-2018 with mostly caucasian patients and Asan dataset with mostly Asian patients. Our results showcase the accuracy of our method compared to several recent competing approaches, and thus presents a classifier that can generalise across patient demographics with fairness and equity, and hence has the potential of practical translation across borders in clinical decision support systems (CDSS).

Chapter 6

Conclusion & Future Directions

Synopsis

This concluding chapter summarizes the key contributions of the thesis and proposes both chapter-specific and overarching research directions for equitable deep learning under sub-population shift. In Section 6.1, we revisit the contributions, limitations, and immediate extensions of each chapter, focusing on targeted improvements in sampling, augmentation, boundary modeling, and fairness-aware optimization. In Section 6.2, we shift attention to broader research opportunities that arise from the thesis as a whole. These include developing theoretically grounded fairness objectives, designing adaptive and continual learning frameworks, and ensuring equity in emerging paradigms such as foundation models and multimodal systems. Together, these discussions outline a long-term vision for building robust, fair, and generalizable AI systems across imbalanced, long-tailed, and demographically diverse environments.

6.1 Contributions and Chapter-Level Outlook

This section briefly summarizes the main contributions and limitations of each chapter in this thesis. For each contributory chapter, we also propose immediate future research directions that emerge from its technical findings. These chapter-level perspectives are meant to guide incremental advances and highlight opportunities for architectural improvement, parameter tuning, or dataset-specific adaptations. These insights provide a foundation for the broader research directions and open challenges presented in Section 6.2. We start with the introductory Chapter 1 which introduces the main focus of the thesis: learning under subpopulation shift, which includes class imbalance, long-tailed distributions, and demographic bias. It explains why building robust and fair learning systems is important by highlighting real-world challenges. The chapter describes different types of subpopulation shift, such as class prior shift, attribute shift, and spurious correlations, and presents long-tailed learning as a key area of study. It also points out the limitations of existing methods and prepares the reader for the new solutions proposed in later chapters. In general, Chapter 1

provides the theoretical background and establishes the motivation for the thesis contributions. The opening chapter ends with an overview of the thesis structure and a brief summary of the key contributions presented in the subsequent chapters.

Chapter 2 addressed binary and multi-class imbalance by *distribution-calibrated oversampling*. We proposed a parametrized oversampling method that estimates the true feature distribution of a minority class by borrowing statistical moments (mean, covariance) from semantically "close" majority classes. By tuning hyperparameters, our method generates synthetic minority samples whose distribution closely approximates the original minority distribution. In experiments on synthetic and real datasets, this approach outperformed standard oversampling techniques (e.g. SMOTE variants) across multiple metrics, especially improving performance on highly skewed classes. The key technical advance is the *minority class statistic estimation* framework, which provides more principled synthetic data generation than prior convex-combination schemes. A limitation of our current approach is its reliance on fractional-norm distances, which poses challenges for high-dimensional feature spaces such as those encountered in image data. As a promising future direction, incorporating feature selection, manifold learning, or embedding-based neighborhood estimation could make the framework more broadly applicable to high-dimensional domains.

Chapter 3 presents **STTP-Net**, a Sampling-Tailored Two-Pronged Network designed to overcome challenges in long-tailed visual classification. The core idea is to divide learning responsibilities across two expert branches—one dedicated to head and medium classes and another specializing in underrepresented tail classes—while maintaining a shared feature extractor. The proposed **Hybrid-Mixup** strategy combines augmentation and sampling adaptations that improve data diversity and class representation, especially for rare classes. Furthermore, the chapter introduces **Effective Balanced Softmax (EBS)**, which recalibrates decision boundaries by accounting for class prior frequency during training. STTP-Net achieves consistent performance gains over strong baselines and prior state-of-the-art methods across multiple datasets, including real-world imbalanced datasets like NIH-CXR-LT. Importantly, its success in the healthcare domain emphasizes its translational potential beyond academic benchmarks. Despite its promising performance, STTP-Net currently relies on hand-crafted sampler configurations and static expert divisions. These fixed boundaries may limit adaptability to dynamically changing data distributions or evolving class hierarchies. Additionally, while more lightweight than multi-expert or two-stage methods, STTP-Net still introduces moderate architectural complexity, which may impact deployment in low-resource settings. Moving forward, future research can focus on integrating *adaptive expert assignment* based on online difficulty estimation or uncertainty modeling. Exploring *meta-learning strategies* for dynamic sampler tuning and *self-distillation mechanisms* for merging expert knowledge could further streamline the architecture. Finally, combining the STTP framework with fairness-aware learning objectives could unlock new capabilities in bias-sensitive domains such as medical imaging and autonomous systems.

Chapter 4 advances the long-tailed recognition paradigm introduced in Chapter 3 by focusing on a critical but underexplored aspect: the fidelity of decision boundaries across imbalanced classes. Traditional models often learn either overly hard boundaries that overfit rare classes or overly soft ones that blur distinctions between frequent and infrequent categories. In response, this chapter proposes a dual-branch framework rooted in the Goldilocks Principle—achieving decision boundaries that are "neither too sharp nor too smooth, but just right." This framework integrates **DBSGM**—a sampler-guided augmentation strategy using instance, median, and reverse samplers—with **ACFR**, a feature-space regularization method based on the **TASCL** loss. TASCL addresses the pitfalls of applying standard supervised contrastive learning to long-tailed data by dynamically modulating temperature parameters based on class and confidence, ensuring appropriate repulsion and attraction forces between features. Experimental evaluations on CIFAR-LT, ImageNet-LT, and iNaturalist demonstrate substantial gains across many-shot, medium-shot, and few-shot regimes. Despite its robustness, the proposed method relies on several tunable hyperparameters (e.g., γ, $\tau_{\min}$, $\tau_{\max}$) and heuristic decisions in sampler design. Furthermore, the contrastive loss's reliance on pairwise comparisons increases computational burden, especially on large-scale datasets like iNaturalist. Also, the framework assumes class labels are trustworthy; performance may degrade under label noise or weak supervision. Future work may explore *automated sampler tuning* using meta-learning or reinforcement learning to reduce reliance on fixed sampling strategies. Additionally, *label-agnostic contrastive objectives* and *noise-aware boundary regularization* could extend the applicability to weakly labeled or noisy datasets. Finally, bridging the decision boundary fidelity framework with fairness-aware learning objectives—particularly for sensitive domains like medical imaging—presents a compelling avenue to ensure equitable model performance across underrepresented subgroups.

Chapter 5 focuses on enhancing the fairness and robustness of medical image classifiers under dual sub-domain shifts: class imbalance and demographic bias. Building on earlier chapters, which addressed frequency-aware sampling (Chapter 3) and contrastive boundary shaping (Chapter 4), this chapter pivots to high-impact clinical applications where real-world biases pose significant challenges. Two key contributions are made. First, the **Mo2E framework** employs dual CNN-based experts trained with tailored MixUp strategies and diverse sampling mechanisms to explicitly model decision boundaries for both frequent and rare disease categories. This architecture significantly improves classification performance, especially for tail classes, across diverse clinical datasets like Hyper-Kvasir (GI lesions) and Eyepacs (Diabetic Retinopathy Grading) datasets. Second, the chapter introduces a **meta-adaptive mixup-based augmentation framework** that addresses racial and skin-tone bias in dermatological datasets. By using a meta-learned heuristic to control adaptive sampling probabilities during MixUp, the approach promotes equitable learning across demographic groups. Evaluations on the ISIC-2018 and Asan datasets demonstrate improved fairness metrics (e.g., Equalized Odds and Opportunity), while maintaining high classification accuracy.

Although Mo2E and the fairness-aware MixUp methods show competitive empirical performance with the existing methods but still large gap exists where the work can be done to improve the classification and fairness performance. Moreover, both frameworks rely on carefully tuned hyper-parameters (e.g., α in MixUp, sampling probabilities), and still require explicit specification of class domains (e.g., head vs. tail or demographic groupings). Moreover, while the meta-learning framework improves fairness, it assumes access to a demographically diverse meta-set, which may not always be available in real-world settings. Future work may explore fully *adaptive expert modeling* where expert assignments and sample weights are learned jointly in an end-to-end fashion. Another promising direction lies in unifying demographic and class-aware fairness objectives through joint optimization. Finally, expanding these frameworks to handle multi-modal data (e.g., radiology + EHR), and validating across diverse datasets will further enhance the generalizability and fairness of clinical AI systems.

6.2 Broader Future Research Directions and Open Challenges

Building on the chapter-wise contributions and localized future directions discussed in Section 6.1, this section outlines broader research directions. These open challenges aim to advance equitable learning under sub-population shift in both theory and practice. We organize these research opportunities around key themes such as adaptive optimization, representation learning under distribution shift, continual and federated learning, fairness in foundation and multimodal models, and causal and invariant modeling. Together, these directions reflect a long-term research agenda that extends the methodologies proposed in this thesis to more general, scalable, and socially responsible machine learning systems. Below we outline these key directions and challenges in detail:

- **Theoretical Foundations for Fairness under Shift.** A rigorous understanding of the accuracy–fairness tradeoff under sub-population shifts remains limited. Recent theory shows that enforcing fairness constraints does not always improve target-domain performance and may even be detrimental in some cases (Maity et al., 2021). It would be valuable to characterize when and how fairness-regularized training yields Bayes-optimal predictors under realistic shifts (Maity et al., 2021). Methods such as Uniform Risk Minimization (URM) have shown that training on uniformly distributed feature representations can provably improve worst-group performance (Krishnamachari et al., 2024). Extending such analysis to multi-class and continuous-group settings could guide the design of loss functions or sampling procedures that guarantee fairness without sacrificing overall accuracy. Developing analytical tools (e.g. generalization bounds under label-shifted and group-shifted distributions) would clarify the limits of fair learning in imbalanced regimes.
- **Adaptive Reweighting and Meta-Learning.** Importance-weighting techniques for sub-population shift (e.g. group re-weighting or DRO (Jung et al., 2023)) are promising but often rely

on heuristics. Recent work proposes jointly optimizing instance weights and model parameters via bi-level optimization (Holstege et al., 2025). This improves both average and worst-group accuracy in limited-data regimes by balancing bias–variance trade-offs (Holstege et al., 2025). Future work could incorporate such adaptive weighting into our frameworks: for example, learning the MixUp sampler probabilities or classifier biases in a data-driven way. Meta-learning approaches could also tune the hyperparameters of our dual-branch models (such as sampler temperatures or MixUp ratios) automatically, as sketched in Chapter 3's discussion. More broadly, combining meta-learning with contrastive objectives might adaptively emphasize underrepresented classes or groups, improving robustness.

- **Continual and Federated Long-Tail Learning.** Real-world data streams are non-stationary: class frequencies can evolve, new classes can emerge, and demographics may shift over time. As noted in recent surveys, handling dynamic imbalances in data streams is largely unexplored (Chen et al., 2024). Extending our methods to continual learning scenarios (e.g. class-incremental long-tailed learning) would be valuable. Techniques like elastic weight consolidation (Jarvis et al., 2024), and replay buffers (Rahimi-Kalahroudi et al., 2023) could be integrated with our mixup and contrastive schemes to maintain performance on old classes while learning new ones. Federated learning (FL) is another frontier: FL often suffers from heterogeneous and imbalanced client data (Zhang et al., 2023b). Future work could adapt sampling and loss balancing to the federated setting, ensuring both accuracy and fairness across clients. For example, client-specific mixup policies or inter-client contrastive regularization might mitigate client-level imbalance.
- **Self-Supervised and Semi-Supervised Learning for Imbalance.** Leveraging unlabeled data can help alleviate rare-class scarcity. Semi-supervised techniques that enforce consistency or pseudo-labeling could be augmented with subpopulation-aware strategies. For instance, one could use our sampler-guided mixup even for unlabeled samples (as pseudo-classes) to ensure that generated features remain fair across groups. Self-supervised pretraining on large, diverse datasets (including out-of-distribution samples) may also help learn representations that generalize better to minority classes (Joshi et al., 2025). However, recent work warns that large-scale foundation models may introduce new biases: e.g. medical imaging foundation models achieve high accuracy but have worse subgroup fairness compared to smaller models (Zong et al., 2023). Future research should therefore study how unsupervised pretraining and fine-tuning interact with fairness, possibly devising fairness-aware pretraining objectives or balanced fine-tuning schedules.
- **Fairness in Multimodal and Foundation Models.** As foundation models (FMs) become prevalent in vision and medical imaging, ensuring their equitable behavior is crucial. Very recent benchmarks (e.g. FairMedFM at NeurIPS 2024) show that diverse foundation models exhibit consistent biases and utility–fairness trade-offs across healthcare tasks (Jin et al., 2024). This suggests a pressing need for developing mitigation strategies tailored to large-scale pretrained models. Potential research includes designing in-processing debiasing layers or post-hoc cali-

bration techniques for FMs, as well as curating pretraining datasets that are demographically balanced. Another direction is to extend this thesis's mixup and contrastive ideas to multimodal settings (e.g. image+text diagnostics), where subpopulation shift might occur across modalities. Integrating structured fairness constraints into multimodal fusion could ensure that no subgroup is disproportionately disadvantaged by multi-view learning.

- **Causal and Invariant Methods.** Many subpopulation shifts arise from spurious correlations (e.g. skin color correlating with diagnosis). Future work could employ causal representation learning to disentangle true signal from confounders. For example, invariant risk minimization or contrastive causal objectives might learn features that are stable across group-shifted environments. Combining causal discovery with the adversarial or contrastive regularizers we developed could further improve worst-group robustness. Moreover, domain-adaptation and invariant-feature methods (e.g. CORAL (Sun et al., 2017), IRM (Arjovsky et al., 2019)) could be incorporated to align representations of underrepresented subpopulations without explicit labels.
- **Applicability to Other Modalities and Domains.** While this thesis evaluates the proposed methods on a diverse set of datasets spanning both medical and non-medical imaging domains, the core contributions are fundamentally learning-driven rather than modality-dependent. Specifically, the proposed bias-mitigation and class-imbalance handling strategies operate through (i) distribution-aware sampling mechanisms, (ii) feature-level regularization via supervised contrastive representations, and (iii) temperature-adaptive loss formulations that explicitly rebalance attraction and repulsion forces between majority and minority classes. These mechanisms do not rely on assumptions tied to a particular image acquisition process; instead, they leverage batch-wise statistics, embedding geometry, and class-conditional interactions, making them directly applicable to other imaging modalities. For instance, in MRI and histopathology, where pathological classes are often severely under-represented, the proposed temperature-adaptive loss can prevent minority class collapse by strengthening intra-class cohesion while moderating excessive repulsion from dominant classes. Similarly, in high-resolution microscopic or brain tissue imaging, where annotated samples are scarce and visually subtle, the contrastive feature learning framework enables more discriminative representations under limited supervision. Beyond medical imaging, the same principles naturally extend to non-medical applications such as facial recognition and person authentication, where demographic imbalance and few-shot enrollment are common. In such settings, the proposed methods can be employed to stabilize online learning with limited samples, reduce representation bias across groups, and maintain robust decision boundaries under skewed data distributions.
- **Enhanced Metrics and Evaluation.** Finally, progress requires careful evaluation. We have used balanced accuracy and fairness metrics (equalized odds, etc.), but future benchmarks should consider richer criteria, including calibration, robustness to outliers, and subgroup representa-

tional adequacy. The "Change is Hard" benchmark shows that optimizing worst-group accuracy often degrades overall metrics (Yang et al., 2023). New multi-objective evaluation protocols could guide development toward models that trade off fairness and accuracy in acceptable ways. Along these lines, automated tools that detect dataset shifts and trigger adaptive re-training (for example, alerting when minority classes fall below a performance threshold (Qin et al., 2024)) would be valuable in real deployments.

In conclusion, this thesis lays a foundation for equitable learning under sub-population shifts, but many challenges remain. Bridging theory and practice to guarantee fairness and robustness in dynamic, real-world settings is an open frontier. We anticipate that extending our approaches via meta-learning, robust optimization, causal modeling, and large-scale self-supervised pretraining will be fruitful. Addressing these challenges will advance the deployment of fair and generalizable AI across critical domains.

Appendix A

A.1 Supplementary for Chapter 3

In this appendix we will provide the proof of the theorem in chapter 3.

A.1.1 Proof of Theorem 1

We recall Theorem 1: *Let's consider* $\tilde{\Phi}$ *as the target conditional probability for a balanced dataset, defined as* $\tilde{\Phi}_j = \tilde{p}(y=j|x) = \frac{\tilde{p}(x|y=j)}{\tilde{p}(x)}\frac{1}{k}$. *Similarly, let's define* $\hat{\Phi}$ *as the target conditional probability for an imbalanced training set, given by* $\hat{\Phi}_j = \hat{p}(y=j|x) = \frac{\hat{p}(x|y=j)}{\hat{p}(x)}\frac{1-\beta^{n_j}}{\sum_{i=1}^{k}(1-\beta^{n_i})}$. *When expressing* $\tilde{\Phi}$ *using the standard Softmax function of the model's output* φ, *we can then represent* $\hat{\Phi}$ *as:*

$$\hat{\Phi}_j = \frac{\exp(\varphi_j)(1-\beta^{n_j})}{\sum_{i=1}^{k}\exp(\varphi_i)(1-\beta^{n_i})} \tag{A.1}$$

Proof. Note that the exponential family parameterization can be employed to express the conditional probability of a categorical distribution. This leads to a conventional Softmax function as the parameter-to-probability transformation:

$$\hat{\Phi}_j = \frac{\exp\{(n_j)\}}{\sum_{i=1}^{k}\exp\{(n_i)\}} \tag{A.2}$$

additionally, the canonical link function is expressed as:

$$\varphi_j = \log\left(\frac{\tilde{\Phi}_j}{\hat{\Phi}_k}\right). \tag{A.3}$$

We begin the derivation by adding $-\log\left(\frac{\tilde{\Phi}_j}{\hat{\Phi}_j}\right)$, to both the side of the Eqn. A.3:

$$\varphi_j - \log\left(\frac{\tilde{\Phi}_j}{\hat{\Phi}_j}\right) = \log\left(\frac{\tilde{\Phi}_j}{\hat{\Phi}_k}\right) - \log\left(\frac{\tilde{\Phi}_j}{\hat{\Phi}_j}\right) = \log\left(\frac{\tilde{\Phi}_j}{\hat{\Phi}_k}\right) \tag{A.4}$$

following that,

$$\tilde{\Phi}_k \exp\left\{(\varphi_j - \log\left(\frac{\tilde{\Phi}_j}{\hat{\Phi}_j}\right))\right\} = \hat{\Phi}_j \tag{A.5}$$

$$\tilde{\Phi}_k \sum_{i=1}^{k} \exp\left\{(\varphi_i - \log\left(\frac{\tilde{\Phi}_i}{\hat{\Phi}_i}\right))\right\} = \sum_{i=1}^{k} \hat{\Phi}_i = 1 \tag{A.6}$$

$$\tilde{\Phi}_k = \frac{1}{\sum_{i=1}^{k} \exp\left\{(\varphi_i - \log\left(\frac{\tilde{\Phi}_i}{\hat{\Phi}_i}\right))\right\}} \tag{A.7}$$

substitute Eqn. A.7 in Eqn. A.5

$$\hat{\Phi}_j = \frac{\exp\left\{(\varphi_j - \log\left(\frac{\tilde{\Phi}_j}{\hat{\Phi}_j}\right))\right\}}{\sum_{i=1}^{k} \exp\left\{(\varphi_i - \log\left(\frac{\tilde{\Phi}_i}{\hat{\Phi}_i}\right))\right\}} \tag{A.8}$$

as previously mentioned we have

$$\tilde{\Phi}_j = \tilde{p}(y = j|x) = \frac{\tilde{p}(x|y = j)}{\tilde{p}(x)} \frac{1}{k} \text{ and}$$
$$\hat{\Phi}_j = \hat{p}(y = j|x) = \frac{\hat{p}(x|y = j)}{\hat{p}(x)} \frac{1 - \beta^{n_j}}{\sum_{i=1}^{k} 1 - \beta^{n_i}}.$$

Using these equations we can write,

$$\frac{\tilde{\Phi}_j}{\hat{\Phi}_j} = \frac{\hat{p}(x)}{\tilde{p}(x)} \frac{\sum_{i=1}^{k}(1 - \beta^{n_i})}{k(1 - \beta^{n_j})} \tag{A.9}$$

since, $\hat{p}(x|y) = \tilde{p}(x|y) = p(x|y)$, then taking log on both side we get,

$$\log\left(\frac{\tilde{\Phi}_j}{\hat{\Phi}_j}\right) = \log\left(\frac{\hat{p}(x)}{\tilde{p}(x)}\right) + \log\left(\frac{\sum_{i=1}^{k}(1 - \beta^{n_i})}{k(1 - \beta^{n_j})}\right) \tag{A.10}$$

to avoid any confusion the term $\sum_{i=1}^{k}(1-\beta^{n_i})$, in the above equation can be re-written as: $\sum_{l=1}^{k}(1-\beta^{n_l})$, now equation A.10, can be re-written as:

$$\log\left(\frac{\tilde{\Phi}_j}{\hat{\Phi}_j}\right) = \log\left(\frac{\hat{p}(x)}{\tilde{p}(x)}\right) + \log\left(\frac{\sum_{l=1}^{k}(1 - \beta^{n_l})}{k(1 - \beta^{n_j})}\right) \tag{A.11}$$

put equation A.11 in equation A.8

$$\hat{\Phi}_j = \frac{\exp\left\{(\varphi_j - \log\left(\frac{\hat{p}(x)}{\tilde{p}(x)}\right) - \log\left(\frac{\sum_{l=1}^{k}(1-\beta^{n_l})}{k(1-\beta^{n_j})}\right))\right\}}{\sum_{i=1}^{k} \exp\left\{(\varphi_i - (\log\left(\frac{\hat{p}(x)}{\tilde{p}(x)}\right) + \log\left(\frac{\sum_{l=1}^{k}(1-\beta^{n_l})}{k(1-\beta^{n_i})}\right)))\right\}} \tag{A.12}$$

To simplify the above equation use logarithmic identities $\exp(a-b-c) = \exp(a)/(\exp(b)\exp(c))$ and $\exp(-\log A) = 1/A$, followed by the cancelation of class-independent terms. Expanding Eqn. A.12 gives:

$$\begin{aligned}\hat{\Phi}_j &= \frac{\exp(\varphi_j)\exp\Big(-\log\frac{\hat{p}(x)}{\tilde{p}(x)}\Big)\exp\Big(-\log\frac{\sum_l(1-\beta^{n_l})}{k(1-\beta^{n_j})}\Big)}{\sum_{i=1}^{k}\exp(\varphi_i)\exp\Big(-\log\frac{\hat{p}(x)}{\tilde{p}(x)}\Big)\exp\Big(-\log\frac{\sum_l(1-\beta^{n_l})}{k(1-\beta^{n_i})}\Big)} \\ &= \frac{\exp(\varphi_j)\frac{\tilde{p}(x)}{\hat{p}(x)}\frac{k(1-\beta^{n_j})}{\sum_l(1-\beta^{n_l})}}{\sum_{i=1}^{k}\exp(\varphi_i)\frac{\tilde{p}(x)}{\hat{p}(x)}\frac{k(1-\beta^{n_i})}{\sum_l(1-\beta^{n_l})}}.\end{aligned}$$

The common factors $\frac{\tilde{p}(x)}{\hat{p}(x)}$ and $\frac{k}{\sum_l(1-\beta^{n_l})}$ cancel out, yielding:

$$\hat{\Phi}_j = \frac{\exp\{(\varphi_j)\}(1-\beta^{n_j})}{\sum_{i=1}^{k}\exp\{(\varphi_i)(1-\beta^{n_i})\}}. \tag{A.13}$$

□

A.2 Supplementary for Chapter 4

A.2.1 Proof of Theorem 3

We recall Theorem 3: Consider a dataset comprising N normalized feature embeddings, $L = \{l_1, l_2, ..., l_N\} \subset \mathfrak{L}^N$, where $\mathfrak{L} = \{l \in \mathbb{R}^h : ||l||_2 = 1\}$ is the unit hypersphere in $\mathbb{R}^h$. Associated with each embedding l_i is a class label $y_i \in C$, where C is the set of all classes. Partition the label set into majority and minority subsets, denoted Y_M and Y_m respectively, such that $Y = Y_M \cup Y_m$. The aggregate batch-wise loss, $\mathcal{L}(L; Y, B, \tau)$, is defined as the summation of class-conditional batch-wise losses across the entire class space:

$$\begin{aligned}\mathcal{L}(L;Y,B,\tau) \geq \log\left(\prod_{y\in Y}|\overline{B_y}|^{|B_y|}\right) + \\ \sum_{y\in Y^M}\left(\sum_{i\in B_y}\left(\underbrace{\frac{\sum_{k\in B\setminus B_y}\frac{l_i\cdot l_k}{\tau_M}}{|\overline{B_y}|}}_{\text{Repulsion by Majority}} - \underbrace{\frac{\sum_{j\in B_y\setminus i}\frac{l_i\cdot l_j}{\tau_M}}{|B_y|-1}}_{\text{Attraction by Majority}}\right)\right) \\ + \sum_{y\in Y^m}\left(\sum_{i\in B_y}\left(\underbrace{\frac{\sum_{k\in B\setminus B_y}\frac{l_i\cdot l_k}{\tau_m}}{|\overline{B_y}|}}_{\text{Repulsion by Minority}} - \underbrace{\frac{\sum_{j\in B_y\setminus i}\frac{l_i\cdot l_j}{\tau_m}}{|B_y|-1}}_{\text{Attraction by Minority}}\right)\right),\end{aligned} \tag{A.14}$$

Proof. From the Lemma 2, we have the bound:

$$\mathcal{L}(L;Y,B,\tau,y) \geq \sum_{i\in B_y} \log\left((|B_y|-1)+|\overline{B_y}|exp(\frac{1}{|\overline{B_y}|}\sum_{k\in\overline{B_y}}\frac{l_i.l_k}{\tau}-\frac{1}{|B_y|-1}\sum_{j\in B_y\setminus i}\frac{l_i.l_j}{\tau})\right) \tag{A.15}$$

Convert this to loss over all classes in a batch as:

$$\sum_{y\in Y}\mathcal{L}(L;Y,B,\tau,y) \geq \sum_{y\in Y}\left(\sum_{i\in B_y}\log\left((|B_y|-1)+|\overline{B_y}|exp(\frac{1}{|\overline{B_y}|}\sum_{k\in\overline{B_y}}\frac{l_i.l_k}{\tau}-\frac{1}{|B_y|-1}\sum_{j\in B_y\setminus i}\frac{l_i.l_j}{\tau})\right)\right) \tag{A.16}$$

$$\mathcal{L}(L;Y,B,\tau) \geq \sum_{y\in Y}\left[\sum_{i\in B_y}\left[\log(|B_y|-1)+\log\left(1+\frac{|\overline{B_y}|}{|B_y|-1}exp\left(\frac{1}{|\overline{B_y}|}\sum_{k\in\overline{B_y}}\frac{l_i.l_k}{\tau}-\frac{1}{|B_y|-1}\sum_{j\in B_y\setminus i}\frac{l_i.l_j}{\tau}\right)\right)\right]\right] \tag{A.17}$$

considering the logarithmic bound $log(1+x) \geq log(x)$ for $x > 0$, we can write the above inequality as:

$$\mathcal{L}(L;Y,B,\tau) \geq \sum_{y\in Y}\left[\sum_{i\in B_y}\left[\log(|B_y|-1)+\log\left(\frac{|\overline{B_y}|}{|B_y|-1}\right)+\left(\frac{1}{|\overline{B_y}|}\sum_{k\in\overline{B_y}}\frac{l_i.l_k}{\tau}-\frac{1}{|B_y|-1}\sum_{j\in B_y\setminus i}\frac{l_i.l_j}{\tau}\right)\right]\right] \tag{A.18}$$

$$\geq \sum_{y\in Y}\left[\sum_{i\in B_y}\left[\log\left(|\overline{B_y}|\right)+\left(\frac{1}{|\overline{B_y}|}\sum_{k\in\overline{B_y}}\frac{l_i.l_k}{\tau}-\frac{1}{|B_y|-1}\sum_{j\in B_y\setminus i}\frac{l_i.l_j}{\tau}\right)\right]\right] \tag{A.19}$$

$$\geq \sum_{y\in Y}\left[\sum_{i\in B_y}\log\left(|\overline{B_y}|\right)\right]+\sum_{y\in Y}\left[\sum_{i\in B_y}\left(\frac{\sum_{k\in B\setminus B_y}\frac{l_i.l_k}{\tau}}{|B|-|B_y|}-\frac{\sum_{j\in B_y\setminus i}\frac{l_i.l_j}{\tau}}{|B_y|-1}\right)\right] \tag{A.20}$$

further by splitting the term into majority (Y^M) and minority classes (y^m) such that $Y = Y^M \cup y^m$,

$$\mathcal{L}(L;Y,B,\tau) \geq \log\left(\prod_{y\in Y}\prod_{i\in B_y}|\overline{B_y}|\right)+\sum_{y\in Y^M}\left(\sum_{i\in B_y}\left(\frac{\sum_{k\in B\setminus B_y}\frac{l_i.l_k}{\tau}}{|\overline{B_y}|}-\frac{\sum_{j\in B_y\setminus i}\frac{l_i.l_j}{\tau}}{|B_y|-1}\right)\right)+$$
$$\sum_{y\in Y^m}\left(\sum_{i\in B_y}\left(\frac{\sum_{k\in B\setminus B_y}\frac{l_i.l_k}{\tau}}{|\overline{B_y}|}-\frac{\sum_{j\in B_y\setminus i}\frac{l_i.l_j}{\tau}}{|B_y|-1}\right)\right) \tag{A.21}$$

$$\mathcal{L}(L;Y,B,\tau) \geq \log\left(\prod_{y\in Y}|\overline{B_y}|^{|B_y|}\right) + \sum_{y\in Y^M}\left(\sum_{i\in B_y}\left(\frac{\sum_{k\in B\setminus B_y}\frac{l_i.l_k}{\tau}}{|\overline{B_y}|} - \frac{\sum_{j\in B_y\setminus i}\frac{l_i.l_j}{\tau}}{|B_y|-1}\right)\right)+ \sum_{y\in Y^m}\left(\sum_{i\in B_y}\left(\frac{\sum_{k\in B\setminus B_y}\frac{l_i.l_k}{\tau}}{|\overline{B_y}|} - \frac{\sum_{j\in B_y\setminus i}\frac{l_i.l_j}{\tau}}{|B_y|-1}\right)\right). \quad (A.22)$$

In the above equation, the value of the minority class repulsion term becomes very small due to division by the higher cardinality of the set and also because the number of samples in the minority class is relatively low compared to the majority, which creates repulsion. One way to balance repulsion from the majority and minority classes is to decrease the repulsion term value of the majority and increase the repulsion value of the minority. This creates well-balanced repulsion forces, which can be accomplished by adjusting the temperature value to divide the dot product. Therefore, the final inequality is as follows:

$$\mathcal{L}(L;Y,B,\tau) \geq \log\left(\prod_{y\in Y}|\overline{B_y}|^{|B_y|}\right) + \sum_{y\in Y^M}\left(\sum_{i\in B_y}\left(\frac{\sum_{k\in B\setminus B_y}\frac{l_i.l_k}{\tau_M}}{|\overline{B_y}|} - \frac{\sum_{j\in B_y\setminus i}\frac{l_i.l_j}{\tau_M}}{|B_y|-1}\right)\right)+ \sum_{y\in Y^m}\left(\sum_{i\in B_y}\left(\frac{\sum_{k\in B\setminus B_y}\frac{l_i.l_k}{\tau_m}}{|\overline{B_y}|} - \frac{\sum_{j\in B_y\setminus i}\frac{l_i.l_j}{\tau_m}}{|B_y|-1}\right)\right). \quad (A.23)$$

Also we can see that these changes not only balance the repulsion forces but also tighten the bounds of the loss equation, particularly when the data is imbalanced.

□

References

Aayushman, H. Gaddey, V. Mittal, M. Chawla, and G. R. Gupta. Fair and accurate skin disease image classification by alignment with clinical labels. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 394–404. Springer, 2024. 21

M. M. Ahsan, S. A. Luna, and Z. Siddique. Machine-learning-based disease diagnosis: A comprehensive review. In *Healthcare*, volume 10, page 541. MDPI, 2022. 1

E. S. Aimar, A. Jonnarth, M. Felsberg, and M. Kuhlmann. Balanced product of calibrated experts for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19967–19977, 2023a. 44

E. S. Aimar, A. Jonnarth, M. Felsberg, and M. Kuhlmann. Balanced product of calibrated experts for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19967–19977, 2023b. 16, 17, 84, 102, 104, 107

N. Alipour, T. Burke, and J. Courtney. Skin type diversity in skin lesion datasets: A review. *Current Dermatology Reports*, 13(3):198–210, 2024. 19

F. An and J. Wang. Pedestrian re-identification algorithm based on multivariate manifold metric-anti-noise manifold space learning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 7(1):261–270, 2023. doi: 10.1109/TETCI.2022.3220259. 79

S. Ando and C. Huang. Deep over-sampling framework for classifying imbalanced data. *CoRR*, abs/1704.07515, 2017. URL http://arxiv.org/abs/1704.07515. 46

F. Ansari, S. Das, and P. Shamsolmoali. Handling class imbalance by estimating minority class statistics. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2023. 9, 26

F. Ansari, A. Bhattacharya, B. Saha, and S. Das. Mo2e: Mixture of two experts for class-imbalanced learning from medical images. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024a. 110, 111

F. Ansari, T. Chakraborti, and S. Das. Algorithmic fairness in lesion classification by mitigating class imbalance and skin tone bias. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 373–382. Springer Nature Switzerland Cham, 2024b. 21, 110, 112

F. Ansari, A. Panigrahi, and S. Das. The goldilocks principle: Achieving just right boundary fidelity for long-tailed classification. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2025. 17, 78

T. Araújo, G. Aresta, L. Mendonça, S. Penas, C. Maia, Â. Carneiro, A. M. Mendonça, and A. Campilho. Dr— graduate: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. *Medical Image Analysis*, 63:101715, 2020. 119

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 136

S. Barua, M. M. Islam, X. Yao, and K. Murase. Mwmote–majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on knowledge and data engineering*, 26(2):405–425, 2012. 9, 10, 29, 35

R. Berk and A. A. Elzarka. Almost politically acceptable criminal justice risk assessment. *Criminology and Public Policy*, 19(4):1231–1257, 2020. 1

P. J. Bevan and A. Atapour-Abarghouei. Detecting melanoma fairly: Skin tone detection and debiasing for skin lesion classification. In *MICCAI Workshop on Domain Adaptation and Representation Transfer*, pages 1–11. Springer, 2022. 20

M. S. M. Bhuiyan, M. A. Rafi, G. N. Rodrigues, M. N. H. Mir, A. Ishraq, M. Mridha, and J. Shin. Deep learning for algorithmic trading: A systematic review of predictive models and optimization strategies. *Array*, 26:100390, 2025. ISSN 2590-0056. doi: https://doi.org/10.1016/j.array.2025.100390. URL https://www.sciencedirect.com/science/article/pii/S2590005625000177. 1

H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen, et al. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data*, 7(1):283, 2020. ix, 13, 117

S. Boughorbel, F. Jarray, and M. El-Anbari. Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PloS one*, 12(6):e0177678, 2017. 35

P. Branco, L. Torgo, and R. P. Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM computing surveys (CSUR)*, 49(2):1–50, 2016. 27

G. Bredell et al. Explicitly minimizing the blur error of variational autoencoders. *arXiv:2304.05939*, 2023. 11

M. Buda, A. Maki, and M. A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.*, 106:249–259, 2018a. 43, 45, 46

M. Buda, A. Maki, and M. A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018b. 111, 114, 118, 120

C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In T. Theeramunkong, B. Kijsirikul, N. Cercone, and T.-B. Ho, editors, *Advances in Knowledge Discovery and Data Mining*, pages 475–482, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-01307-2. 46

C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap. Dbsmote: density-based synthetic minority over-sampling technique. *Applied Intelligence*, 36(3):664–684, 2012. 9, 10

J. Cai, Y. Wang, and J.-N. Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 112–121, 2021. 44

K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 43, 47, 65, 66, 68, 69, 72, 100, 102, 104, 107, 111, 114

A. Chaudhary, H. P. Gupta, and K. K. Shukla. Real-time activities of daily living recognition under long-tailed class distribution. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(4):740–750, 2022. doi: 10.1109/TETCI.2022.3150757. 79

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, jun 2002a. doi: 10.1613/jair.953. URL https://doi.org/10.1613\\%2Fjair.953. 8, 10, 43, 46, 79, 82

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002b. 29, 34

J. Chen, L. Ding, Y. Yang, and Y. Xiang. Active diversification of head-class features in bilateral-expert models for enhanced tail-class optimization in long-tailed classification. *Engineering Applications of Artificial Intelligence*, 126:106982, 2023. 16, 17, 44, 50, 67, 69, 84, 101, 104

S. Chen, G. Guo, and L. Chen. A new over-sampling method based on cluster ensembles. In *2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops*, pages 599–604, 2010. doi: 10.1109/WAINA.2010.40. 9, 10

W. Chen, K. Yang, Z. Yu, Y. Shi, and C. P. Chen. A survey on imbalanced learning: latest research, applications and future directions. *Artificial Intelligence Review*, 57(6):137, 2024. 135

H.-P. Chou, S.-C. Chang, J.-Y. Pan, W. Wei, and D.-C. Juan. Remix: Rebalanced mixup. In *Computer Vision – ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI*, page 95–110, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-65413-9. doi: 10.1007/978-3-030-65414-6_9. 15, 17, 44, 48, 83, 111, 114, 122

D. Cieslak, N. Chawla, and A. Striegel. Combating imbalance in network intrusion datasets. In *2006 IEEE International Conference on Granular Computing*, pages 732–737, 2006. doi: 10.1109/GRC.2006.1635905. 9, 10

D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002. doi: 10.1109/34.1000236. 9, 10

J. Cui, S. Liu, Z. Tian, Z. Zhong, and J. Jia. Reslt: Residual learning for long-tailed recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3695–3706, 2022. xii, xiii, 16, 17, 44, 49, 67, 68, 69, 71, 72, 73, 74, 79, 83, 101, 102, 104, 105, 106, 107

Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019a. 111, 114, 118, 120, 121, 122

Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019b. 15, 17, 43, 47, 61, 66, 68, 69, 72, 79, 82, 100, 102, 104

D. Dablain, B. Krawczyk, and N. V. Chawla. Deepsmote: Fusing deep learning and smote for imbalanced data, 2021. 11, 12

D. Dablain, B. Krawczyk, and N. V. Chawla. Deepsmote: Fusing deep learning and smote for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 43, 46

R. Daneshjou, K. Vodrahalli, R. A. Novoa, M. Jenkins, W. Liang, V. Rotemberg, J. Ko, S. M. Swetter, E. E. Bailey, O. Gevaert, et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances*, 8(31):eabq6147, 2022. 19

S. Das, S. Datta, and B. B. Chaudhuri. Handling data irregularities in classification: Foundations, trends, and future challenges. *Pattern Recognition*, 81:674–693, 2018. 26

S. Das, S. S. Mullick, and I. Zelinka. On supervised class-imbalanced learning: An updated perspective and some key challenges. *IEEE Transactions on Artificial Intelligence*, 3(6):973–993, 2022. 27

J. de la Calleja and O. Fuentes. A distance-based over-sampling method for learning from imbalanced data sets. In *FLAIRS conference*, pages 634–635, 01 2007. 9, 10

J. de La Torre, D. Puig, and A. Valls. Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognition Letters*, 105:144–154, 2018. 119

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. 66, 100

J. Derrac, S. Garcia, L. Sanchez, and F. Herrera. Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *J. Mult. Valued Logic Soft Comput*, 17, 2015. 34

P. Domingos. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 155–164, 1999. 28

Y. Dong, T. Gong, H. Chen, S. Song, W. Zhang, and C. Li. How does distribution matching help domain generalization: An information-theoretic analysis. *IEEE Transactions on Information Theory*, 2025. 19

S. Du, B. Hers, N. Bayasi, G. Hamarneh, and R. Garbi. Fairdisco: Fairer ai in dermatology via disentanglement contrastive learning. In *European Conference on Computer Vision*, pages 185–202. Springer, 2022. 20

E. Dugas, Jared, Jorge, and W. Cukierski. Diabetic retinopathy detection. https://kaggle.com/competitions/diabetic-retinopathy-detection, 2015. Kaggle Competition. ix, 13

C. Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001. 28

D. Elreedy, A. F. Atiya, and F. Kamalov. A theoretical distribution analysis of synthetic minority oversampling technique (smote) for imbalanced learning. *Machine Learning*, pages 1–21, 2023. 29

K. Emory, F. Douma, and J. Cao. Autonomous vehicle policies with equity implications: Patterns and gaps. *Transportation Research Interdisciplinary Perspectives*, 13:100521, 2022. 1

M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996. 9, 10

S. Fan, X. Zhang, Z. Song, and W. Shao. Cumulative dual-branch network framework for long-tailed multi-class classification. *Engineering Applications of Artificial Intelligence*, 114:105080, 2022. 16, 17, 44, 49, 67, 69, 71, 84, 101, 104, 105

A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863–905, 2018. 29

K. R. M. Fernando and C. P. Tsokos. Dynamically weighted balanced loss: Class imbalanced learning and confidence calibration of deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2940–2951, 2022. doi: 10.1109/TNNLS.2020.3047335. 80

A. Galdran, J. Dolz, H. Chakor, H. Lombaert, and I. Ben Ayed. Cost-sensitive regularization for diabetic retinopathy grading from eye fundus images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, pages 665–674. Springer, 2020. 119

A. Galdran, G. Carneiro, and M. A. González Ballester. Balanced-mixup for highly imbalanced medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 323–333. Springer, 2021. 111, 114, 122

J. Gauthier. Conditional generative adversarial nets for convolutional face generation. *Class project for Stanford CS231N: convolutional neural networks for visual recognition, Winter semester*, 2014 (5):2, 2014. 11, 12

A. Ghadiri, M. Pagnucco, and Y. Song. Xtranprune: explainability-aware transformer pruning for bias mitigation in dermatological disease classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 749–758. Springer, 2024. 20

A. S. Ghorab, W. M. Ashour, and S. I. Abudalfa. An adaptive oversampling method for imbalanced datasets based on mean-shift and smote. In *Explore Business, Technology Opportunities and Challenges After the Covid-19 Pandemic*, pages 13–23. Springer, 2022. 9, 10

C. González-Gonzalo, B. Liefers, B. van Ginneken, and C. I. Sánchez. Iterative augmentation of visual evidence for weakly-supervised lesion localization in deep interpretability frameworks. *arXiv preprint arXiv:1910.07373*, 2019. 119

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014. 10, 12

F. Graf, C. Hofer, M. Niethammer, and R. Kwitt. Dissecting supervised contrastive learning. In *International Conference on Machine Learning*, pages 3821–3830. PMLR, 2021. 93, 94

M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, A. Koochek, and O. 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*, pages 1820–1828, 2021. 20, 112, 123

T. Guo, X. Zhu, Y. Wang, and F. Chen. Discriminative sample generation for deep imbalanced learning. In *Twenty-Eighth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 2019. 11, 12

H. Han, W.-Y. Wang, and B.-H. Mao. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In D.-S. Huang, X.-P. Zhang, and G.-B. Huang, editors, *Advances in Intelligent Computing*, pages 878–887, Berlin, Heidelberg, 2005a. Springer Berlin Heidelberg. ISBN 978-3-540-31902-3. 8, 10, 46

H. Han, W.-Y. Wang, and B.-H. Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005b. 29, 35

S. S. Han, M. S. Kim, W. Lim, G. H. Park, I. Park, and S. E. Chang. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *Journal of Investigative Dermatology*, 138(7):1529–1538, 2018. ix, 18, 126

H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.*, 21(9): 1263–1284, 2009. 15, 17, 43, 45, 79, 82

H. He, Y. Bai, E. A. Garcia, and S. Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE, 2008. 9, 10, 29, 35

G. Holste, S. Wang, Z. Jiang, T. C. Shen, G. Shih, R. M. Summers, Y. Peng, and Z. Wang. Long-tailed classification of thorax diseases on chest x-ray: A new benchmark study. In *MICCAI Workshop on Data Augmentation, Labelling, and Imperfections*, pages 22–32. Springer, 2022. 66, 71

F. Holstege, B. Wouters, N. V. Giersbergen, and C. Diks. Optimizing importance weighting in the presence of sub-population shifts. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=j4gzziSUr0. 135

Y. Hong, S. Han, K. Choi, S. Seo, B. Kim, and B. Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6626–6636, 2021. 51, 67, 68, 69, 101, 102, 104, 105, 107

Z. hong, J. Cui, S. Liu, and J. Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16489–16498, 2021. 16, 17, 44, 51, 67, 68, 69, 73, 79, 85, 101, 102, 104, 105, 107

G. V. Horn, O. M. Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie. The inaturalist species classification and detection dataset, 2018. 100

C. Huang, Y. Li, C. C. Loy, and X. Tang. Learning deep representation for imbalanced classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016a. 15, 79, 82

C. Huang, Y. Li, C. C. Loy, and X. Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016b. 43, 47, 122

A. Iscen, A. Araujo, B. Gong, and C. Schmid. Class-balanced distillation for long-tailed visual recognition. In *Proc. BMVC'21*, 2021. 67, 69, 71, 101, 105

N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study1. *Intell. Data Anal.*, 6(5):429–449, 2002a. 15, 17, 43, 45, 79, 82

N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002b. 111, 114

D. Jarvis, G. N. Tasse, S. Lee, S. Zwane, R. Klein, B. Rosman, and A. M. Saxe. Full elastic weight consolidation via the surrogate hessian-vector product, 2024. URL https://openreview.net/forum?id=IyRQDOPjD5. 135

R. Jin, Z. Xu, Y. Zhong, Q. Yao, Q. Dou, S. K. Zhou, and X. Li. Fairmedfm: Fairness benchmarking for medical imaging foundation models. *arXiv preprint arXiv:2407.00983*, 2024. 135

Y. Jin, M. Li, Y. Lu, Y.-m. Cheung, and H. Wang. Long-tailed visual recognition via self-heterogeneous integration with knowledge excavation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23695–23704, 2023. 44

J. M. Johnson and T. M. Khoshgoftaar. Survey on deep learning with class imbalance. *J. Big Data*, 6(1), 2019. 45

S. Joshi, J. Ni, and B. Mirzasoleiman. Dataset distillation via knowledge distillation: Towards efficient self-supervised pre-training of deep networks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=c61unr33XA. 135

S. Jung, T. Park, S. Chun, and T. Moon. Re-weighting based group fairness regularization via classwise robust optimization. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Q-WfHzmiG9m. 134

N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel. Large language models struggle to learn long-tail knowledge, 2023. URL https://arxiv.org/abs/2211.08411. 16, 17, 85

B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020. xiii, 16, 17, 44, 50, 66, 67, 70, 71, 72, 73, 74, 79, 83, 84, 100, 101, 105, 106, 107

P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33: 18661–18673, 2020. 92

A. Kim and I. Jung. Optimal selection of resampling methods for imbalanced data with high complexity. *Plos one*, 18(7):e0288540, 2023. 80

B. Kim and J. Kim. Adjusting decision boundary for class imbalanced learning. *IEEE Access*, 8: 81674–81685, 2020. doi: 10.1109/ACCESS.2020.2991231. 80

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 10, 12

N. M. Kinyanjui, T. Odonga, C. Cintas, N. C. Codella, R. Panda, P. Sattigeri, and K. R. Varshney. Estimating skin tone and effects on classification performance in dermatology datasets. *arXiv preprint arXiv:1910.13268*, 2019. 19

A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby. Big transfer (bit): General visual representation learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 491–507, 2020. 118, 126

F. Koto. Smote-out, smote-cosine, and selected-smote: An enhancement strategy to handle imbalance in data level. In *2014 international conference on advanced computer science and information system*, pages 280–284. IEEE, 2014. 111, 114

J. Kozerawski, V. Fragoso, N. Karianakis, G. Mittal, M. Turk, and M. Chen. BLT: Balancing long-tailed datasets with adversarially-perturbed images. In *Computer Vision – ACCV 2020*, pages 338–355. Springer International Publishing, Cham, 2021. 47

K. Krishnamachari, S.-K. Ng, and C.-S. Foo. Uniformly distributed feature representations for fair and robust learning. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=PgLbS5yp8n. 134

A. Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL https://api.semanticscholar.org/CorpusID:18268744. 65, 100

M. Kubat, S. Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *Icml*, volume 97, page 179. Citeseer, 1997. 28, 111, 114

J.-H. Lee, C. Lee, and C.-S. Kim. Learning multiple pixelwise tasks based on loss scale balancing. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5087–5096, 2021. doi: 10.1109/ICCV48922.2021.00506. 80

D.-C. Li, S.-Y. Wang, K.-C. Huang, and T.-I. Tsai. Learning class-imbalanced data with region-impurity synthetic minority oversampling technique. *Information Sciences*, 607:1391–1407, 2022a. 29

J. Li, Z. Tan, J. Wan, Z. Lei, and G. Guo. Nested collaborative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6949–6958, 2022b. 44

M. Li, Y.-m. Cheung, and Y. Lu. Long-tailed visual recognition via gaussian clouded logit adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6929–6938, June 2022c. 15, 17, 79, 82, 101, 102, 104, 105, 107

M. Li, Y.-m. Cheung, and Y. Lu. Long-tailed visual recognition via gaussian clouded logit adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6929–6938, June 2022d. 67, 68, 69, 73

S. Li, K. Gong, C. H. Liu, Y. Wang, F. Qiao, and X. Cheng. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5212–5221, 2021. 15, 17, 44, 48, 67, 68, 69, 70, 71, 79, 83, 101, 102, 104, 105

T. Li, P. Cao, Y. Yuan, L. Fan, Y. Yang, R. S. Feris, P. Indyk, and D. Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6918–6928, 2022e. 16, 17, 50, 67, 68, 69, 84, 101, 102, 104, 105, 107

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 111, 114, 118, 120, 121, 122

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection, 2018. 48, 71

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. doi: 10.1109/TPAMI.2018.2858826. 105

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. ix, 2

Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019a. 100, 101, 105, 107

Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019b. 43, 66, 67, 70, 71

L. Ma and S. Fan. Cure-smote algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. *BMC bioinformatics*, 18(1):1–18, 2017. 9, 10

D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018. 118, 120, 121, 122

S. Maity, D. Mukherjee, M. Yurochkin, and Y. Sun. Does enforcing fairness mitigate biases caused by subpopulation shift? In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=6mUrD5rg-UU. 134

G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, and C. Malossi. Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655*, 2018. 11, 12

A. K. Mondal, L. Singhal, P. Tiwary, P. Singla, and P. AP. Minority oversampling for imbalanced data via class-preserving regularized auto-encoders. In F. Ruiz, J. Dy, and J.-W. van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 3440–3465. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/mondal23a.html. 11

P. Mondal, F. Ansari, and S. Das. Cco: A cluster core-based oversampling technique for improved class-imbalanced learning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, pages 1–13, 2024a. doi: 10.1109/TETCI.2024.3407784. 15, 82

P. Mondal, F. Ansari, and S. Das. Cco: A cluster core-based oversampling technique for improved class-imbalanced learning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024b. 46

S. S. Mullick, S. Datta, and S. Das. Generative adversarial minority oversampling. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1695–1704, 2019. 11, 12

S. S. Mullick, S. Datta, and S. Das. Generative adversarial minority oversampling, 2020. 43, 46

I. Nekooeimehr and S. K. Lai-Yuen. Adaptive semi-unsupervised weighted oversampling (a-suwo) for imbalanced datasets. *Expert Systems with Applications*, 46:405–416, 2016. 111, 114

H. M. Nguyen, E. W. Cooper, and K. Kamei. Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 3(1): 4–21, 2011. 9, 10, 29, 35

C. N. Nwafor, O. Nwafor, and S. Brahma. Enhancing transparency and fairness in automated credit decisions: an explainable novel hybrid machine learning approach. *Scientific Reports*, 14 (1):25174, 2024. 1

J. Oh and C. Yun. Provable benefit of mixup for finding optimal decision boundaries. In *International Conference on Machine Learning*, pages 26403–26450. PMLR, 2023. 92

S. Park, J. Lim, Y. Jeon, and J. Y. Choi. Influence-balanced loss for imbalanced visual classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 735–744, October 2021. 67, 68, 69, 72, 101, 102, 104, 107

S. Park, Y. Hong, B. Heo, S. Yun, and J. Y. Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1234–1243. IEEE, 2022. 15, 44, 48, 52, 54, 67, 68, 69, 70, 71, 73, 79, 83, 101, 102, 104, 105, 107, 122

P. Qin, S. Li, X. Liu, Z. Zheng, and S. Y. Chong. Threshold moving for online class imbalance learning with dynamic evolutionary cost vector. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=EIPnUofed9. 137

A. Rahimi-Kalahroudi, J. Rajendran, I. Momennejad, H. van Seijen, and S. Chandar. Replay buffer with local forgetting for adaptive deep model-based reinforcement learning, 2023. URL https://openreview.net/forum?id=uWpq1-rQbV. 135

J. Ren, C. Yu, X. Ma, H. Zhao, S. Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020a. 48, 67, 68, 69, 71, 122

J. Ren, C. Yu, S. Sheng, X. Ma, H. Zhao, S. Yi, and H. Li. Balanced meta-softmax for long-tailed visual recognition. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020b. Curran Associates Inc. ISBN 9781713829546. 15, 17, 79, 82, 101, 102, 104, 105, 107

M. Shao, D. Li, C. Zhao, X. Wu, Y. Lin, and Q. Tian. Supervised algorithmic fairness in distribution shifts: A survey. *arXiv preprint arXiv:2402.01327*, 2024. 19

L. Shen, Z. Lin, and Q. Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 467–482. Springer, 2016. 122

J.-X. Shi, T. Wei, Y. Xiang, and Y.-F. Li. How re-sampling helps for long-tail learning? In *Advances in Neural Information Processing Systems*, volume 36, pages 75669–75687. Curran Associates, Inc., 2023. 15, 17, 83

B. Sun, J. Feng, and K. Saenko. Correlation alignment for unsupervised domain adaptation. *Domain adaptation in computer vision applications*, pages 153–171, 2017. 136

A. Sánchez, E. Morales, and J. Gonzalez. Synthetic oversampling of instances using clustering. *International Journal of Artificial Intelligence Tools*, 22, 01 2013. doi: 10.1142/S0218213013500085. 9, 10

K. Tang, J. Huang, and H. Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in Neural Information Processing Systems*, 33:1513–1524, 2020a. 67, 68, 69, 70, 71, 73, 101, 105

K. Tang, J. Huang, and H. Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*, 2020b. 102, 104

P. Tschandl, C. Rosendahl, and H. Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. ix, 18, 20, 126

J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano. Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th international conference on Machine learning*, New York, NY, USA, 2007a. ACM. 15, 17, 79, 82

J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano. Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th international conference on Machine learning*, pages 935–942, 2007b. 43, 45, 46

J. Vanschoren, J. N. Van Rijn, B. Bischl, and L. Torgo. Openml: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014. 34

B. Wang, P. Wang, W. Xu, X. Wang, Y. Zhang, K. Wang, and Y. Wang. Kill two birds with one stone: Rethinking data augmentation for deep long-tailed learning. In *International Conference on Learning Representations*, 2024. URL https://api.semanticscholar.org/CorpusID:271691573. 16, 17, 83, 104, 105, 107

J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. S. Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 2022a. 18, 19

P. Wang, K. Han, X.-S. Wei, L. Zhang, and L. Wang. Contrastive learning based hybrid networks for long-tailed image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 943–952, 2021a. 16, 17, 50, 67, 68, 69, 84, 101, 102, 104, 107

T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020. 95

X. Wang, Y. Lyu, and L. Jing. Deep generative model for robust imbalance classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14124–14133, 2020. 11, 12

X. Wang, L. Lian, Z. Miao, Z. Liu, and S. Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021b. 16, 17, 44, 49, 67, 69, 70, 73, 84, 101, 104, 107

X. Wang, L. Jing, Y. Lyu, M. Guo, J. Wang, H. Liu, J. Yu, and T. Zeng. Deep generative mixture model for robust imbalance classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2897–2912, 2022b. 11, 12

X. Wang, L. Lian, Z. Miao, Z. Liu, and S. X. Yu. Long-tailed recognition by routing diverse distribution-aware experts, 2022c. 79

Y. Wang, X. Pan, S. Song, H. Zhang, G. Huang, and C. Wu. Implicit semantic data augmentation for deep networks. *Advances in Neural Information Processing Systems*, 32, 2019a. 15, 17, 83

Y. Wang, X. Pan, S. Song, H. Zhang, G. Huang, and C. Wu. Implicit semantic data augmentation for deep networks. *Advances in Neural Information Processing Systems*, 32, 2019b. 48

Y.-X. Wang, D. Ramanan, and M. Hebert. Learning to model the tail. *Advances in neural information processing systems*, 30:7032–7042, 2017. 15, 43, 47, 79, 82

S. Weisberg. Yeo-johnson power transformations. *Department of Applied Statistics, University of Minnesota. Retrieved June*, 1:2003, 2001. 31

L. Xiang, G. Ding, and J. Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 247–263. Springer, 2020. 44

G. Xu, Y. Duan, Z. Liu, X. Li, M. Jiang, M. Lemmon, W. Jin, and Y. Shi. Incorporating rather than eliminating: Achieving fairness for skin disease diagnosis through group-specific expert. *arXiv preprint arXiv:2506.17787*, 2025. 20, 21

W. Xu, P. Wang, Z. Zhao, B. Wang, X. Wang, and Y. Wang. When imbalance meets imbalance: Structure-driven learning for imbalanced graph classification. In *The Web Conference 2024*, 2024a. URL https://openreview.net/forum?id=zyWwZrItIH. 16, 17, 85

Z. Xu, D. Shen, T. Nie, Y. Kou, N. Yin, and X. Han. A cluster-based oversampling algorithm combining smote and k-means for imbalanced medical data. *Information Sciences*, 572:574–589, 2021. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2021.02.056. 9, 10

Z. Xu, S. Zhao, Q. Quan, Q. Yao, and S. K. Zhou. Fairadabn: Mitigating unfairness with adaptive batch normalization and its application to dermatological disease classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 307–317, Cham, 2023. Springer Nature Switzerland. 21

Z. Xu, J. Li, Q. Yao, H. Li, M. Zhao, and S. K. Zhou. Addressing fairness issues in deep learning-based medical image analysis: a systematic review. *npj Digital Medicine*, 7(1):286, 2024b. 20

S. Yang, L. Liu, and M. Xu. Free lunch for few-shot learning: Distribution calibration. *arXiv preprint arXiv:2101.06395*, 2021. 27, 30

W. Yang, C. Pan, and Y. Zhang. An oversampling method for imbalanced data based on spatial distribution of minority samples sd-kmsmote. *Scientific Reports*, 12(1):16820, 2022. 29

Y. Yang and Z. Xu. Rethinking the value of labels for improving class-imbalanced learning. *Advances in neural information processing systems*, 33:19290–19301, 2020. 67, 71, 73, 101, 105

Y. Yang, R. Khanna, Y. Yu, A. Gholami, K. Keutzer, J. E. Gonzalez, K. Ramchandran, and M. W. Mahoney. Boundary thickness and robustness in learning models. *Advances in Neural Information Processing Systems*, 33:6223–6234, 2020. 86, 92

Y. Yang, H. Zhang, D. Katabi, and M. Ghassemi. Change is hard: A closer look at subpopulation shift. In *International Conference on Machine Learning*, 2023. 2, 3, 4, 14, 137

Y. Yang, H. Zhang, J. W. Gichoya, D. Katabi, and M. Ghassemi. The limits of fair medical imaging ai in real-world generalization. *Nature Medicine*, 30(10):2838–2848, 2024. 20

S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6022–6031, 2019a. doi: 10.1109/ICCV.2019.00612. 15, 17, 44, 48, 52, 54, 67, 68, 69, 73, 79, 83, 101, 102, 104

S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019b. 122

C. Zhang, G. Almpanidis, G. Fan, B. Deng, Y. Zhang, J. Liu, A. Kamel, P. Soda, and J. Gama. A systematic review on long-tailed learning. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2025. doi: 10.1109/TNNLS.2025.3539314. 13

H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017a. 44, 48, 111, 114, 122

H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *CoRR*, abs/1710.09412, 2017b. 15, 17, 51, 52, 67, 68, 69, 73, 79, 83, 85, 101, 102, 104

H. Zhang, L. Zhu, X. Wang, and Y. Yang. Divide and retain: A dual-phase modeling for long-tailed visual recognition. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12, 2023a. doi: 10.1109/TNNLS.2023.3269907. 44, 67, 69, 71, 73, 101, 104, 105

J. Zhang, L. Liu, P. Wang, and C. Shen. To balance or not to balance: A simple-yet-effective approach for learning with long-tailed distributions. *arXiv preprint arXiv:1912.04486*, 2019. 67, 71, 73, 101, 105

J. Zhang, A. Li, M. Tang, J. Sun, X. Chen, F. Zhang, C. Chen, Y. Chen, and H. Li. Fed-CBS: A heterogeneity-aware client sampling mechanism for federated learning via class-imbalance reduction. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41354–41381. PMLR, 23–29 Jul 2023b. URL https://proceedings.mlr.press/v202/zhang23y.html. 135

S. Zhang, Z. Li, S. Yan, X. He, and J. Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2361–2370, 2021. 67, 70, 71, 73, 101, 105, 107

Z. Zhao, P. Wang, H. Wen, W. Xu, S. Lai, Q. Zhang, and Y. Wang. Two fists, one heart: Multi-objective optimization based strategy fusion for long-tailed learning. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=MEZydkOr3l. 16, 17, 85, 104, 105, 107

B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, pages 1–8, 2020. xiii, 16, 17, 44, 49, 50, 67, 68, 69, 70, 73, 79, 83, 84, 101, 102, 104, 107, 114, 118, 120

J. Zhu, Z. Wang, J. Chen, Y.-P. P. Chen, and Y.-G. Jiang. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6908–6917, 2022. 44, 98, 101

Y. Zong, Y. Yang, and T. Hospedales. Medfair: Benchmarking fairness for medical imaging. In *International Conference on Learning Representations (ICLR)*, 2023. 135