

Politecnico di Torino

Politecnico di Torino

Corso di Laurea Magistrale in Ingegneria Informatica (Computer Engineering)

Orientamento in Artificial Intelligence and Data Analytics

A.a. 2025/2026

Sessione di Laurea Settembre 2026

Curriculum Learning Methodologies for Deep Learning in Medical Image Analysis

Relatrici:

Eliana Pastor

Eleonora Poeta

Candidato:

Matteo Di Gregorio

Abstract

Deep learning-based segmentation models can exhibit inconsistent performance across patient subgroups and acquisition settings, limiting their reliability when applied to data that differ from the training distribution. Standard training paradigms, such as Empirical Risk Minimization (ERM), primarily optimize overall predictive performance and do not explicitly account for differences in difficulty or performance between subpopulations. As a result, performance disparities may persist for less represented or intrinsically challenging groups. Alternative training strategies, such as Curriculum Learning (CL), have been proposed to improve model optimization by organizing training samples according to their difficulty and progressively exposing the model to more challenging examples. Despite the substantial progress enabled by CL, conventional approaches typically define difficulty using model-derived quantities, such as training loss, which may lack clinical interpretability. Consequently, the resulting difficulty levels are not necessarily associated with clinically meaningful patient characteristics or acquisition-related factors that systematically influence segmentation performance. This thesis investigates different training methodologies for medical image segmentation, using standard ERM training and loss-based curriculum learning as baselines, and introduces a novel approach: Preview-based Divergence Curriculum Learning (Preview Div-CL). Preview Div-CL is a metadata-aware framework that defines sample difficulty based on divergence in segmentation performance, quantified as the difference between subgroup performance and the average performance across clinically or acquisition-relevant subgroups, and combines this formulation with a preview-based training strategy. Rather than introducing difficulty-based buckets sequentially, as in conventional CL, Preview Div-CL exposes the model to samples from all buckets from the beginning of training while progressively adjusting their contribution through a mathematical weighting function based on difficulty. To evaluate the proposed approach, we conducted experiments on two heterogeneous multi-institutional segmentation benchmarks (UCSD-PTGBM and MAMA-MIA) using multiple architectures. Our results demonstrate that Preview Div-CL significantly reduces computational cost in terms of the number of epochs, compared to both loss-based CL and standard ERM training, while maintaining comparable or even improved Dice score performance. These results suggest that Preview Div-CL can be a computationally efficient alternative to existing strategies, without compromising segmentation performance. In addition, its divergence-based construction provides the curriculum with clinical meaning, making the paradigm interpretable.

Acknowledgements

Computational resources provided by HPC@PoliTO (www.hpc.polito.it)

Table of Contents

List of Figures IV

1 Introduction 1
- 1.1 Context and motivation . . . 1
- 1.2 Contributions . . . 2
- 1.3 Outline of the Thesis . . . 3
- 1.4 Research questions . . . 4

2 Related Works 5
- 2.1 Background . . . 5
  - 2.1.1 Explainability and Interpretability . . . 5
  - 2.1.2 Preview Strategy Curriculum Learning . . . 6
- 2.2 Curriculum Learning . . . 6
- 2.3 Curriculum Learning in Medical . . . 8
- 2.4 Annealing and Ramp . . . 9
- 2.5 Mitigation . . . 10
- 2.6 Subgroup-Aware Training . . . 12

3 Methodology 14
- 3.1 Problem setup . . . 14
- 3.2 Subgroup divergence estimation . . . 15
  - 3.2.1 DivExplorer . . . 15
- 3.3 Curriculum Learning construction . . . 16
- 3.4 Preview strategy implementation . . . 17

4 Experimental Analysis 21
- 4.1 Experimental setup . . . 21
  - 4.1.1 Datasets . . . 21
  - 4.1.2 Pre-Processing . . . 22
  - 4.1.3 Backbone architectures . . . 22
  - 4.1.4 Metrics . . . 24

4.2 Final method: Preview Div-CL . . . 25
4.3 Comparison with baseline and alternatives . . . 25
4.4 Subgroup-level benefits: Comparison with random generated . . . 28
4.5 Loss curves . . . 33
4.6 Ablation and Sensitivity analysis . . . 37
4.6.1 Sensitivity analysis of ramp values . . . 37
4.6.2 U-Net and SegResNet comparison . . . 40
4.6.3 Preview strategy: Single vs. Three-ramp approach . . . 43
4.6.4 Preview strategy: Static vs. Dynamic ramps . . . 44
4.7 Alternative approaches and exploratory surveys . . . 46
4.7.1 Adding other metrics in DivExplorer . . . 46
4.7.2 Variant of preview strategies . . . 49
4.8 Discussion . . . 51
4.8.1 Results interpretation . . . 52
4.8.2 Comparison with related works . . . 55

5 Conclusion 56
5.1 Summary and main results . . . 56
5.2 Practical applications . . . 57
5.3 Limitation . . . 57
5.4 Future works . . . 58
5.4.1 Subgroup extraction . . . 58
5.4.2 Revisiting Curriculum difficulty assignment . . . 58
5.4.3 Testing domain-shift on Preview Curriculum Learning . . . 58
5.4.4 Extending Preview Curriculum Learning to other tasks . . . 59
5.4.5 Dynamic partitioning in Curriculum scheduling . . . 59
5.4.6 Pruning redundant subgroups . . . 59

6 Appendix 60
6.1 Appendix A: Loss curves and Dice score curves . . . 60
6.2 Appendix B: SegResNet analysis . . . 67

Bibliography 80

List of Figures

3.1 Weight scheduling strategies over 50 epochs. Following the preview strategy, the curves weight the different difficulty buckets using a non-cyclical sigmoid ramp. The *Hard* (solid red) function exhibiting a smoother transition (sharpness = 12.0) compared to the *Medium* (solid orange) function (sharpness = 16.0). Conversely, the *Easy* ramp (dashed green) decreases from 1.0 to 0.0. The vertical dashed line at epoch 30 marks the center of the sigmoid function, indicating the point of steepest change. 18
3.2 Schematic summary of the methodology chapter. 20

4.1 Dice coefficient progression over a subset of training epochs (including the optimal epoch per setup) on the MAMA-MIA dataset. The global average is shown in blue, while the mean trajectories for the Bottom-5 and Top-5 subgroups are displayed in red and green, respectively. Group assignments for Top-k and Bottom-k rely on the performance recorded at epoch n_0 30
4.2 Dice coefficient progression over a subset of training epochs (including the optimal epoch per setup) on the UCSD-PTGBM dataset. The global average is shown in blue, while the mean trajectories for the Bottom-5 and Top-5 subgroups are displayed in red and green, respectively. Group assignments for Top-k and Bottom-k rely on the performance recorded at epoch n_0 31
4.3 Dice coefficient progression over a subset of training epochs (including the optimal epoch per setup) on the UCSD-PTGBM and MAMA-MIA dataset using randomly shuffled buckets. The global average is shown in blue, while the mean trajectories for the Bottom-5 and Top-5 subgroups are displayed in red and green, respectively. Group assignments for Top-k and Bottom-k rely on the performance recorded at epoch n_0 32

4.4 Training and validation loss curves on the UCSD-PTGBM and MAMA-MIA datasets using a standard Empirical Risk Minimization (ERM) approach. The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. 33
4.5 Training and validation loss curves on the UCSD-PTGBM and MAMA-MIA datasets using a Loss-based Curriculum Learning approach. The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. The vertical dashed lines indicate the epochs at which medium and hard samples are introduced. 34
4.6 Training and validation loss curves on the UCSD-PTGBM and MAMA-MIA datasets using divergence-based curriculum learning with preview strategy (Preview Div-CL). The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. 34
4.7 Comparison of validation loss curves on the UCSD-PTGBM and MAMA-MIA datasets across three different training strategies. The plots illustrate the performance of the Preview Div-CL method (solid blue line) against the standard Div-CL approach (dashed green line) and the baseline ERM training (dashed red line). 35
4.8 Comparison of validation loss curves on the UCSD-PTGBM and MAMA-MIA datasets for different Curriculum Learning strategies. The plots contrast the performance of Preview Div-CL (solid blue line) against two loss-based approaches: Preview Loss-CL (dashed red line) and standard Loss-CL (dashed green line). 36
4.9 Pareto frontier of Dice Score vs. Training Epochs. The red dashed lines indicate optimal configurations, with elbow points identifying the best performance-to-cost trade-off at Center = 25 for the MAMA-MIA dataset and Center = 30 for the UCSD-PTGBM dataset. . . . 39
4.10 Cyclical linear ramp function over three cycles, initially evaluated for weight management within the preview strategy. 50
4.11 Cyclical cosine ramp function over three cycles, tested as a periodic variant for the learning process. 50
4.12 Cyclical sigmoid ramp function over three cycles. The sharpness parameter is set to 12 to provide independent control over the transition steepness. 51

6.1 Training and validation loss curves on the UCSD-PTGBM dataset using Divergence-based Curriculum Learning approach. The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. The vertical dashed lines indicate the epochs at which medium and hard samples are introduced. 60
6.2 Training and validation loss curves on the MAMA-MIA dataset using a Divergence-based Curriculum Learning approach. The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. The vertical dashed lines indicate the epochs at which medium and hard samples are introduced. . . 61
6.3 Training and validation loss curves on the UCSD-PTGBM dataset using loss-based curriculum learning with preview strategy (Preview Loss-CL). The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. . . . 61
6.4 Training and validation loss curves on the MAMA-MIA dataset using loss-based curriculum learning with preview strategy (Preview Loss-CL). The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. . . . 62
6.5 Comparison of Dice score trends on the UCSD-PTGBM dataset across three different training strategies. The plot illustrates the performance of the Div-CL method (solid blue line) and the Loss-CL approach (dashed green line) against the baseline ERM training (dashed red line). 63
6.6 Comparison of Dice score trends on the MAMA-MIA dataset across three different training strategies. The plot illustrates the performance of the Div-CL method (solid blue line) and the Loss-CL approach (dashed green line) against the baseline ERM training (dashed red line). 64
6.7 Comparison of Dice score trends on the UCSD-PTGBM dataset across preview-based training strategies. The plot illustrates the performance of the Preview Div-CL method (solid blue line) and the Preview Loss-CL method (dashed green line) against the baseline ERM training (dashed red line). 65
6.8 Comparison of Dice score trends on the MAMA-MIA dataset across preview-based training strategies. The plot illustrates the performance of the Preview Div-CL method (solid blue line) and the Preview Loss-CL method (dashed green line) against the baseline ERM training (dashed red line). 66

6.9 Dice coefficient progression over a subset of training epochs (including the optimal epoch per setup) on the MAMA-MIA dataset. The global average is shown in blue, while the mean trajectories for the Bottom-5 and Top-5 subgroups are displayed in red and green, respectively. Group assignments for Top-k and Bottom-k rely on the performance recorded at epoch n_0 67
6.10 Dice coefficient progression over a subset of training epochs (including the optimal epoch per setup) on the UCSD-PTGBM dataset. The global average is shown in blue, while the mean trajectories for the Bottom-5 and Top-5 subgroups are displayed in red and green, respectively. Group assignments for Top-k and Bottom-k rely on the performance recorded at epoch n_0 68
6.11 Training and validation loss curves on the UCSD-PTGBM dataset using a standard Empirical Risk Minimization (ERM) approach. The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. 69
6.12 Training and validation loss curves on the MAMA-MIA dataset using a standard Empirical Risk Minimization (ERM) approach. The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. 69
6.13 Training and validation loss curves on the UCSD-PTGBM dataset using Loss-based Curriculum Learning approach. The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. The vertical dashed lines indicate the epochs at which medium and hard samples are introduced. . . . 70
6.14 Training and validation loss curves on the MAMA-MIA dataset using a Loss-based Curriculum Learning approach. The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. The vertical dashed lines indicate the epochs at which medium and hard samples are introduced. . . 70
6.15 Training and validation loss curves on the UCSD-PTGBM dataset using divergence-based curriculum learning with preview strategy (Preview Div-CL). The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. 71
6.16 Training and validation loss curves on the MAMA-MIA dataset using divergence-based curriculum learning with preview strategy (Preview Div-CL). The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. 71

6.17 Comparison of validation loss curves on the UCSD-PTGBM dataset across three different training strategies. The plot illustrates the performance of the Preview Div-CL method (solid blue line) against the standard Div-CL approach (dashed green line) and the baseline ERM training (dashed red line). 72
6.18 Comparison of validation loss curves on the MAMA-MIA dataset across three different training strategies. The plot illustrates the performance of the Preview Div-CL method (solid blue line) against the standard Div-CL approach (dashed green line) and the baseline ERM training (dashed red line). 72
6.19 Comparison of validation loss curves on the UCSD-PTGBM dataset for different Curriculum Learning strategies. The plot contrasts the performance of Preview Div-CL (solid blue line) against two loss-based approaches: Preview Loss-CL (dashed red line) and standard Loss-CL (dashed green line). 73
6.20 Comparison of validation loss curves on the MAMA-MIA dataset for different Curriculum Learning strategies. The plot contrasts the performance of Preview Div-CL (solid blue line) against two loss-based approaches: Preview Loss-CL (dashed red line) and standard Loss-CL (dashed green line). 73
6.21 Training and validation loss curves on the UCSD-PTGBM dataset using Divergence-based Curriculum Learning approach. The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. The vertical dashed lines indicate the epochs at which medium and hard samples are introduced. 74
6.22 Training and validation loss curves on the MAMA-MIA dataset using a Divergence-based Curriculum Learning approach. The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. The vertical dashed lines indicate the epochs at which medium and hard samples are introduced. . . 74
6.23 Training and validation loss curves on the UCSD-PTGBM dataset using loss-based curriculum learning with preview strategy (Preview Loss-CL). The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. . . . 75
6.24 Training and validation loss curves on the MAMA-MIA dataset using loss-based curriculum learning with preview strategy (Preview Loss-CL). The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. . . . 75

6.25 Comparison of Dice score trends on the UCSD-PTGBM dataset using a SegResNet backbone across three different training strategies. The plot illustrates the performance of the Div-CL method (solid blue line) and the Loss-CL approach (dashed green line) against the baseline ERM training (dashed red line). 76
6.26 Comparison of Dice score trends on the MAMA-MIA dataset using a SegResNet backbone across three different training strategies. The plot illustrates the performance of the Div-CL method (solid blue line) and the Loss-CL approach (dashed green line) against the baseline ERM training (dashed red line). 77
6.27 Comparison of Dice score trends on the UCSD-PTGBM dataset using a SegResNet backbone across preview-based training strategies. The plot illustrates the performance of the Preview Div-CL method (solid blue line) and the Preview Loss-CL method (dashed green line) against the baseline ERM training (dashed red line). 78
6.28 Comparison of Dice score trends on the MAMA-MIA dataset using a SegResNet backbone across preview-based training strategies. The plot illustrates the performance of the Preview Div-CL method (solid blue line) and the Preview Loss-CL method (dashed green line) against the baseline ERM training (dashed red line). 79

Chapter 1

Introduction

1.1 Context and motivation

In the current context, Deep Neural Networks represent the dominant architecture in a wide range of applications, ranging from speech recognition, text classification, computer vision and medical imaging.

Many recent studies achieve state-of-the-art performance by designing progressively deeper neural networks, exposing a series of severe issues.

The most widely adopted strategy is called Empirical Risk Minimization (*ERM*). Since it is impossible to calculate a model's true error across the entire universe of possible data, it is approximated by minimizing the average loss across a representative dataset, the training set.

Over the years and throughout the literature, ERM has proven to be a solid principle, and its qualities have been well established. In the pursuit of improvement, research has moved toward a simple but very limited strategy, consisting of building deeper networks. However, scaling up architectures under the ERM paradigm exposes a series of severe issues.

Leading among these is the exponential increase in computational requirements and memory footprint. Without the application of any optimization technique, training and deploying such networks becomes prohibitive in terms of both time and computational costs. Deeper networks require a large pool of data for the training phase, which is not easily attainable for a lot of fields such as medical imaging. This architectural complexity makes it remarkably harder for the model to achieve robust generalization, making it difficult to prevent overfitting.

Beyond computational and data constraint, the scale of these architectures critically undermines their interpretability. They effectively operate as "black boxes", where highly complex, non-linear transformations distributed across millions of parameters obscure the reasoning underlying any prediction made by the model.

In this context, Explainable AI (XAI) emerges as a solution to the limitations of conventional AI by providing greater transparency and interpretability. Within the scope of the present work, this transparency allows healthcare professional to better understand and trust AI-driven decisions. [1]

Nevertheless, interpretability alone is not enough to be considered sufficient for ensuring fair and reliable predictions. In this regard, a critical aspect that needs to be accounted data imbalance across subgroups. Most of the datasets reflect the majority of the population, demographic or linguistic minorities are often severely underrepresented. Since traditional algorithms are design to maximize global accuracy, they compromise the performance of minority subgroups, as their misclassifications has minimal impact on the final evaluation.

Addressing these multiple challenges, several alternative training strategies have been developed as alternatives to traditional ERM framework. By changing the main focus to the training data instead of the network architecture, it is possible to ensure a more efficient and interpretable model.

Curriculum Learning strategy was based on the idea proposed by Elman [2], that neural networks in the learning phase, similar to human, benefit significantly from "starting small", meaning that the progression of complexity during the training could lead to a more efficient and robust learning. Louradour et al. [3] formalize the method known as Curriculum Learning (CL). Unlike the Empirical Risk Minimization, which samples the data uniformly at random order from the entire dataset from the first epoch, Curriculum Learning structures the training process by progressing from simpler to more complex concepts once these are acquired.

Since the formalization, this methodology has proved to be extremely effective across various fields, particularly in healthcare and medical imaging sector. Medical datasets are often considered challenging due to their limited size, caused by data scarcity and class imbalances.

Although Curriculum Learning has demonstrated promising performance improvements, the strategy typically relies on mathematical or loss-based difficulty that lack of clinical relation with patient, failing to provide meaningful information for clinicians and thus compromising interpretability.

Given the above-mentioned challenges of the medical domain, there is a demand more efficient solutions that reduce computational costs while maintaining the benefits of interpretability and performance.

1.2 Contributions

This thesis concentrates on curriculum learning in the context of medical image segmentation, providing a comprehensive review of existing approaches alongside a

novel methodological contribution. The main contributions are as follows.

1. **A comprehensive comparative analysis**. A systematic comparison of different curriculum learning methodologies is presented, including Empirical Risk Minimization (ERM) as baseline, *Loss-CL*, a loss-based curriculum learning approach and *Div-CL*, a divergence-based curriculum learning framework that uses metadata from the dataset to define samples difficulty using the divergence at subgroup-level[1].

2. **A Novel Curriculum Learning Methodology**. Building upon divergence and Div-CL, *Preview Div-CL* (as well as *Preview Loss-CL*) is introduced, a novel approach that combines the subgroup extraction using DivExplorer[4] combined with the preview strategy proposed by Ding et al.[5], introduced to accelerates convergence while improving interpretability and mitigates the subgroups disparities.

3. **Experimental Validation**. Preview Div-CL is evaluated on the *UCSD-PTGBM* and *MAMA-MIA* datasets on image segmentation task, demonstrating improved convergence speed, computed as the number of epochs, compared to the ERM baseline and the other curriculum learning methodologies analyzed in the comparative study in Chapter 4.

4. **Subgroup Fairness Analysis and Interpretability**. A fairness analysis across population subgroups is conducted on the UCSD-PTGBM and MAMA-MIA datasets, comparing *Preview Div-CL* against the ERM baseline and the other curriculum learning methodologies considered in this thesis. The results show that the proposed method reduces the performance gap between different subgroups, mitigating potential bias. It is further shown that using *divergence* as a difficulty metric provides an additional benefit, as it constitutes an interpretable metric.

1.3 Outline of the Thesis

This thesis is structured as follows. Chapter 2 provides the necessary background and reviews related work, focusing on Curriculum Learning, Explainability and Interpretability, Annealing, Mitigation and the application of Curriculum learning in medical image segmentation. Chapter 3 illustrates the methodological aspects

[1]The original idea of *Div-CL* was developed by Eliana Pastor and Eleonora Poeta and is currently being prepared for publication; it was not yet publicly available at the time this thesis was written

of *Div-CL* and *Preview Div-CL*, with greater emphasis placed on the construction of the curriculum and on the implementation of preview strategy. The core of this thesis, Chapter 4, presents the experimental setup, including details on the datasets and the architecture used. It compares all the methodologies and evaluates the efficacy of Preview Div-CL, alongside other preliminary experiments. Finally, Chapter 5 concludes the thesis by summarizing the main contributions, discussing the limitations of the current experimental setup, and suggesting directions for future research.

1.4 Research questions

- Is Preview Divergence-Based Curriculum Learning effective for medical imaging segmentation?
- Does Preview strategy applied to Divergence-Based Curriculum Learning works?
- Does Preview strategy applied to Divergence-Based Curriculum Learning improve performances?
- Is performing a subgroup aware strategy beneficial to the problem?

Chapter 2

Related Works

2.1 Background

2.1.1 Explainability and Interpretability

Over the past few years, deep learning models have achieved unparalleled level of accuracy across a wide range of fields. However, this outstanding performance increase has been complemented by a dramatic decline in transparency and interpretability.

These state-of-art models operates as true "black boxes", producing precise output while making impossible to comprehend the fundamental decision-making progress

In many critical domains for instance the healthcare, where a single decision has a direct impact on the patience's safety, relaying purely on a statistical prediction without any understanding of the reasoning process is not only insufficient but could also lead to potentially irreversible consequences for the individuals involved. Clinical personnel must be able to trust the tool, and this trust is inherently dependent on the system's ability to provide logically coherent and clinically meaningful explanations for its predictions

To overcome these ethical implications, the scientific literature has developed a substantial line of research. In this context, is essential to clearly distinction between the terms "Explainability" and "Interpretability", which are mistakenly and very frequently used as synonyms.

Following the definition made by Barredo Arrieta et al. and Lipton. *Explainability* refers to an explanation acting as an interface between humans and a decision-maker. It must accurately reflect the system's logic while remaining easy for humans to understand while *Interpretability* is defined as the ability to explain or to provide the meaning in understandable terms to a human [6]

Given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand. Interpretability refers to

a passive characteristic of a model referring to the level at which a given model makes sense for a human observer, an interpretable model is made by design.

Based on these definitions, many techniques have been made to mitigate the opaque nature of most of Deep Learning systems. Two of the most famous techniques to explain machine learning models black-box models are SHAP [8] and LIME [9].

LIME (Local Interpretable Model-Agnostic Explanations)[9] is an explanation technique that explains the predictions of any classifier in an interpretable and faithful manner. It provides interpretable local explanations by training a linear, interpretable model on the vicinity of an individual prediction. It scores the samples in the neighbourhood with the black-box system, and it uses the newly obtained labeled data as the training dataset.

Conversely, SHAP (SHapley Additive exPlanations) [8] leverages game theory to understand the contribution of each feature to local predictions. It considers all combinations and subsets of features, which makes it computationally heavy, thus an approximated version of SHAP, KernelSHAP [8], has been developed. Kernel SHAP performs as a model-agnostic surrogate, using weighted linear regression to approximate Shapley values by sampling feature rather than computing every possible combination.

In consideration of this distinction, the present work places itself in the current landscape by focusing specifically on interpretability rather than active explainability.

2.1.2 Preview Strategy Curriculum Learning

Shifting from the original formulation of the Curriculum Learning [10], multiple approaches have been proposed in the literature. "Preview strategy", presented in [5], aims to avoid the exclusion of harder samples at the beginning of training.

As a result, the model that sees this 'preview' is better prepared to handle complex patterns as training progresses, leading to a more stable learning process and accelerating convergence, thus reducing the total number of training epochs required.

This works employs a deviation from [5] "preview strategy" rethinking the weight's assignment that controls the loss weighting based on difficulty, a concept central to curriculum learning and discussed in detail below.

2.2 Curriculum Learning

The idea of curriculum learning was formally introduced by Bengio et al. as a strategy inspired on the method of how humans and animals acquire knowledge. Empirical observations of biological learning demonstrate a fundamental principle:

structuring the acquisition of concepts in order of progressively increasing complexity significantly enhances learning efficacy.

Applying this intuition to a mathematical model, the approach starts by assigning a difficulty score to each sample based on a certain metric before training and the grouping them into distinct buckets. During the training, the model is initially trained only on the easiest bucket, progressively expanding the training data pool to include more complex samples.

Mitchell proposed the following definition of machine learning:

Definition 2.2.1 *A model M is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.*

Most of the studies on curriculum learning follow this natural approach [3, 10] by increasing the experience E during the training process. However, some studies propose to apply curriculum with respect to the other components in Mitchell's 1997 definition. For instance, a series proposed to gradually increase the modeling capacity Training machine learning models in a meaningful order, from the easy samples to the hard ones, using curriculum learning can provide performance improvements over the standard training approach based on random data shuffling, without any additional computational costs. The integration of curriculum learning has been proved to bring a notable improvement in accuracy while simultaneously reducing convergence time

According to [12], the founding work by Bengio et al. is referred to as **Vanilla CL**, and since then, a wide range of curriculum forms have been explored. In Self-paced learning (SPL) [13] there is not a priori order of the samples that the model is fed with. Instead, the difficulty is measured repeatedly during training, dynamically adjusting the sequence of samples in the process. In Kumar et al., the sequence order is decided on the fly during the training, the model iteratively selects samples whose training loss falls below a progressively increasing threshold λ. Jiang et al. extends this approach by adding a diversity of regularization term to the objective function. This allows the model to dynamically selects samples with low loss (easy) and with high diversity, ensuring a balanced learning sequence across different groups from the beginning

Self-paced curriculum learning (SPCL) is a paradigm that combines both prior knowledge known before training and the with the model's empirical loss evaluated during training. The original paper introducing SPCL, Jiang et al., developed a framework to solve the problem of multimedia event detection. An extension of this concept for Reinforcement Learning (RL) was proposed by Klink et al., allowing the agent to dynamically adjust task difficulty during training It has also been exploited such as recommender systems [17] or weakly-supervised learning [18]

In **Progressive Curriculum Learning (PCL)** data samples are presented in a traditional random order, the curriculum itself is implemented at the task level. This means that difficulty is not managed by filtering or reordering the input data, but rather by progressively increasing the complexity of the objective the model have to solve In Karras et al., the task difficulty progression is achieved by incrementally growing both the generator and the discriminator. Training starts at a low resolution and new layers are gradually added to handle increasingly higher resolutions. Another example is Morerio et al., where instead using a fixed dropout probability throughout training, a time-based schedule was introduced to gradually raise the number of dropped neurons, slowly increasing the difficulty and leading to better generalization in image classification

Teacher-Student Curriculum Learning (TSCL) splits training into two roles: a Student, whose goal is to learn a complex target task, and a Teacher, that has to automatically select which subtask assign at the Student at that training point. In Kim and Choi, the Teacher-Student framework is implemented by attaching a secondary network, directly to the main network. This network has the job to assign a weight to the loss of the main network rather than excluding or including samples

In machine translation Chen et al. present a framework where a Teacher model, pre-trained on a pivot-to-target language pair transfers its knowledge to a Student model source-to-target over a source-pivot corpus to achieve zero-resource translation eliminating the need for direct parallel data.

Soviany et al. refers also to a situation when a CL is applied without any explicit curriculum building, this is called *Implicit Curriculum* (IC). In [23], curriculum is not explicitly designed but emerges implicitly from the self-play dynamics. At the beginning of training, the agent faces only weak opponents, an early version of itself, and the difficulty of the encountered situations scales as the agent improves. Another example of IC can be found in [24], where convolutional activation maps are gradually sharpened during training, effectively increasing the complexity of the representations the network have to learn

2.3 Curriculum Learning in Medical

Applying Curriculum Learning in medical fields provides significant advantages, since this enables models to mimic the training process for doctors [25]. By following the "easy-to-hard" process, the model improves its ability to detect lesions and its diagnostic accuracy, while also mitigating the problem of noisy labels.

Previous works explored some application for medical imaging:

In the context of mammogram classification, Lotter et al. proposed a multiscale convolutional neural network (CNN) trained using a curriculum learning strategy

with the goal of effectively handle the "needle in a haystack" nature of full-image mammogram classification. The authors employ a two stage curriculum learning approach. Starting from the easy phase, training classifiers on high precision lesion segmentation masks. Further on in this paper, they switched on a harder phase where the model employs these pre-learned features to analyze the entire, uncropped mammogram

In 2022, Jiménez-Sánchez et al. designed a curriculum learning approach that improves over the basic CNNs performance under such conditions. This paper make use of three curriculum strategies: individually weighting training samples, reordering the training set, and sampling subsets of data. Two scoring functions have been defined: one from domain-specific prior knowledge and an original self-paced uncertainty score. The first one assign a difficulty score based on the hierarchical structure of the AO classification system, that intrinsically reflects clinical complexity of the different fracture types and subtypes. The second one assigns a difficulty score based on an original self-paced uncertainty measure, that is dynamically estimated from the model's predictions during training. This was proved to be particularly useful in the absence of reliable domain-specific prior knowledge

An interesting application of CL was proposed by Jiménez-Sánchez et al. where we can see joint application of Curriculum Learning and Federated Learning [29]. The work's focus is improving the consistency of the local models penalizing inconsistent predictions, i.e., forgotten samples. The proposed curriculum learning approach orders the training samples prioritizing those that are forgotten after the deployment of the global model. This is combined with unsupervised domain adaptation to handle with domain shift while preserving data privacy

Another form of Curriculum Learning was presented in [30], that introduced a convolutional neural network (CNN) based attention-guided curriculum learning (AGCL) framework. This study objective is to jointly classify and weakly localize thoracic diseases in chest radiographs, using only image-level labels coupled with disease severity-level (DSL) data for a specific subset The images are divide in groups based on different severity-levels and used for training the CNN to learn gradually. In addition, highly confident samples and their corresponding class-conditional heatmaps, generated by the CNN, are extracted and passed into the AGCL framework to guide the learning of more discriminative convolutional features in the next iteration.

2.4 Annealing and Ramp

One of the major issues with optimization functions is that they tend to converge too early to a local minimum, resulting in suboptimal solutions. To address this

issue, adaptive scheduling strategies for both regularization parameters and learning rates have proven essential. These strategies are rooted in from the thermodynamic concept of *annealing*. Originally developed in metallurgy as a controlled cooling process to reduce defects in materials, the idea was later formalized as an optimization framework Kirkpatrick et al., who extended the Metropolis acceptance criterion [32] with the introduction of a temperature cooling schedule to control the stochastic search.

Following this thermodynamic analogy, when a molten metal is cooled too rapidly, the atoms lack sufficient time to rearrange themselves into an ordered structure, resulting in a very defective structure. On the other hand, a slow and controlled cooling process allows the atoms to gradually settle into their optimal configuration, having a nearly perfect crystalline lattice. Kirkpatrick et al. [31] translated this process into mathematics, the algorithm controls the optimization thought a "virtual temperature". In the initial phase, the systems accept more entropy, transitions temporarily worsen the objective function to better explore the loss landscape and avoid local minima. As the temperature goes down, mimicking the physical stability, the system become more rigid restricting its updates until it "freezes" to a converged state.

While this single monotonic descent is proved to be very effective smooth and simpler optimization problems, it often proves to be too rigid for modern deep learning non-convex loss. Once the standard simulated annealing freezes, it cannot be recovered if it settles in a suboptimal solution. In modern applications, such as training Variational Autoencoders (VAEs), this rigidity often leads to the well-known "KL Vanishing" problem. Instead of single cooling phase, Fu et al. [33] proposed a cyclical strategy when, after reaching the "freeze", the regulation term β is reset to a high value and then the cooling process is repeated. Another example of cyclical annealing is the Stochastic Gradient Descent with Warm Restarts (SGDR) introduced by Loshchilov and Hutter [34]. Instead of keeping a monotonic decaying learning rate, SGDR multiply the LR for by cosine function that periodically resets to its maximum value, allowing the "warm restart" that helps the model to escape the local minima.

The idea of using a scheduling function to control the loss during training, as introduced in [31] and [34], was adopted in this work to improve the convergence of the model.

2.5 Mitigation

In recent years, machine learning models have achieved unprecedented levels of accuracy across many different applications. Despite this, a critical challenge that necessitates in-depth investigation has been identified: the significant disparities

in performance outcomes between subgroups. Within the scope of this thesis, a subgroup is defined as a set of metadata attributes. These sets include patient demographic characteristics (e.g., age, ethnicity, and gender), as well as external variables related to the acquisition hardware, such has been identified as the scanner's model name and manufacturer.

The presence of these performance gaps raises concerning issues for ethical and safety reasons. Following the standard paradigm of minimizing the Empirical Risk (ERM) tends to inherently prioritize majority groups, failing to extract robust features for underrepresented subgroups. While these models typically perform reliably on standard or majority populations, they frequently exhibit significant performance degradation when applied to minority or underrepresented subgroups. This phenomenon is called *algorithmic unfairness* and has been investigated in recent academic literature [35] [36]

To mitigate this unfairness and ensure equitable performance across demographics, many mitigation techniques have been proposed in recent literature to address algorithmic bias and ensure fairness across different demographic subgroups. These approaches can be divided into three categories based on the stage of the machine learning pipeline at which they are applied: pre-processing, in-processing, post-processing

According to the definition presented in [37],

- Pre-processing techniques operate at data-level, changing the dataset to remove intrinsic biases before any learning occurs. This is typically done by modifying class label and features, or by adjusting samples weights to nullify misleading correlations between sensitive attributes. Calmon et al. introduced a probabilistic framework that learns randomize mapping of the training data through convex optimization designed to reduce the influence of protected attributes, such as race, age or gender, on the predicted outcome. In [39], two different data preprocessing techniques have been proposed: "massaging" and "reweighting". Massaging flips the labels of a few instances near the decision boundary, assigning positive outcomes to the discriminated group and negative ones to the non-discriminated, while reweighting assigns a weight to each instance to compensate for the underrepresentation or overrepresentation of certain groups, without altering features or labels.

- In-processing approaches intervene directly in the learning algorithm by forcing fairness constraints directly into the training loop. Instead of preprocessing the training data, these methods integrate fairness requirements into the model's objective function by adding fairness-aware regularization terms. In [40], adversarial learning was used to mitigate the biases. A predictor who aims to maximize the accuracy on the target task and an adversary has to infer the sensitive attribute from the predictor's outputs. If the inference is

successful, it means that the model still carries the information about the sensitive attribute, and the predictor is penalized accordingly. Kamishima et al. introduces a regularization term, called "prejudice remover", that penalizes the statistical dependency between the sensitive attribute and target variable, measured through a "prejudice index" based on mutual information.

- Post-processing methods came at the end of the ML pipeline, they are applied by recalibrating the final predictions of and classifier that is already trained. Since these models treat the underlying as a complete black box, leaving the internal logic it completely unchanged, they rely on adjusting thresholds for each demographic subgroup. As a result, this post-processing techniques are essential solution for proprietary or legacy systems where the algorithm cannot be modified to meet the fairness criteria In [42], the authors propose a methodology to recalibrate the prediction using decision thresholds for each group designed to equalize the true positive rates across different demographic groups. Pleiss et al. shows that calibration and equilization odds cannot be both satisfied when the base rates of the groups are different. In response to this limitation, a post-processing method that, based the output scores of an trained model, finds an optimal randomized thresholding policy for each group, increasing fairness without retraining the model.

Some works combine multiple mitigation strategies in a single framework, rather than relying on a single approach. Koudounas et al. [44] propose a divergence-aware dual mitigation strategy to address subgroup disparities in speech models. Relying on the notion of divergence to quantify the performance gap between a subgroup and the overall behavior, this notion is used to identify and prioritize underperforming subgroups. The authors propose three parallel approaches: a post-processing strategy that guides the acquisition of new data samples toward underperforming subgroups, and two in-processing approaches consisting of a regularization term that assigns greater weight to subgroups with higher divergence during training, and a data augmentation technique that increases the representation of challenging subgroups by applying audio transformations. This combined method makes the framework more versatile and applicable to both constrained and unconstrained models, offering flexibility for across different real-world scenarios.

2.6 Subgroup-Aware Training

Recent advances in deep learning have produced state-of-the-art models that achieve striking accuracy in healthcare tasks. However, these models may exhibit inherent biases, leading to unfair and disproportionate performance gaps across patient subgroups defined by characteristics such as ethnicity, sex, age, or socioeconomic status

A specific branch of mitigation strategies addressing this problem is Subgroup-Aware Training, which accounts for patient subgroups during the training process in order to reduce such disparities

For instance, Huti et al. [45] demonstrate that it is not only deep and complex "black box" deep learning models that suffer from racial discrepancies; even interpretable "white box" models like Random Forests show similar biases. This suggests that the biases are inherently dependent to the training data and the overall methodological approach, as algorithms trained on imbalanced datasets naturally tend to favor the overrepresented demographic, creating unfairness that standard methods struggle to mitigate.

A promising approach to address this problem involves the use of in-processing techniques, which incorporate fairness constraints directly into the model's optimization process, rather than modifying the input data or manipulating the resulting predictions. However, traditional in-processing methods often focus on a single demographic attribute at a time (e.g., race or gender in isolation), thereby excluding possible intersections among different attributes.

Poeta et al. [46] proposed a solution to overcome this limitation: FairMedSeg, an in-processing mitigation strategy adapted for medical image segmentation. FairMedSeg automatically identifies subgroups using divergence as a performance metric, defined as the deviation of each subgroup's performance from the population-level average, and reduces such disparities during training.

In [47], the authors propose a framework named GEORGE (Generative Optimization for Robustness to Generalization Errors) that allows to both measure and mitigate hidden stratification even when subclass labels are unknown. Starting with a standard training based on ERM, the model learns to classify only the known classes using the mean accuracy. Afterward, GEORGE uses the latent representation of the model to identify the underlying classes using cluster algorithms such as Gaussian Mixture Models (GMMs). Then, using the resulting clusters as a pseudo-label for re-training using Group DRO [48]. Instead of minimizing the average training loss across the entire dataset, Group DRO minimizes the worst-case loss of these clusters, thus ensuring a robust performance across all the subgroups, including hidden ones.

Chapter 3

Methodology

This chapter presents the theoretical foundations of the proposed methodology, *Preview Div-CL*. It begins by formalizing the problem setup, where the dataset structure, the composition of clinical metadata and the warm-up phase are defined. Then, the focus shifts to subgroup definition, detailing the evaluation metrics and the role of the *DivExplorer* framework. Following this, the methodology outlines the design of the curriculum learning strategy and its scheduling policy. The chapter concludes by detailing how the dynamic weight scheduling is integrated into the loss function (Preview strategy).

3.1 Problem setup

Let $D = (x_i, y_i, m_i)_{i=1}^{N}$ the extensive training datasets utilized in this study, where N represents the total number of available patient samples. Within these datasets, each individual triplet is composed of three distinct types of data. In particular $x_i \in \mathcal{X}$ is a medical image, while $y_i \in \mathcal{Y}$ is the corresponding ground-truth segmentation mask. Finally, $m_i \in \mathcal{M}$ is a vector that contains various clinical features and attributes extracted directly from the patient's metadata. This metadata includes demographic and clinical variables such as age, sex, and ethnicity, combined with acquisition-related factors derived from the scanner, including equipment manufacturer and model name. Both categories of variables are widely recognized as known sources of distributional heterogeneity in medical imaging datasets [49].

Let $f_\theta : \mathcal{X} \to \mathcal{Y}$ be a segmentation model parameterized by θ and trained via a loss function $\ell(\cdot)$. To evaluate the model's performance for each sample, we compute the score:

$$s_i = M(f_\theta(x_i), y_i)$$

where $M(\cdot, \cdot)$ indicates a performance metric that increases monotonically with respect to quality (i.e., a higher value indicates better performance).

As initial step, a warm-up model $\theta^{(0)}$ is obtained by training for n_0 epochs via standard empirical risk minimization (ERM)

$$\theta^{(0)} = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \ell(f_\theta(x_i), y_i)$$

During the initial warm-up phase, the system generates baseline performance estimates by applying a uniform sampling strategy across the entire dataset to compute the global mean performance $\bar{M}$. The metrics acquired from this unbiased evaluation are subsequently analyzed against $\bar{M}$ to identify and quantify performance disparities that exist among various subgroups. Eventually, these disparity mappings provide the foundation for building an interpretable metadata-aware curriculum.

3.2 Subgroup divergence estimation

A *subgroup* g is formally defined as a logical conjunction of metadata predicates, allowing for the precise partitioning of a dataset into distinct, analyzable sets (for example, $\{\text{sex} = \text{male}, \text{age} = [50,60]\}$). To ensure consistency, any continuous metadata variables are discretized into fixed intervals when necessary. Starting from the prediction made by the warm-up model, a global performance mean $\bar{\mathcal{M}}$ is computed. For each $g \in \mathcal{G}_\tau$, a mean performance is estimated as

$$\bar{\mathcal{M}}_g = \frac{1}{|g|} \sum_{i \in g} s_i$$

For each subgroup, its divergence is defined as a difference between the global mean performance and the subgroup performance:

$$\Delta_g = \bar{\mathcal{M}}_g - \bar{\mathcal{M}}$$

If Δ_g is a negative value ($\Delta_g < 0$), it means that g performs below the average global performance, exposing a systematic performance decrease in accuracy linked to a specific combination of metadata.

Most metadata configurations yield an insufficient number of samples, making them statistically insignificant. Therefore, exploring all possible attribute combinations is not only computationally inefficient but also essentially uninformative, since rare patterns are unreliable due to statistical fluctuation.

3.2.1 DivExplorer

To derive the subgroups, Div-CL relies on DivExplorer [50], a tool for analyzing subgroup performance within a dataset and efficiently identifying anomalous data

subgroups. Given a dataset, DivExplorer extract subgroups that exhibit statistically significant deviations in specific attributes compared to the global average, revealing where these values are substantially higher or lower compared to the global average. It is based on Frequent Pattern Mining (FPM), a technique used to identify combinations of attributes, or patterns, that occur frequently within a dataset. For the initial extraction of subsets (i.e., subgroups), DivExplorer can be paired with any FPM technique, such as *Apriori* [51] or *FP-growth* [52]. Since FPM algorithms operate on discrete attributes, every continuous attribute needs to be discretized beforehand. As the number of possible itemsets grows exponentially with the number of dataset attributes, a two phase selection problem is involved. The first phase is excluding any itemset which has support under a certain threshold, allowing to maintain only statistical significant patterns and avoid any possible noise related fluatiation. In the second phase, the entire pattern is pruned if it contains an item with a marginal contribution under *alpha*, indicating the attribute has no significant influence. However, for the purpose of this study, this second phase was not applied. DivExplorer introduces the concept of "corrective" attribute values, which reduce divergence when added to an itemset, indicating that divergence (Δ) is not monotonic. To assess the impact of a single item, DivExplorer compares two approaches: Individual Divergence, which is simple but misses out on how various items work together, and Global Item Divergence, which uses Shapley Values to determine each item's significant contribution across all potential combinations.

3.3 Curriculum Learning construction

The fundamental metric used in this work to build the curriculum is divergence Δ_g. Starting from Δ_g, the curriculum difficulty score is defined as $d_g = -\Delta_g$. A subgroup with a difficulty below average has a negative Δ_g which means that the difficulty score d_g is positive; this can be perceived as a more difficult subgroup. It typically occurs that a sample belongs to multiple subgroups, thus sample level difficulty is assigned by following the worst case scenario: the difficulty score of each sample is decided according to the highest difficulty among those of its subgroups.

Let $\mathcal{G}(s)$ denote the set of discovered subgroups containing sample i. The sample-level difficulty is defined as:

$$d_i = \begin{cases} \max_{g \in \mathcal{G}(i)} d_g & \text{if } \mathcal{G}(i) \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

After sorting samples in ascending order based on difficulty, they are divided equally into three buckets: easy, medium, and hard. Following the finalization of the curriculum schedule, the proper CL training loop begins at epoch n_0+1, immediately after the warm-up phase. In a standard Curriculum Learning framework, the

scheduler starts by providing only the easy bucket, Training proceeds on this subset epoch after epoch, while a plateau check is performed at the end of every epoch. The progression condition require that the metric should not plateau (i.e., accuracy improvement drops below a minimum delta value `plateau min delta`) for a longer period than the certain *patience* value. Once a plateau is detected, the scheduler expands the training subset by adding the medium bucket, keeping together with the easy bucket. The training continue with this combined set until the same patterns is repeated, triggering the transition that incorporate the hard bucket. Once the dataset is fully integrated, an early stopping strategy is adopted to prevent resource waste.

This approach interrupts the process if no improvements are observed after reaching a predefined patience value.

3.4 Preview strategy implementation

Despite the Divergence Curriculum Learning (Div-CL) framework, as described earlier, provides a robust foundation for training, there remains for a significant acceleration in its convergence speed. Therefore, the methodology presented in this work expands upon Div-CL, further refining the learning process to reduce training time and optimize training efficiency.

The *Preview Strategy* introduced by Ding et al. [5] is based on the intuition of dynamically assigning weights to the samples according to their difficulty. Instead of completely excluding hard samples, which is usually done in traditional curriculum learning, it assigns them a reduced weight, thus allowing the model to be exposed to a "preview" of the entire dataset and enhancing the learning process.

In this work, inspired from the idea of Simulated Annealing [31] where an exponential function controls and adapts the system's behavior during training, the Preview Strategy is implemented using a sigmoid function to regulate the weights assigned to samples of different difficulty levels.

Let s denote the current training step, c represents the center of the sigmoid function (i.e., the midpoint of the transition), and k the parameter that controls the sharpness of the curve. The ramp function is defined as:

$$W(s) = \begin{cases} 1 & \text{if } c \leq 0 \vee s \geq 2c \\ \frac{1}{1+e^{-\frac{k}{2}\left(\frac{s-c}{c}\right)}} & \text{otherwise} \end{cases} \tag{3.2}$$

After dividing the samples into difficulty buckets, each bucket is assigned a ramp characterized by a specific center and sharpness. Following the division into "easy-medium-hard" that was introduced in the previous section, three distinct scheduling ramps are employed.

The initial ramp, utilized for "easy" samples, follows an inverted sigmoid shape. As a result, the weights given to these samples decrease gradually, making them progressively less informative as the number of training epochs increases. The idea behind this design is that easy samples lose their informational value as the model approaches convergence, and gradually eliminating them avoids unnecessary gradient updates [53]. The second and third ramps, used for medium and hard samples, follow a standard sigmoid shape but are initialized with two different sharpness values, introducing hard samples more gradually than medium ones to ensure that they are effectively assimilated without destabilizing the learning dynamics.

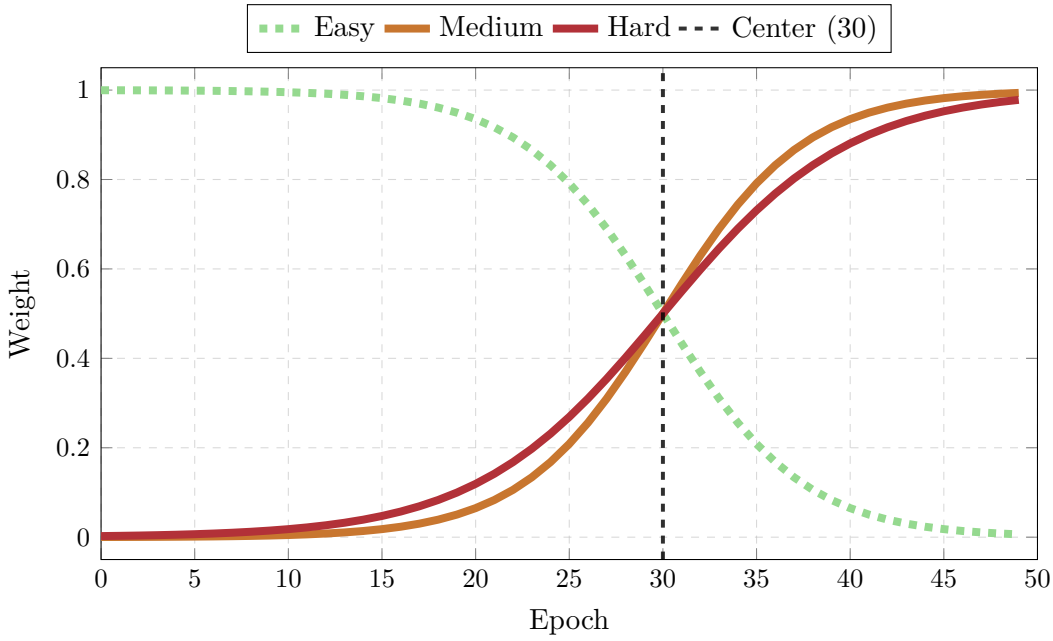


Figure 3.1: Weight scheduling strategies over 50 epochs. Following the preview strategy, the curves weight the different difficulty buckets using a non-cyclical sigmoid ramp. The *Hard* (solid red) function exhibiting a smoother transition (sharpness = 12.0) compared to the *Medium* (solid orange) function (sharpness = 16.0). Conversely, the *Easy* ramp (dashed green) decreases from 1.0 to 0.0. The vertical dashed line at epoch 30 marks the center of the sigmoid function, indicating the point of steepest change.

To further ensure robust optimization and avoid premature convergence, this thesis introduces a *Dynamic Ramping Strategy* as one of its major contributions. Standard static ramps strictly constrain the model's adaptability by imposing a rigid schedule, regardless of whether the loss is actually improving. The dynamic ramp shifts its center backwards by an empirically defined percentage decrement whenever

the patience threshold is exceeded (i.e., when the loss shows no improvement for a predefined number of epochs). The decrement is computed and then rounded to the nearest integer.

Let C_t be the ramp center at step t, p the patience threshold, and n the counter of consecutive steps without metric improvement. The dynamic ramp update is defined as:

$$C_{t+1} = \begin{cases} \lfloor C_t \cdot (1 - \rho) \rceil & \text{if } n \geq p \\ C_t & \text{otherwise} \end{cases} \tag{3.3}$$

Where ρ represents the percentage decrement step size. Upon a ramp update, the counter n is reset to zero.

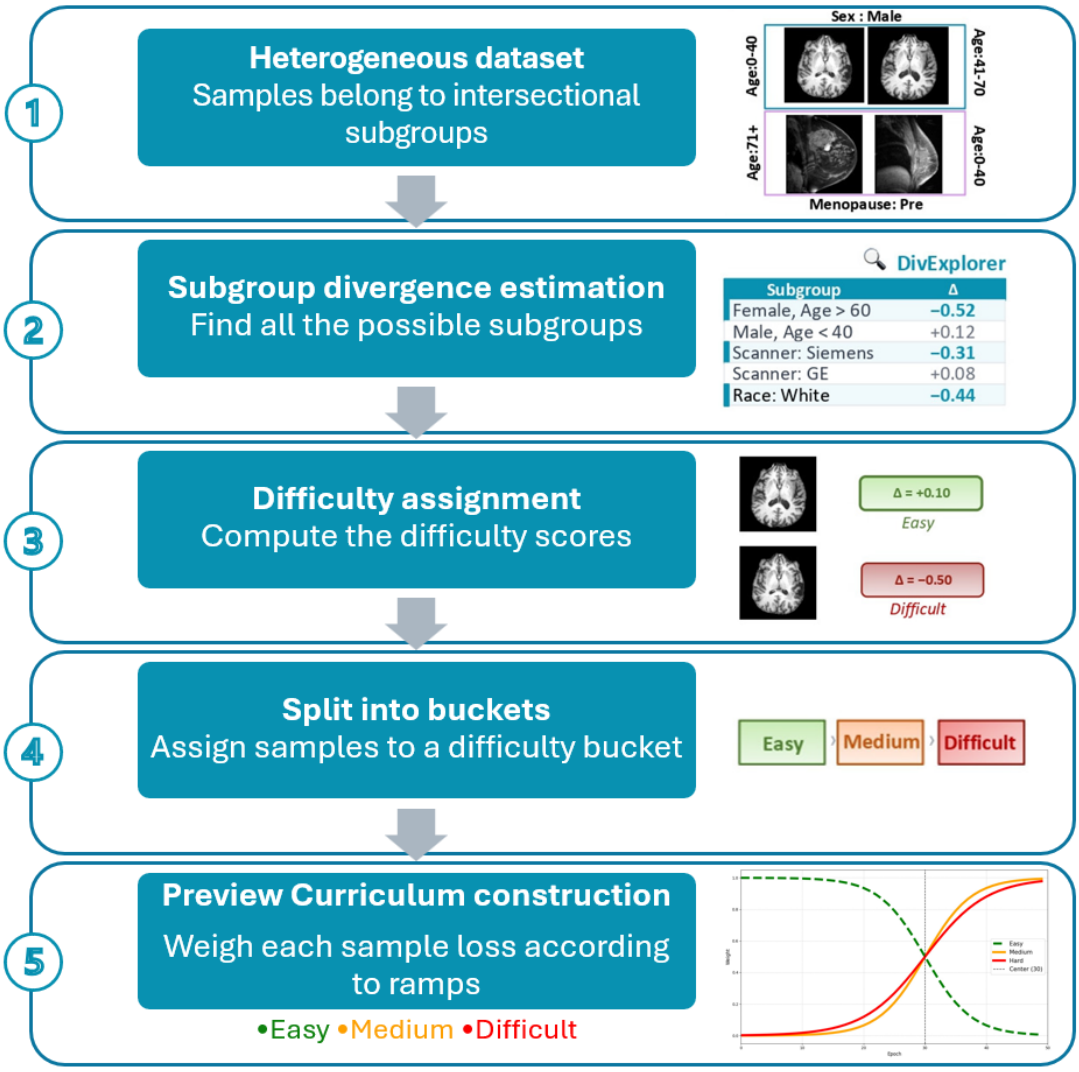


Figure 3.2: Schematic summary of the methodology chapter.

Chapter 4

Experimental Analysis

4.1 Experimental setup

4.1.1 Datasets

The experimental evaluation in this thesis relies on two public medical imaging datasets: MAMA-MIA and UCSD-PTGBM. A summarized comparison of their characteristics is provided in Table 4.1.

MAMA-MIA [54] [55] is a large-scale, multi-institutional dataset comprising 2,080 pre-treatment T1-weighted Dynamic Contrast-Enhanced MRI (DCE-MRI) cases The Training set includes 1,506 patients from the United States, sourced from I-SPY 1 (171 cases) and I-SPY 2 (980 cases) clinical trials (MRI-based treatment monitoring), the Duke University breast cancer group (291 cases), and an extra neoadjuvant chemotherapy NACT collection (64 cases). Meanwhile, the validation and testing cohorts include 574 private, independently collected cases from three European clinical centers: 30 from the Medical University of Gdansk (GUM, Poland), 232 from Kauno Klinikos (KAU, Lithuania), and 312 from Hospital Clínic Barcelona (HCB, Spain). Collecting diverse data from multiple institutions and countries ensures that the training dataset achieves high representativeness and robustness. For every case, the dataset includes high-quality segmentations of the primary tumor. These masks were initially predicted by a deep learning model and later validated by a group of 16 radiologists averaging nine years of clinical experience. In addition, the MAMA-MIA dataset has a wide range of clinical and demographic annotations for each patient. These metadata include variables like patient age, menopausal status, ethnicity, and breast density, but also acquisition variables such as the scanner manufacturer. **UCSD-PTGBM** [56] is a multimodal MRI dataset composed of 243 post-treatment timepoints from 178 patients with histopathologically proven glioblastoma (GBM). The dataset includes 192 timepoints featuring residual or recurrent tumors, and 51 timepoints exhibiting

only post-treatment changes (for example pseudoprogression and radiation necrosis), sourced entirely from the University of California San Diego (UCSD). The images were obtained using a 3 Tesla MRI scanner following a very comprehensive protocol. It includes standard 3D anatomical sequences (pre- and post-contrast T1, T2, FLAIR) as well as advanced imaging techniques such as multishell diffusion (Restricted Spectrum Imaging, RSI) and perfusion imaging (Arterial Spin Labelling, ASL, and Dynamic Susceptibility Contrast, DSC). The dataset provides, for each case, multicompartment tumor segmentations following the standard BraTS protocol [57] [58]. These masks were initially predicted by a deep learning model and then manually corrected by trained radiologists. Annotations were also produced by radiologists and include both standard BraTS-style tumor masks and separate cellular tumor segmentations. The UCSD-PTGBM metadata cover five areas: demographic and acquisition data (age, sex, race, ethnicity, scanner), diagnosis (glioma type, WHO 2021 classification, grade), molecular biomarkers (MGMT, IDH, 1p19q, ATRX), and treatment (surgery, radiation, chemotherapy, including Avastin®). For this thesis, only the demographic and acquisition variables: age, sex, race, ethnicity, manufacturer, and scanner model.

4.1.2 Pre-Processing

Beyond the dataset-specific preprocessing already described in the original MAMA-MIA and UCSD papers, additional preprocessing was performed to ensure consistency and homogeneity in the metadata from the two sources. At the beginning, the preprocessing step selected and filtered the features of interest for the study, handling missing or invalid values through the assignment of a default neutral category ("unknown"). The continuous age variable was discretized into five predefined age brackets (0-40, 41-50, 51-60, 61-70, 71+). A semantic normalization was performed on dataset-specific categorical variables: in the breast imaging data, menopausal status was consolidated into two macro-classes, "pre" (merging pre- and peri-menopausal cases) and "post". In UCSD-PTGBM, patient sex was standardized into "male" and "female" formats. Whitespaces and empty strings in fields detailing patient demographics and technical equipment (ethnicity, race, manufacturer, and model) were sanitized. To conclude, the entire dataframe was converted to a string format to avoid data type inconsistencies during the subsequent stages of the experiment.

4.1.3 Backbone architectures

U-Net

In order to perform the image segmentation task on the previously mentioned MAMA-MIA and UCSD-PTGBM datasets, the U-Net architecture [59] has been

Table 4.1: Comparison between the dataset used in this work

Characteristic	**MAMA-MIA**	**UCSD-PTGBM**
Clinical domain	Breast cancer	Glioblastoma (post-treatment)
Number of cases	1506	243 timepoints / 178 subjects
Imaging modality	DCE-MRI	Structural + multishell diffusion (RSI) + ASL/ DSC perfusion, 3T
Source	4 TCIA cohorts (multi-center)	UCSD (single center), via TCIA
Segmentation	Primary tumor + non-mass enhancement	Standard (BraTS-style) + cellular tumor segmentation
Acquisition-related metadata	scanner model; manufacturer	scanner model; manufacturer
Patient-related metadata	Age; menopause; ethnicity; breast density	Age; sex; ethnicity; race

adopted as one of the backbone architectures for this thesis. This choice is driven by the architecture's proven data efficiency, which is particularly advantageous for medical imaging where the dataset are often very limited. Unlike other deep learning architectures that require massive amounts of annotated data to converge properly, U-Net can achieve highly accurate segmentations even with datasets of limited size. Moreover, it acts as a well-established benchmark in the literature due to the encoder-decoder structure, paired with skip connections, ensuring solid performance.

SegResNet

Following the ablation studies conducted with the preliminary U-Net model, which were used to identify the optimal configuration for the methods requiring hyperparameter tuning, the SegResNet architecture was applied across all the investigated methods on both datasets. In the configuration used in this study, SegResNet comprises approximately 18.8M parameters, compared to 4.8M for U-Net. This model was selected to test whether a larger network with a different architecture, using residual connections, outperforms the baseline U-Net, allowing us to determine whether the potential increase in computational cost translates into meaningful performance gains. Furthermore, employing two different models also provides evidence of the model-agnostic nature of the evaluated approaches.

4.1.4 Metrics

To evaluate the performance of the model, extract subgroups, and compute their divergence using DivExplorer [4] we select a set of metrics. In the present study, the Dice score is employed as the primary performance metric and baseline result.

DICE

The Dice score [60] measures the spatial overlap between predicted and ground truth segmentations, providing an indicator of overall segmentation precision, where higher values (closer to 1) indicate better performance. Formally, the Dice Similarity Coefficient (DSC) is defined as follows:

$$DSC = \frac{2\,|A \cap B|}{|A| + |B|} = \frac{2\,TP}{2\,TP + FP + FN} \tag{4.1}$$

Where A and B represent the sets of voxels in the predicted and ground-truth segmentations. This metric is widely used in medical image analysis to quantify the similarity between segmented regions and reference annotations.

As stated previously, this metric is used as the primary indicator of the model's performance and while also been used to as constant metric for computing the

divergence score across all comparative analyzes. Moreover, it acts as the constant metric, which is paired with additional metrics to compute the sample difficulty (divergence score) for all the samples.

4.2 Final method: Preview Div-CL

Out of all the approaches and configuration tested, the implementation of Divergence Curriculum Learning combined with Preview Strategy (referred as Preview Div-CL) has been found to produce the best overall results. This configuration achieves a significant reduction in computational costs while maintaining global performance comparable to and, in some instances, surpassing the more expensive baselines. Particularly, Preview Div-CL significantly improved performance across specific subgroups, effectively narrowing the performance gap and ensuring a more fair and unbiased model. (See 4.4 Subgroups benefit)

Using the optimal parameters derived from the ablation study presented in Section 4.6, the setup was configured as follows. Starting with the ramp configurations for both the MAMA-MIA and UCSD-PTGBM datasets, the ramp sharpness parameters were set to 6 for the easy ramp ($ramp_slope_e$), 6 for the medium ramp ($ramp_slope_m$), and 4 for the hard ramp ($ramp_slope_h$). For the ramp centers, the values were fixed at 25 for MAMA-MIA and 30 for UCSD-PTGBM. For the dynamic decrease of the centers, the optimal values were determined to be a 20% decrease ($decr_m$) for the medium ramp and a 10% decrease ($decr_h$) for the hard ramp, applied after 5 epochs of no improvement ($ramp.patience$). Meanwhile, the inverted ramp, assigned to the easy bucket, was excluded from this dynamic system and kept static. Hereafter, this configuration will be referred to as **Config-Ref**.

On these setups, Preview Div-CL scored a global Dice score of 0.7704 on MAMA-MIA dataset in 44 epochs, and reached a Dice score of 0.8433 in 46 epochs on the UCSD-PTGBM dataset.

4.3 Comparison with baseline and alternatives

To carefully evaluate the proposed method, Preview Div-CL was benchmarked against the standard Empirical Risk Minimization (ERM) and three Curriculum Learning variations: Loss-CL, Loss-CL with Preview, and standard Div-CL. The comparative analysis focused on global Dice score and computational efficiency (epochs to convergence). For a fair comparison across all methods and architectures, all experiments were conducted using the same ramp configurations, $\texttt{Config-Ref}_{4.2}$, as introduced in the previous section.

To provide a comprehensive assessment of the trade-off between segmentation performance and computational cost, the relative percentage changes in both

metrics with respect to the ERM baseline (%ΔDice and %ΔEpochs) are reported.

As context for the comparative analysis, the baselines under evaluation are briefly described below:

- **ERM - Empirical Risk Minimization**: Serving as baseline for computational cost and performance, ERM is the standard training paradigm in Deep Learning where the model processes the entire dataset uniformly without any curriculum strategy.
- **Div-CL (without Preview Strategy)**: A curriculum learning approach that divides the data into subgroups based on the computation of divergence from the Dice score.
- **Loss-CL**: A variation inspired by self-paced curriculum learning [61], where the difficulty of a training sample is determined by its loss value rather than its divergence score.
- **Loss-CL (with Preview Strategy):** An extension of the standard Loss-CL method that integrates the Preview Strategy. This configuration allows us to evaluate the impact of the preview when applied to a traditional loss-based curriculum, helping to determine whether its benefits can be generalized. This baseline will be also referred later in these work as 'Preview Loss-CL'

The results for both datasets are summarized in 4.2.

As anticipated, the standard training using ERM has been proved to be a solid baseline in terms of segmentation accuracy, achieving a Dice score of 0.7881 on MAMA-MIA and 0.8609 on UCSD-PTGBM. Converging in 65 and 60 epochs, it was also found to be relatively efficient.

The introduction of Curriculum Learning strategies produces very mixed results across the configuration and suggests a possible trade-off in terms of number of epochs and Dice score. While Loss-CL reaches the highest Dice score of 0.8750 on UCSD-PTGBM, surpassing the ERM baseline, this CL strategy comes at a significant computational cost, requiring over 90 epochs to converge

Conversely, integrating the Preview mechanism greatly accelerates the training process. Considering Preview Loss-CL, the impact on computational efficiency varies significantly between the two datasets.

On the MAMA-MIA dataset, although it achieves a 14% reduction in training time compared to normal Loss-CL (from 94 to 81 epochs) and a 23% reduction compared to Div-CL (105 epochs), Preview Loss-CL performs approximately 25% slower than the standard ERM baseline (65 epochs). On the UCSD-PTGBM dataset, the preview technique proves to be highly efficient, dropping the required training time to just 47 epochs. This produces a massive 51% reduction compared to normal Loss-CL (96 epochs), a nearly 58% reduction compared to Div-CL

Table 4.2: Comparison of global Dice score and computational efficiency across the MAMA-MIA and UCSD-PTGBM datasets. Results include absolute Dice score, convergence epochs, and their respective relative changes (%Δ) compared to the ERM baseline. The proposed Preview Div-CL method is displayed in bold.

Method	**Dice Score (↑)**	$\Delta_{\textbf{Dice}}$	**Epochs (↓)**	$\Delta_{\textbf{Epoch}}$
Dataset: MAMA-MIA				
ERM (Baseline)	0.7881	-	65	-
Loss-CL	0.7760	-1.54%	94	+44.6%
Preview Loss-CL	0.7781	-1.27%	81	+24.6%
Div-CL	0.7770	-1.41%	105	+61.5%
Preview Div-CL	**0.7705**	**-2.23%**	**44**	**-32.3%**
Dataset: UCSD-PTGBM				
ERM (Baseline)	0.8609	-	60	-
Loss-CL	0.8750	+1.63%	96	+60.0%
Preview Loss-CL	0.8426	-2.13%	47	-21.6%
Div-CL	0.8610	+0.01%	111	+85.0%
Preview Div-CL	**0.8433**	**-2.04%**	**46**	**-23.3%**

(111 epochs), and a positive outcome over ERM baseline with a 22% reduction in computational cost (60 epochs).

The proposed Preview Div-CL successfully overcomes the trade-off between computational cost and dice score. The integration of the preview strategy into the divergence-based curriculum allows the model to combine the advantages from both approaches. This method achieves convergence in only 44 epochs on MAMA-MIA and 46 epochs on UCSD-PTGBM, a nearly 32% and 23% reduction compared to the ERM baseline, respectively, and a striking 58% and 59% reduction compared to standard Div-CL. This acceleration reduces the segmentation accuracy on the UCSD-PTGBM dataset. Compared to the ERM baseline, the drop was around 2% for both Preview Div-CL, up to 0.017, and Preview Loss-CL, losing 0.0183 in Dice score. On MAMA-MIA dataset, the accuracy drop is insignificant, under 3% considering Div-CL approach, with Preview Loss-CL, which reaches the highest Dice score (0.7781) losing less than $\sim$1.2%.

After the analysis, while maintaining comparable Dice scores with respect to the other method, Preview Div-CL emerges as the most computationally efficient strategy.

4.4 Subgroup-level benefits: Comparison with random generated

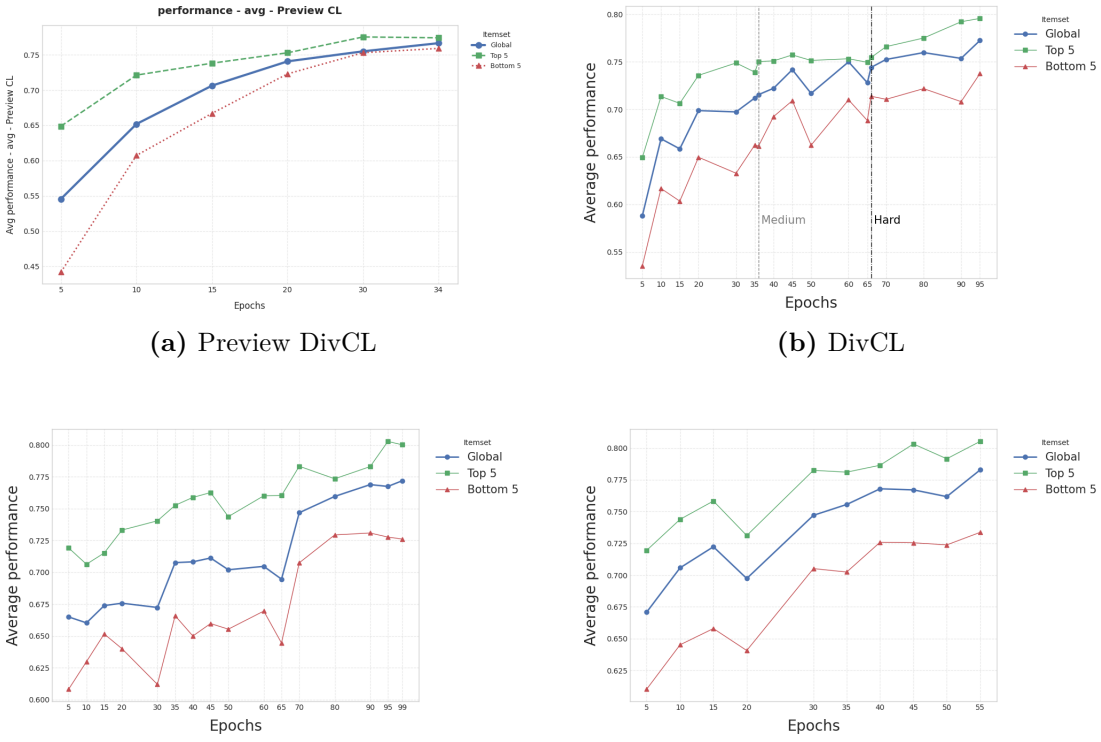
One of the issues discussed in Section 1 is the tendency of deep learning models to present variance and disparities in performances when applied in different subgroups, likely caused by underlaying biases. This section examines several methodologies for assessing performance improvements in this domain and determines whether the disparities have been successfully mitigated. The analysis evaluates whether the performance across different subgroups aligns by the end of training. In Figures 4.1 and 4.2, we consider the top-5 and bottom-5 subgroups in terms of initial performance, in parallel with the overall average.

As shown by the plots (Figures 4.1and 4.2), performance variability between subgroups changes significantly across different approaches. UCSD-PTGBM dataset exhibits greater disparities between the top-5 and bottom-5 subgroups, showing a Dice score gap of 0.3 across all evaluated methods, while MAMA-MIA shows a better initial baseline, with only a 0.15 difference in Dice score between the top and bottom subgroups.

Across both MAMA-MIA and UCSD-PTGBM datasets, Empirical Risk Minimization (ERM) is the worst performing training method, as expected, producing the largest gap between Top-5 and Bottom-5 subgroups in the final Dice Score. This behavior can be traced back to the nature of the algorithm: by minimizing

the average error across the entire dataset, ERM implicitly prioritizes simpler subgroups or subgroups with higher representation, sacrificing the less significant ones. Figures 4.1d and 4.2d reveal that most of the learning occurs between epochs 20 and 30, during which marked increase in performance can be observed. In contrast to ERM, when analyzing the UCSD-PTGBM dataset, the remaining approaches present highly similar performance, managing to reduce the gaps between subgroups. Turning to the MAMA-MIA dataset, divergence-based strategies emerge as significantly more effective. In particular, with Preview-DivCL, the results for all subgroups are very close the global average by the end of the training process.

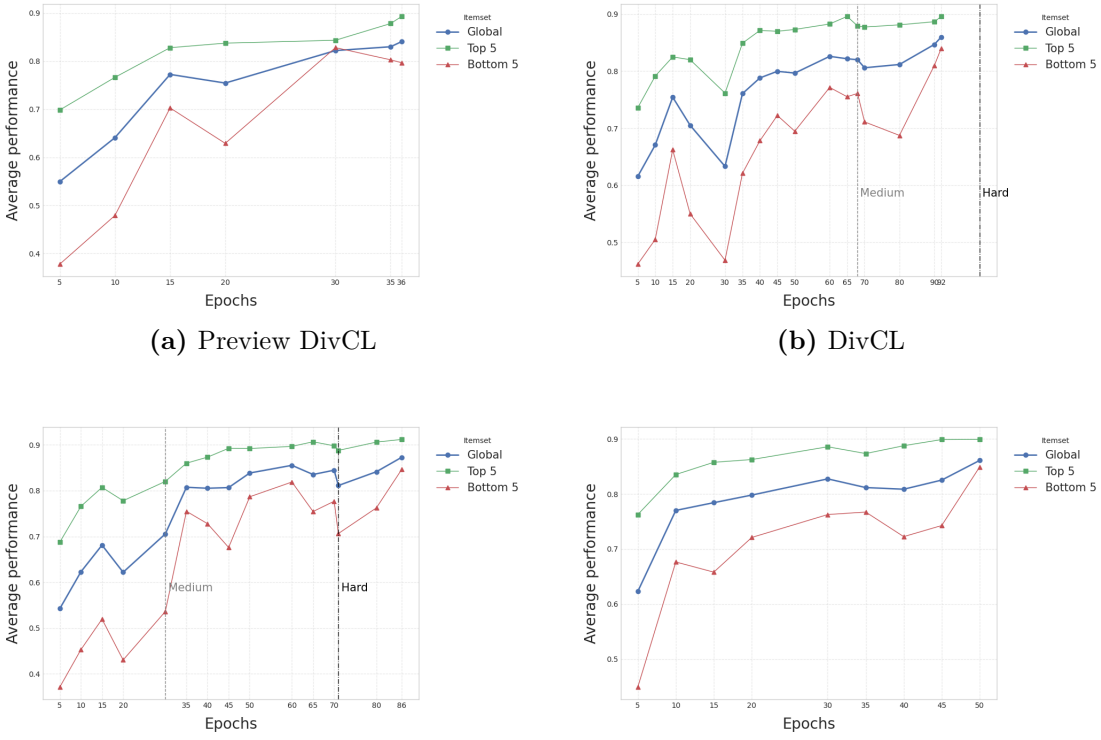
It is interesting to note that both Div-CL (Figures 4.1b and 4.2b) and Loss-CL (Figures 4.2c and 4.1c) exhibit two performance drops for the bottom-5 groups, which are not observed in Preview-DivCL. These instabilities seem to be linked to the curriculum scheduling strategy. As training progresses, Medium and Hard difficulty buckets are introduced to the model, when this happens the model's performance temporarily declines. However, since the Top-5 subset consists of the easiest subgroups, which are available for the model from the beginning, their trend remains stable throughout these transitions. Unlike the previously discussed methods, Preview Div-CL is unaffected by this phenomenon. When observing the UCSD-PTGBM dataset, the performance curve (Figure 4.2a) follows an asymptotic growth, reaching a plateau in the final epochs without experiencing any drop.



(a) Preview DivCL **(b)** DivCL

(c) Loss CL **(d)** ERM

Figure 4.1: Dice coefficient progression over a subset of training epochs (including the optimal epoch per setup) on the MAMA-MIA dataset. The global average is shown in blue, while the mean trajectories for the Bottom-5 and Top-5 subgroups are displayed in red and green, respectively. Group assignments for Top-k and Bottom-k rely on the performance recorded at epoch n_0

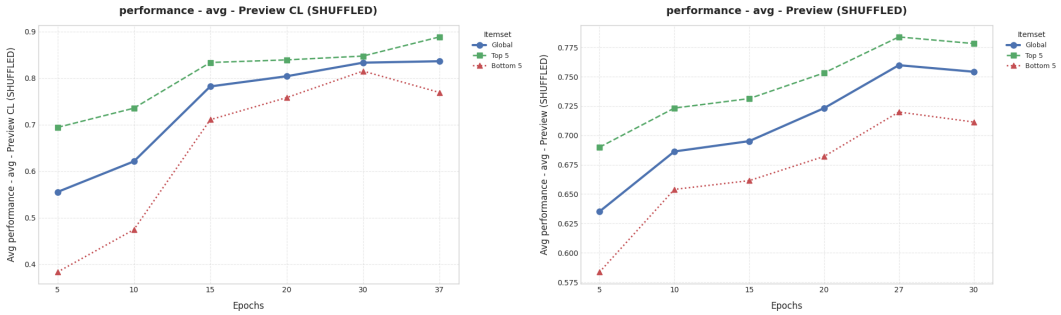


(a) Preview DivCL **(b)** DivCL

(c) Loss CL **(d)** ERM

Figure 4.2: Dice coefficient progression over a subset of training epochs (including the optimal epoch per setup) on the UCSD-PTGBM dataset. The global average is shown in blue, while the mean trajectories for the Bottom-5 and Top-5 subgroups are displayed in red and green, respectively. Group assignments for Top-k and Bottom-k rely on the performance recorded at epoch n_0

To isolate and validate the contribution of Curriculum Learning within Preview-DivCL framework, an experiment was conducted by randomly shuffling the samples before splitting them into buckets, creating homogeneous buckets. The results, in Figures 4.3a and 4.3b, prove that disrupting the difficulty-level presentation order has a negative impact on segmentation performance. When the buckets are shuffled, the gaps between the top performing subgroups and the least performing subgroups become wider. This experiment underlines the contribution of the curriculum order in Preview Div-CL strategy.



(a) Preview DivCL with shuffled difficulty buckets

(b) Preview DivCL with shuffled difficulty buckets

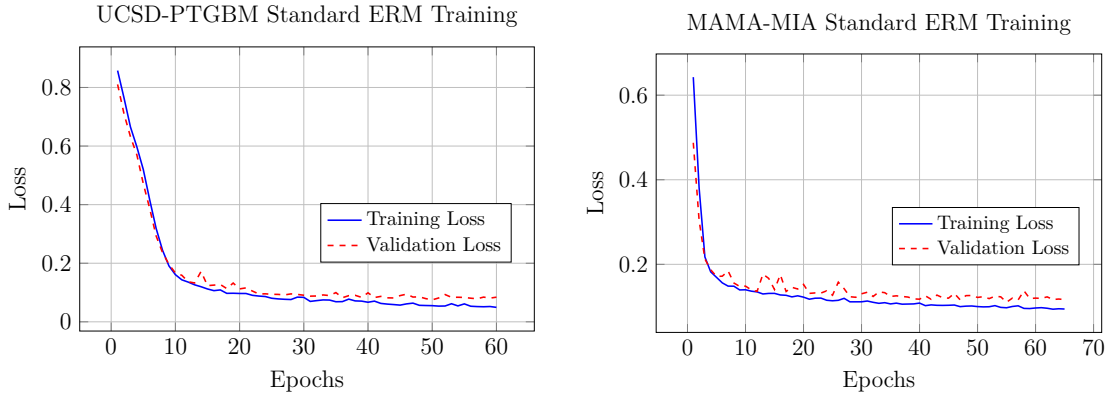
Figure 4.3: Dice coefficient progression over a subset of training epochs (including the optimal epoch per setup) on the UCSD-PTGBM and MAMA-MIA dataset using randomly shuffled buckets. The global average is shown in blue, while the mean trajectories for the Bottom-5 and Top-5 subgroups are displayed in red and green, respectively. Group assignments for Top-k and Bottom-k rely on the performance recorded at epoch n_0

In summary, the experimental evaluation showcased the limitation of ERM in mitigating performance disparities among subgroups. The results demonstrate that Curriculum learning paradigms can overcome this aspect, in particular implementing Divergence-based variant. Preview Div-CL has been proved to not only improve the computational cost and the general performance, but also outperform Div-CL in equalizing performance across all population subsets, preventing potential demographic biases within the model.

Additional investigations on this topic were conducted by Mantione [62].

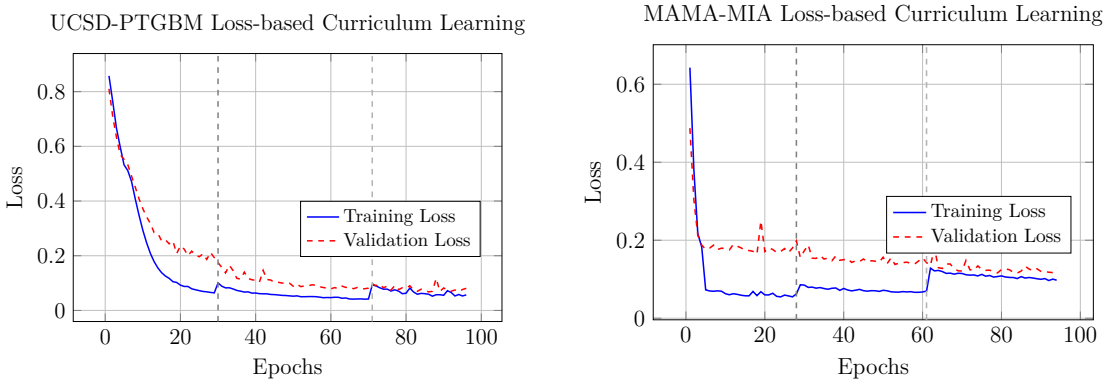
4.5 Loss curves

In this section, for each model discussed 4.3, the loss curves obtained during training are analyzed. The objective is to evaluate the convergence of the model, analyze its speed and when it occurs, and identify any instability or overfitting phenomena during training. For each experiment, the training loss and the validation loss are reported as a function of epochs, for all the tested methodologies. The configuration for the methodologies Preview Div-CL and Preview Loss-CL is `Config-Ref`$_{4.2}$.



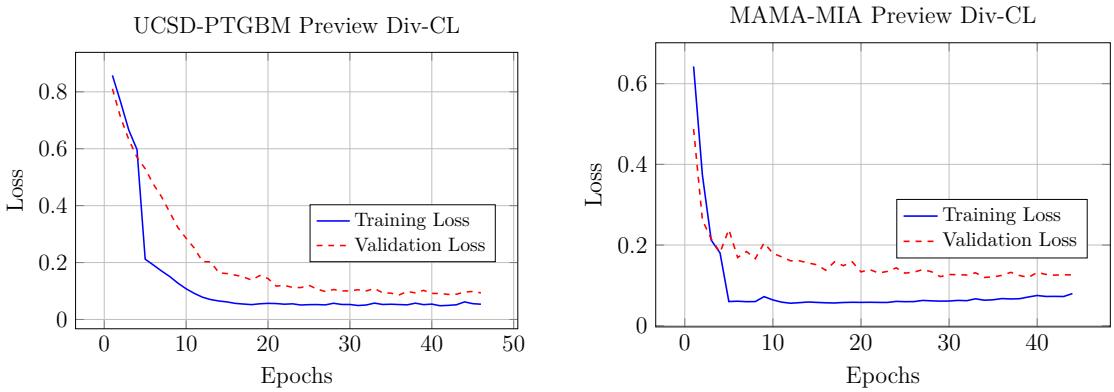
(a) UCSD-PTGBM **(b)** MAMA-MIA

Figure 4.4: Training and validation loss curves on the UCSD-PTGBM and MAMA-MIA datasets using a standard Empirical Risk Minimization (ERM) approach. The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs.



(a) UCSD-PTGBM **(b)** MAMA-MIA

Figure 4.5: Training and validation loss curves on the UCSD-PTGBM and MAMA-MIA datasets using a Loss-based Curriculum Learning approach. The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. The vertical dashed lines indicate the epochs at which medium and hard samples are introduced.

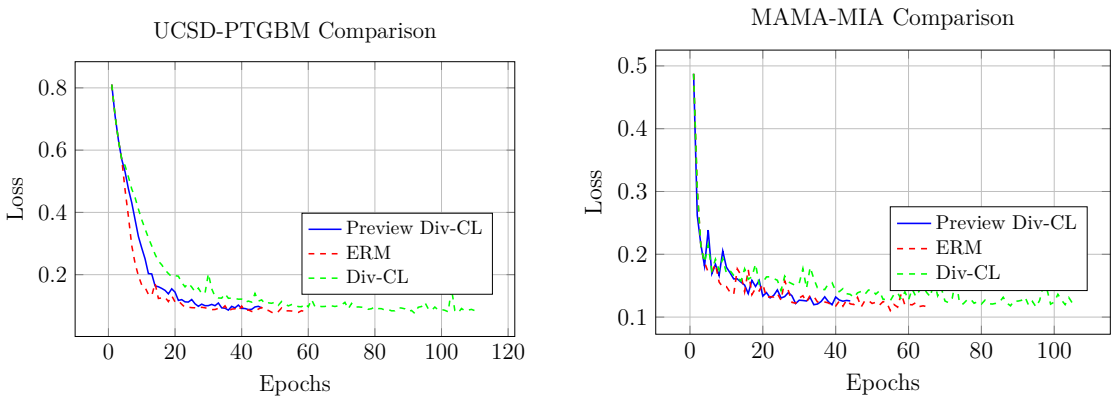


(a) UCSD-PTGBM **(b)** MAMA-MIA

Figure 4.6: Training and validation loss curves on the UCSD-PTGBM and MAMA-MIA datasets using divergence-based curriculum learning with preview strategy (Preview Div-CL). The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs.

Considering Figure 4.7a, different methods present different optimization dynamics. Standard training via Empirical Risk Minimization (ERM) starts aggressively, reducing errors rapidly within the first 15 to 20 epochs. It is plausible to assume that this sudden drop at the start introduces instability and micro-oscillations during the plateau phase, preventing the early stopping mechanism from activating and leading the model to stabilize early in suboptimal local minima. In contrast, Preview Div-CL converges to a similar loss in fewer epochs, despite having a slower initial descend. This suggests that the gradient of Preview Div-CL drives optimization towards more favorable loss landscape regions, guaranteeing greater stability and avoiding premature convergences into suboptimal minima.

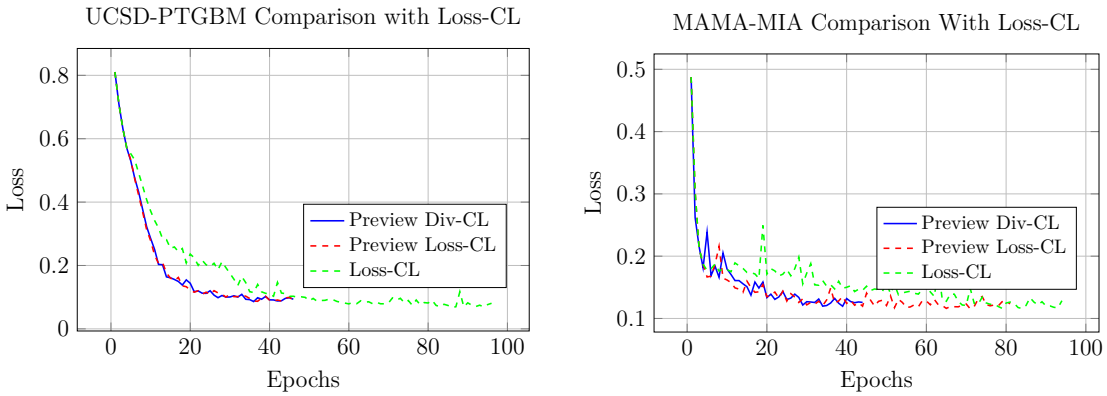
For the MAMA-MIA dataset, Figure 4.7b, both Preview Div-CL and ERM curves seems to be comparable, both reaching small loss value in few epochs, stabilizing themselves around epoch 25. Focusing on the Div-CL curve, even though it descends rapidly at the start, it settles to a higher loss compared to the other two methods (between 0.13 and 0.18) and showing evident instability lasting until the end of the graph.



(a) UCSD-PTGBM **(b)** MAMA-MIA

Figure 4.7: Comparison of validation loss curves on the UCSD-PTGBM and MAMA-MIA datasets across three different training strategies. The plots illustrate the performance of the Preview Div-CL method (solid blue line) against the standard Div-CL approach (dashed green line) and the baseline ERM training (dashed red line).

To better analyze the impact of the weighted ramps, Figure 4.8a and Figure 4.8b present a comparison of the training loss for methods using the Preview strategy against those without it. Compared to standard Loss-CL, it is evident how the integration of this approach considerably accelerates the early stages of training and guarantees convergence to a lower and more stable minima. The adoption of this strategy allowed both Divergent-based and Loss-based curriculum learning to overcome the severe oscillations of the standard methods, achieving faster convergence and a more generalized learning process for both considered datasets.



(a) UCSD-PTGBM **(b)** MAMA-MIA

Figure 4.8: Comparison of validation loss curves on the UCSD-PTGBM and MAMA-MIA datasets for different Curriculum Learning strategies. The plots contrast the performance of Preview Div-CL (solid blue line) against two loss-based approaches: Preview Loss-CL (dashed red line) and standard Loss-CL (dashed green line).

4.6 Ablation and Sensitivity analysis

4.6.1 Sensitivity analysis of ramp values

This section analyzes the sensitivity of the model to hyperparameter variations, specifically focusing on ramp configurations for the preview strategy: sharpness (slope), center and center decrements when the patience value is reached. The objective is to evaluate the robustness of the system and identify the optimal configurations for the two datasets under consideration: MAMA-MIA and UCSD-PTGBM.

A two-stage sequential optimization strategy was adopted. In the first phase, a standard configuration for ramp slopes was kept fixed (Ramp config 1) in order to analyze the sensitivity of the model to the positioning of the Center parameter. The aim was to identify the best trade-off between segmentation performance (Dice Score) and computational cost (number of training epochs). In the second phase, by fixing the identified optimal center, a study was carried out to improve the ramp sharpness and the dynamic decrements.

The investigation for the centers of the weighted ramps (Tables 4.3 e 4.4) shows a strong positive correlation between the center and the number of epochs required to reach convergence. To measure the benefit of each configuration, we evaluated the marginal gain performance against the computational cost increment.

For the UCSD-PTGBM dataset, going from Center = 30 (0.8433, 46 epochs) to Center = 35 leads to the peak performance (0.8750). However, the absolute increase of 3.7% in the Dice score requires more than double the training epochs (from 46 to 96). Considering this trade-off, a center value of approximately 30 was identified as the optimal choice, given the associated low computational cost.

For the MAMA-MIA dataset, *center* = 25 reach a competitive Dice score (0.7705) in 44 epochs Adjusting the center parameter to 35 results in a marginal improvement in Dice score (+0.004), but it comes at a considerable computational cost (55 epochs). The best performance (0.776) is reached when the center parameter is set to 20, but the associated 61 epochs of training make it not optimal.

The analysis of Figure 4.9 shows the relationship between performance (Dice Score) and computational cost (Training Epochs) for both datasets, with the Pareto-optimal configurations marked by the red dashed frontier lines. By applying the elbow method to these frontiers, is possible to identify configurations that offer an optimal balance.

In order to analytically locate the elbow point and determine the optimal trade-off, we compute the **Marginal Gain (GM)** between adjacent configurations using the following equation:

$$GM = \frac{\Delta\text{Dice Score}}{\Delta\text{Epochs}} = \frac{\text{Dice Score}_{i+1} - \text{Dice Score}_i}{\text{Epochs}_{i+1} - \text{Epochs}_i} \tag{4.2}$$

For the MAMA-MIA dataset, the elbow point is at Center = 25. Similarly, for the UCSD-PTGBM dataset, the optimal compromise is achieved at Center = 30. At these specific configurations, the best balance between maximizing the Dice score and minimizing computational resources is observed. Beyond these points, any additional gain in performance decreases significantly and does not justify rapid computational cost.

Table 4.3: Sensitivity studies on the MAMA-MIA dataset for ramp centers, evaluated using $\texttt{Config-Ref}_{4.2}$. Configurations are ordered by increasing training Epochs to emphasize the performance (Dice Score) trade-off. The chosen elbow point is indicated in bold text.

Ramp centers	Epochs	Dice Score	Marginal Gain (GM)
15	32	0.7523	1.52×10^{-3}
25	**44**	**0.7705**	$\mathbf{9.10 \times 10^{-4}}$
45	53	0.7787	-2.12×10^{-3}
35	55	0.7745	1.03×10^{-4}
30	58	0.7748	6.97×10^{-4}
20	61	0.7769	∞
40	61	0.7783	-

Table 4.4: Sensitivity studies on the UCSD-PTGBM dataset for ramp centers, evaluated using $\texttt{Config-Ref}_{4.2}$. Configurations are ordered by increasing training Epochs to emphasize the performance (Dice Score) trade-off. The chosen elbow point is indicated in bold text.

Ramp centers	Epochs	Dice Score	Marginal Gain (GM)
15	27	0.8197	1.24×10^{-3}
30	**46**	**0.8433**	$\mathbf{5.55 \times 10^{-4}}$
25	59	0.8505	6.96×10^{-4}
45	68	0.8568	-1.62×10^{-4}
20	77	0.8553	2.03×10^{-3}
40	80	0.8614	8.99×10^{-4}
35	96	0.8758	-

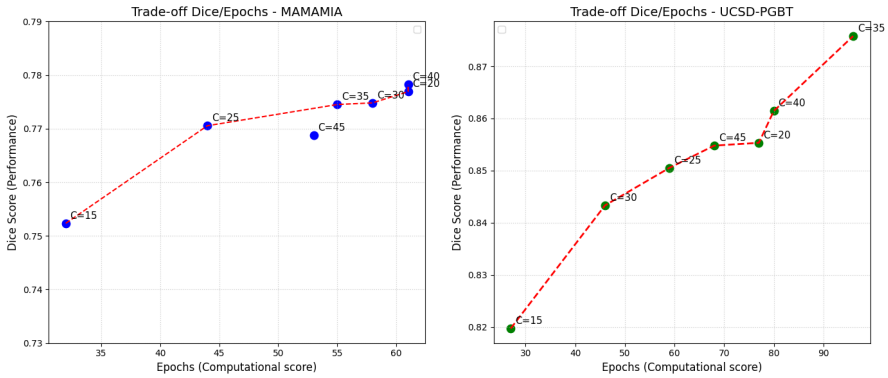


Figure 4.9: Pareto frontier of Dice Score vs. Training Epochs. The red dashed lines indicate optimal configurations, with elbow points identifying the best performance-to-cost trade-off at Center = 25 for the MAMA-MIA dataset and Center = 30 for the UCSD-PTGBM dataset.

The values of 25 for MAMA-MIA and 30 for UCSD-PTGBM are selected as centers to balance performance and computational efficiency. Once the optimal center is identified, the second stage evaluates the sensitivity of the model to variations of the ramp slope and the decrement parameters (Tables 4.5 e 4.6). In this fine-tuning phase, different behaviors were observed across datasets.

MAMA-MIA has been found to be more sensitive to sudden transitions: using steeper ramps (e.g., Configuration 12-12-8) degrades performance compared to smoother transitions (e.g., Configuration 6-6-4). In contrast, UCSD-PTGBM proves to have a greater stability, maintaining very similar Dice scores (≤ 0.005 fluctuations) when the ramp slope is moderate (e.g., `ramp_slope_e` = 6 and `ramp_slope_e` = 12), while showing slight variations in convergence times. Nevertheless, introducing a more extreme slope (e.g., `ramp_slope_e` = 16) compromises the training duration, leading to a higher number of required epochs, and combining it with higher decrements reduces the dice score.

As before, the configuration that better handles the trade-offs is chosen by selecting the setup right before the Marginal Gain GM drops, ensuring high accuracy without excessive computational cost.

Table 4.5: Sensitivity studies on the MAMA-MIA dataset for different ramp configurations, keeping the ramp center fixed at 25. Configurations are ordered by increasing training Epochs to emphasize the performance (Dice Score) trade-off. The chosen setup is indicated in bold text.

Ramp slope e	Ramp slope m	Ramp slope h	Decr m	Decr h	Dice score	Epochs	Marginal Gain (GM)
12	12	8	34	25	0.7541	35	2.15×10^{-3}
12	12	8	20	10	0.7626	39	1.58×10^{-3}
6	**6**	**4**	**20**	**10**	**0.7705**	**44**	$\mathbf{-4.01 \times 10^{-3}}$
6	6	4	34	25	0.7665	45	3.72×10^{-4}
16	16	12	20	10	0.7762	71	9.60×10^{-5}
16	16	12	34	25	0.7767	76	-

Table 4.6: Sensitivity studies on the UCSD-PTGBM dataset for different ramp configurations, keeping the ramp center fixed at 30. Configurations are ordered by increasing training Epochs to emphasize the performance (Dice Score) trade-off. The chosen setup is indicated in bold text.

Ramp slope e	Ramp slope m	Ramp slope h	Decr m	Decr h	Dice score	Epochs	Marginal Gain (GM)
16	16	12	34	25	0.8363	44	3.50×10^{-3}
6	**6**	**4**	**20**	**10**	**0.8433**	**46**	$\mathbf{2.07 \times 10^{-4}}$
12	12	8	34	25	0.8446	52	5.94×10^{-4}
6	6	4	34	25	0.8475	57	5.40×10^{-4}
12	12	8	20	10	0.8486	59	-1.11×10^{-3}
16	16	12	20	10	0.8408	66	-

4.6.2 U-Net and SegResNet comparison

To evaluate the impact of network structures on technique efficacy, comparative experiments were performed on both *U-Net* and *SegResNet*. Both models were trained and tested under the same methodology, including Empirical Risk Minimization, Div-CL, Loss-CL, Preview Div-CL, and Preview Loss-CL, with identical experiments conducted for each.

As shown by the results in Table 4.7, SegResNet proves to be the best architecture when considering Dice score, outperforming U-Net across majority of the evaluated configurations. Starting from ERM, SegResNet achieves a Dice Score of 0.78938

compared to U-Net's 0.7757 on MAMA-MIA dataset. This performance gap becomes even more evident on the UCSD-PTGBM dataset where SegResNet reaches a Dice Score of 0.8840, surpassing U-Net (0.8609) by a margin of over 2.3%. Furthermore, SegResNet achieves the highest absolute segmentation performance reported in our experiments, reaching a peak Dice Score of 0.88403 with Loss-CL on the UCSD-PTGBM dataset.

Conversely, when evaluating convergence efficiency in terms of training epochs, U-Net demonstrates a clear advantage. For the majority of the tested methodologies, U-Net requires significantly fewer epochs to converge compared to SegResNet. On average, U-Net reduces training time by roughly 22% compared to SegResNet. The reduction gap peaks under ERM on MAMA-MIA dataset, where U-Net reaches convergence in just 65 epochs compared to 99 for SegResNet. A comparable result appears with Loss-CL on the UCSD-PTGBM dataset, where 96 epochs are needed against 129, corresponding to a 25% decrease.

Div-CL on the UCSD-PTGBM dataset and Preview Loss-CL on MAMA-MIA dataset can be considered the only exceptions on this trend. In the first example SegResNet converges in just 83 epochs, roughly 25% faster than the 111 epochs required by U-Net. Similarly, for Preview Loss-CL on MAMA-MIA, SegResNet reaches convergence in 68 epochs compared to the 81 epochs of U-Net, a $\sim$16% reduction, while maintaining a nearly identical Dice Score.

Table 4.7: Comparison of performance (Dice Score) and convergence (Epochs) between U-Net and SegResNet on two datasets, across different applied methods. Bold values indicate the best Dice Score for each row.

Dataset	**Method**	**U-Net**		**SegResNet**	
		Dice Score	**Epochs**	**Dice Score**	**Epochs**
MAMA-MIA	ERM	0.7757	65	0.7893	99
	Loss-CL	0.7760	94	0.7831	124
	Preview Loss-CL	0.7780	81	0.7752	68
	Div-CL	0.7770	105	0.7809	125
UCSD-PTGBM	ERM	0.8609	60	0.8840	79
	Loss-CL	0.8750	96	0.8840	129
	Preview Loss-CL	0.84255	47	0.8518	51
	Div-CL	0.861	111	0.8698	83

The analysis of the data reported in 4.8, conducted using the best centers and configuration $\texttt{Config-Ref}_{4.2}$ for each respective dataset, shows a significant benefit provided by the *Preview Div-CL* approach compared to the baseline, notable in both models tested. Considering absolute performance, SegResNet achieves a higher Dice Score in both datasets. The use of residual connections leads to a consistent

improvement in Dice Score compared with *U-Net*, with an absolute increase of approximately 0.01 on the MAMA-MIA dataset, from 0.7705 to 0.7803, and of 0.011 on the UCSD-PTGBM dataset, from 0.8433 to 0.8547. However, the most significant finding emerges observing the reduction of training epochs compared to the baseline configurations (Δ_{Epoch}). While U-Net achieves substantial reductions in the number of training epochs, 32.3% on MAMA-MIA and 23.3% on UCSD-PTGBM, SegResNet demonstrates an even greater reduction in computational cost, reaching 36.4% on MAMA-MIA and an exceptional 41.8% on UCSD-PTGBM. The gain in training efficiency comes with a minor trade-off (Δ_{Dice}), with both models showing limited Dice Score decreases ranging from -0.67% to -3.4%. Finally, analyzing the absolute values, it becomes evident that the computational cost varies across architectures and datasets. On UCSD-PTGBM, both models converge at the same epoch (46), whereas on MAMA-MIA, SegResNet requires 19 more epochs to reach convergence (63 vs. 44 epochs).

Table 4.8: Performance comparison of the Preview Div-CL approach using U-Net and SegResNet on the MAMA-MIA and UCSD-PTGBM datasets. The table reports the Dice Score, number of training epochs, and their relative percentage variations (Δ) compared to the standard Empirical Risk Minimization (ERM). Ideal behavior corresponds to a large positive Δ_{Dice} and a large negative Δ_{Epoch}.

	MAMA-MIA				**UCSD-PTGBM**			
Model	**Dice**	$\Delta_{\textbf{Dice}}$	**Epochs**	$\Delta_{\textbf{Epoch}}$	**Dice**	$\Delta_{\textbf{Dice}}$	**Epochs**	$\Delta_{\textbf{Epoch}}$
U-Net	0.7705	-0.67%	44	-32.3%	0.8433	-2.00%	46	-23.3%
SegResNet	0.7803	-1.17%	63	-36.4%	0.8547	-3.40%	46	-41.8%

Regarding subgroup fairness, Table 4.9 compares the U-Net and SegResNet architectures by showing how the gap between the average Dice score of Top-5 groups and that of the Bottom-5 changes as training progresses. The results demonstrate that, while both models can effectively reduce this performance disparity, they differ in their consistency.

SegResNet proves to be overall more robust and consistent across configuration changes, the gap reduction never drops below 62% in any of the tested scenarios. In contrast, U-Net proves to be significantly less stable across different configurations, achieving a gap reduction of only 29% with ERM on the MAMA-MIA dataset.

However, the analysis reveals a complex dynamic in which standard ERM approaches and those based on Curriculum Learning (CL) are effective in different and complementary scenarios. While ERM achieves the best gap reductions in specific configurations, U-Net on UCSD-PGTBM (-84%) and SegResNet on MAMA-MIA (-78%), CL methods prove indispensable for mitigating severe architectural vulnerabilities. When ERM underperforms significantly, as in the case of U-Net on

MAMA-MIA (29%), strategies such as Preview Div-CL bring the final gap to the lowest absolute value recorded in all tests (0.034).

Although the Dice score is a standard reference metric, relying solely on it provides a limited view of the system's actual capabilities. It is therefore necessary to perform a comprehensive evaluation considering metrics such as global Shapley values and subgroup identification. This analysis was conducted in «Analysis of Subgroup-Aware Curriculum Learning Strategies for Medical Image Segmentation» [62]

Table 4.9: Comparison of the performance gap reduction between top-5 and bottom-5 subgroups from the beginning to the end of training. Values are reported in the format *Initial Gap → Final Gap (Percentage Reduction)*. The best results for each architecture are displayed in bold.

Architecture	Method	MAMA-MIA	UCSD-PGTBM
U-Net	ERM	0.100 → 0.071 (-29%)	0.314 → **0.051** (**-84%**)
	Loss-CL	0.111 → 0.044 (-60%)	0.316 → 0.065 (-79%)
	Div-CL	0.113 → 0.062 (-45%)	0.274 → 0.056 (-80%)
	Preview Div-CL	0.206 → **0.034** (**-83%**)	0.320 → 0.096 (-70%)
SegResNet	ERM	0.179 → **0.039** (**-78%**)	0.205 → 0.071 (-65%)
	Loss-CL	0.184 → 0.070 (-62%)	0.370 → 0.049 (**-87%**)
	Div-CL	0.216 → 0.067 (-69%)	0.285 → **0.043** (-85%)
	Preview Div-CL	0.167 → 0.064 (-62%)	0.300 → 0.091 (-70%)

4.6.3 Preview strategy: Single vs. Three-ramp approach

The underlying hypothesis behind this experimental setup is that (1) employing a distinct pacing ramps for the easy bucket, where the weights this samples is steadily decreased over time, and (2) designing the sigmoid slope for hard samples to be less steep than the medium counterpart, can facilitate the learning process by introducing the most complex data at a slower rate.

Using Preview Div-CL as base method with $\texttt{Config-Ref}_{4.2}$ configuration, two evaluations were conducted to validate this assumption. The first experiment evaluates the effect of progressively decreasing the weight of easy samples, comparing it against a single-ramp baseline that does not distinguish between sample difficulties. The second experiment isolates the second part of the hypothesis, comparing the proposed approach against a configuration where the medium and hard ramps share the same steepness (equal slopes), designed to determine whether the slower introduction of hard samples is genuinely effective.

When comparing the use of three uniform ramp slopes with different ramp sharpness values across difficulty levels, an important distinction emerged. Using a lower initial ramp sharpness combined with a smaller dynamic decrement for the "high" difficulty ramp resulted in the most substantial reduction in the number of epochs for both the MAMA-MIA and UCSD-PTGBM datasets. As shown in Table 4.10, while this introduced to a very minor trade-off in the Dice score for UCSD-PTGBM compared to the uniform configuration, it produced a notable improvement in the Dice score on the MAMA-MIA dataset (from 0.7681 to 0.7705). These results support the initial hypothesis, demonstrating that progressively reducing the weight of easy samples while introducing hard samples at a slower rate optimizes the learning process. Compared to traditional training (ERM), this setup leads to a faster convergence reducing epochs from 65 to 44 for MAMA-MIA, and from 60 to 46 for UCSD-PTGBM, while maintaining highly competitive segmentation performance with a marginal Dice score loss of less than 0.35% and 1.55%, respectively.

Although single-ramp and uniform three-ramp configurations perform worse than the final method, the study demonstrates that even these simpler configurations significantly improve upon standard curriculum strategies that lack these mechanisms, such as Div-CL. Implementing a simple weight ramp drastically reduces the computational cost compared to the Div-CL baseline 4.2, dropping the number of epochs from 115 to 62 for MAMA-MIA and from 111 to 63 for UCSD-PTGBM. Furthermore, adopting a uniform three-ramp configuration accelerates convergence even further to 48 and 58 epochs, respectively, proving that the fundamental principle of progressively reducing sample weights optimizes the learning process far better than standard Div-CL.

These results support the initial hypothesis, demonstrating that progressively reducing the weight of easy samples while introducing hard samples at a slower rate optimizes the learning process. Consequently, this strategy achieves significantly faster convergence while maintaining, and potentially even improving, segmentation performance.

4.6.4 Preview strategy: Static vs. Dynamic ramps

Table 4.11 presents a comparison between static and dynamic ramps on the MAMA-MIA and UCSD-PTGBM datasets, allowing us to analyze the impact of dynamic centers on the setup. Both tests were conducted using the `Config-Ref`$_{4.2}$ configuration. For the dynamic ramps, the "patience" parameter (i.e., the number of epochs before a decrement occurs) was set to 5, while the center decrease values (`decr_m` and `decr_h`) were 20% for the medium ramp and 10% for the easy ramp.

For the MAMA-MIA dataset, the dynamic setup outperforms the static one on both metrics considered, achieving a higher Dice score (0.7705 vs. 0.7645) and

Table 4.10: Comparison of ramp configurations in terms of model performance (Dice score) and convergence speed (Epochs) for the MAMA-MIA and UCSD-PTGBM datasets. The 3-ramp setups are evaluated using both different slopes (hard and medium have distinct steepness) and equal slopes.

Dataset	**Configuration**	**Dice Score**	**Epochs**
MAMA-MIA	1 ramp	0.7732	62
	3 ramps (uniform slopes)	0.7681	48
	3 ramps (different slopes)	0.7705	44
UCSD-PTGBM	1 ramp	0.8563	63
	3 ramps (uniform slopes)	0.8479	58
	3 ramps (different slopes)	0.8433	46

converging one epoch earlier (44 vs. 45).

On the UCSD-PTGBM dataset, the effect of the dynamic strategy is more striking. While costing around ∼1.3% on the Dice score performance, the reduction of the number of epochs needed by the model to reach convergence was nearly 25% (46 vs. 60)

These results shows that the dynamic ramp strategy benefits both datasets in terms of faster convergence, and the higher Dice score achieved for MAMA-MIA may suggest that this dynamic adjustment could contribute to a more effective learning process for this dataset.

Table 4.11: Comparison of dynamic and static ramp configurations on MAMA-MIA and UCSD-PTGBM

	MAMA-MIA		**UCSD-PTGBM**	
Configuration	**Dice Score**	**Epochs**	**Dice Score**	**Epochs**
Static	0.7645	45	**0.8570**	60
Dynamic	**0.7705**	**44**	0.8433	**46**

4.7 Alternative approaches and exploratory surveys

4.7.1 Adding other metrics in DivExplorer

To evaluate the performance of the model, extract subgroups, and compute their divergence using `DivExplorer` [4], different sets of metrics were selected. During this study, the Dice score is employed as the primary performance metric and fixed baseline for results comparison. In this experimental setup, we introduced a secondary metric which was provided to DivExplorer along with Dice score in order to guide the identification of relevant subgroups.

Confidence

Let there be a neural network-based classifier trained on a set of classes and a dataset composed of observations. For each input, the model computes a probability distribution over the possible classes, typically using a softmax activation function. The Confidence [63] for a single pixel is defined as the maximum predicted probability among all available classes:

$$C(x) = \max_{c \in \mathcal{C}} P(\hat{y} = c \mid x) \tag{4.3}$$

The Mean Max Confidence (MMC) is the average of those peak probabilities, computed across all pixels (or spatial elements) of a single sample (image):

$$MMC = \frac{1}{N} \sum_{i=1}^{N} \max_{c \in \mathcal{C}} P(\hat{y} = c \mid x_i) \tag{4.4}$$

Where N denotes the total number of pixels in the image, $\mathcal{C}$ is the set of all classes predicted by the model, and $P(\hat{y} = c \mid x_i)$ is the probability predicted by the model for class c given pixel x_i.

A value of MMC near 1 indicates that the model is very confident in its prediction and the uncertainty is low.

Throughout this work, the term *confidence* will be used to refer specifically to the Mean Max Confidence.

HD95

The Hausdorff distance [64] is a metric that measures the similarity between two sets of points by quantifying the maximum distance between any point in one set and its nearest neighbor in the other set. It is particularly useful in image comparison and shape matching, as it does not require explicit point correspondences and is robust to small perturbations. Hausdorff distance is defined as the maximum of two directed distances, one from the first set to the second, and vice versa.

Formally, given two non-empty sets of points A and B, the distance is defined as the maximum of two directed distances: one from the first set to the second, and vice versa:

$$d_H(A, B) = \max(h(A, B), h(B, A)) \tag{4.5}$$

Where $h(A, B)$ is the directed distance from set A to set B, expressed as:

$$h(A, B) = \max_{a \in A} \min_{b \in B} d(a, b)$$

And $d(a, b)$ represents the underlying distance metric (e.g., Euclidean distance) between points a and b.

The standard Hausdorff distance is inherently highly sensitive to outliers. To mitigate this issue, the 95th percentile Hausdorff Distance (HD95) is widely adopted in medical image segmentation.

Instead of taking the absolute maximum distance, HD95 evaluates the 95th percentile of the nearest-neighbour distances, effectively excluding the top 5% of the largest positional errors. This makes the metric more robust to small perturbations and noisy predictions, providing a more reliable estimation.

Entropy

The Entropy, or Shannon Entropy, is a concept that stems from information theory that measures the uncertainty of a given variable. In this work, entropy is computed for every pixel directly from the neural network's raw outputs (logits). Given the estimated probability p_c that a spatial pixel x belongs to class $c \in \{1, 2, \ldots, C\}$, the Shannon entropy $H(x)$ is calculated as:

$$H(x) = -\sum_{c=1}^{C} p_c \log(p_c + \epsilon) \tag{4.6}$$

Where $\epsilon = 10^{-8}$ is a near-zero constant added to the computation to ensure numerical stability, preventing undefined logarithmic operations when a class probability is exactly zero. This method produces a spatial entropy map, with $H(x) \approx 0$ values denoting high confidence and higher scores indicating degrees of ambiguity in prediction across multiple classes. To extract a single global value for each image, this pixel-wise map is averaged into a scalar value H_img:

H_{img}:

$$H_{img} = \frac{1}{N} \sum_{i=1}^{N} H(x_i)$$

Where N is the total number of pixels.

Results

Loss Function	Dice Score	Epochs
Dice	0.7774	105
Dice + Confidence	0.7794	133
Dice + Hausdorff	0.7785	104
Dice + Entropy	0.7650	74

Table 4.12: Comparison of Dice Score and training epochs for different metrics in subgroup extraction.

As shown in 4.12, the baseline standard, using only the Dice Score, performed at 0.7774 across 105 epochs. Adding the *Confidence* metric led the model to achieve the best Dice score in this testing configuration (0.7794); however this addition comes with an important computation cost increase of 28 epochs (133 epochs) The inclusion of the *Hausdorff distance* (HD95) resulted in minimal improvements in performance without affecting convergence time (104 epochs).

The most convincing result in terms of training efficiency was the combination of Dice score and *Entropy*, cutting the training time to 74 epochs (∼30% less than

the baseline). Nonetheless, this notable improvement in speed led to a minor drop in overall segmentation performance, resulting in a Dice score of 0.7650.

This study demonstrates that incorporating uncertainty-based metrics, like *Entropy*, into the subgroup identification process can effectively bring the model toward a faster convergence, but it also results in a slight degradation of the overall Dice score, reaching the lowest 0.7650.

While the introduction of Entropy alongside the Dice score demonstrated a notable 30% reduction in training epochs, this exploratory approach was not implemented in the final method. During the development phase, the research shifted to a different strategy that was demonstrated to be more efficient in terms of computational cost and performance. Although the two solutions may be compatible, their integration would require extensive hyperparameter tuning and an additional set of experiments that were not feasible due to time constraints.

4.7.2 Variant of preview strategies

Other ramps: cyclical linear and cosine .

During the design and the development phase of the Preview Div-CL method, other form of ramps for weight management were investigated. In particular, a cyclical approach was explored, inspired from by the Cyclical Annealing Schedules [33]. The initial hypothesis was that following a periodic trend characterized by a cycle of repeated phases of growth and a plateau, could favor the model's performance. Based on this paradigm, different type of mathematical functions for the ramp were designed and evaluated.

Initial experiments involved the use of a cyclic linear ramp, followed by a cosine variant. However, these experiments did not show significant improvements: the cyclic linear ramp achieved a score of 0.7752 on MAMA-MIA in 91 epochs, only slightly better than to the Div-CL baseline without preview strategy. Given these suboptimal results, it was proposed that one possible constraint might originate from the inherent rigidity of these functions, preventing the steepness of the growth phase from being independently adjusted. To test this hypothesis, a cyclic sigmoid ramp was introduced. The addition of a sharpness parameter offers more control over the steepness of the transition, allowing for a better fit the learning process. Although the sigmoid proved to be the most geometrically flexible solution among those proposed, the testing phases revealed a structural limitation of the cyclic approach itself. In the sigmoid case, cyclic solutions consistently returned far worse results than a non-periodic approach. It was then decided to discard the cyclical setting, maintaining only the geometric formulation of the sigmoid declined in a monotone ramp. This methodological approach clarifies and justifies why the quantitative analysis, presented in the first part of this chapter, focused on the non-cyclic exclusively configuration.

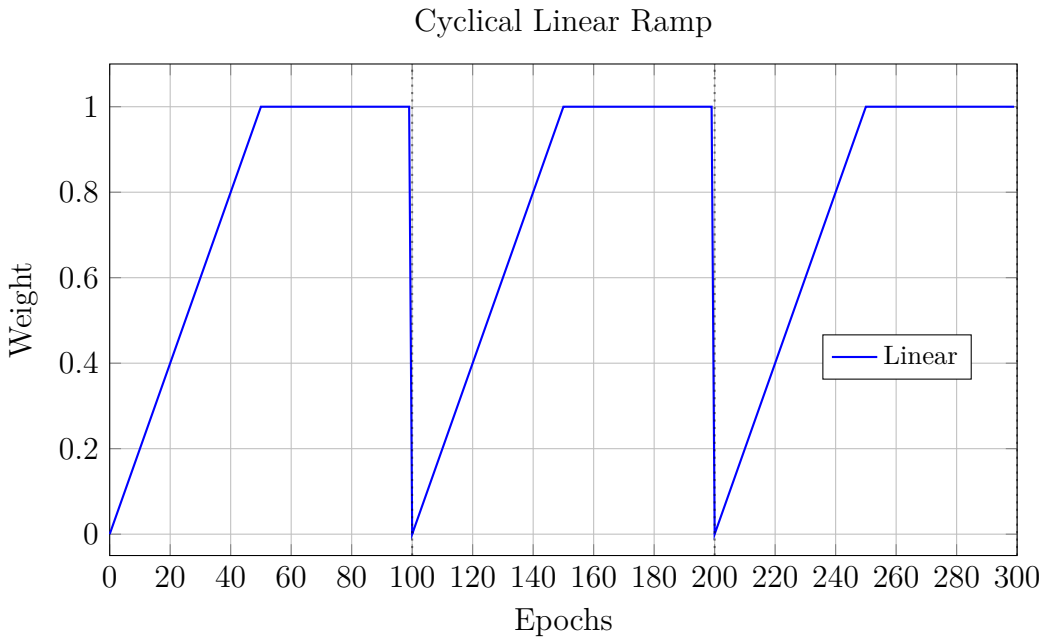


Figure 4.10: Cyclical linear ramp function over three cycles, initially evaluated for weight management within the preview strategy.

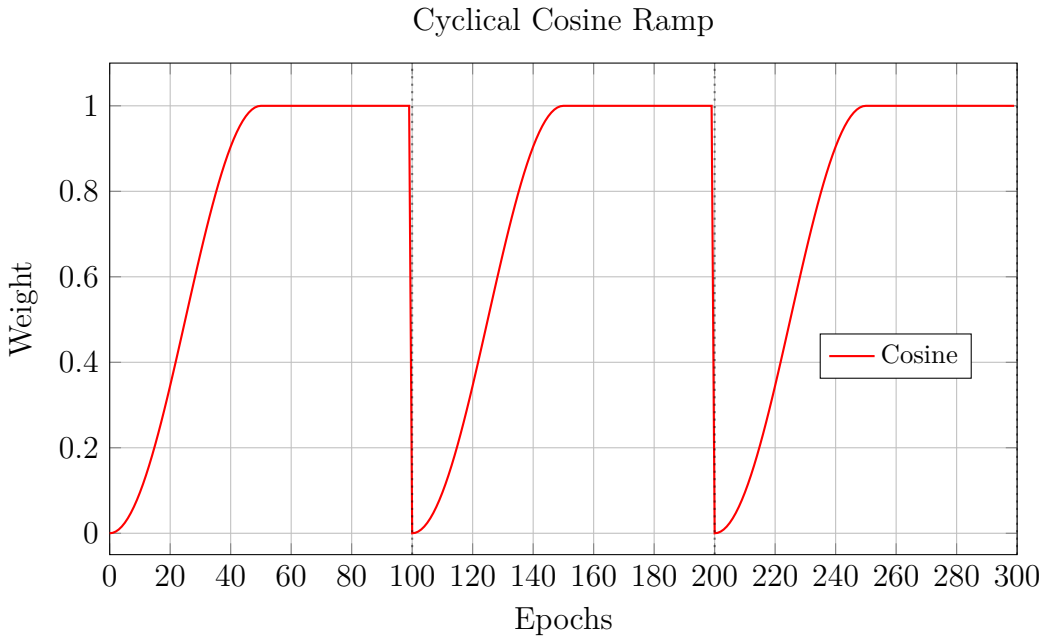


Figure 4.11: Cyclical cosine ramp function over three cycles, tested as a periodic variant for the learning process.

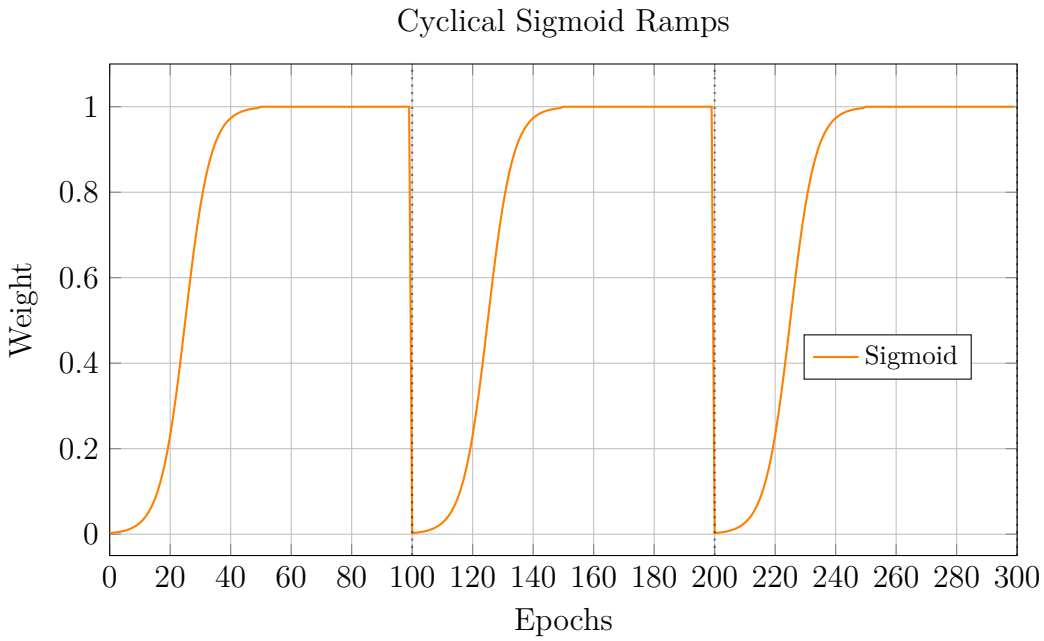


Figure 4.12: Cyclical sigmoid ramp function over three cycles. The sharpness parameter is set to 12 to provide independent control over the transition steepness.

4.8 Discussion

The results presented in the previous sections confirm that the analyzed methodologies, based on Curriculum Learning, represent a competitive alternative for image segmentation techniques applied to annotated datasets such as MAMA-MIA and UCSD-PTGBM. In particular, the introduction of the preview strategy with the weighted ramps during training, allowed the model to achieve faster convergence while maintaining comparable performance, with only a marginal drop in Dice score in specific scenarios. Another important objective of this thesis was to verify if the proposed methodologies solutions could reduce the unfairness caused by performance discrepancies across different data subgroups. The experimental evaluation demonstrated that these approaches successfully narrowed the performance gap compared to the baseline model. The combination of reduced training time and improved robustness makes the proposed approach particularly promising for domains such as medical imaging, where computational resources and time-consuming fine-tuning phases are very frequent limiting factors. The following sections will present a detailed analysis of these results.

4.8.1 Results interpretation

Curriculum Learning approaches . The analysis of the Curriculum Learning (CL) methods reveals a marked trade-off between the quality prediction (Dice Score) and the computational cost (Number of Epochs). Although introducing a progressive approach to training theoretically facilitates model learning, the experiment results show a huge difference based on the difficulty metrics, and the dataset on which this model is applied. Considering performance into account alone, Loss-CL was the only method that scored a higher result compared to the ERM baseline, by $\sim$1.6%, but at the cost a more over 60% increase in computational cost, and this result was only achieved on UCSD-PTGBM. When evaluating the MAMA-MIA dataset, both Div-CL and Loss-CL reached slightly better or comparable results, at the cost of a significant increase in computational cost.

As expected, Loss-based curriculum learning performs worse than the baseline in terms of epochs and better in performance with a higher Dice Score. This could be related to the regularizing effect of the curriculum pacing. With the exposure to easier samples at the start, the model builds up a robust foundation that allows it to quickly identify the simpler morphological features of the samples. However, with the introduction of medium and hard samples as shown in Figure 4.5a and Figure 4.5b, the model learning trajectory gets distorted, forcing the network to spend additional epochs recovering. On the MAMA-MIA dataset, when more difficult data are forced into the batches (e.g., at epochs 30 and 61) it triggers violent spikes in the loss. A similar phenomenon occurs on the UCSD-PTGBM dataset, where noticeable turbulences can be detected around epochs 30 and 71.

For divergence-based curriculum learning, the marginal gains in performance may not be caused by the ineffectiveness of the difficulty measure, but rather from the static approach used to divide samples into categories. The current approach, which consists in dividing the N samples equally, is based on the incorrect assumption that the data are uniformly distributed. As originally proposed by Mantione[1], an analysis of the subgroups after the difficulty assignment step reveals the presence of multiple subgroups with similar or identical difficulty. Therefore, a fixed bucket size risks arbitrarily splitting homogeneous samples across different difficulty categories. This approach can undermine the gradual progression that characterizes curriculum scheduling. Future works should focus on the adoption on a data-driven dynamic partition.

Applying Preview-strategy . In order to improve the computational criticalities previously mentioned, the introduction of a preview strategy was evaluated.

[1]Silvia Mantione. «Analysis of Subgroup-Aware Curriculum Learning Strategies for Medical Image Segmentation». In preparation. MA thesis. Torino, Italy: Politecnico di Torino, 2026.

Incorporating this strategy into CL methods yielded substantial improvements in computational cost, achieving a strong reduction in training epochs compared to the standard Loss-CL and Div-CL approaches. In particular, when applied on Div-CL, it allowed the model to converge in just 44 and 46 epochs, cutting the training time by 59%, on both the MAMA-MIA and UCSD-PTGBM datasets, surpassing even the ERM baseline in terms of speed. While this accelerated convergence introduced a minor accuracy drop on the UCSD-PTGBM dataset (2% for Preview Div-CL and 3.7% for Preview Loss-CL compared to the ERM baseline), the performance is still competitive, having a favorable trade-off considering the computational cost reduction. In addition, this performance drop seems to be dataset-specific. When evaluated on the MAMA-MIA dataset, Preview Div-CL performance decreased by less than 1%, while Preview Loss-CL Dice score actually surpassed the ERM baseline, reaching the highest Dice score (0.7781), but cost 16 epochs more. The success of the method can be attributed to the controlled pacing introduced by the weighted ramps. Gradually exposing the model to samples, weighted on the basis of their difficulty using ramps of different sharpness (as shown in the previous experiments 4.3), overcomes the loss spikes observed before in standard curriculum learning and caused by the sudden introduction of harder samples, resulting in smoother and faster convergence. The relevance of divergence as a difficulty measure is confirmed by Preview Div-CL, which achieves the best overall result when considering the performance–computational cost trade-off.

Mitigating unfairness across subgroups .

A critical aspect mentioned before was the fairness related to discrepancy between demographic subgroups. When evaluating the U-Net architecture on the MAMA-MIA dataset, the baseline struggles to close the performance gap in some configurations, both Loss-CL and Div-CL exhibit a notable narrowing of this disparity by convergence time, despite some fluctuations in the initial phases. The results suggest the potential contribute of curriculum scheduling in mitigating bias, by discouraging the reliance on spurious 'shortcuts', forcing the model to learn causal features that are inherently more generalizable and complex. In particular, methods that used divergence as a metric showed even higher mitigation. This is likely due to divergence which, unlike standard loss, defines difficulty specifically based on the model's interaction with specific subgroups within the dataset, resulting in a more realistic measure that better reflects the true underlying data distribution. Experimental results also proved that *Preview Div-CL* enhances the mitigation properties of Div-CL.

Based on Figures 4.2 and 4.1, while both datasets reach a similar performance gap between the best five and worst five subgroups at the end of training, their initial scenario differs significantly. From the beginning, MAMA-MIA presents a relatively narrow gap, when UCSD-PTGBM starts with a much wider disparity.

Since both datasets converge to a similar final gap, UCSD-PTGBM seems to have benefit more from the curriculum learning strategy. This behavior can be attributed to the inherent characteristics of the two datasets. The MAMA-MIA dataset has more than 1200 clinical cases and presents a high multi-centric heterogeneity, including differences in acquisition protocols, scanners and breast tissue density. This variability may have prevented the model from over-optimizing on specific characteristics, keeping performance more uniform across subgroups. The UCSD-PTGBM dataset is smaller, with only 184 cases for training, so the model may initially favor "easier" cases, leading to greater difficulty on more complex ones.

The improvement and the stable convergence of the subgroups toward a uniform Dice score when preview strategy is applied (Figures 4.1a and 4.2a) suggest that scaling the loss promotes a more gradual optimization. As a result, this mathematically controlled learning mechanism effectively mitigates subgroup biases.

Generalization across architectures . Examining architectural differences allows us to assess the degree of generalizability of the proposed methods. Consistent results across both architectures would suggest that the observed performance originates from the method itself, rather than being a product of a particular network.

As shown in Table 4.7, the experiment confirms this hypothesis. Despite the distinct structural nature of the two networks, with SegResNet systematically outperforming U-Net in terms of Dice Score across both datasets, the effect of the proposed methods follows a uniform pattern across both architectures. Furthermore, it is noticeable that, in the majority of cases, such as ERM and Loss-CL on both datasets, SegResNet requires more epochs to converge than U-Net to achieve its higher Dice score. It is noteworthy to observe that across both architectures, the application of the Preview strategy resulted in a decrease in the number of training epochs required for convergence compared to the respective baseline method. The interaction between the architecture and method is not entirely consistent. Preview Loss-CL on MAMA-MIA, for example, can be considered a partial exception, as U-Net slightly outperforms SegResNet in performance despite converging more slowly; a similar pattern emerges for Div-CL on UCSD-PTGBM where SegResNet converge faster than U-Net. Focusing on Preview Div-CL (4.8), on the UCSD-PTGBM dataset both architectures converge in the exact same number of epochs, which is considerably lower than both the ERM baseline and standard Div-CL, with SegResNet maintaining a superior Dice Score. Moreover, on the MAMA-MIA dataset, SegResNet maintains this efficiency advantage, requiring marginally fewer epochs to converge compared to the ERM baseline, while remaining faster than the standard Div-CL.

With respect to mitigation effectiveness, according to Table 4.9, although the SegResNet architecture proves to be more stable than the U-Net, the minimum final

gaps between the Top-5 and Bottom-5 performing subgroups are largely comparable across the two architectures. On the MAMA-MIA dataset, the best result for U-Net (0.034, achieved with Preview Div-CL) closely matches that of SegResNet (0.039, achieved with ERM); similarly, on UCSD-PGTBM, the optimal values are 0.051 and 0.043, respectively.

In conclusion, these results suggest that the proposed methods are fundamentally architecture-agnostic. However, the absolute results in Dice Score and epochs denote a dependency on the network's topology, hence a broader range of neural networks is required to fully confirm the generalizability of these approaches.

4.8.2 Comparison with related works

.

The idea of a ramping schedule, derived from the preview strategy in [5], proved highly effective in the context of this study. Integrating this schedule into the core principle of curriculum learning approach, which divides the training samples according to their level of difficult, significantly reduced convergence time without compromising the advantages, such as model robustness and performance gains across subgroups.

In contrast to the advantages described by the approach proposed in [33], the results from Experiment 4.7.2 revealed that the cyclical schedule, as implemented within the context of this study, did not yield any significant improvements in performance when compared to a standard non-cyclical baseline. However, this comparison must be interpreted with caution, as the cyclical approach was employed following the standard bucket-based introduction schedule, where data was progressively fed to the model in discrete subsets, rather than exposing all the dataset from the start as preview strategy implies. Future investigations should isolate cyclical annealing from the curriculum scheduling to better evaluate its impact.

Chapter 5

Conclusion

5.1 Summary and main results

This thesis presented an extensive analysis of training strategies for medical image segmentation models. Starting from established methodologies such as ERM training and loss-based curriculum learning, the research also analyzed more experimental approaches, such as divergence-based curriculum learning. By employing these as baselines, a novel approach, Preview Curriculum Learning, was developed specifically designed to overcome the structural limitations exhibited by conventional methods. The experimental analyses conducted confirmed the validity and robustness of the proposed methodology on the tested dataset MAMA-MIA and UCSD-PTGBM. Considering the trade-off between computational cost reduction and performance loss, *Preview Div-CL* proved to be the most effective method when compared to ERM and standard curriculum learning configurations. Preview Div-CL reduced the number of epochs required for convergence by approximately 33% with respect to Empirical Risk Minimization, while maintaining comparable segmentation performance, with a marginal decrease of 2% on the UCSD-PGBT dataset and 0.66% on MAMA-MIA. In addition to this favorable trade-off, the curriculum learning approaches tested in this work successfully reduced disparities among subgroups associated with demographics and acquisition parameters, with all methods achieving strong results. This narrowing of the performance gap was most noticeable for ERM and Preview Div-CL. Nevertheless, Preview Div-CL is inherently interpretable because it relies on divergence as a metric of difficulty. This property qualifies this technique as the optimal choice for the medical field and other domains that strictly require interpretable machine learning models. Preliminary analyses, conducted on the most important components, revealed that the optimal configuration involves the use of dynamic ramps based on the difficulty level. These factors proved essential for faster convergence of the model.

5.2 Practical applications

The practical applications of the proposed method, Preview Div-CL, are found primarily in the field of medical image computing and the automatic segmentation of clinical images, where preserving interpretability alongside low computational costs is crucial.

The proposed approach achieved a 30% acceleration in convergence with only a marginal variation in the Dice Score (around 2 to 4%) when evaluated on the MAMA-MIA and UCSD-PTGBM datasets. This important trade-off creates opportunities for further research, as it could potentially allow the processing of massive datasets and the deployment of high-parameter models in resource-constrained environments, such as real-world healthcare facilities. Medical research centers and clinical institutions can rapidly train and validate their models on newly acquired patients without incurring in the usual high expenses related to large GPU clusters.

Beyond the computational efficiency and robustness provided by Curriculum Learning, the method possesses an inherent layer of interpretability, as it is built upon DivExplorer [4] and the concept of divergence. Consequently, it can potentially be applied to any dataset across different domains, as long as it is included with the metadata required by the algorithm.

As demonstrated in the original DivExplorer study [4], typical application examples include the evaluation of algorithmic fairness on socio-demographic data (such as the Adult Census dataset) and bias detection in judicial predictive models (e.g., the COMPAS dataset).

5.3 Limitation

- **Reliance on metadata availability**: The proposed methodologies inherently rely on the availability of rich population metadata, which are rarely available in public medical datasets.
- **Conservative worst-case difficulty assignment**: In cases where a sample is present in multiple overlapping subgroups, it is assigned the difficulty score of the most challenging subgroup. This pessimistic approach could lead to a distorted and inconsistent samples distribution across the curriculum stages.
- **Limited sensitivity analysis**: The experimental evaluation features a constrained sensitivity analysis due to time limits. For Preview Div-CL, The parameter tuning followed a sequential approach: first identifying the best ramp center, and subsequently optimizing the other parameters to define a dataset-specific ideal setup (*Conf-Ref*). This configuration was then kept

static for the remaining experiments. Despite its effectiveness, a exhaustive grid search across all parameters was not computationally feasible.

- **Lack of extensive benchmarking**: Due to resource constraints, the experiments and parameter evaluations were limited to UNet and SegResNet architectures. Although using a second architecture validates the method's adaptability, future research should incorporate additional state-of-the-art backbones (e.g., nnU-Net, Swin UNETR) into a multi-architecture benchmark. This extensive testing is required to fully prove that the method's benefits are completely independent of the underlying backbone architecture.

5.4 Future works

5.4.1 Subgroup extraction

A natural and promising direction for future research lies in the integration of multiple metrics in *DivExplorer* [4] for subgroups extraction, specifically using Entropy alongside the Dice score, combined with Preview Div-CL. Our exploratory surveys demonstrated that incorporating uncertainty-based metrics can accelerate the convergence time (reducing training epochs by approximately 30%). Since these two approaches are theoretically complementary, their combination could lower even more the computational cost while keeping the benefit for the subgroups and the competitive performance. Future studies should focus on conducting experiments to evaluate this combined architecture, as well as exploring other potentially relevant metrics that could be integrated in this setup.

5.4.2 Revisiting Curriculum difficulty assignment

As noted in Section 5.3, resolving subgroup overlaps choosing the worst-case assumption can lead to inconsistent curriculum. To address this, future work could replace the maximum difficulty assignment with an average difficulty score computed across all overlapping subgroups.

5.4.3 Testing domain-shift on Preview Curriculum Learning

Although the current study demonstrates the effectiveness of Preview Curriculum Learning on the MAMA-MIA and UCSD-PTGBM datasets, the evaluation was conducted using standard data splits, representing an in-distribution setting. While these datasets feature high multi-centric heterogeneity, this setup does not explicitly measure the model's robustness against entirely novel domains. In the

medical field, unknown hardware, unknown clinical protocols, and shifting patient demographics frequently lead to performance degradation [49, 65]. A possible direction for future work is the evaluation of the proposed strategy under strict Out-of-Distribution (OOD) conditions, such as cross-dataset validation or Leave-One-Center-Out (LOCO) protocols, to test its ability to generalize in entirely unknown medical settings.

5.4.4 Extending Preview Curriculum Learning to other tasks

The proposed Preview Curriculum Learning methodology and dynamic ramp scheduling are fundamentally task-agnostic. While current validation focuses on medical image segmentation, future research will explore the extension of this framework beyond medical imaging to other domains, such as image classification, object detection, and predictive modeling on tabular Electronic Health Records (EHR) [4]. Furthermore, the preview strategy and its weighted ramp system could be integrated into other Curriculum Learning frameworks that rely on alternative metrics for difficulty assignment.

5.4.5 Dynamic partitioning in Curriculum scheduling

Currently, our curriculum relies on a uniform, rigid three-way split, a clear division that does not reflect the actual distribution of sample difficulties. A potential direction to improve the system would be defining dynamic thresholds based on difficulty percentiles (e.g., the 33rd and 66th percentiles). Following the calculation of divergence and the assignment of subgroups, another path that could be evaluated is the one proposed in the original preview formulation [5], where a continuous weighted function (e.g., $w_i = e^{-\gamma \cdot difficulty}$) transforms the difficulty of each individual sample into a soft weight, eliminating the need for rigid discrete buckets.

5.4.6 Pruning redundant subgroups

In the current methodology, the subgroup extraction process did not include the redundancy pruning step based on marginal contribution. Although the specific behavior of individual subgroups was not analyzed in this thesis (an aspect carefully examined by Mantione [62]), future work could integrate this feature to determine whether filtering redundant patterns leads to possible improvements in overall model performance.

Chapter 6

Appendix

6.1 Appendix A: Loss curves and Dice score curves

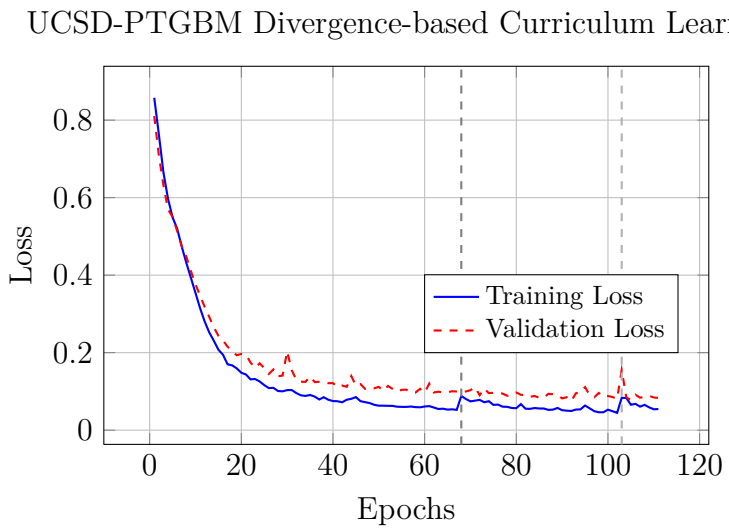


Figure 6.1: Training and validation loss curves on the UCSD-PTGBM dataset using Divergence-based Curriculum Learning approach. The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. The vertical dashed lines indicate the epochs at which medium and hard samples are introduced.

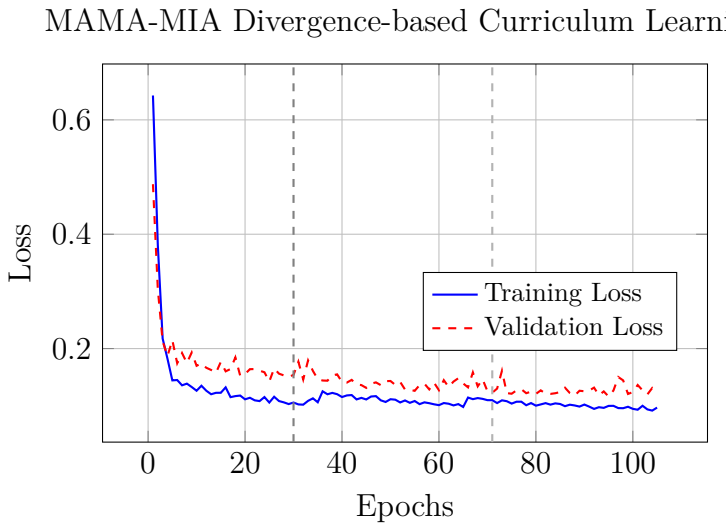


Figure 6.2: Training and validation loss curves on the MAMA-MIA dataset using a Divergence-based Curriculum Learning approach. The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. The vertical dashed lines indicate the epochs at which medium and hard samples are introduced.

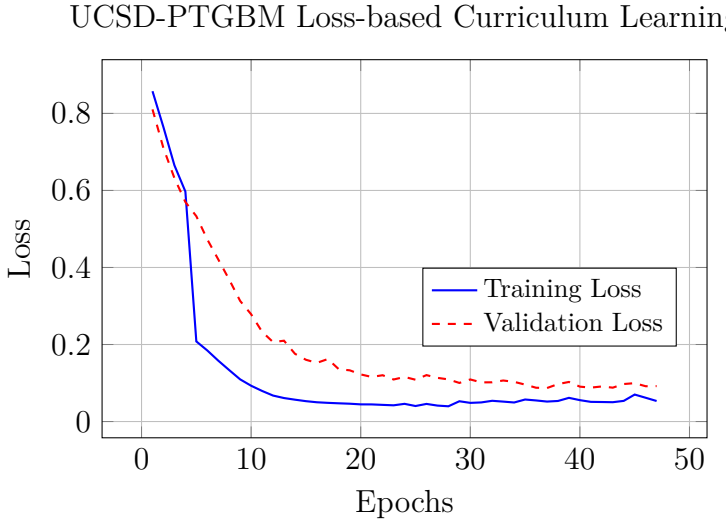


Figure 6.3: Training and validation loss curves on the UCSD-PTGBM dataset using loss-based curriculum learning with preview strategy (Preview Loss-CL). The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs.

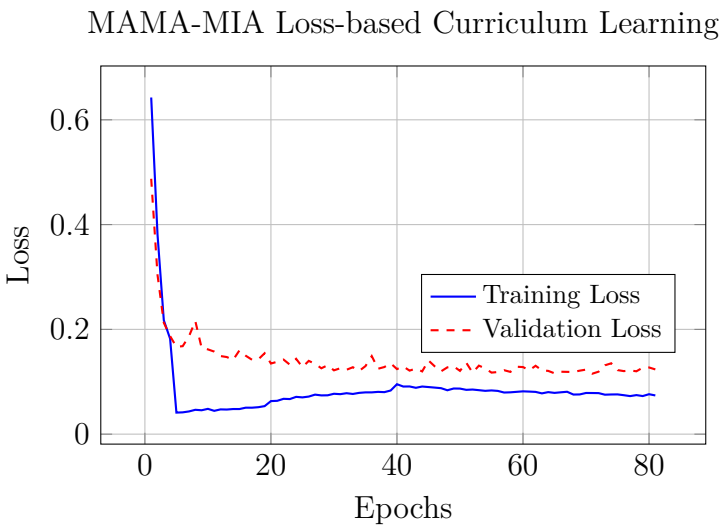


Figure 6.4: Training and validation loss curves on the MAMA-MIA dataset using loss-based curriculum learning with preview strategy (Preview Loss-CL). The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs.

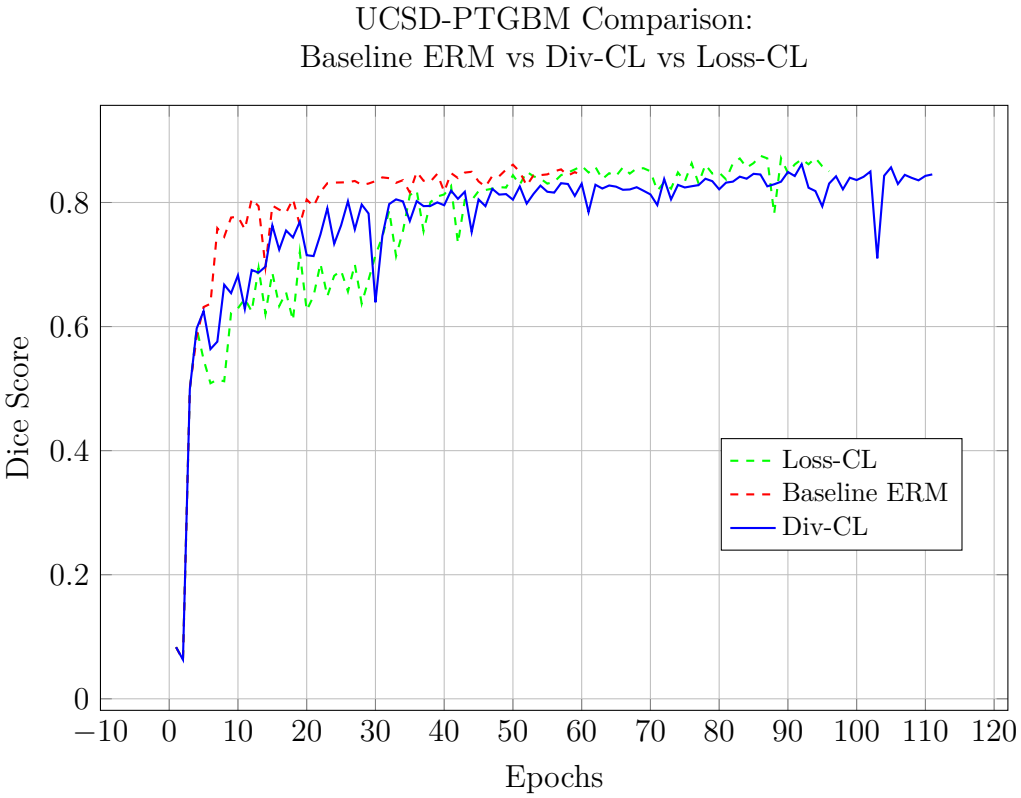


Figure 6.5: Comparison of Dice score trends on the UCSD-PTGBM dataset across three different training strategies. The plot illustrates the performance of the Div-CL method (solid blue line) and the Loss-CL approach (dashed green line) against the baseline ERM training (dashed red line).

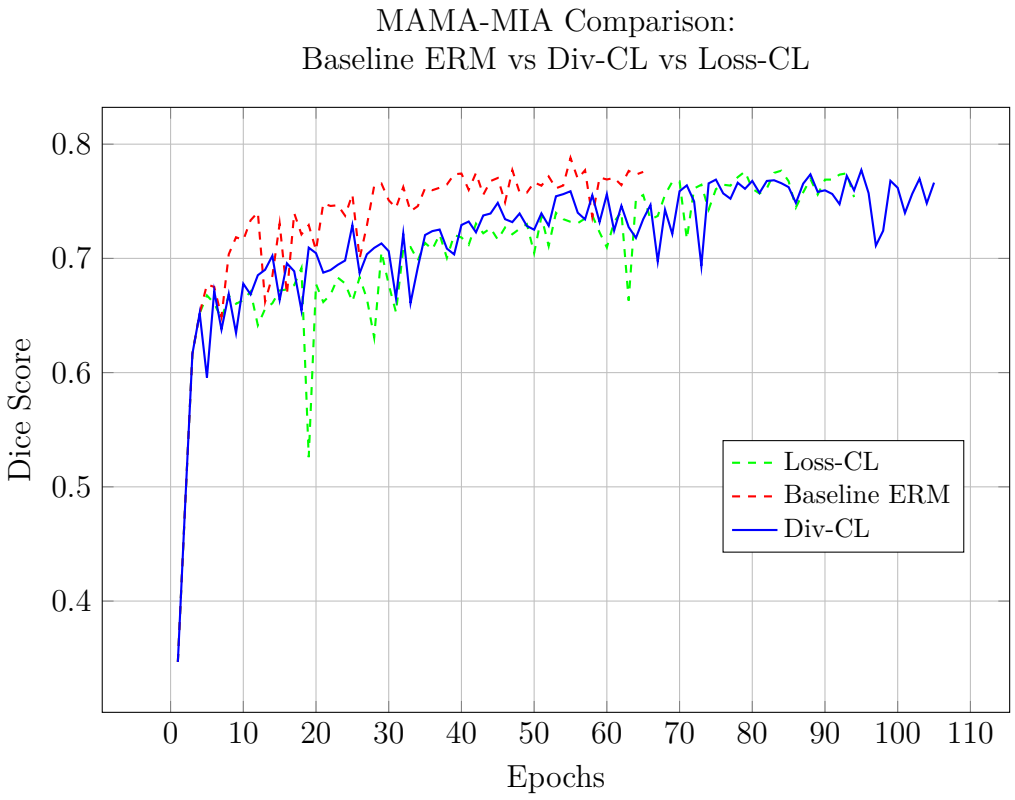


Figure 6.6: Comparison of Dice score trends on the MAMA-MIA dataset across three different training strategies. The plot illustrates the performance of the Div-CL method (solid blue line) and the Loss-CL approach (dashed green line) against the baseline ERM training (dashed red line).

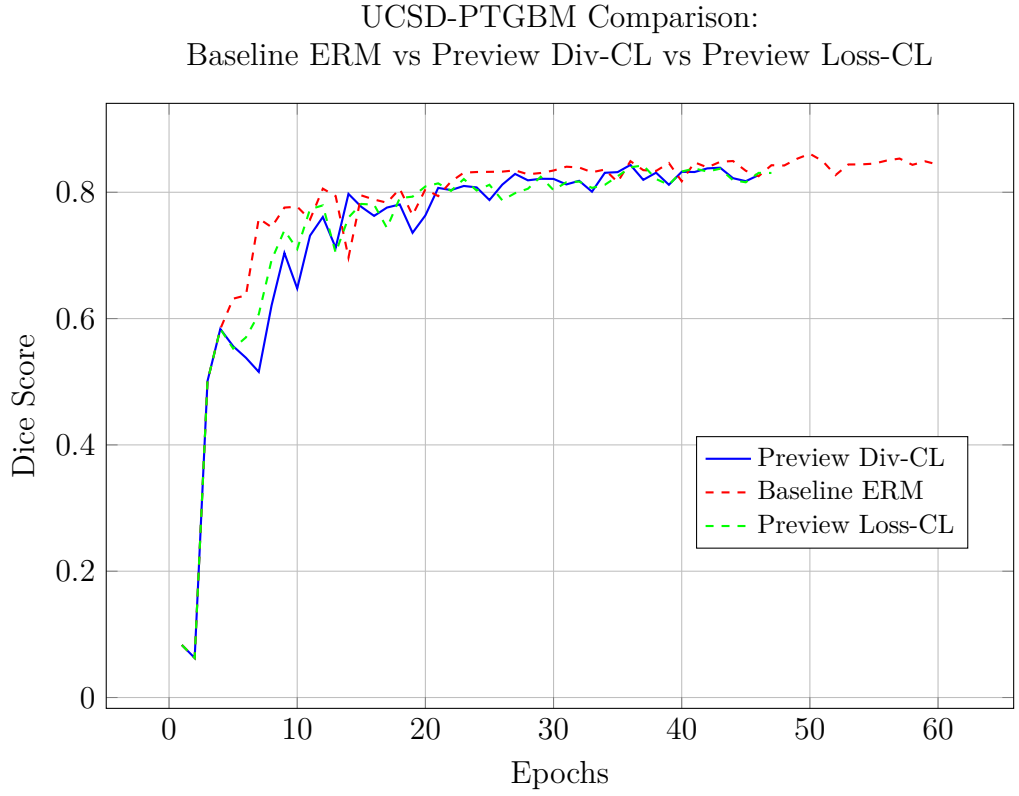


Figure 6.7: Comparison of Dice score trends on the UCSD-PTGBM dataset across preview-based training strategies. The plot illustrates the performance of the Preview Div-CL method (solid blue line) and the Preview Loss-CL method (dashed green line) against the baseline ERM training (dashed red line).

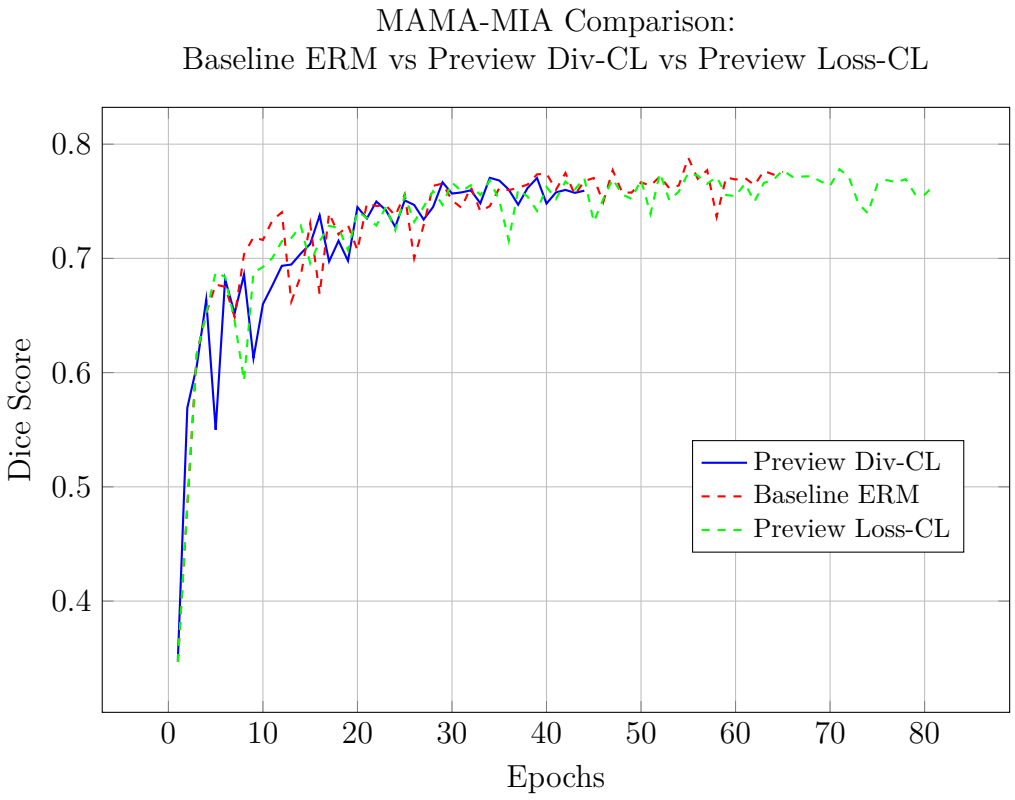
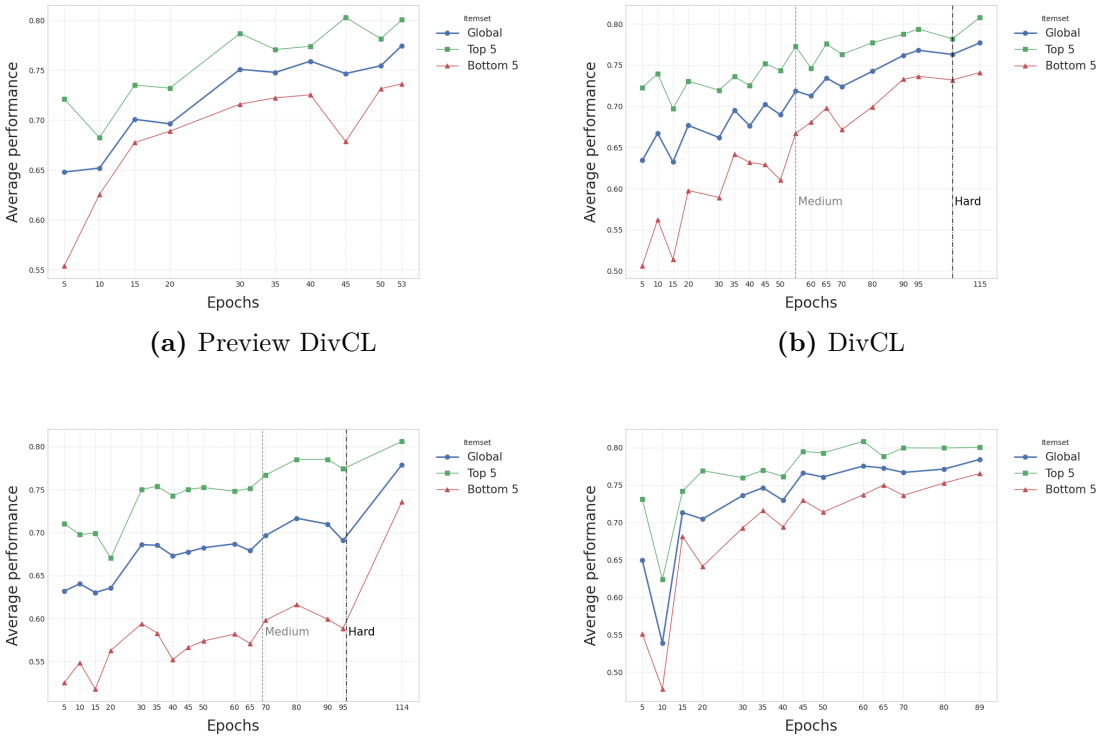


Figure 6.8: Comparison of Dice score trends on the MAMA-MIA dataset across preview-based training strategies. The plot illustrates the performance of the Preview Div-CL method (solid blue line) and the Preview Loss-CL method (dashed green line) against the baseline ERM training (dashed red line).

6.2 Appendix B: SegResNet analysis



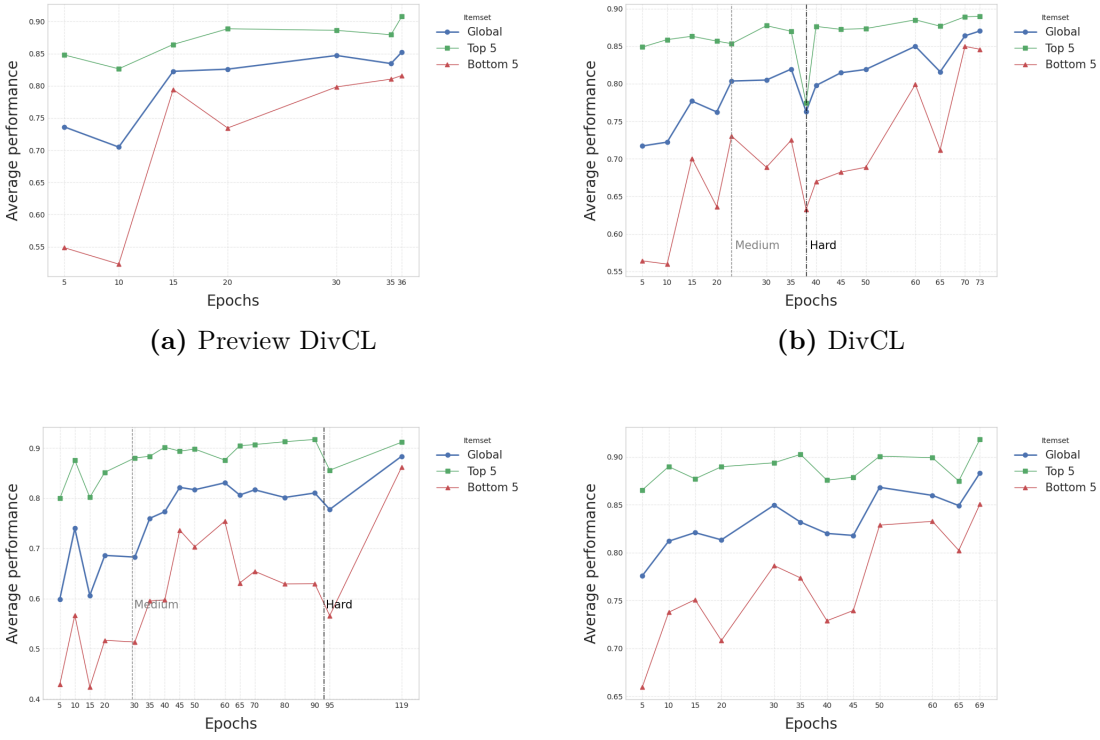
(a) Preview DivCL

(b) DivCL

(c) Loss CL

(d) ERM

Figure 6.9: Dice coefficient progression over a subset of training epochs (including the optimal epoch per setup) on the MAMA-MIA dataset. The global average is shown in blue, while the mean trajectories for the Bottom-5 and Top-5 subgroups are displayed in red and green, respectively. Group assignments for Top-k and Bottom-k rely on the performance recorded at epoch n_0



(a) Preview DivCL

(b) DivCL

(c) Loss CL

(d) ERM

Figure 6.10: Dice coefficient progression over a subset of training epochs (including the optimal epoch per setup) on the UCSD-PTGBM dataset. The global average is shown in blue, while the mean trajectories for the Bottom-5 and Top-5 subgroups are displayed in red and green, respectively. Group assignments for Top-k and Bottom-k rely on the performance recorded at epoch n_0

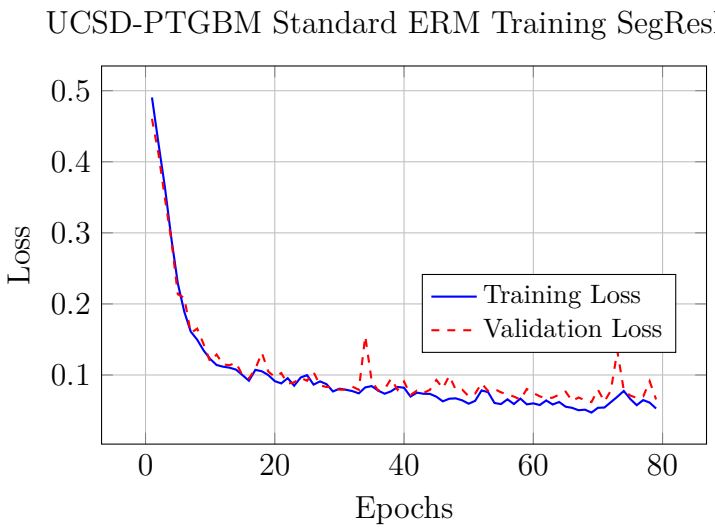


Figure 6.11: Training and validation loss curves on the UCSD-PTGBM dataset using a standard Empirical Risk Minimization (ERM) approach. The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs.

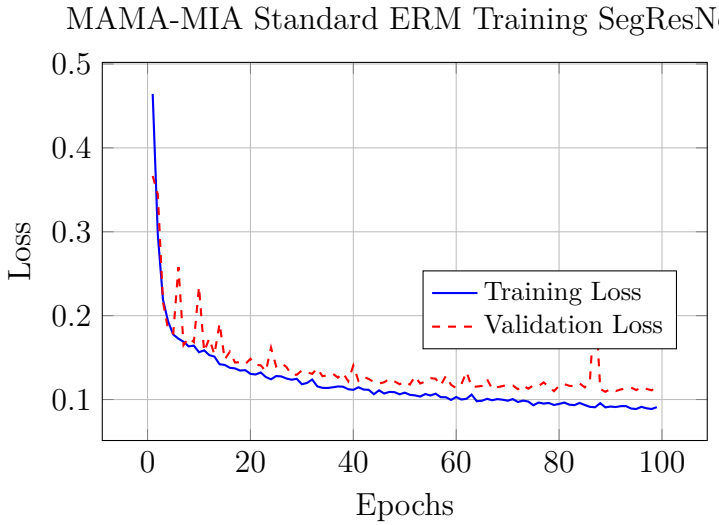


Figure 6.12: Training and validation loss curves on the MAMA-MIA dataset using a standard Empirical Risk Minimization (ERM) approach. The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs.

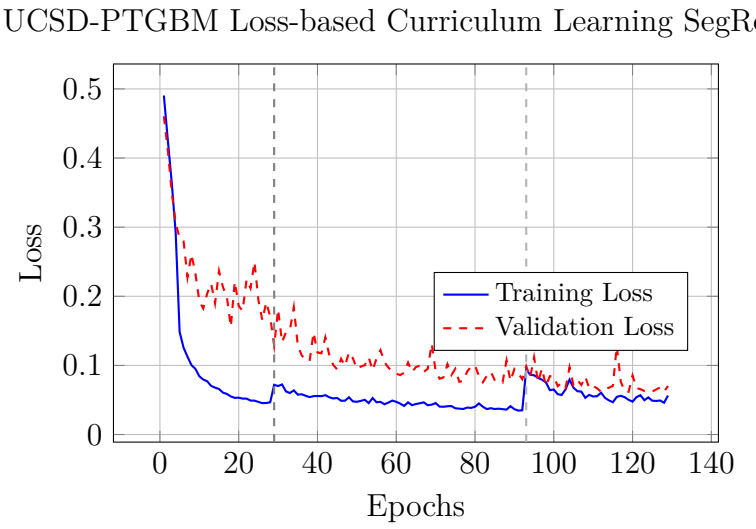


Figure 6.13: Training and validation loss curves on the UCSD-PTGBM dataset using Loss-based Curriculum Learning approach. The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. The vertical dashed lines indicate the epochs at which medium and hard samples are introduced.

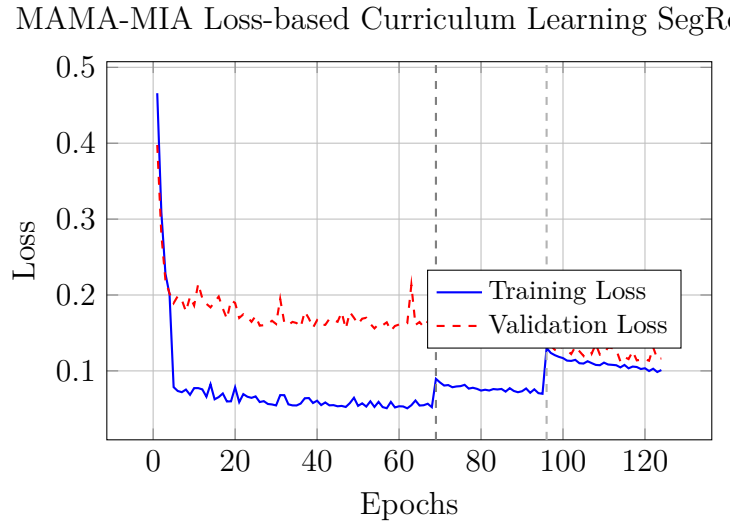


Figure 6.14: Training and validation loss curves on the MAMA-MIA dataset using a Loss-based Curriculum Learning approach. The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. The vertical dashed lines indicate the epochs at which medium and hard samples are introduced.

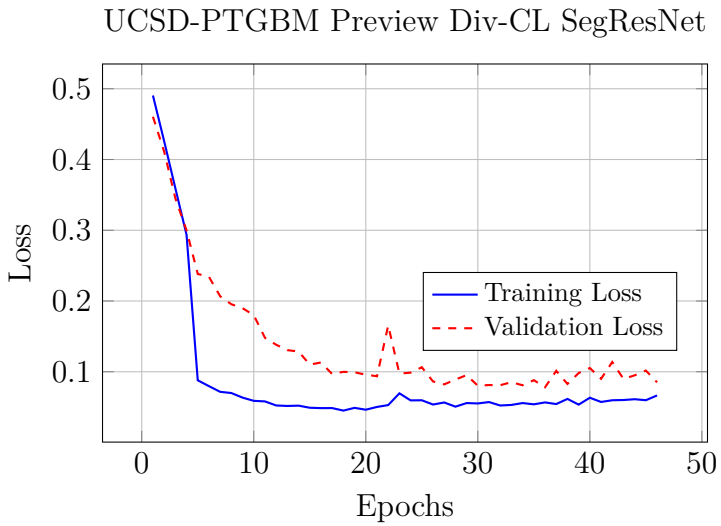


Figure 6.15: Training and validation loss curves on the UCSD-PTGBM dataset using divergence-based curriculum learning with preview strategy (Preview Div-CL). The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs.

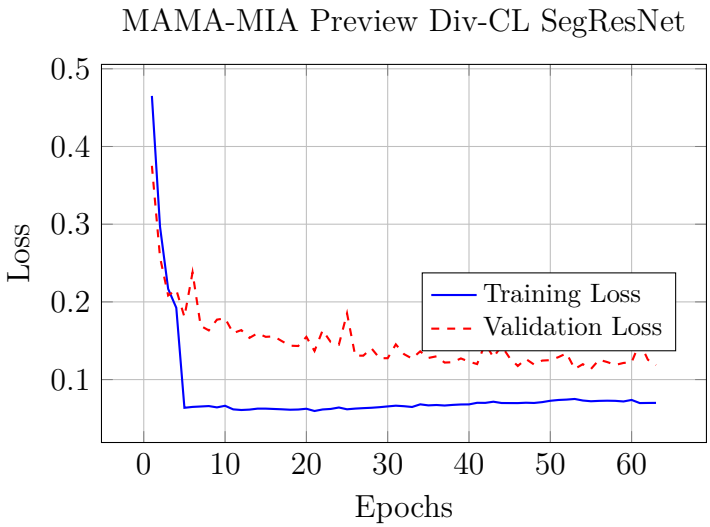


Figure 6.16: Training and validation loss curves on the MAMA-MIA dataset using divergence-based curriculum learning with preview strategy (Preview Div-CL). The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs.

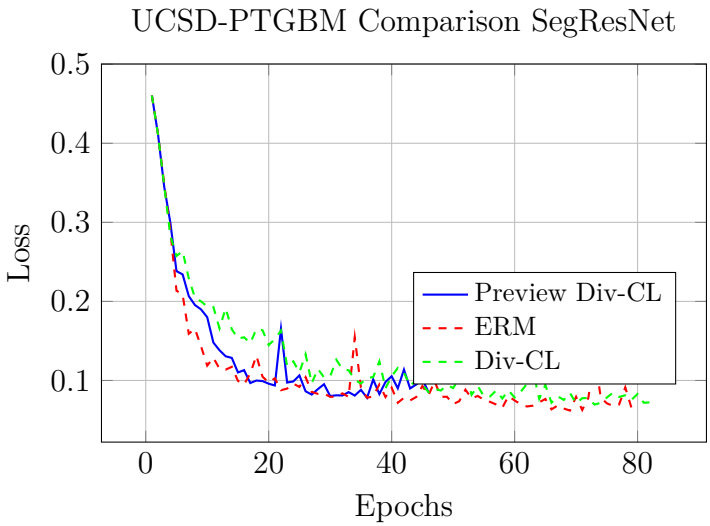


Figure 6.17: Comparison of validation loss curves on the UCSD-PTGBM dataset across three different training strategies. The plot illustrates the performance of the Preview Div-CL method (solid blue line) against the standard Div-CL approach (dashed green line) and the baseline ERM training (dashed red line).

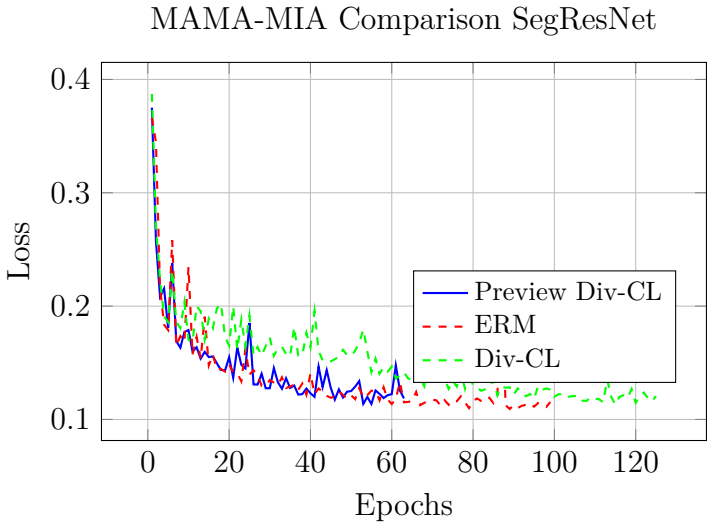


Figure 6.18: Comparison of validation loss curves on the MAMA-MIA dataset across three different training strategies. The plot illustrates the performance of the Preview Div-CL method (solid blue line) against the standard Div-CL approach (dashed green line) and the baseline ERM training (dashed red line).

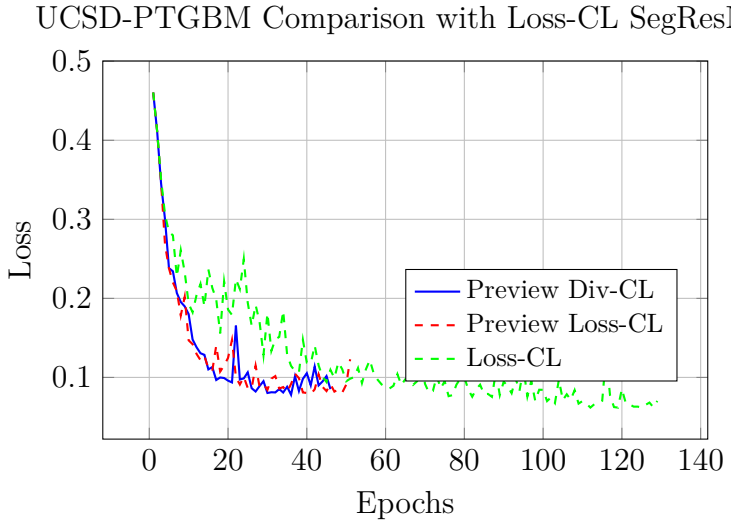


Figure 6.19: Comparison of validation loss curves on the UCSD-PTGBM dataset for different Curriculum Learning strategies. The plot contrasts the performance of Preview Div-CL (solid blue line) against two loss-based approaches: Preview Loss-CL (dashed red line) and standard Loss-CL (dashed green line).

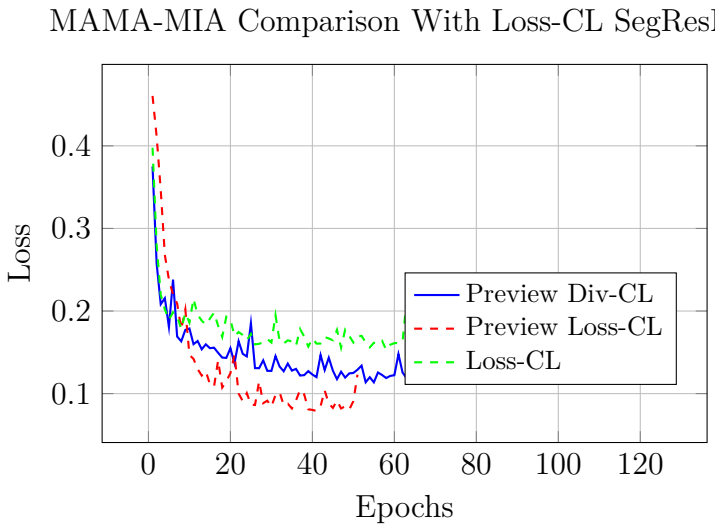


Figure 6.20: Comparison of validation loss curves on the MAMA-MIA dataset for different Curriculum Learning strategies. The plot contrasts the performance of Preview Div-CL (solid blue line) against two loss-based approaches: Preview Loss-CL (dashed red line) and standard Loss-CL (dashed green line).

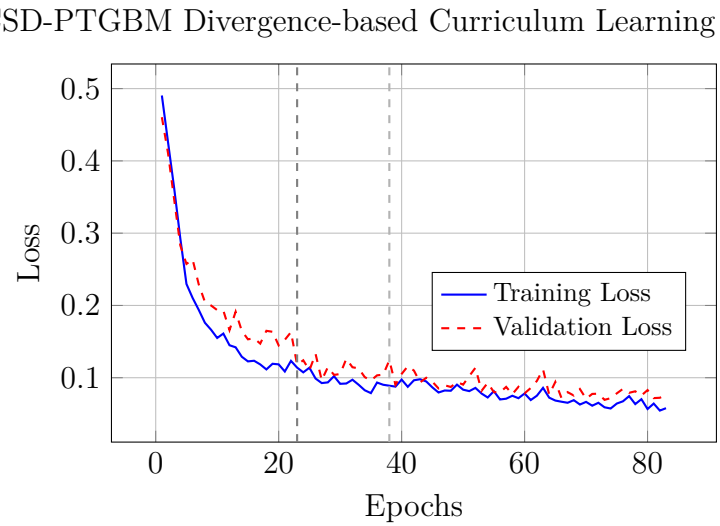


Figure 6.21: Training and validation loss curves on the UCSD-PTGBM dataset using Divergence-based Curriculum Learning approach. The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. The vertical dashed lines indicate the epochs at which medium and hard samples are introduced.

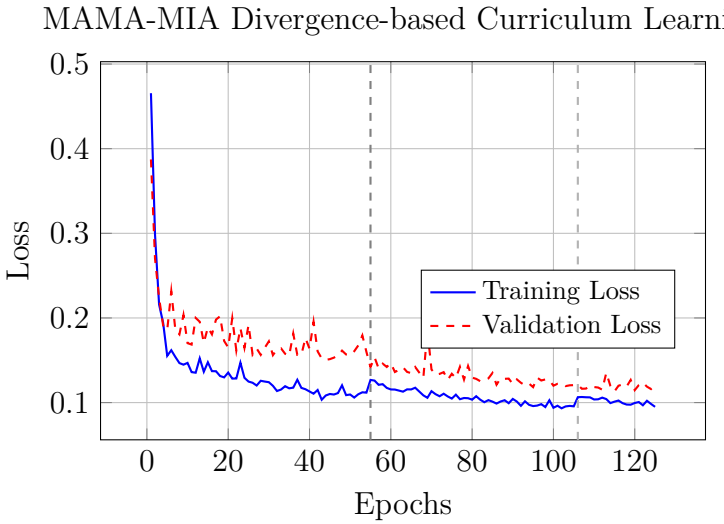


Figure 6.22: Training and validation loss curves on the MAMA-MIA dataset using a Divergence-based Curriculum Learning approach. The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs. The vertical dashed lines indicate the epochs at which medium and hard samples are introduced.

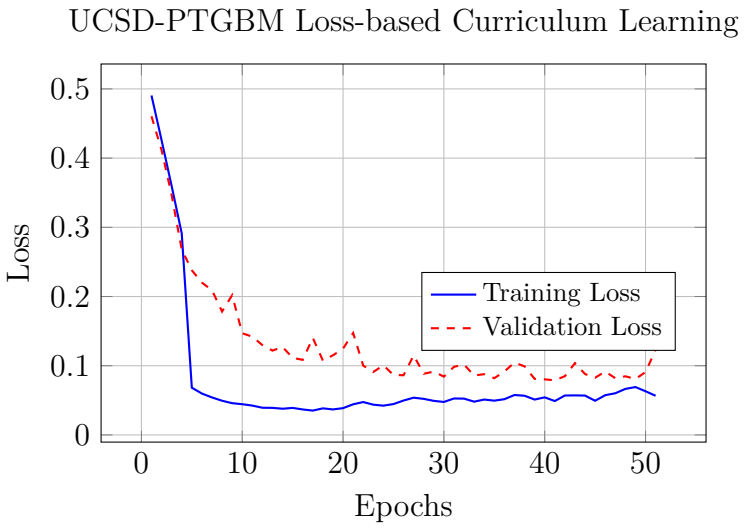


Figure 6.23: Training and validation loss curves on the UCSD-PTGBM dataset using loss-based curriculum learning with preview strategy (Preview Loss-CL). The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs.

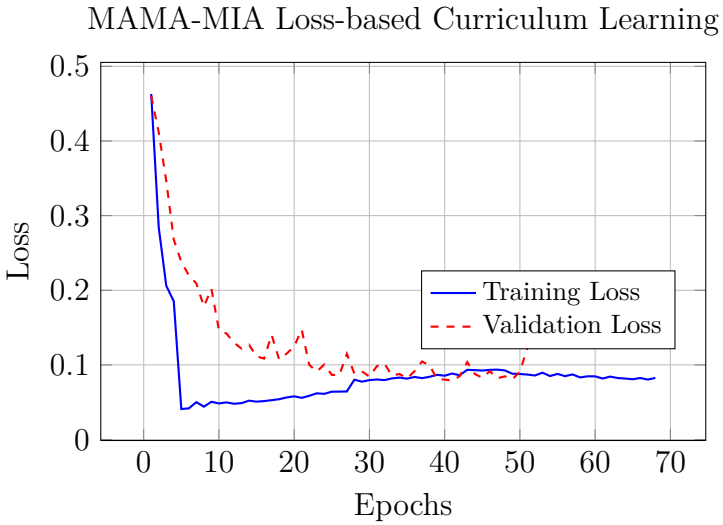


Figure 6.24: Training and validation loss curves on the MAMA-MIA dataset using loss-based curriculum learning with preview strategy (Preview Loss-CL). The solid blue line indicates the training loss, while the dashed red line represents the validation loss over the epochs.

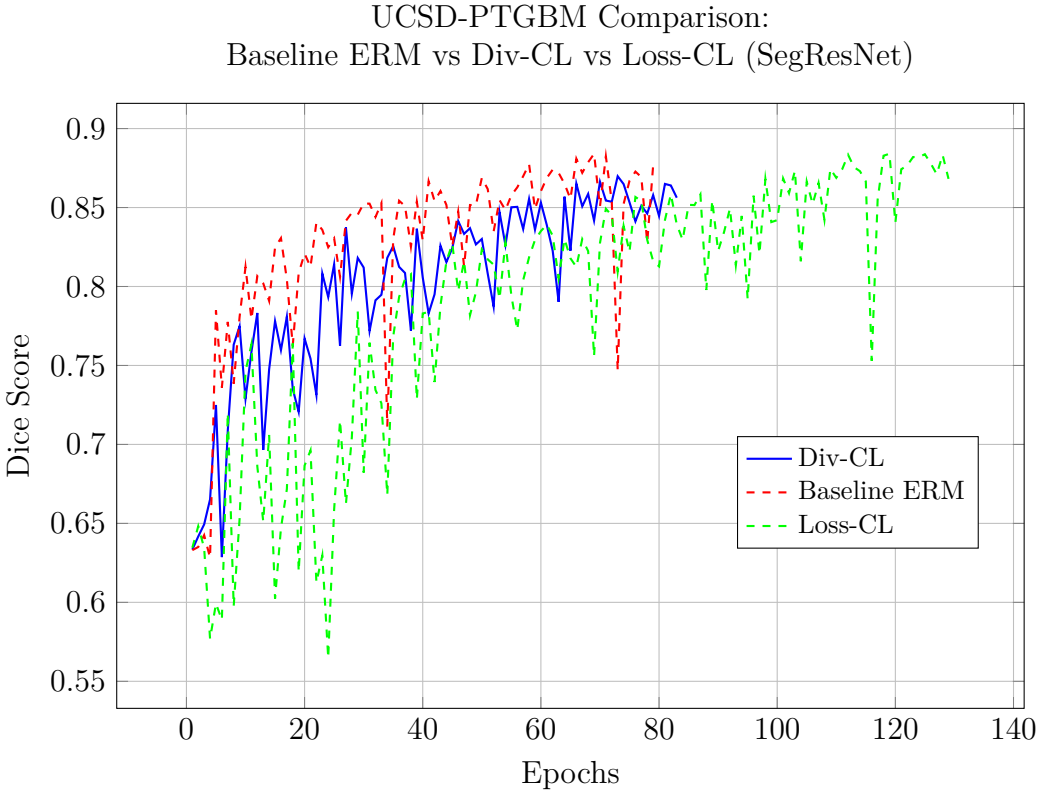


Figure 6.25: Comparison of Dice score trends on the UCSD-PTGBM dataset using a SegResNet backbone across three different training strategies. The plot illustrates the performance of the Div-CL method (solid blue line) and the Loss-CL approach (dashed green line) against the baseline ERM training (dashed red line).

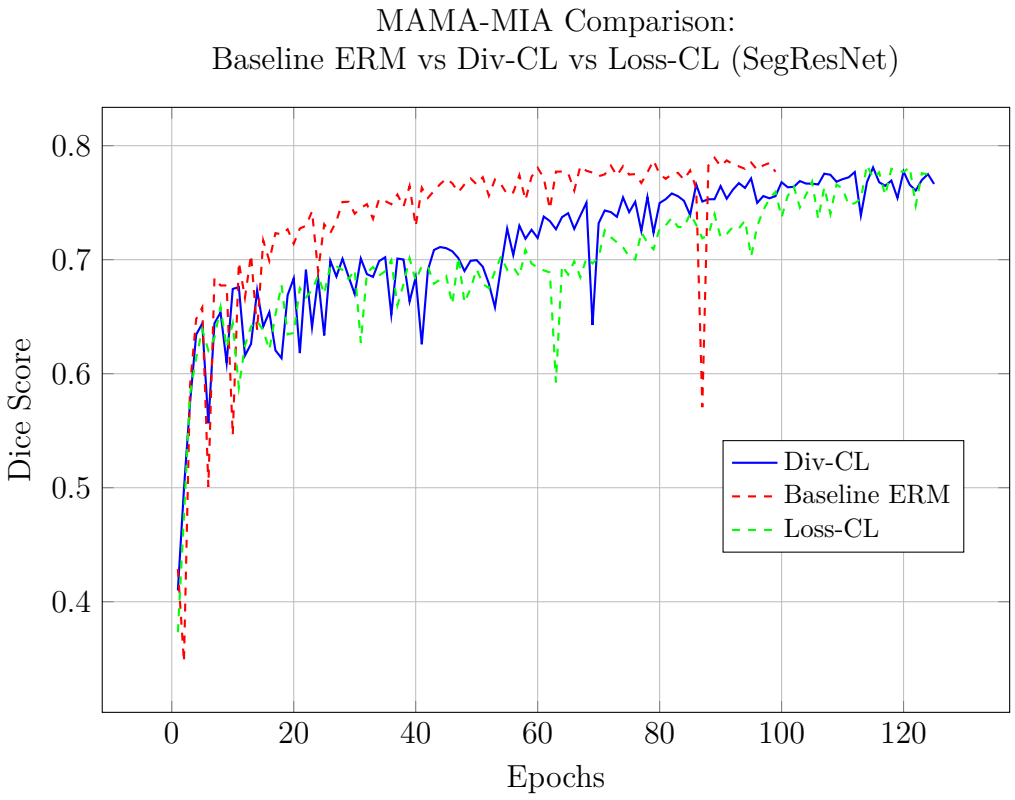


Figure 6.26: Comparison of Dice score trends on the MAMA-MIA dataset using a SegResNet backbone across three different training strategies. The plot illustrates the performance of the Div-CL method (solid blue line) and the Loss-CL approach (dashed green line) against the baseline ERM training (dashed red line).

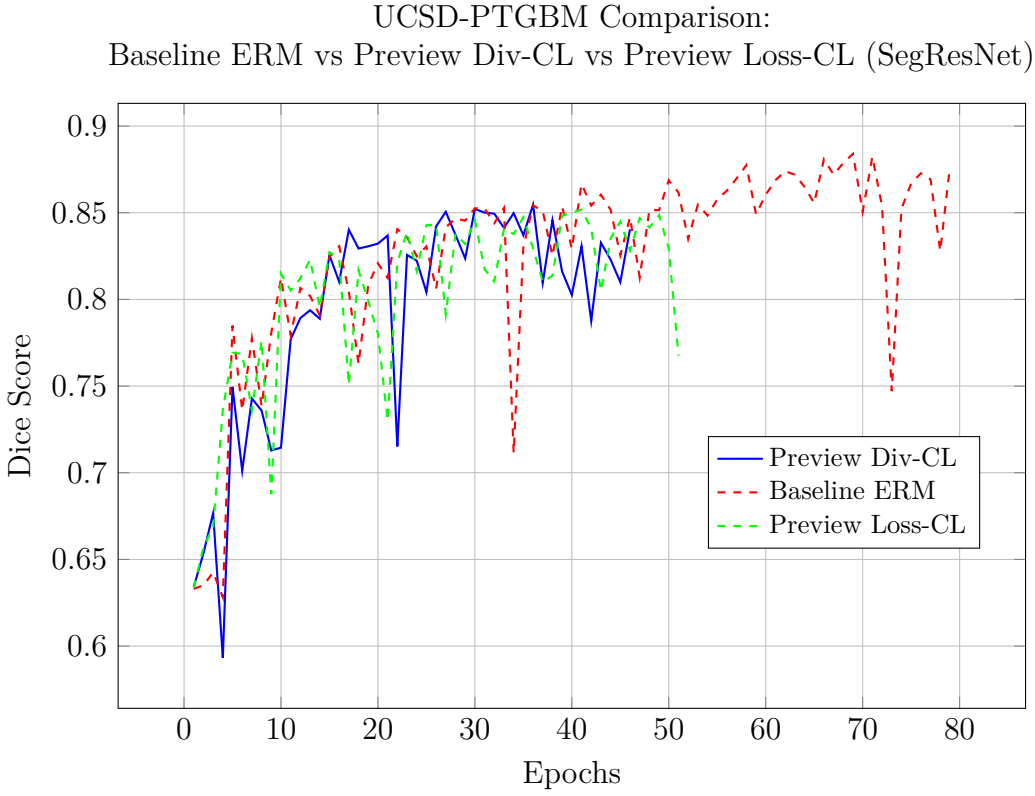


Figure 6.27: Comparison of Dice score trends on the UCSD-PTGBM dataset using a SegResNet backbone across preview-based training strategies. The plot illustrates the performance of the Preview Div-CL method (solid blue line) and the Preview Loss-CL method (dashed green line) against the baseline ERM training (dashed red line).

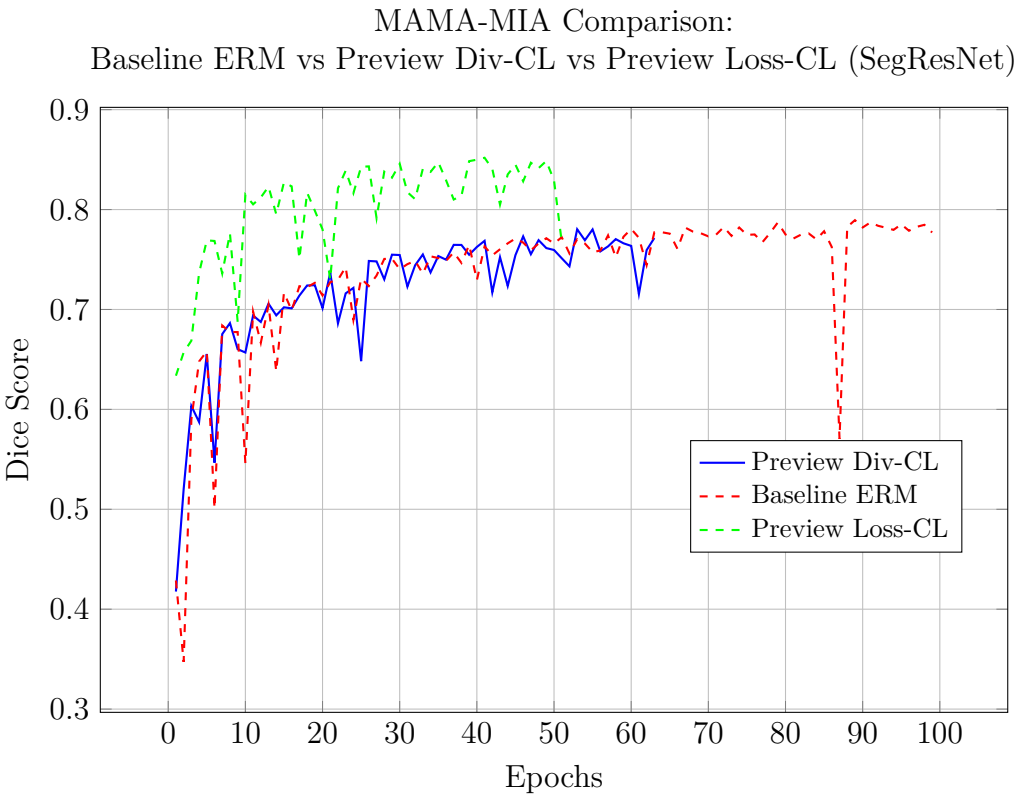


Figure 6.28: Comparison of Dice score trends on the MAMA-MIA dataset using a SegResNet backbone across preview-based training strategies. The plot illustrates the performance of the Preview Div-CL method (solid blue line) and the Preview Loss-CL method (dashed green line) against the baseline ERM training (dashed red line).

Bibliography

[1] Zixuan Teng et al. «A literature review of artificial intelligence (AI) for medical image segmentation: from AI and explainable AI to trustworthy AI». In: *Quantitative Imaging in Medicine and Surgery* 14.12 (Dec. 2024), pp. 9620–9652. DOI: 10.21037/qims-24-723 (cit. on p. 2).

[2] Jeffrey L Elman. «Learning and development in neural networks: The importance of starting small». In: *Cognition* 48.1 (1993), pp. 71–99 (cit. on p. 2).

[3] Jérôme Louradour, Ronan Collobert, Yoesoep Rachmad, and Jason Weston. «Curriculum learning». In: *Journal of the American Podiatry Association* 60 (June 2009), p. 6. DOI: 10.1145/1553374.1553380 (cit. on pp. 2, 7).

[4] Eliana Pastor, Luca de Alfaro, and Elena Baralis. «Looking for Trouble: Analyzing Classifier Behavior via Pattern Divergence». In: *Proceedings of the 2021 International Conference on Management of Data*. SIGMOD '21. Virtual Event, China: Association for Computing Machinery, 2021, pp. 1400–1412. ISBN: 9781450383431. DOI: 10.1145/3448016.3457284. URL: https://doi.org/10.1145/3448016.3457284 (cit. on pp. 3, 24, 46, 57–59).

[5] Muhe Ding, Jianlong Wu, Xue Dong, Xiaojie Li, Pengda Qin, Tian Gan, and Liqiang Nie. «Preview-Based Category Contrastive Learning for Knowledge Distillation». In: *IEEE Transactions on Circuits and Systems for Video Technology* 35.7 (2025), pp. 6712–6725. DOI: 10.1109/TCSVT.2025.3541067 (cit. on pp. 3, 6, 17, 55, 59).

[6] Alejandro Barredo Arrieta et al. «Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI». In: *Information Fusion* 58 (2020), pp. 82–115. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2019.12.012. URL: https://www.sciencedirect.com/science/article/pii/S1566253519308103 (cit. on p. 5).

[7] Zachary C Lipton. «The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.» In: *Queue* 16.3 (2018), pp. 31–57 (cit. on p. 5).

[8] Scott M. Lundberg and Su-In Lee. «A unified approach to interpreting model predictions». In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 4768–4777. ISBN: 9781510860964 (cit. on p. 6).

[9] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. «"Why Should I Trust You?": Explaining the Predictions of Any Classifier». In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1135–1144. ISBN: 9781450342322. DOI: 10.1145/2939672.2939778. URL: https://doi.org/10.1145/2939672.2939778 (cit. on p. 6).

[10] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. «Curriculum learning». In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09. Montreal, Quebec, Canada: Association for Computing Machinery, 2009, pp. 41–48. ISBN: 9781605585161. DOI: 10.1145/1553374.1553380. URL: https://doi.org/10.1145/1553374.1553380 (cit. on pp. 6, 7).

[11] Tom M. Mitchell. *Machine Learning*. Vol. 1. New York, NY, USA: McGraw-Hill, 1997 (cit. on p. 7).

[12] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. «Curriculum Learning: A Survey». In: *International Journal of Computer Vision* 130.6 (June 2022), pp. 1526–1565. ISSN: 1573-1405. DOI: 10.1007/s11263-022-01611-x. URL: https://doi.org/10.1007/s11263-022-01611-x (cit. on pp. 7, 8).

[13] M. Kumar, Benjamin Packer, and Daphne Koller. «Self-Paced Learning for Latent Variable Models». In: *Advances in Neural Information Processing Systems*. Ed. by J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta. Vol. 23. Curran Associates, Inc., 2010. URL: https://proceedings.neurips.cc/paper_files/paper/2010/file/e57c6b956a6521b28495f2886ca0977a-Paper.pdf (cit. on p. 7).

[14] Lu Jiang, Deyu Meng, Shoou-I Yu, Zhenzhong Lan, Shiguang Shan, and Alexander G. Hauptmann. «Self-Paced Learning with Diversity». In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger. Vol. 27. Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/8786a6172486a9582ff0b94492e8f06a-Paper.pdf (cit. on p. 7).

[15] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander Hauptmann. «Self-paced curriculum learning». In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. 1. 2015 (cit. on p. 7).

[16] Pascal Klink, Hany Abdulsamad, Boris Belousov, and Jan Peters. «Self-Paced Contextual Reinforcement Learning». In: *Proceedings of the Conference on Robot Learning*. Ed. by Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura. Vol. 100. Proceedings of Machine Learning Research. PMLR, Oct. 2020, pp. 513–529. URL: https://proceedings.mlr.press/v100/klink20a.html (cit. on p. 7).

[17] Qian Zhao, Deyu Meng, Lu Jiang, Qi Xie, Zongben Xu, and Alexander Hauptmann. «Self-paced learning for matrix factorization». In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 29. 1. 2015 (cit. on p. 7).

[18] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. «Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels». In: *International conference on machine learning*. PMLR. 2018, pp. 2304–2313 (cit. on p. 7).

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. «Progressive growing of gans for improved quality, stability, and variation». In: *arXiv preprint arXiv:1710.10196* (2017) (cit. on p. 8).

[20] Pietro Morerio, Jacopo Cavazza, Riccardo Volpi, René Vidal, and Vittorio Murino. «Curriculum dropout». In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 3544–3552 (cit. on p. 8).

[21] Tae-Hoon Kim and Jonghyun Choi. «Screenernet: Learning self-paced curriculum for deep neural networks». In: *arXiv preprint arXiv:1801.00904* (2018) (cit. on p. 8).

[22] Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. «A Teacher-Student Framework for Zero-Resource Neural Machine Translation». In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Regina Barzilay and Min-Yen Kan. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 1925–1935. DOI: 10.18653/v1/P17-1176. URL: https://aclanthology.org/P17-1176/ (cit. on p. 8).

[23] David Silver et al. *Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm*. 2017. arXiv: 1712.01815 [cs.AI]. URL: https://arxiv.org/abs/1712.01815 (cit. on p. 8).

[24] Samarth Sinha, Animesh Garg, and Hugo Larochelle. «Curriculum by smoothing». In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 21653–21664 (cit. on p. 8).

[25] Gabriel Maicas, Andrew P. Bradley, Jacinto C. Nascimento, Ian Reid, and Gustavo Carneiro. «Training Medical Image Analysis Systems like Radiologists». In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Ed. by Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger. Cham: Springer International Publishing, 2018, pp. 546–554. ISBN: 978-3-030-00928-1 (cit. on p. 8).

[26] William Lotter, Greg Sorensen, and David Cox. «A Multi-scale CNN and Curriculum Learning Strategy for Mammogram Classification». In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Ed. by M. Jorge Cardoso et al. Cham: Springer International Publishing, 2017, pp. 169–177. ISBN: 978-3-319-67558-9 (cit. on p. 8).

[27] Amelia Jiménez-Sánchez, Diana Mateus, Sonja Kirchhoff, Chlodwig Kirchhoff, Peter Biberthaler, Nassir Navab, Miguel A. González Ballester, and Gemma Piella. «Curriculum learning for improved femur fracture classification: Scheduling data with prior knowledge and uncertainty». In: *Medical Image Analysis* 75 (2022), p. 102273. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2021.102273. URL: https://www.sciencedirect.com/science/article/pii/S1361841521003182 (cit. on p. 9).

[28] Amelia Jiménez-Sánchez, Mickael Tardy, Miguel A. González Ballester, Diana Mateus, and Gemma Piella. «Memory-aware curriculum federated learning for breast cancer classification». In: *Computer Methods and Programs in Biomedicine* 229 (2023), p. 107318. ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2022.107318. URL: https://www.sciencedirect.com/science/article/pii/S016926072200699X (cit. on p. 9).

[29] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. «Communication-Efficient Learning of Deep Networks from Decentralized Data». In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Ed. by Aarti Singh and Jerry Zhu. Vol. 54. Proceedings of Machine Learning Research. PMLR, Apr. 2017, pp. 1273–1282. URL: https://proceedings.mlr.press/v54/mcmahan17a.html (cit. on p. 9).

[30] Yuxing Tang, Xiaosong Wang, Adam P. Harrison, Le Lu, Jing Xiao, and Ronald M. Summers. «Attention-Guided Curriculum Learning for Weakly Supervised Classification and Localization of Thoracic Diseases on Chest Radiographs». In: *Machine Learning in Medical Imaging*. Ed. by Yinghuan Shi, Heung-Il Suk, and Mingxia Liu. Cham: Springer International Publishing, 2018, pp. 249–258. ISBN: 978-3-030-00919-9 (cit. on p. 9).

[31] Scott Kirkpatrick, C Daniel Gelatt Jr, and Mario P Vecchi. «Optimization by simulated annealing». In: *science* 220.4598 (1983), pp. 671–680 (cit. on pp. 10, 17).

[32] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. «Equation of state calculations by fast computing machines». In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092 (cit. on p. 10).

[33] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. «Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing». In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 240–250. DOI: 10.18653/v1/N19-1021. URL: https://aclanthology.org/N19-1021/ (cit. on pp. 10, 49, 55).

[34] I Loshchilov and F SGDR Hutter. «Stochastic gradient descent with warm restarts. 2016». In: *DOI: https://doi. org/10.48550/arXiv* 1608 (2016) (cit. on p. 10).

[35] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. «A Survey on Bias and Fairness in Machine Learning». In: *ACM Comput. Surv.* 54.6 (July 2021). ISSN: 0360-0300. DOI: 10.1145/3457607. URL: https://doi.org/10.1145/3457607 (cit. on p. 11).

[36] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. «Dissecting racial bias in an algorithm used to manage the health of populations». In: *Science* 366.6464 (2019), pp. 447–453. DOI: 10.1126/science.aax2342. eprint: https://www.science.org/doi/pdf/10.1126/science.aax2342. URL: https://www.science.org/doi/abs/10.1126/science.aax2342 (cit. on p. 11).

[37] Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. «Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey». In: *ACM J. Responsib. Comput.* 1.2 (June 2024). DOI: 10.1145/3631326. URL: https://doi.org/10.1145/3631326 (cit. on p. 11).

[38] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. «Optimized Pre-Processing for Discrimination Prevention». In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL:

https://proceedings.neurips.cc/paper_files/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf (cit. on p. 11).

[39] Faisal Kamiran and Toon Calders. «Data preprocessing techniques for classification without discrimination». In: *Knowledge and information systems* 33.1 (2012), pp. 1–33 (cit. on p. 11).

[40] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. «Mitigating unwanted biases with adversarial learning». In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pp. 335–340 (cit. on p. 11).

[41] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. «Fairness-aware classifier with prejudice remover regularizer». In: *Joint European conference on machine learning and knowledge discovery in databases*. Springer. 2012, pp. 35–50 (cit. on p. 12).

[42] Moritz Hardt, Eric Price, and Nati Srebro. «Equality of opportunity in supervised learning». In: *Advances in neural information processing systems* 29 (2016) (cit. on p. 12).

[43] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. «On fairness and calibration». In: *Advances in neural information processing systems* 30 (2017) (cit. on p. 12).

[44] Alkis Koudounas, Eliana Pastor, Luca de Alfaro, and Elena Baralis. «Mitigating Subgroup Disparities in Speech Models: A Divergence-Aware Dual Strategy». In: *IEEE Transactions on Audio, Speech and Language Processing* 33 (2025), pp. 883–895. DOI: 10.1109/TASLPRO.2025.3539429 (cit. on p. 12).

[45] Mohamed Huti, Tiarna Lee, Elinor Sawyer, and Andrew P. King. «An Investigation into Race Bias in Random Forest Models Based on Breast DCE-MRI Derived Radiomics Features». In: *Clinical Image-Based Procedures, Fairness of AI in Medical Imaging, and Ethical and Philosophical Issues in Medical Imaging*. Ed. by Stefan Wesarg et al. Cham: Springer Nature Switzerland, 2023, pp. 225–234. ISBN: 978-3-031-45249-9 (cit. on p. 13).

[46] Eleonora Poeta, Luisa Vargas, Daniele Falcetta, Vincenzo Marciano', Eliana Pastor, Tania Cerquitelli, Elena Baralis, and Maria A. Zuluaga. «Divergence-Aware Training with Automatic Subgroup Mitigation for Breast Tumor Segmentation». In: *Artificial Intelligence and Imaging for Diagnostic and Treatment Challenges in Breast Care*. Ed. by Tianyu Zhang, Oliver Lester Saldanha, Luyi Han, Nika Rasoolzadeh, Lidia Garrucho Moras, Jarek van Dijk, Tao Tan, Jakob Nikolas Kather, and Ritse Mann. Cham: Springer Nature Switzerland, 2026, pp. 52–62. ISBN: 978-3-032-05559-0 (cit. on p. 13).

[47] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. «No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classification Problems». In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 19339–19352. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/e0688d13958a19e087e123148555e4b4-Paper.pdf (cit. on p. 13).

[48] Xenia Konti, Yi Shen, Zifan Wang, Karl Henrik Johansson, Michael J. Pencina, Nicoleta J. Economou-Zavlanos, and Michael M. Zavlanos. *Group Distributionally Robust Machine Learning under Group Level Distributional Uncertainty*. 2025. arXiv: 2509.08942 [cs.LG]. URL: https://arxiv.org/abs/2509.08942 (cit. on p. 13).

[49] Brian Guo, Darui Lu, Gregory Szumel, Rongze Gui, Pengyu Li, Tingyu Wang, Nicholas Konz, and Maciej A Mazurowski. «The impact of scanner domain shift on deep learning performance in medical imaging: an experimental study». In: *International Journal of Computer Assisted Radiology and Surgery* (2026), pp. 1–11 (cit. on pp. 14, 59).

[50] Eliana Pastor, Luca de Alfaro, and Elena Baralis. *DivExplorer project page*. https://divexplorer.github.io/. 2021. URL: https://divexplorer.github.io/ (cit. on p. 15).

[51] Rakesh Agrawal and Ramakrishnan Srikant. «Fast Algorithms for Mining Association Rules in Large Databases». In: *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, pp. 487–499 (cit. on p. 16).

[52] Jiawei Han, Jian Pei, and Yiwen Yin. «Mining frequent patterns without candidate generation». In: *ACM sigmod record* 29.2 (2000), pp. 1–12 (cit. on p. 16).

[53] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. «An empirical study of example forgetting during deep neural network learning». In: *arXiv preprint arXiv:1812.05159* (2018) (cit. on p. 18).

[54] Lidia Garrucho et al. «A large-scale multicenter breast cancer DCE-MRI benchmark dataset with expert segmentations». In: *Scientific Data* 12.1 (2025), p. 453. DOI: 10.1038/s41597-025-04707-4 (cit. on p. 21).

[55] Lidia Garrucho et al. *The MAMA-MIA Challenge: Advancing Generalizability and Fairness in Breast MRI Tumor Segmentation and Treatment Response Prediction*. 2026. arXiv: 2603.01250 [cs.CV]. URL: https://arxiv.org/abs/2603.01250 (cit. on p. 21).

[56] Louis Gagnon et al. «The University of California San Diego Post-Treatment Glioblastoma (UCSD-PTGBM) annotated multimodal MRI Dataset». In: *Scientific Data* 13.1 (2026), p. 183. ISSN: 2052-4463. DOI: 10.1038/s41597-025-06499-z. URL: https://doi.org/10.1038/s41597-025-06499-z (cit. on p. 21).

[57] Bjoern H Menze et al. «The multimodal brain tumor image segmentation benchmark (BRATS)». In: *IEEE transactions on medical imaging* 34.10 (2014), pp. 1993–2024 (cit. on p. 22).

[58] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. «Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features». In: *Scientific data* 4.1 (2017), p. 170117 (cit. on p. 22).

[59] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. «U-net: Convolutional networks for biomedical image segmentation». In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241 (cit. on p. 22).

[60] Lee R Dice. «Measures of the amount of ecologic association between species». In: *Ecology* 26.3 (1945), pp. 297–302 (cit. on p. 24).

[61] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander Hauptmann. «Self-paced curriculum learning». In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 29. 1. 2015 (cit. on p. 26).

[62] Silvia Mantione. «Analysis of Subgroup-Aware Curriculum Learning Strategies for Medical Image Segmentation». In preparation. MA thesis. Torino, Italy: Politecnico di Torino, 2026 (cit. on pp. 32, 43, 52, 59).

[63] Dan Hendrycks and Kevin Gimpel. «A baseline for detecting misclassified and out-of-distribution examples in neural networks». In: *arXiv preprint arXiv:1610.02136* (2016) (cit. on p. 46).

[64] Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. «Comparing images using the Hausdorff distance». In: *IEEE Transactions on pattern analysis and machine intelligence* 15.9 (2002), pp. 850–863 (cit. on p. 47).

[65] Aminu Musa, Rajesh Prasad, Peter Onwualu, and Monica Hernandez. «A Systematic Review of Cross-Population Shifts in Medical Imaging Analysis with Deep Learning». In: *Big Data and Cognitive Computing* 10.3 (2026). ISSN: 2504-2289. DOI: 10.3390/bdcc10030076. URL: https://www.mdpi.com/2504-2289/10/3/76 (cit. on p. 59).