UNISINOS

Programa de Pós-Graduação em

Computação Aplicada

Mestrado/Doutorado Acadêmico

João Batista Rodrigues Neto

Continual Knowledge Distillation for Histopathology

São Leopoldo, 2024

João Batista Rodrigues Neto

CONTINUAL KNOWLEDGE DISTILLATION FOR HISTOPATHOLOGY

Dissertação apresentada como requisito parcial para a obtenção do título de Mestre pelo Programa de Pós-Graduação em Computação Aplicada da Universidade do Vale do Rio dos Sinos — UNISINOS

Advisor:
Prof. Dr. Gabriel de Oliveira Ramos

São Leopoldo
2024

R696c Rodrigues Neto, João Batista.
Continual Knowledge Distillation for Histopathology / João Batista Rodrigues Neto – 2024.
79 f. : il. color ; 30 cm.

Dissertação (mestrado) – Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Computação Aplicada, São Leopoldo, 2024.
"Advisor: Prof. Dr. Gabriel de Oliveira Ramos."

1. Aprendizagem contínua. 2. Segmentação. 3. Cancêr. 4. Histopatologia. 5. Patologia. I. Título.

CDU 004.7

Dados Internacionais de Catalogação na Publicação (CIP)
(Bibliotecária: Bruna Sant'Anna – CRB 10/2360)

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior Brasil (CAPES) - Código de Financiamento 001
This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001

JOÃO BATISTA RODRIGUES NETO

CONTINUAL KNOWLEDGE DISTILLATION FOR HISTOPATHOLOGY

Dissertação apresentada como requisito parcial para a obtenção do título de Mestre pelo Programa de Pós-Graduação em Computação Aplicada da Universidade do Vale do Rio dos Sinos — UNISINOS

Aprovado em 7 de Outubro de 2024

BANCA EXAMINADORA

Prof. Dr. Gabriel de Oliveira Ramos (Orientador) - UNISINOS

Prof. Dr. Cristiano André da Costa (Avaliador) - UNISINOS

Prof. Dr. Cláudio Rosito Jung (Avaliador) - UFRGS

ABSTRACT

Com o surgimento da patologia computadorizada, muitos datasets e competições foram publicados para incentivar pesquisadores a desenvolverem soluções que auxiliem nas tarefas da área da patologia. A análise de segmentos histopatológicos, conduzida por patologistas para detectar células cancerígenas ou metástases em imagens de tecido, é uma dessas tarefas, para a qual, visão computacional foi aplicada com sucesso e até superou o desempenho de especialistas. Apesar dos excelentes resultados na literatura, a maioria das abordagens é dependente do dataset usado e carecem de generalização, fazendo com que até os melhores modelos desempenhem mal quando apresentados a tecidos diferentes. Neste trabalho, nós desenvolvemos um novo método de aprendizagem contínua, que alavanca a generalização do modelo nos datasets usando uma destilação melhorada de conhecimento. Verificamos, através de profundos e extensos experimentos em 19 datasets, uma melhoria geral de **15,66%** em comparação dos métodos comuns da literatura, e métricas superiores em relação a modelos com total disponibilidade de datasets. Além disso, nosso método foi o único a atingir índices **positivos** de transferência de conhecimento para frente (FWT) e para trás (BWT), mitigando consideravelmente o efeito de esquecimento catastrófico.

Keywords: Aprendizagem Contínua. Segmentação. Câncer. Histopatologia. Patologia.

ABSTRACT

With the emergence of computational pathology, many datasets were made public and challenges were published to encourage researches into developing assistant frameworks for pathology tasks. The analysis of histopathological slides, made by pathologists to detect tumorous cells or metastasis in tissue images, is one of such tasks, for which, computer vision had been successfully applied and even outperformed human expert levels. Despite the excellent results in the literature, the majority of approaches are dataset-dependent and lack generalization, making even the best documented models perform poorly when presented with different tissues. In this work, we designed a novel continuous learning method, that leverages the model generalization across datasets using enhanced knowledge distillation. We verified, through deep and extensive experimentation on 19 datasets, an overall improvement of **15,66%** in comparison to common literature methods, and superior metrics in relation to models with full dataset availability. Also, our method was the only one to achieve **positive** forward (FWT) and backward (BWT) knowledge transfer indexes, considerably mitigating the catastrophic forgetting effect.

Keywords: Continual Learning. Segmentation. Cancer. Histopathology. Pathology.

LIST OF FIGURES

Figure 1: Model predictions after being trained on the adrenal glands dataset. The first column presents the ground truths, the second one the predictions, and the third one presents the overlap of predictions over the ground truth. 19

Figure 2: Model predictions after being trained on the colon tissue dataset. The first column presents the ground truths, the second one the predictions, and the third one presents the overlap of predictions over the ground truth. 19

Figure 3: Example of a preparation protocol applied to a biological sample to produce the corresponding histological slide. 24

Figure 4: Illustration of the generalization process in a model using a gradient-based method. Tasks are $T0$, $T1$ and $T2$, while L is the total loss function, λ is the model loss function, and the as are the regularization terms. 30

Figure 5: Illustration of the generalization process in a model using a architecture-based method. The tasks are $T0$, $T1$ and $T2$. 31

Figure 6: Illustration of the generalization process in a model using a memory-based method. The tasks are $T0$, $T1$ and $T2$. 31

Figure 7: Illustration of the generalization process in a model using a meta learning method. The tasks are $T0$, $T1$ and $T2$, while the $Integration$ steps are the model-fuse operations. 32

Figure 8: Illustration of our method generalization process, the model knowledge is generalized using the regularization terms extracted from the previous version of this same model. The tasks are $T0$, $T1$ and $T2$. 44

Figure 9: Architecture of our UNet segmentation model. The red layers are Convolutional layers, the blue layers are activations (Relu or Sigmoid), the green layers are Batch Normalizations, the yellow layers are dimension reduction (MaxPool) or expansion (ConvTranspose), and the gray layers are the concatenations. 45

Figure 10: Illustration of the generalization process in a model using the Learning without Forgetting. The tasks are $T0$, $T1$ and $T2$, and α is the regularization weight. 46

Figure 11: Illustration of the continual loss masks and values mid training. The last three images represent each, in order, the compound loss function terms, the titles on top of the images are referring to their term form. Since these terms are functions (Dice) comparing the truth with a value, the dark red areas represent the truth, and the green areas represent the values. 50

Figure 12: Breast gland tissue samples (on the left) and their respective annotations (on the right) from the Pannuke dataset. The annotations colors are red for tumors, pink for inflammatory tissue and green for connective/soft tissue. . . 56

Figure 13: From top to bottom: breast, liver and head, tissue samples (on the left) and their respective annotations (on the right) from the Pannuke dataset. The annotations colors are red for tumors, pink for inflammatory tissue and green for connective/soft tissue. 57

Figure 14: Training and validation model loss across the epochs for all tasks independently trained. 60

Figure 15: Segmentation metrics of all tasks independently trained. The horizontal lines represent the average metric values. 62

Figure 16: Final IoU for all datasets, segmented by CL method. The horizontal lines represent the average IoU values. 64
Figure 17: Continual IoU metric during the training of each dataset, the lines represent each a CL baseline, the final dot represents the FULL baseline (trained at once). 69
Figure 18: Samples of the selected datasets, segmented by models produced by each CL method. Datasets are top-bottom ordered by their difficulty, being the top the easiest. 70

LIST OF TABLES

Table 1: Initial Works 36
Table 2: Snowballing process 36
Table 3: Literature Review 40

Table 4: Experiment parameters 54
Table 5: Datasets 55
Table 6: Model Segmentation Metrics 61
Table 7: Continual Learning: IoU Results 63
Table 8: Continual Learning: FWT and BWT Results (based on IoU) 65

LIST OF ACRONYMS

AI	Artificial Intelligence
AUC	Area Under the ROC Curve
BWT	Backward Transfer
CL	Continual Learning
CNN	Convolutional Neural Network
CPATH	Computational Pathology
CPU	Central Processing Unit
EWC	Elastic Weight Consolidation
FWT	Forward Transfer
GAN	Generative Adversarial Networks
GB	Gigabyte
GHz	Gigahertz
GPU	Graphics Processing Unit
IOU	Intersection Over Union
KD	Knowledge Distillation
LWF	Learning Without Forgetting
MAS	Memory Aware Synapses
MR	Magnetic Resonance
MRI	Magnetic Resonance Imaging
RAM	Random Access Memory
RELU	Rectified Linear Unit
ROI	Region of Interest
SGD	Stochastic Gradient Descent

CONTENTS

1 INTRODUCTION **17**
1.1 Context 17
1.2 Problem 18
1.3 Research Question 20
1.4 Objective 21
1.5 Contributions 21

2 THEORETICAL FOUNDATION **23**
2.1 Histopathology 23
2.2 Artificial Intelligence 25
2.2.1 Machine Learning 25
2.2.2 Deep Learning 26
2.2.3 Continual Learning 28
2.3 Discussion 33

3 LITERATURE REVIEW **35**
3.1 Methodology 35
3.2 Results 37
3.3 Opportunities 39

4 METHOD **43**
4.1 Overview 43
4.2 Model 44
4.3 Learning without Forgetting 45
4.4 Enhancement 47
4.5 Discussion 49

5 EXPERIMENTAL EVALUATION **53**
5.1 Methodology 53
5.1.1 Datasets 54
5.1.2 Baselines 55
5.1.3 Metrics 58
5.2 Results 59
5.2.1 Model 60
5.2.2 Continual Learning 61
5.2.3 Output Analysis 66
5.3 Discussion 68

6 CONCLUSION **71**

REFERENCES **73**

1 INTRODUCTION

1.1 Context

Since the feasibility of Whole-Slide-Image (WSI) scanners, many medical image datasets were made public and challenges were published to encourage researches into developing tools that integrate Artificial Intelligence (AI) into medical imaging tasks (KUMAR; GUPTA; GUPTA, 2020). Since then, the literature evolved focusing on the development of frameworks that achieve expert level performance at medical tasks, with the main goal of assisting medical time-consuming tasks, prognostics, analysis, and mitigating human error (BáNDI et al., 2023). In this context, computer vision has been successfully applied to many medical image analysis problems, sometimes even outperforming expert level metrics. One of these many successful cases is the analysis of histopathological slides, conducted by pathologists to segment tumorous cells, detect tumor metastasis in tissue images, and other image related tasks (MUSUMECI, 2014).

The documented works vary on their approach, but its noticeable that they share the same solution methodology: Deep Learning models trained or fine-tuned to specific tasks over specific types of tissues (ZHOU et al., 2021; JANOWCZYK; MADABHUSHI, 2016; XING et al., 2017; LAAK; LITJENS; CIOMPI, 2021). Despite the excellent results documented in the literature, these approaches are very site-dependent, meaning that the models were trained specifically for a single task over a specific dataset tissue considering a specific staining/enhancing method (KOTHARI et al., 2013; ZHOU et al., 2021; XING et al., 2017; LAAK; LITJENS; CIOMPI, 2021; BáNDI et al., 2023). This means that even the best documented models would perform poorly when presented with different data, like tissues from different body sites (JAHANIFAR et al., 2023a; BáNDI et al., 2023). This non-generalization characteristic transforms the present literature into a vast ocean of very specific high-performance models of the same task trained for different body tissues.

In a real-world scenario, there is a plethora of protocols, machines, microscopes and staining techniques, that cause a huge heterogeneity in the available datasets, making difficult even the generalization of models within the same task domain (JAHANIFAR et al., 2023a; LAAK; LITJENS; CIOMPI, 2021). Ultimately, the literature endorses that to really make Artificial Intelligent solutions available to pathologists effectively, a system with a strong generalization capacity is necessary (KOTHARI et al., 2013; ZHOU et al., 2021; LAAK; LITJENS; CIOMPI, 2021; JAHANIFAR et al., 2023b), and also proposes that a generic deep model trained with very distinctive image types is of urgent need to deal with the wide variations in pathology imaging (KOTHARI et al., 2013; ZHOU et al., 2021; XING et al., 2017; BáNDI et al., 2023).

The non-generalization problem of deep neural networks is called Catastrophic Forgetting and it roots from the very nature of neural networks training process (FRENCH, 1999). Searching in the literature, one may find many works attacking catastrophic forgetting with different

strategies, but recent surveys and their results suggest that Continual Learning (CL), also known as Lifelong Learning, methods are more suitable and perform better at generalization (BáNDI et al., 2023; AL-THELAYA et al., 2023; GONZALEZ; SAKAS; MUKHOPADHYAY, 2020a). Continual Learning implementations are vastly documented in the literature, with multiple variations of similar algorithms, therefore, many works prefer to group the implementations into classes based in their solution methodology, such as Gradient, Architecture or Meta-learning solutions (HADSELL et al., 2020). Even so, at the time of this present work, no solution could be found to satisfactorily ease the effects of catastrophic forgetting and promote proper generalization, this is mainly due to limitations of each technique, such as heuristic efficiency, and intrinsic complications of the problem, such as data availability (GONZALEZ; SAKAS; MUKHOPADHYAY, 2020b; BáNDI et al., 2023; HADSELL et al., 2020).

In this work we proposed an enhanced method, we improved the Learning without Forgetting (LwF) method, specifically its Knowledge Distillation (KD) strategy, to resolve a histopathology segmentation problem considering multiple datasets. We investigated and documented our problem in details, revised the modern literature and highlighted its present limitations, then, we designed our method with these foundings in mind. Later, we numerically and semantically evaluated our method performance, documenting an overall improvement of **15,66%** in comparison to common state-of-the-art baselines, positive transfer indexes and superior metrics in relation to models with full dataset availability.

1.2 Problem

In order to produce a generalist system, able to be integrated into pathologists day-to-day work (KOTHARI et al., 2013; ZHOU et al., 2021), a solution to ease catastrophic forgetting should be developed.

Catastrophic forgetting is an inherit effect when training neural networks, the training algorithms update the model to perform better at the most recent data seen by the model, and by doing so, forgetting the knowledge acquired from older data. Which means that, when a model is learning a new task (different tumors, different tissues, different staining technique), it is also forgetting the older one (FRENCH, 1999; HADSELL et al., 2020).

So, in this work, the problem that we tackled is catastrophic forgetting on histopathology segmentation models. To better illustrate the effects of forgetting on the histopathology area, we trained a tumor segmentation model considering just two datasets, both representing the same problem, tumorous cells segmentation, but from different tissue samples, adrenal glands tissue and colon tissue. The model was trained on both datasets sequentially, first in the adrenal glands one, then in the colon tissues one. The model used here is described in Section 4.2, it is the same we later used for our full approach.

Figures 1 and 2 present the result of the predictions made on samples of both datasets after the training, its possible to notice how the model completely forgets how to segment tumors in

adrenal glands after being trained in colon tissues, and how precise are the predictions for this last one.

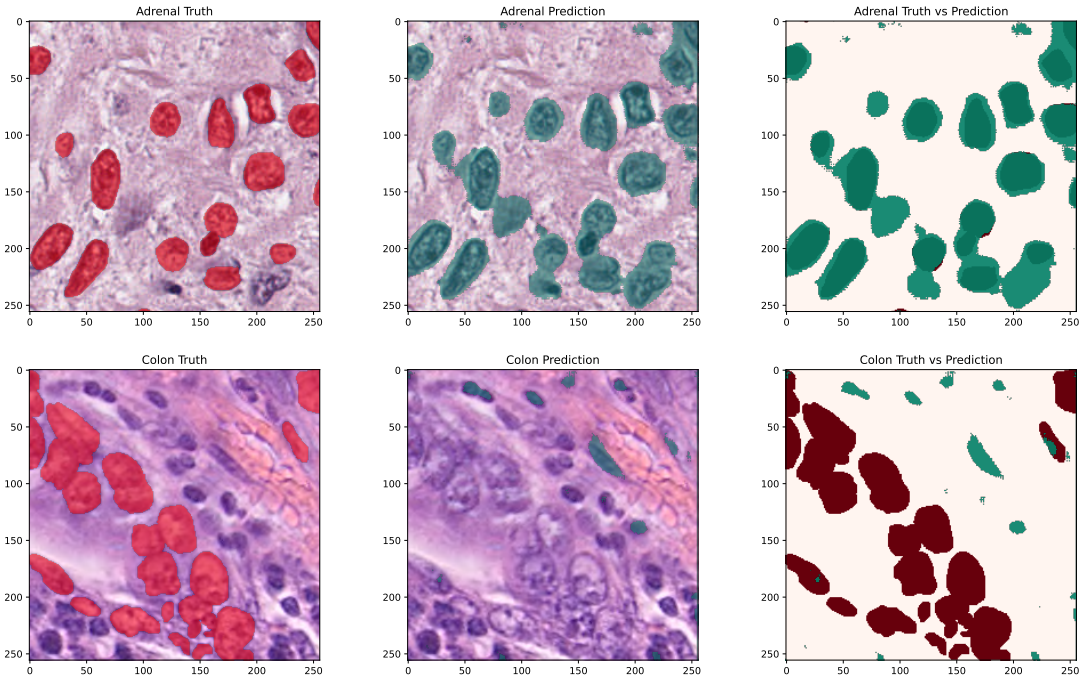

Figure 1: Model predictions after being trained on the adrenal glands dataset. The first column presents the ground truths, the second one the predictions, and the third one presents the overlap of predictions over the ground truth.

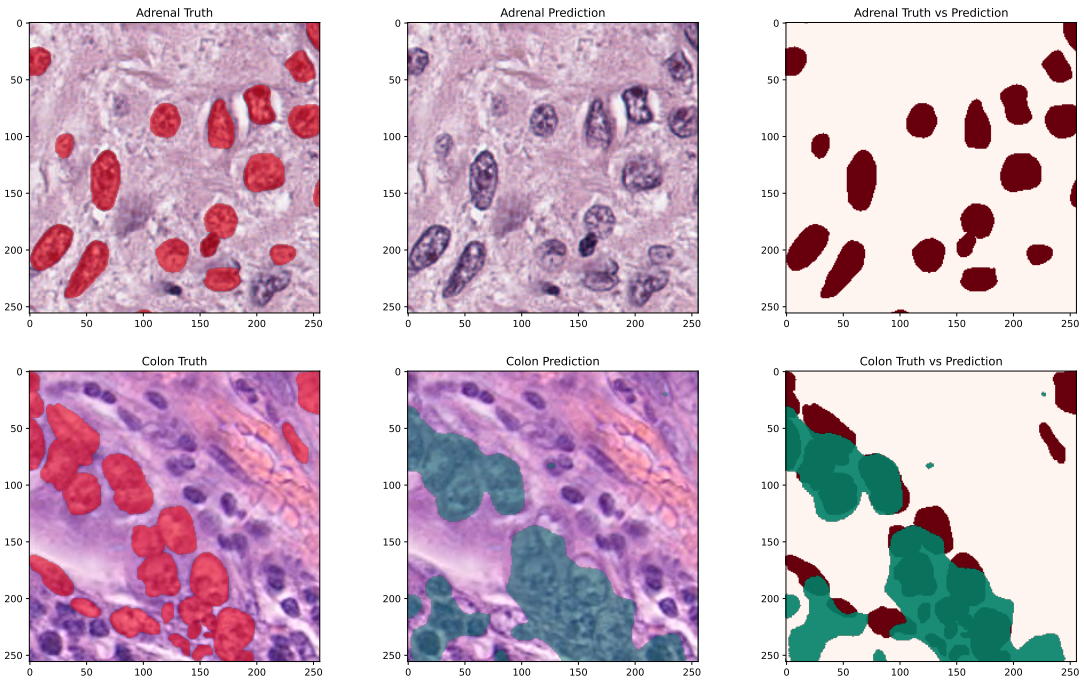

Figure 2: Model predictions after being trained on the colon tissue dataset. The first column presents the ground truths, the second one the predictions, and the third one presents the overlap of predictions over the ground truth.

The forgetting effect poses a hard barricade on the adoption of computer vision solutions for histopathology teams, this domain works with data with a highly variability and low availability, renting most solutions useless in a practical way (JAHANIFAR et al., 2023a; BáNDI et al., 2023).

Although there are plenty of works in the literature addressing the catastrophic forgetting problem, unfortunately, they address it only partially, while there are important nuances to the method that should be considered in order fulfill pathology related requirements:

- **New tasks**: the method should allow the model to incorporate new knowledge and achieve a good performance in new tasks (forward transfer) (BáNDI et al., 2023; KOTHARI et al., 2013; ZHOU et al., 2021; HADSELL et al., 2020).

- **Old tasks**: the method should not only preserve the learned knowledge, but also increase the performance of older tasks (backward transfer) (GONZALEZ; SAKAS; MUKHOPADHYAY, 2020b; HADSELL et al., 2020).

- **Old data**: there are strong regulations and limitations to the access of patient data, and also limitations about memory usage in some environments, so the method should not take for granted the access to data from previous tasks (BáNDI et al., 2023; GONZALEZ; SAKAS; MUKHOPADHYAY, 2020a; KAUSTABAN et al., 2022; GONZALEZ; SAKAS; MUKHOPADHYAY, 2020b; HADSELL et al., 2020).

- **New data**: given the sheer variety of the data acquisition processes, the method should also generalize the knowledge to different distributions of new data for the same old task (Domain Shift) (BáNDI et al., 2023; JAHANIFAR et al., 2023a; GONZALEZ; SAKAS; MUKHOPADHYAY, 2020a).

Summarizing, there are not major works in the histopathology literature considering the generalization problem, and there are not complete solutions in the catastrophic forgetting literature that could perform the generalization with the specific requirements of the pathology area.

1.3 Research Question

Given the presented context and the related problematics, this work proposes to answer to the following question:

*How can we leverage the **generalization** of deep learning histopathology segmentation models considering the data **variability** and the **requirements** for real-world clinic applications?*

From this question, we seek to investigate the catastrophic forgetting problem in digital pathology models, and propose a method to deal with this effect considering the intrinsic problematics of the pathology area. We raise the hypothesis of improving an existing method to

alleviate forgetting in segmentation models considering the problematics in Section 1.2, if such improvement could provide enough leverage for the segmentation model to be suitable in a real-world clinic application.

1.4 Objective

The main objective of this work is to study a heuristic method capable of preserving and enhancing knowledge that is common to multiple tasks learned over sequential training steps. This method should grant neural models the ability to accumulate knowledge over the time without relying on previous data replay, and generalize the learned representations to new data, improving performance on already learned tasks while learning new ones. So, we can leverage the learning abilities of deep neural models, improving their performance over time and accumulating domain knowledge. Models with this ability could be trained to generalize knowledge over multiple instances of the same problem or different problems, also it could be used as pre-trained feature extractors to more robust models (knowledge transfer).

As complementary goals of our work, we aim to study the catastrophic forgetting problem itself, its solutions and details, review the modern literature for approaches of dealing with the forgetting problem in the pathology area, design our method using insights form the literature and experimentation, validate and evaluate our approach, given multiple metrics and datasets, and document our findings and comparisons with state-of-the-art baselines.

1.5 Contributions

For the sake of simplicity, we listed bellow all the contributions of this work, together with our main numerical and qualitative results in comparison to the present literature.

- As far as we know, our work is the first one to perform a literature review on generalization methods for computational pathology.
- Our work presents an enhanced version of the LwF approach, with an additional regularization term to improve generalization.
- Our method, unlike most regularization techniques, does not rely of any hyperparameter optimization.
- Our work uses more datasets that the previously documented works.
- Our work is the first to perform a deep dive comparative analysis of the catastrophic forgetting effects on multiple CL methods.
- Our results outperformed established literature baselines in 15,66%, surpassing the results of a joint-training approach, and even overcoming the metrics of a multiple-models scenario.

- Our method is the only approach able to produce positive knowledge transfers in both cases: forward (0.02) and backward (0.03).
- Our method produced more semantically accurate segmentation masks than the baselines.

2 THEORETICAL FOUNDATION

This chapter presents the fundamental concepts that are important do fully understand the method we implemented in this work, the next sections aim to cover each topic related to our work with a detailed overview of the area. Section 2.1 dissects the Histopathology field, its techniques, limitations, and improvements in the recent years. Section 2.2 exposes the AI artifacts that are pertinent to our method, along with their explanations and motivations in a more general way. The Sub-section 2.2.3 describes the problem of Continual Learning in computation and the documented methodologies that attack the catastrophic forgetting problem. Finally, Section 2.3 discusses the learning achievements of this chapter, its contribution to the problem understanding, and to our method design.

2.1 Histopathology

Histopathology is the clinical study of tissues and cells with their characteristics and abnormalities that could be the causes or the results of diseases, it consists of the microscopical analysis of surgical samples taken from patients for the diagnosis and screening of various tumors. It is the main clinical examination technique done by pathologists to detect cancer (biopsy). (MUSUMECI, 2014)

Histological techniques are very important in this context, they are employed by histologists or histopathologists as a pre-processing stage before the histopathological examination of a tissue. There are a series of preparations that have to be applied to the tissue, depending on the kind of tissue, the kind of structures on it, and the available laboratory chemicals, thus creating a high variability of non-standardized processes and protocols that are employed. This variability can produce bad samples with artifacts (contaminants) that could led to misinterpretations in the tissue (CHATTERJEE, 2014; SALVI et al., 2021).

Generally, there is a consensus about some preparations techniques: the tissue must be fixed to prevent cellular death, dehydrated and cleared using chemicals, then, frozen and sliced into thin histological slices, these slices would then be stained in order to make the sample ready for observation with a microscope or digitalization (MESCHER, 2018; SALVI et al., 2021). Figure 3 presents an example of protocol used to handle a sample. The common histology staining technique is the H&E stain (Hematoxylin and Eosin), Hematoxylin stains the nuclei within cells in blue and Eosin stains the cytoplasm of cells in pink. The H&E stain became so common because its suitable for light microscope examination, and light microscopes are preeminent in cytological and histological diagnosis on a daily basis, thus making H&E the gold standard for diagnosis (MUSUMECI, 2014).

At the year 1999, Whole-Slide-Image (WSI) scanners were introduced to provide the opportunity of digitally converting an entire tissue on glass slide into a high-resolution virtual slide. With the hardware limitation of the epoch, it was not possible to convert or store thousands of

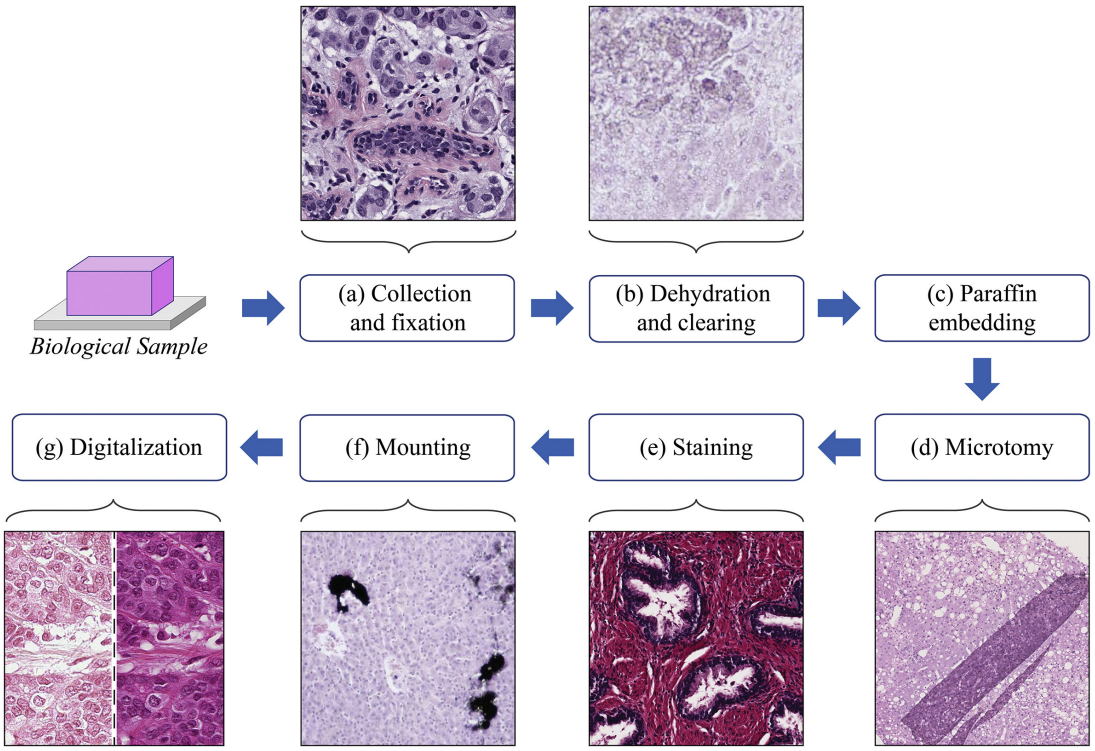

Figure 3: Example of a preparation protocol applied to a biological sample to produce the corresponding histological slide.

Source: (SALVI et al., 2021)

high-resolution images (KUMAR; GUPTA; GUPTA, 2020). Since then, there was an exponential growth of technology, and now, datasets of WSI images are available publicly, with recent trends on Computational Pathology (CPATH) evolving towards a more digital science. Digital pathology is promised to be of great help for mitigating human error in diagnosis and automating time-consuming pathology tasks, this, of course, is dependent of the solution of some computational challenges (KUMAR; GUPTA; GUPTA, 2020; KOTHARI et al., 2013; ZHOU et al., 2021; LAAK; LITJENS; CIOMPI, 2021).

Among these challenges, the capability of domain generalization is one of the most important, given it solves major problems in the pathology domain. From a operational perspective, the generalization decreases the need for strict protocols for sample preparations and slide staining. From a regulatory perspective, the performance of a single model model is more easily evaluated and explained. From an ethical perspective, generalizing the model to different populations removes a possible bias toward certain ethnicities. Finally, from a cultural perspective, the clinical society tends to trust more in a consistent and generalist solution to help clinical decision making (ZHOU et al., 2021; GONZALEZ; SAKAS; MUKHOPADHYAY, 2020a; LAAK; LITJENS; CIOMPI, 2021; JAHANIFAR et al., 2023b).

2.2 Artificial Intelligence

Artificial intelligence is a multidisciplinar research area that comprehends the human efforts of building intelligent entities, given that "intelligence is concerned mainly with rational action" and so "an intelligent agent takes the best possible action in a situation" (RUSSELL; NORVIG, 2016). Also, folks can understand AI as a form of knowledge refinement, where "a reliability and competence of codification can be produced which far surpasses the highest level that the unaided human expert has ever, perhaps even could ever, attain.". (CRAWFORD, 2021)

Even though there are many definitions, the core idea of automatically learning a task remains the same. Initially, the AI field focused on learning problems that were intellectually difficult for human beings but relatively straightforward for computers, like learning a formal mathematical ruleset. Nowadays, the true challenge to AI is on learning the tasks that are easy for people to perform but hard for people to describe formally, like recognizing spoken words or faces in images, and for these problems, we rely on modern machine learning techniques to improve the learning capabilities of AI. (GOODFELLOW; BENGIO; COURVILLE, 2016)

2.2.1 Machine Learning

Machine learning is the research field dedicated to study algorithms with the ability of learning a function through iterative data exposition, the function is, usually, an optimization of a certain objective over the given data. The learning process, called **training**, iteratively observes the function performance on the data and adjust the function parameters according to the deviation from the objective, the repetition of observations and adjustments end up optimizing the function parameters, and by so, the algorithm learned an optimized function for that certain objective. So, essentially, machine learning algorithms are a form of applied statistics, where computers are used to statistically estimate complicated functions and a decreased emphasis on proving confidence intervals around these functions. (GOODFELLOW; BENGIO; COURVILLE, 2016)

Although exist deviations from this general form of learning, there are three main classes of machine learning algorithms, that differ on the way that the learning process perceives the feedback from its adjustments: supervised learning, unsupervised learning and reinforcement learning.

Supervised learning algorithms use the data in pairs input – output, learning function parameters that could best match the input to the output using the objective function as the performance indicator. The name supervised comes from the idea that the learning algorithm has a teacher, that is, the correct output is available, this is not always the case given that, for some problems, the correct output for a certain input must be manually **annotated**, thus turning the data gathering an expensive process. For example, most histopathology datasets were manually annotated by specialists with the Region of Interest (ROI) for each image. (BROWNLEE, 2016)

Unsupervised learning algorithms use only the input data, learning function parameters that best model the data underlying structures to the objective function. The name unsupervised so denotes that there is no correct answer, the algorithms should discover structures in the data by themselves. The main example of this class is **clustering** algorithms, they are used to segregate and group data into clusters by their structural similarities. (BROWNLEE, 2016)

Reinforcement learning algorithms use context limited data, learning function parameters that best model a **behavior policy** into the context, using the objective function as a policy performance indicator. The name comes from the nature of the learning, the algorithm iteratively interacts with an environment and receives feedback for this interaction, the interactions are optimized as a behavior function, so the most rewardable interaction policy is learned. (GOODFELLOW; BENGIO; COURVILLE, 2016)

Machine learning algorithms make possible to machines to grasp statistical representations from data an learn a function to a certain task, but sometimes, the task complexity requires a more sophisticated function, that uses more powerful representations to understand the problem, in this scenario deep learning emerges.

2.2.2 Deep Learning

Deep learning techniques allows machines to learn more complex and sophisticated concepts by stacking learnable parameters on top of each other, and by so, learning representations of representations, this depth allows the algorithms to draw complex concepts out of the inputs and learn compound functions. Deep learning is specially useful for highly complex problems where the patterns are not easy to identity like, for example, our case study, the segmentation of tumorous cells may depend on the cell color, shape, structure or other characteristics that we may even do not notice. (GOODFELLOW; BENGIO; COURVILLE, 2016)

Deep learning takes a huge step from machine learning by implementing multi-layer models that allow the composition of representations, each layer is connected to the next, and the layers are composed of **perceptrons**, a computational abstraction of the learning method of a biological neuron (ROSENBLATT, 1958). This architecture creates a network of neurons divided into layers, hence the name **neural networks**, and, since each layer learns from the previous one, the composition of layers allows the composition of representations. (GOODFELLOW; BENGIO; COURVILLE, 2016) Equation 2.1 presents the notation of the processing of a perceptron, here, w and b are the learnable parameters weights and bias respectively, and x is the input. (HAYKIN, 2009)

$$y = \sum wx + b \tag{2.1}$$

Unfortunately, the simple composition of layers of perceptrons is not enough to learn complex concepts, this is due to the fact that perceptrons are only able to learn linear representations, as one can notice by looking at the perceptron linear Equation 2.1, and so does hap-

pens that complex problems have many nonlinear relationships between its representations. To implement this nonlinear characteristic were developed **activation functions**, these functions are attached between the perceptron layers and map the perceptron outputs into a nonlinear distribution. The most famous example of activation is the rectified linear unit (**ReLU**) (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), this activation function introduces nonlinearity by simply clipping the values bellow zero, Equation 2.2 presents the ReLU function. (GOODFELLOW; BENGIO; COURVILLE, 2016)

$$a = max(0, y) \tag{2.2}$$

Usually, the perceptrons and activations are implemented together into the same layer called **Dense** layer, for simplicity, also, this layer abstracts the logic behind the connections between layers, dense layers use a fully connected architecture, to all neurons from the previous layer are connected to all neurons of the current layer.

Deep learning made possible that we could construct models to learn complex functions made of compound representations of a problem, this allows us to tackle really complex problems, but even so, considering in our study case, image analysis, we would need hundreds or thousands of parameters per layer to learn representations of each pixel, thankfully, when it comes about deep learning for image analysis, there are more specific solutions.

2.2.2.1 Convolutional Neural Network

Convolutional neural networks (CNN) are deep learning networks that use a special type of perceptron, the convolutional **kernel**, to process grid-like inputs, such as images. Convolutional kernels can extract spatial knowledge of the inputs, in the case of images, the relationships that adjacent pixels have, using shared trainable parameters that are convoluted over the input, as a sliding window grid filtering a bigger grid. The idea behind convolutions is based on neuroscientific discoveries about how the mammalian brain perceives images on different layers, and that, at earlier layers, the neurons are more focused on identifying simple spatial patterns, like lines, edges and curves, and at later layers, the neurons react more to complex patterns. (GOODFELLOW; BENGIO; COURVILLE, 2016)

The Equation 2.3 describes how the convolution is applied, here, I and K are respectively the image and kernel (trainable parameters), i and j are coordinates in the image, and m and n are the dimensions of the kernel. Through this equation one can notice how the same kernel parameters are applied to different regions of the image, given by the coordinates, that when iterated over the image, produce a mapping the kernel pattern over the whole image. Supposing, for instance, that the kernel parameters identify vertical bars, this convolution will produce as output a feature mapping of the same dimensions of the image, highlighting all vertical bars on it. (GOODFELLOW; BENGIO; COURVILLE, 2016)

$$s(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i+m, j+m) K(m,n) \tag{2.3}$$

By analyzing Equation 2.3 we can also notice that i, j, m and n are parameters of the convolution, and so happens that they actually are hyperparameters of convolutional layers. The pair i and j are the coordinates of the region to be convoluted in the image, the iteration over these parameters convolutes the kernel over the whole image, the hyperparameter **stride** defines the step size at each iteration, controlling how much overlapping should the convolutions have. And the pair m and n are defined by the hyperparameter **kernel size**, that defines the dimensions of the kernel, usually, the kernels are squared, so $m = n$. (GOODFELLOW; BENGIO; COURVILLE, 2016)

The convolutional operation has the ability of learning patterns in grid-like inputs, but we must also consider that the angle of these patterns are not going to be the same for all inputs, that in some inputs it may be spatially translated, this problem makes training way more difficult, since some kernels will learn the normal input, and others the translated input, and later both will compete to detect the same pattern with slightly variations. (GOODFELLOW; BENGIO; COURVILLE, 2016)

Pooling layers help us aid this problem, a pooling operation consists in creating an output where every region represents some summary statistic of that same region over all the kernel outputs, this way, it alleviates the effects of some specific kernels by summarizing all kernel outputs. This produces an output that is a robust representation of the whole convolutional layer, being approximately invariant the spatial translations of the input. The common example is the max pooling operation, that produces an output with the maximum value at all regions over all the kernels. (GOODFELLOW; BENGIO; COURVILLE, 2016)

Pooling layers also have the same stride and kernel size hyperpermaters, but here we have an interesting feature with the stride. Since the pooling operation summarizes the kernels over a statistic, if some use a stride $s > 1$ with less pooling units than convolutional kernels, than the pooling operation actually produces an output that is roughly s times smaller than the input. This feature is called **downsampling** and is commonly used to reduce the dimension of the representations, this behavior consumes less memory and usually improves the overall performance. (GOODFELLOW; BENGIO; COURVILLE, 2016)

With pooling we can learn representations that are robust when dealing with different spatial translations of the inputs, but for the context of our work, we still we have to consider more representation deviations over the inputs, we must produce an approach that is robust to a whole new distribution of data, a whole new dataset.

2.2.3 Continual Learning

Continual learning, or lifelong learning, is the class of methods that enhance models with the ability of knowledge generalization when facing not one, but a collection of tasks over

its lifetime, using the knowledge acquired in the previous $n-1$ tasks to bias the learning of the task n (THRUN, 1998). In performance terms, the generalization of previous knowledge should allow the model to perform better at each subsequent task n (forward transfer) and also to perform better on the previous $n-1$ learned tasks (backward transfer), when revisited, given the accumulation of knowledge from the most recent task (HADSELL et al., 2020).

In the specific context of neural networks, CL is a family of techniques suited to solve the catastrophic forgetting problem, but ultimately, these techniques are only capable of smoothing the forgetting effect, and since this effect is so deeply rooted in the mechanics of the training process of neural networks, getting completely rid of the forgetting effect has become a very complex problem (HADSELL et al., 2020).

Essentially, **catastrophic forgetting** is an inherit effect of training neural networks using **stochastic gradient descent** (SGD) on sequential data, the SGD algorithm makes small updates on the weights of the network to minimize the error in a given input data, meaning that, the algorithm is prone to fit the model better to the last batches of data inputted, so when an already trained network is retrained in a new dataset, it tends to learn the new patters while it forgets the older ones (BOTTOU, 1998; FRENCH, 1999; HADSELL et al., 2020). Since catastrophic forgetting is a well known problem of neural networks, during the decades, the number of works implementing solutions of this problem grew a lot, fortunately, the recent literature prefers to classify the works into groups based on the scenarios of task relations and the methodology of the solution.

We can divide the CL problem into three common task scenarios depending on how the tasks relate to each other: **Task-incremental learning** (Task-IL), the model has to incrementally learn a sequence of distinct tasks (different problems); **Class-incremental learning** (Class-IL), the model has to incrementally learn to discriminate between a growing number classes (same problem, different classes); and **Domain-incremental learning** (Domain-IL), the model has to incrementally learn a sequence of distinct domains of the task (same problem, different distributions) (VEN; TUYTELAARS; TOLIAS, 2022). Given our problem context, generalization of histopathology segmentation across multiple datasets, we are actually dealing with a domain-incremental CL problem.

For the solutions, we can divide them by the paradigm adopted to deal with the problem, **Gradient Based solutions** aims to alleviate the forgetting effect focusing on the root of the problem, the gradients, by controlling the gradients, one can control the updates and prevent the weights from drifting away of present distributions, thus limiting the plasticity of the network on important parameters of previous tasks. One weakness of these methods is that they are based in approximations, so in challenging settings or complex domains they fail to achieve satisfactory performance (HADSELL et al., 2020; WANG et al., 2022). Examples of these methods are Elastic Weight Consolidation (EWC) (KIRKPATRICK et al., 2017) and Memory Aware Synapses (MAS) (ALJUNDI et al., 2018), both use heuristics to approximate of the importance of each weight of the network and penalize in the loss function when the optimizer

updates the important weights, while LwF (LI; HOIEM, 2017a) and KD (HINTON; VINYALS; DEAN, 2015a), update the loss function to optimize the outputs of one model to approximate the outputs of another.

Figure 4 illustrates the pipeline of training a generalized model over multiple tasks using a gradient-based method to perform the generalization. Notice how it only uses one model, and updates it with a new task style, while maintaining some of the old tasks style.

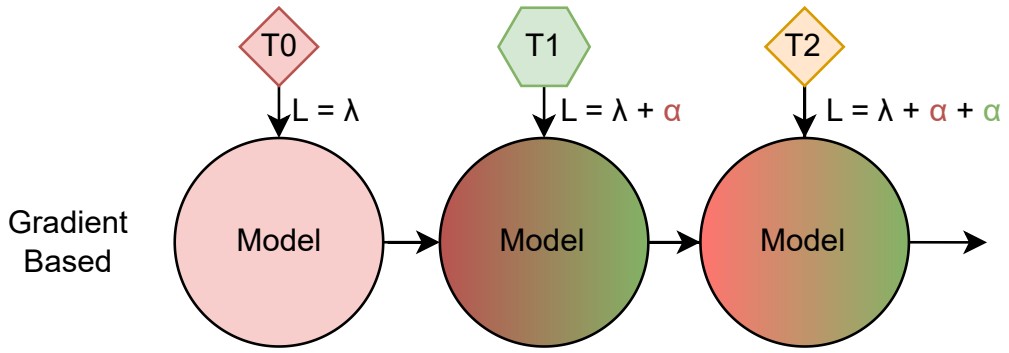

Figure 4: Illustration of the generalization process in a model using a gradient-based method. Tasks are $T0$, $T1$ and $T2$, while L is the total loss function, λ is the model loss function, and the αs are the regularization terms.

While **Architecture Based solutions** use modularity to prevent forgetting, these methods are focused on having disjoint network components for each task, by dividing the network into modules, with generalist reusable modules and task-specific modules, weights of some parts of the network (task-specific) can be updated without affecting the others modules. Since task-specific components are identified by expanding the network, the learned weights of previous tasks are separated in their own modules, non-affected by updates in the current task, thus preventing the forgetting. The downside of this paradigm is that, since new parameters are being added to the network as new tasks are encountered, the computational requirements of these methods is a constraint, and additional mechanics are required to manage the growing capacity of the network (HADSELL et al., 2020; WANG et al., 2022). Important examples of this paradigm are the LwF (LI; HOIEM, 2017b), that uses a KD sub-routine to ease the sparsity of the model, and DEN (YOON et al., 2017), that uses a heuristic with a threshold to control the expansion of the network.

Figure 5 illustrates the pipeline of training a generalized model over multiple tasks using a architecture-based method to perform the generalization. Notice how it expands the model size, adding a new layer on top of the older ones for each new task.

Memory Based solutions preserve old tasks knowledge by replaying samples (memory) of the these tasks along with the current task, these methods keep a buffer of old tasks data and, during the training of the current task, sample old examples to replay along the current task data. The main problem related to this paradigm is the buffer of old data, it is not scalable for scenarios with many tasks, and constraining the buffer buffer size deteriorates the performance, also,

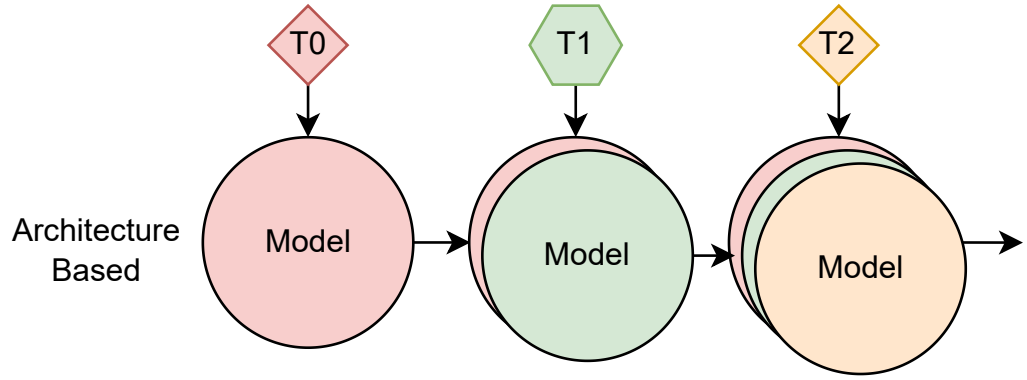

Figure 5: Illustration of the generalization process in a model using a architecture-based method. The tasks are $T0$, $T1$ and $T2$.

these methods are not applicable to scenarios where data from old tasks is not available (HADSELL et al., 2020; WANG et al., 2022). Three main examples of methods in this paradigm are rehearsal (ROBINS, 1995), episodic memory (CHAUDHRY et al., 2019) and ICarl (REBUFFI et al., 2017), all make use of replay to strengthen memory retention, but episodic memory and ICarl use heuristics to choose representative samples to store in the buffer, rather than saving all the old data. To deal with the buffer size problem there are also generative methods, that train generative models to produce rehearsal data as needed (SHIN et al., 2017).

Figure 6 illustrates the pipeline of training a generalized model over multiple tasks using a memory-based method to perform the generalization. Notice how it expands the training dataset every time a new task appears, adding a sample of the older tasks together with the new task data.

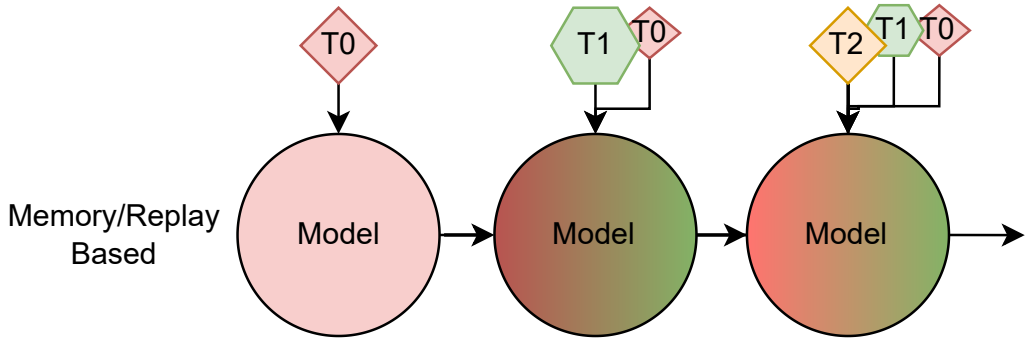

Figure 6: Illustration of the generalization process in a model using a memory-based method. The tasks are $T0$, $T1$ and $T2$.

And **Meta-Learning solutions** enable CL through optimizing the learning process itself, the optimization is data-driven, thus getting rid of the hand-engineered mechanisms of the previous solutions that creates an inductive bias towards trade-off solutions instead of a more automatic learning approach. Solutions in this paradigm usually focus in two loops, an inner-loop that performs a fine-grained training in the meta-learning problems, and an outer-loop that integrates

the coarse-grained knowledge at each iteration to optimize the future inner-loops. One characteristic of the methods in this paradigm is the fast adaption and recovery, meaning that the focus of this solutions is not to produce a model with good performance in all tasks, but instead a model with the capability of fast recovering the good performance (remembering) everytime an old task is forgotten. Given this aspect, meta-learning solutions are not suitable for scenarios where a task can never be forgotten (lose performance), besides this, meta-learning solutions are also very computationally demanding and require careful design of the task distribution (HADSELL et al., 2020). Implementations of meta-learning can vary a lot on the algorithms used in the inner and outer loops, but usually the goals are the same, the inner-loop focus on learning the task, while the outer-loop focus on speeding up the learning of the inner-loop, it could be implemented using gradient descent in both loops (ANDRYCHOWICZ et al., 2016), or probabilistic inference in the inner-loop (FINN; XU; LEVINE, 2018), or any mixture of methods.

Figure 7 illustrates the pipeline of training a generalized model over multiple tasks using a meta learning method to perform the generalization. Notice that the key operation is the model integration, an heuristic that produces a model by fusing the characteristics of other models.

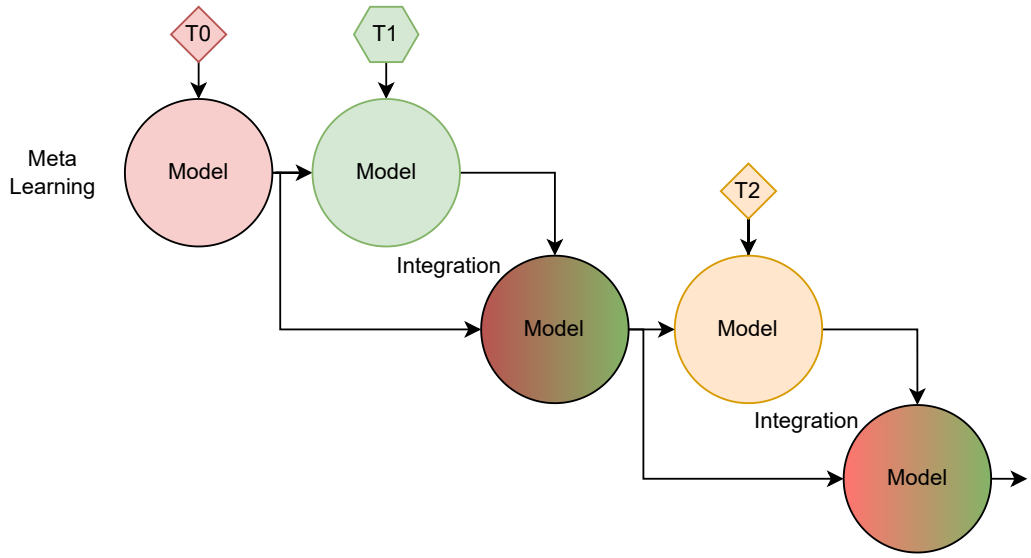

Figure 7: Illustration of the generalization process in a model using a meta learning method. The tasks are $T0$, $T1$ and $T2$, while the $Integration$ steps are the model-fuse operations.

With regard of the histopathology problem, the literature suggests that gradient-based solutions are more suitable, since they cons of the other approaches are too heavy on the real-world scenario of this problem, such as no data availability of old tasks, limited computational resources available in clinics, and the huge variability of data. Recently, reviewing works are confirming the capability of the gradient-based approaches to enforce generalization along multiple histopathological instances, suggesting that these methods are promising candidates for real-world implementations of CL in histopathology (BáNDI et al., 2023; AL-THELAYA et al., 2023; GONZALEZ; SAKAS; MUKHOPADHYAY, 2020a).

2.3 Discussion

In this chapter we researched the theoretical foundation of the fields involved in this work, we quickly reviewed the histopathology analysis process, the requirements and limitations for the future of this area, we also highlighted the importance of our drive in this work, the generalization of CPATH solutions.

we also made an in-depth review of the topics regarding AI that are relevant to our work, we covered the general techniques and architecture details about deep learning models, specificities of the applications involving image processing and finally the state-of-the-art methods known to deal with our object of study: model generalization.

With all that being said, we can better recognize the importance of all these aspects when building our approach, the analysis of histopathological images is a complex task containing many spatial variances between the images and non clear relationships between the elements of these images. Hence the need of our approach to use deep learning to find complex representations in the images, convolutions and poolings to efficiently draw these representations based on the spatial properties of the inputs. On top of that, we are dealing with a domain-incremental CL problem, so we also need an optimized CL method to generalize the learned representations across the tasks, and grant robustness to our model when it deals with new datasets.

The conclusions we can draw from this chapter are that Digital Histopathology using AI is far from a clinical application due to AI limitations, AI approaches mainly lack the generalization ability necessary to implement AI systems to augment the performance of clinicians. The analysis of this chapter draws a clear frontier of the current state of model generalization and where our efforts can be put into action to expand this frontier, the generalization is a long known problem in the core of the learning process of AI models, and, despite many efforts, this problem was not yet been solved.

In the next chapter we aim to study the application of generalization techniques in different domains of CPATH, analyzing the pros and cons of each implementation to grasp insights for our method.

3 LITERATURE REVIEW

In this chapter we performed a literature review in the area of pathology model generalization, we are going to use the highlights of this review to better understand the state-of-the art of the catastrophic forgetting solutions for digital pathology, we aim to gather insights and knowledge of modern methods to guide the design of our solution. In Section 3.1 we established our review methodology and criterias to select the works, then, in Section 3.2, we discussed the findings and interesting aspects of every work reviewed, and finally, in Section 3.3 we summarized the review, discussing our insights and the major decisions drawn.

3.1 Methodology

The goal of this review is to search on the literature for modern works that address the catastrophic forgetting problem in digital pathology, highlighting works that stand out by its contributions. The methodology used in this review is **Snowballing**, that consists of searching for relevant works in the references and citations of a list of works. Defined as snowballing backward, the analysis of the references cited by the selected works, and snowballing forward, the analysis of citations of the selected works (WOHLIN, 2014).

The first step is to select the list of works that would be used to perform the Snowballing, this is a very important step since the lack of relevant works may lead the review to miss important works (WOHLIN, 2014). We performed this initial search on **Google Scholar**. To increase our range of search, we collected surveys that cover the topics of current digital pathology challenges, we analyzed and filtered the surveys using the **Survey Criterias** list bellow, then selected the remaining works based on their novelty and relevance, the initial list of works and were documented in Table 1.

Survey Criterias

1. Should match on the query string: *deep learning AND digital pathology AND survey trends challenges*.
2. Should address or cite the problem of lack of generalization in the present methods.
3. Should be a relevant work (number of citations or novelty).
4. Should not be older than 2016.

Table 1: Initial Works

Work	Year	Citations
(GONZALEZ; SAKAS; MUKHOPADHYAY, 2020a)	2020	15
(KAUSTABAN et al., 2022)	2022	29
(LAAK; LITJENS; CIOMPI, 2021)	2021	395
(BAWEJA; GLOCKER; KAMNITSAS, 2018a)	2018	53
(BáNDI et al., 2023)	2023	5

Defined the initial set of works, Snowballing also requires a stopping criteria, in our case, we considered the state-of-the-art reached when no more works could be found given our **Article Criterias** list bellow, this way we can ensure that the recent literature was completely considered for reviewing. With these definitions we performed the Snowballing process over the initial set of works and used the new found works as the new initial set to perform the next iteration, going on until we reached the stopping criteria. Table 2 summarizes the article selection and filtering process, it took us 3 iterations to reach the stopping criteria.

Article Criterias

C1. Should match on the query string: *continual learning OR generalization*.

C2. Should not have been already reviewed in previous iterations.

C3. Should have a title or abstract focusing on generalization solutions for digital pathology.

C4. Should address the catastrophic forgetting problem.

C5. Should have at least one solution that does not take into account availability of old data, since this is a major constraint in the histopathology domain.

C6. Should be a relevant work (not surveys comparing off-the-shelf methods).

Table 2: Snowballing process

Iteration	Initial	Snowballing	C1	C2+C3+C4	C5+C6
Iteration 1	5	786	158	32	9
Iteration 2	9	511	195	16	1
Iteration 3	1	84	23	5	0
Totals:		1381	376		10

Defined the state-of-the-art we can finally perform a in-depth analysis of the final 10 works and a overall discussion of the whole process, the next section dissects the results of our literature analysis.

3.2 Results

The first aspect we would like to highlight is the disparity of works that use relevant and suitable methods for the pathology domain, we located 1381 recent works in the pathology area, from which, 376 papers approach the generalization problem, revising these papers, only 10 did not fail our criterias. So roughly **2.65%** of the works in the area approach the problem in a realistic way, by considering the old data unavailable, this scenario evidences the lack of concrete and robust works in the area, a major reason for the development of our method. With that been said, we now review each selected works according to its characteristics and implementations, for a overview of the findings we produced the Table 3 with some summary information of the works.

Continuous pathology classification in X-Ray chest images was done by (LENGA; SCHULZ; SAALBACH, 2020) using a DenseNet121 model, the pathologies such as: Cardiomegaly, Edema, Pneumonia and Pneumothorax were classifiend in two datasets: ChestX-ray14 and MIMIC-CXR. The model was trained in the first and fine-tuned in the second one, the generalization was performed by EWC and LWF, and the main metric was AUC. The experiments indicated that regularization techniques allow the model generalization while preserving performance on the original domain. Unfortunately, this method have the drawbacks of using EWC, such as fine-tuning the hyper-parameter and limited generalization in some contexts.

Another pathology classification approach was implemented in (YANG et al., 2023) work, the model used was a ResNet18, and the task was to perform pathology classification in the medical images of a famous medical dataset, the MedMNIST. MedMNIST MNIST-like 18 groups of 2D and 3D biomedical images of different pathologies. The generalization was performed using a domain constrained distillation-like loss, the main idea is to use pseudo-task data together with the actual first task to force optimize using separated domain spaces, then, incrementally replace the pseudo-tasks by real tasks. The main metrics are accuracy and performance dropping rate, the experiments concluded that this method can outperform related methods in the literature. Unfortunately, this method requires a explicit task enumeration before model training and a fine-tuned adjusting of the tasks used for the first training.

Glioma segmentation in brain MR images was performed by (GARDEREN et al., 2019) with a 3D Unet model, two datasets were used: 2018 BraTS and an in-house dataset, and the main metric was mean dice score. The generalization was performed with EWC and the experiments concluded that EWC alleviated catastrophic forgetting but also the restrained ability to the model to adapt to the a new domain. Also, training the model on a similar dataset and then generalizing it turns to be very effective, but in some cases the performance is not optimal as the same as two separated models.

Another 3D model was implemented by (BAWEJA; GLOCKER; KAMNITSAS, 2018b) to also perform a continuous segmentation of brain MR images, but this time it was not for glioma, but a semantic segmentation. The model used was the 3D DeepMedic and it was trained in two

datasets that were extracted from the UK Biobank. The generalization was performed by EWC and the main metric was dice score. The experiments concluded that EWC is a promising method for alleviating catastrophic forgetting. Unfortunately, this method have the drawbacks of using EWC, such as fine-tuning the hyper-parameter and limited generalization in some contexts.

Brain segmentation is a common topic given the availability of MR images, (ÖZGÜN et al., 2020) used a QuickNAT model to also perform a continuous MRI brain segmentation, three datasets were used: Child and Adolescent NeuroDevelopment Initiative, Multi-Atlas Labelling Challenge, and Alzheimer's Disease Neuroimaging Initiative. The generalization was performed using MAS and two MAS variations developed by this work, one controlling the learning rate based in the importances and one freezing some parameters based on their importance, the main metrics used are dice score, FWT (Forward Transfer) and BWT (Backward Transfer). The experiments concluded that both MAS variation methods performed better than MAS, specially the learning rate one, since it restrained changes to important network parameters in a more smooth manner. Also, this work noticed that normalizing outliers in the importance matrix leads to a more stable training and higher performance.

A more complicated approach was implemented by (MCCLURE et al., 2018) using a MeshNet to perform a continuous semantic segmentation, again in brain MR images, five datasets were used: Human Connectomme Project, Nathan Kline Institute, Buckner Laboratory, Washington University (WU120), and ABIDE Project. This work differs from the usual literature, it implemented a Variational Bayesyan Inference neural network, in which the weights are not fixed but actually distributions. This implementation allowed this work to create its method: Distributed Weight Consolidation (DWC), which turn the generalization problem into a distributed learning problem, and used Bayesian Inference to consolidate multiple networks into one. The main metric used was dice score, and the experiments concluded that this implementation can achieve a similar performance on all tasks that the separated models. Unfortunately, this method is very complex to implement, and since it uses Variational Bayesyan Inference, some problems can not have an optimal performance in this kind of model.

The adversarial paradigm was explored in the work of (MEMMEL; GONZALEZ; MUKHOPADHYAY, 2021), using a Image2Image Generative Adversarial Networks (GAN) and an Unet model to perform hippocampal segmentation on brain MR images, in this work, three datasets were used: 2018 Medical Segmentation Decathlon, Scientific Data and Alzheimer's Disease Neuroimaging Initiative. The generalization was performed by its own method called The Adversarial Continual Segmenter (ACS), that uses a GAN to regularize and encode the inputs into a variational space, making the Unet apply the segmentation over a generalized representation of the problem, the Unet output is then re-encoded by the GAN. The main metrics are Intersection Over Union (IoU) and dice score, the experiments documented drastic improvements in the performance compared to the state-of-the-art. Unfortunately, this method is very complex to implement.

Still in the adversarial approaches, but deviating from the brain focused works, a GAN and an Unet model were used in (CHEN et al., 2023) to perform two tasks: segmentation of optic discs to detect glaucoma, three datasets were used: Retinal Fundus Glaucoma Challenge, Indian Diabetic Retinopathy Image Dataset and Retinal Image database for Optic Nerve Evaluation for Deep Learning; and cardiac segmentation of MRI, with the multi-vendor multi-disease dataset from Cardiac Segmentation (M&Ms) challenge. The method used to perform the generalization was implemented din this work, Generative Appearance Replay for continual Domain Adaptation (GarDA), basically trains the GAN to output images similar to the ones from previous tasks and inserts these generated images into the training batches (replay), also KD was used in the GAN and Unet to prevent catastrophic forgetting. The main metric used was dice score, and the experiments concluded that the GAN can produce images of the old tasks with enough representability to avoid forgetting them. Unfortunately, this method has a complex implementation and can take additionally many hours to train the GAN.

Given the great potential of adversarial approaches, but also great complexity of implementation, (THANDIACKAL et al., 2023) implemented a relatively more simple strategy. The models were a GAN and ResNet, the task was to perform continual classification of colorectal biopsies with WSI, using three datasets: K-16, K-19 and CRC-TP, the images were cut into patches. The generalization was performed using multi-scale feature aggregation in the GAN and KD in the ResNet. The main metric is F1-Score, and the experiments concluded classifications of patches of histopathological images is feasible, and that the implemented method can outperforms other existing methods. Besides the more simple implementation of this method compared to other GAN-based approaches, unfortunately, it still requires a fine-tuning of the model capacity of the components of the GAN (feature extractor and domain discriminator).

Lastly, we have a transformer-based approach, the work of (RANEM; GONZÁLEZ; MUKHOPADHYAY, 2022) used a ViT (Visual Transformer) UNet to perform hippocampus segmentation, three datasets were used: Harmonized Hippocampal Protocol dataset, Dryad, e Medical Decathlon Challenge. The generalization was performed using EWC and RWalk, and the main metrics are dice score, FWT and BWT. The experiments concluded that transformers can be used together with regularization-based CL methods to preserve knowledge. Unfortunately, this method has limitations: the regularization could interfere in the self-attention mechanism over the training of more and more tasks, thus allowing it to forget the previous knowledge.

3.3 Opportunities

From this review we can notice the coverage of the generalization problem over the pathology literature, there are works considering classification (LENGA; SCHULZ; SAALBACH, 2020; THANDIACKAL et al., 2023; YANG et al., 2023) and segmentation (GARDEREN et al., 2019; BAWEJA; GLOCKER; KAMNITSAS, 2018b; ÖZGÜN et al., 2020; MCCLURE et al., 2018; MEMMEL; GONZALEZ; MUKHOPADHYAY, 2021; RANEM; GONZÁLEZ;

Table 3: Literature Review

Work	Model	Domain	Problem	Datasets	Generalization	Metric
(LENGA; SCHULZ; SAALBACH, 2020)	DenseNet	Chest	Pathology classification	2	EWC+LWF	AUC
(GARDEREN et al., 2019)	3D Unet	Brain	Glioma segmentation	2	EWC	Dice
(BAWEJA; GLOCKER; KAMNITSAS, 2018b)	3D DeepMedic	Brain	Semantic segmentation	2	EWC	Dice
(ÖZGÜN et al., 2020)	QuickNAT	Brain	Semantic segmentation	3	MAS	Dice
(MCCLURE et al., 2018)	MeshNet	Brain	Semantic segmentation	5	DWC	Dice
(MEMMEL; GONZALEZ; MUKHOPADHYAY, 2021)	I2I GAN+Unet	Brain	Hippocampus segmentation	3	ACS	IoU
(RANEM; GONZÁLEZ; MUKHOPADHYAY, 2022)	ViT Unet	Brain	Hippocampus segmentation	3	EWC+RWalk	Dice
(CHEN et al., 2023)	GAN+Unet	Eye	Glaucoma detection	9	GarDA	Dice
(THANDIACKAL et al., 2023)	GAN+ResNet	Colorectal	Biopsy classification	3	Custom+KD	F1-Score
(YANG et al., 2023)	ResNet	Multiple	Pathology classification	18	KD	Accuracy

MUKHOPADHYAY, 2022; CHEN et al., 2023), with domains ranging from optic discs (CHEN et al., 2023) to almost full-body (LENGA; SCHULZ; SAALBACH, 2020), and approaches as simple as off-the-shelf methods (LENGA; SCHULZ; SAALBACH, 2020; GARDEREN et al., 2019; BAWEJA; GLOCKER; KAMNITSAS, 2018b; ÖZGÜN et al., 2020) or complex domain-specific implementations (MCCLURE et al., 2018; MEMMEL; GONZALEZ; MUKHOPADHYAY, 2021).

We can also draw major decisions made by the literature, that can help us to develop our own solution, for instance, the Unet architecture is a common model to implement pathology segmentation (GARDEREN et al., 2019; MEMMEL; GONZALEZ; MUKHOPADHYAY, 2021; CHEN et al., 2023), regularization-based methods are a promising line of study to deal with forgetting (LENGA; SCHULZ; SAALBACH, 2020; GARDEREN et al., 2019; BAWEJA; GLOCKER; KAMNITSAS, 2018b; ÖZGÜN et al., 2020; RANEM; GONZÁLEZ; MUKHOPADHYAY, 2022) and, considering that KD is a regularization-like method, because it operates in the loss function, we can also see the importance of the loss-based mechanics and heuristics, even outside of the common regularization approaches (CHEN et al., 2023; THANDIACKAL et al., 2023; YANG et al., 2023). Also, in the matter of metrics we can see common metrics for classification works, but for segmentation, we noticed a preference for the metrics that measure of the intersection area, Dice score and IoU.

There are clear gaps in the modern literature works, such as high complexity, either done by specific methods (MCCLURE et al., 2018; MEMMEL; GONZALEZ; MUKHOPADHYAY, 2021; CHEN et al., 2023) or the customization of off-the-shelf methods (ÖZGÜN et al., 2020; RANEM; GONZÁLEZ; MUKHOPADHYAY, 2022), the lack of generalization capacity (LENGA; SCHULZ; SAALBACH, 2020; GARDEREN et al., 2019; BAWEJA; GLOCKER; KAMNITSAS, 2018b; ÖZGÜN et al., 2020; RANEM; GONZÁLEZ; MUKHOPADHYAY, 2022; THANDIACKAL et al., 2023), that is a reflex of limitations of the generalization methods, and a small

dataset list (LENGA; SCHULZ; SAALBACH, 2020; GARDEREN et al., 2019; BAWEJA; GLOCKER; KAMNITSAS, 2018b; ÖZGÜN et al., 2020; MEMMEL; GONZALEZ; MUKHOPADHYAY, 2021; RANEM; GONZÁLEZ; MUKHOPADHYAY, 2022), which does not really stress the generalization capability of the network.

These limitations work against our **Research Question** in Section 1.3, specifically when we consider the requirements of a real-world scenario that we enforced through this work. For instance, the high computational complexity of some methods prevent them from being used in common clinical hardware, also, the methods that lack generalization capacity or were validated in small datasets do not present strong indications that they could handle a higher variability of data or a greater dataset number.

With these insights we can better understand the complications of the CL problem in the histopathology settings, and evolve the literature on top of what was already discovered. In our work, we built against these limitations, countering these gaps with a novel approach based on the findings of this review.

4 METHOD

This chapter presents the algorithms and ideas used to implement our generalization method for histopathology models, in the next sections are going to discuss the aspects and details of the each mechanism within our method.

First, we briefly introduce our framework with an Overview (Section 4.1), where we present a general resume of the approach along with the connections between the functionalities that would be expanded in further details later. While in Model (Section 4.2), we presented the segmentation model that is going to be generalized over the datasets. Then we introduce our base Learning without Forgetting (Section 4.3) approach, used to shape the generalization methodology of our framework. And finally, we present our enhanced version of the base LwF and the core idea behind its design (Section 4.4).

In the final section of this chapter (Section 4.5), we digress over our implementation, with its theoretical strengths and weaknesses, bringing the necessity of practical experimentation to attest its validity.

4.1 Overview

In this work we implemented an enhanced method, based on the LwF approach, to deal with catastrophic forgetting in histopathology models. The specific family of catastrophic forgetting effect we are dealing with is the domain-incremental problem, where the task have multiple representational datasets that are significantly different from each other, thus, even though being the same task, one dataset induces the forgetting of the other. On top of that, these datasets should be learned sequentially since the arrival of a new dataset excludes the access of data from the previous datasets seen.

Our enhanced method improves the KD function of the off-the-shelf LwF approach, focusing on preserving the common knowledge between tasks through a specific regularization term, rather than simply regularizing the whole old task knowledge. In Figure 8 we illustrated the our method architecture, from the old model we derived two regularization terms, the direct old task knowledge term, and the common knowledge term.

This framework encapsulates an UNet segmentation model, initially, the model is trained on the first task using only its loss function, in the next tasks, this function is turn into a compound KD function, to add the regularization terms during the subsequent training loops. As a result, the final model produced after applying this framework across all tasks, summarizes the common knowledge of the multiple tasks and is generalist enough to perform the segmentation on any given patch of any dataset.

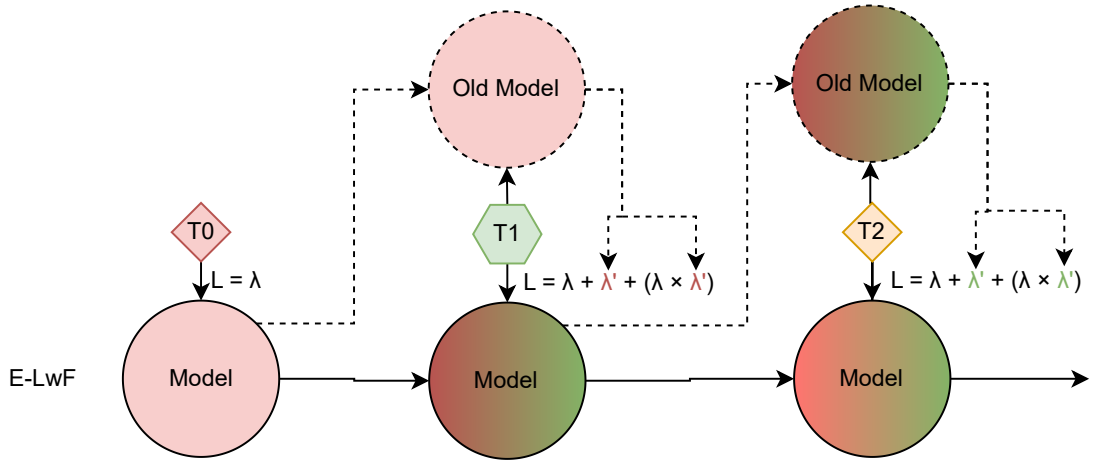

Figure 8: Illustration of our method generalization process, the model knowledge is generalized using the regularization terms extracted from the previous version of this same model. The tasks are $T0$, $T1$ and $T2$.

4.2 Model

To perform the segmentation of the tissues from multiple datasets we implemented a single **UNet CNN**, that is trained considering inputs x as image tissue patches of fixed size $256x256px$, and y as the $256x256px$ annotated segmentation masks for that patches, the model will then output a segmentation mask $\hat{y}$ of the same size $256x256px$. Since the generalization is done by routines outside the model, we have to make sure that the model can be trained in any dataset, for any dimension size, and with the fixed patch size we do not need to consider the original resolution of the image, and by so, train the model on any kind of dataset.

Our segmentation model follows an Unet architecture, this choice was made based in the strong Unet literature documenting remarkable results on medical problems, specially for segmentation problems (GARDEREN et al., 2019; MEMMEL; GONZALEZ; MUKHOPADHYAY, 2021; CHEN et al., 2023). Essentially, an Unet is an encoder-decoder architecture shaped like the letter *"U"*, that disposes the network into two towers of multiple levels, with **skip connections** linking layers in the same level from both sides of the network. The performance of the Unet is explained by its design, the encoder extracts local features from the image through max pooling, and downsamples the image until bottom layer of the encoder, while the decoder upsamples the image from the bottom to the his top layer, while propagating the local features from the encoder half in its layers (OLAF; PHILIPP; THOMAS, 2015).

Figure 9 presents the architecture of our segmentation model. The encoder levels are composed by 2 blocks of layers *Conv+ReLU+BatchNorm* and a final *MaxPool*, while the decoder levels are composed by a *ConvTranspose*, a concatenation with the skip-connection, and 2 blocks of layers *Conv+ReLU+BatchNorm*. Conv layers in the same level have the same number of kernels and the kernel size is top-bottom duplicated at each new level. All Conv layers use a *kernel size* of 3 and *stride* of 1, all MaxPooling use a *pool size* of 2 and a *stride* of 2, and all

ConvTranspose use a *kernel size* of 2 and a *stride* of 2. All Conv weights were initialized using *Uniform Xavier* initialization. Finally, the last encoder block does not have the final MaxPool layer, since information should flow upwards in the decoder, and the final decoder block is a *Conv* layer that outputs the segmentation mask, with a *kernel size* of 1, *stride* 1 and activation *sigmoid*.

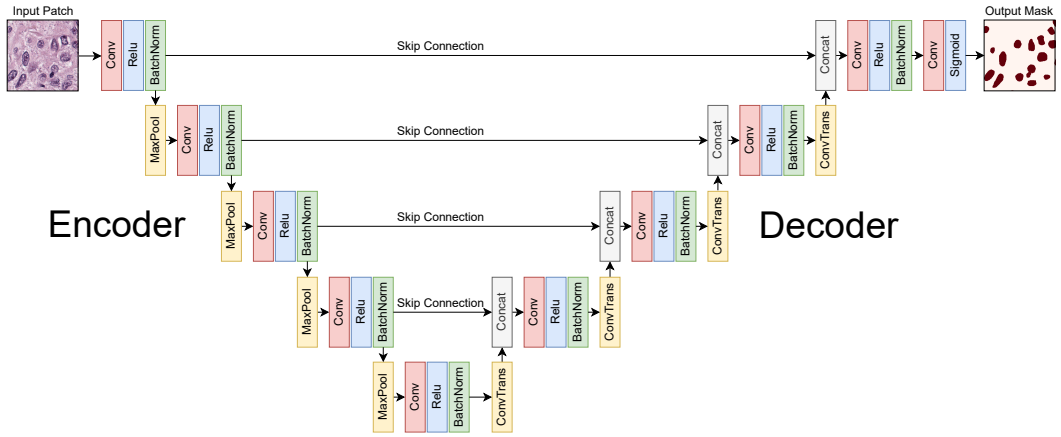

Figure 9: Architecture of our UNet segmentation model. The red layers are Convolutional layers, the blue layers are activations (Relu or Sigmoid), the green layers are Batch Normalizations, the yellow layers are dimension reduction (MaxPool) or expansion (ConvTranspose), and the gray layers are the concatenations.

To train our model we used the common **Adam** optimizer, with a learning rate of 0.0001, that will optimize our weights to the cost function **Dice loss**. The Dice loss is a variation form of the Dice metric, simply by subtracting the metric from one, $1 - Dice$, both the metric and the loss function are commonly used in the segmentation literature, further details of the Dice metric are discussed in Section 5.1.3. Equation 4.1 presents the formula of our loss function, where X represents the model predicted output and Y the actual ground truth.

$$Dice\_Loss = 1 - \frac{2 \times |X \cap Y|}{|X \cup Y|} \tag{4.1}$$

4.3 Learning without Forgetting

To generalize the segmentation model knowledge we development an enhanced version of the common Learning without Forgetting (LI; HOIEM, 2017a) approach. The LwF method is a regularization-based CL technique that uses the predictions made by an older version of the model, previous to the current training task, to preserve the knowledge of the old tasks (LI; HOIEM, 2017a).

Instead of directly preventing the model weights from deviating of their current distributions, like most regularization-based approaches does, LwF focus in preserving the old model outputs during the new task training, thus allowing the weights to change freely. This strategy is inspired on the *modus operandis* of the Knowledge Distillation (HINTON; VINYALS;

DEAN, 2015b) method, where the outputs of a pre-trained teacher model are distilled, through loss function composition, into the the predictions of a student model, improving its learning capacity and, usually, outperforming a model trained independently.

For simplicity, LwF could be understood as a weighted hybrid of KD and fine-tuning (LI; HOIEM, 2017a). In Figure 10 we illustrated the pipeline of training a generalized model over multiple tasks using the LwF method, notice how, before training in the new task, the model is copied and its predictions over the new task data are used to regularize the training of the original model.

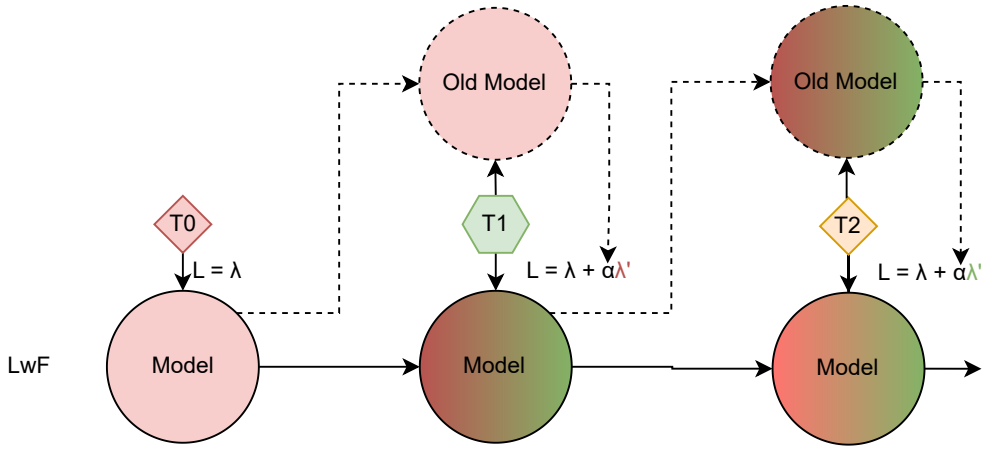

Figure 10: Illustration of the generalization process in a model using the Learning without Forgetting. The tasks are $T0$, $T1$ and $T2$, and α is the regularization weight.

Using the distillation of the output, LwF permits the learning of parameters that are discriminative for the new task while preserving outputs for the original tasks on the training data, converging for different parameter spaces that produce similar outputs to the ones produced before the new task arrival. Even though this approach is very different from most regularization techniques, it is still a form of regularization, since the distillation is performed by composing the loss function. So the LwF also has a common limitation of regularization methods, the vanishing representations of older tasks (HADSELL et al., 2020; WANG et al., 2022).

This effect is a result of the knowledge retaining mechanics of the regularization methods, the representations of the older tasks knowledge tend to fade away from the current training loops. In a sense, LwF alleviates the effects of catastrophic forgetting, but it can not completely avoid it, so as more and more tasks arrive, the accumulation of small catastrophic forgetting effects from the endless tug-of-war dynamics between stability and plasticity causes a gradually knowledge fading effect, that could be worsened by the lack of similarity between tasks (FENG et al., 2022).

4.4 Enhancement

The improvement we made on the LwF method was designed after many experiments and investigations on its KD function, to understand the weight that the regularization has on the general performance, and also the impact of each dataset arrival on it.

The original distillation function has a hyperparameter λ on the regularization term, which is a fixed value scaling the importance of the old task knowledge on the new training loop. According to the original KD paper, this hyperparameter should preferably not be a small value, supported by performance experiments with some variations of λ (HINTON; VINYALS; DEAN, 2015b). Likewise, the LwF original work, as of implementations of it, follow the same rule, using bigger λ values (LI; HOIEM, 2017a). To attest this effect, we performed manual experiments with multiple λ values. In summary, we noticed the same effect for our problem, the overall model performance peaks around $\lambda = 2$, hardly improves after this threshold and eventually decreases with greater values. Also, we experienced a surprisingly good performance on some individual experiments when $\lambda = 1$, with results that did not appear in a lower or greater λ values, but they were sporadic and the overall performance considering multiple runs was not superior to $\lambda = 2$.

We and concluded that the key logic behind an efficient knowledge retention is to shift the trade-off, between old and new knowledge, slightly into the direction of the old knowledge. Unfortunately, this means that we are sacrificing the convergence of new tasks, a proper generalization of the problem, and condemning the model to get stuck into the old representations learned, which, given the random nature of the weight initialization of neural networks, may not always be the most flexible representation to generalize. Also, in the topic of randomness, the sporadic good performances then $\lambda = 1$ seem to be explained by the random weights initialization and/or the first task training. We concluded that, if a model has learned very good representations already, whether is by good initialization, and/or by the first training loop, the stochastic gradient optimization algorithm can solve the generalization trade-off by its own with good results and without any hyperparamter.

Considering this findings, we designed a dynamic distillation function, in which, the scale factor λ, was replaced by a secondary regularization term. We did not simply turned λ into a calculated value because the **multiplication** of the scale factor is a weighting operation that is too harsh to by dynamically approximated, it would have to be fine-tuned with careful precision given that small changes on it produce big effects on the model knowledge. The secondary term is an **addition** to the regularization part of the loss, which is more robust to mistakes because it distributes better the regularization importance between two terms. Comparing the Figure 10 and Figure 9 in contrast, one can notice the similarities the differences between the base LwF and our enhanced version.

$$
\begin{aligned}
Continual\_Loss = & \; Dice(gt,\ pred) \\
& + Dice(old\_pred,\ pred) \\
& + Dice(gt \cap old\_pred,\ gt \cap old\_pred \cap pred)
\end{aligned}
\tag{4.2}
$$

Consequently, our approach has a compound loss function (Equation 4.2) with three terms: convergence term, retention term (first regularization) and generalization term (secondary regularization). The **convergence term** is the default loss function, it compares the predictions against the new task ground truth, its purpose is to fit the weights considering the new task knowledge.

While the **retention term** is the default LwF loss regularization, it compares the predictions against the old task predictions (soft ground truth), its purpose is to force the weights optimization to also consider how the model used to respond to the new task data, before being trained on it. This term then creates the trade-off between convergence and knowledge retention.

Finally, the **generalization** term is introduced by our approach, it compares the intersection between the predictions, the ground truth and the soft ground truth, against the intersection between the ground truth and the soft ground truth. Its purpose is to leverage the common knowledge between the old and new task knowledge, and also use their similarity as a strengthening regularization factor, like the λ hyperparameter.

The Algorithm 1 presents the method implementation used in our approach, the green part of the code represents our main contribution in this work, the generalization term of the compound loss function.

Algorithm 1 Enhanced-LwF

Given τ tasks, $\mathcal{L}$ loss, α learning rate
$\theta \leftarrow segmentaion\_model()$
for i in τ **do**
  if $i = 1$ **then**
    for x, y in τ_i **do**
      $y^{'} \leftarrow \theta(x)$
      $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}(y, y^{'})$
    end for
  else
    $\phi \leftarrow \theta$
    for x, y in τ_i **do**
      $y^{'} \leftarrow \theta(x)$
      $y^{''} \leftarrow \phi(x)$
      $\theta \leftarrow \theta - \alpha \nabla_\theta (\mathcal{L}(y, y^{'}) + \mathcal{L}(y^{''}, y^{'}) + \boldsymbol{\mathcal{L}(y \times y^{''}, y \times y^{''} \times y^{'})})$
    end for
  end if
end for=0

The leverage of common knowledge in the generalization term happens because the extra

loss value from these common areas between tasks forces the gradient optimization to seek representations that are common to both tasks, thus producing more generalist representations at each new task.

And the strengthening factor happens because the generalization term is controlling the trade-off between the old and new knowledge, its value is directly proportional to the tasks similarity and inversely proportional to the prediction intersection. That is, when the tasks are completely different (no intersecting areas), the term value is zero, so the optimizer solves the trade-off by himself. Whereas if the tasks are similar (have intersecting areas), the term value can be high, if the predictions do not fit into the areas, or low, if otherwise, so the optimizer tends to focus on the loss on the common areas, which is usually more easy to fit than the whole trade-off. Also, having two regularization terms tends to increase the overall regularization factor (old knowledge importance), which produce better final results, supported by the literature (LI; HOIEM, 2017a; HINTON; VINYALS; DEAN, 2015b).

In summary, the key idea behind our strategy is to force the model to learn the shared representations first, when the generalization term has a high value, because they are **easier** to optimize than the whole trade-off. Later, when the term value is lower, the optimization has more freedom the wander between the knowledge trade-off, but now, having learned the shared knowledge first, there is a better chance that these shared representations have **smoothed** the trade-off problem. Usually, the optimization would have an overall regularization loss greater than the convergence loss, and, in the worst case scenario, where tasks are completely different, the optimization would have to deal with the whole trade-off.

To better illustrate the loss computation we mounted the Figure 11, that contains the segmentation masks retrieved when a new task arrived during a CL experiment, the three figures at the bottom represent the masks, in order, for each term in our compound loss function, with their dice score bellow. Notice how the tasks have high similarity, since the ground truth and old predictions are very similar, and the dice score for the first term (convergence) image is way bigger than the others, which allows the optimizer to focus more on the new task knowledge. Also, notice how the third term (retention) image represents a middle ground between both tasks, which is easier to train into than each task independently.

4.5 Discussion

In this chapter we presented in details the implementation and modeling of our approach, that, in resume, consists of an enhanced version of LwF method applied over an UNet segmentation model trained across multiple tasks sequentially.

Through literature investigation and manual experimentation we designed a new regularization term for the KD function of the LwF approach, this term is responsible for leveraging the common knowledge between tasks, improving knowledge retention, and dynamically strengthening the regularization factor, improving task convergence, both effects having a positive impact

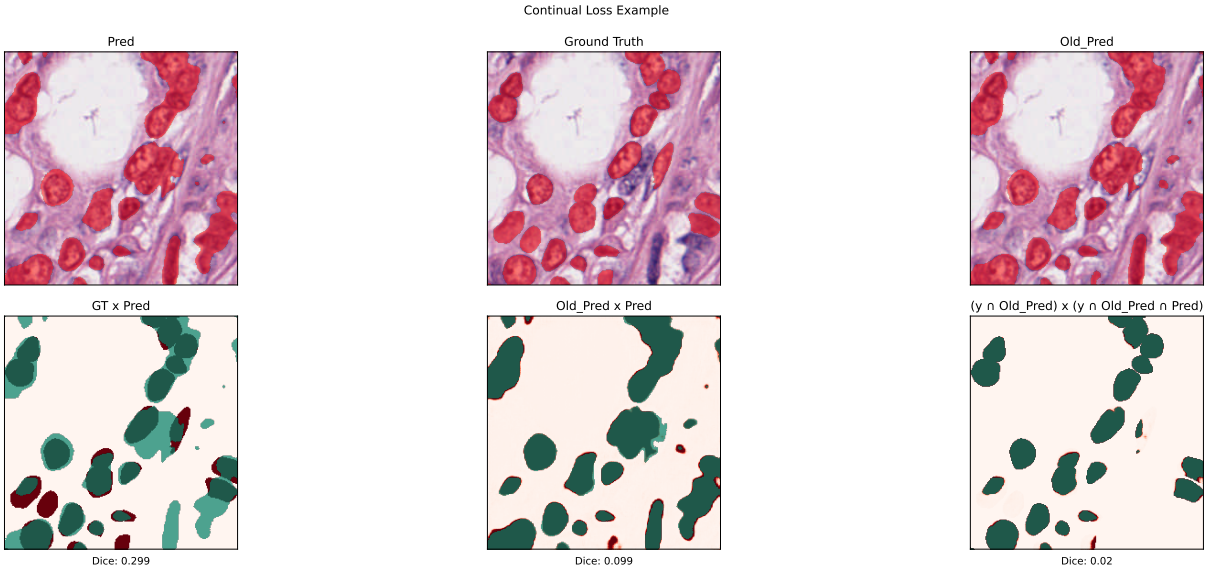

Figure 11: Illustration of the continual loss masks and values mid training. The last three images represent each, in order, the compound loss function terms, the titles on top of the images are referring to their term form. Since these terms are functions (Dice) comparing the truth with a value, the dark red areas represent the truth, and the green areas represent the values.

on the generalization capacity of the model.

Since our method works as an improved loss regularization to distill knowledge, we are relying on the stochastic optimizer to pickup our loss *"hints"* and behave the way we intended. Despite preliminary experiments of method design showing promising results, the stochastic nature of the approach also comprehends the possibility of the optimizer behave differently. For instance, if the present representations learned would be way more easily drifted to a new task, than to keep aligned with the previous ones, that is, the changes in the first term reduce the total loss way lower than the regularization terms, the optimizer would obviously induce the forgetting. Moreover, depending of the difficulty of the tasks, the model could also fail to reach a balance between two tasks, causing lack of convergence in one, and catastrophic forgetting in the other. Additionally, the order of the tasks also influences on the generalization process, if too many tasks with low similarity arrive consecutively, it is probable that each tug-of-war would then take a bite off the generalized knowledge, eventually corrupting the already learned representations.

In summary, since our method depends on the **imbalance** between the convergence and regularization terms, if they, somehow, tend to balance each other without a good knowledge representation learned yet, our new regularization term would be ignored and the optimizer would then try to tackle the whole trade-off problem, probably inducing catastrophic forgetting.

In this chapter we presented our approach with mathematical notations and detailed modeling explanations, justified our improvements in comparison to the literature, highlighted our insights and derived the limitations of our design. Although theoretically efficient, we barely validated our approach in this chapter, performing only preliminary experiments. In the next

chapter we are going to numerically validate our assumptions and evaluate our method performance with real-world histopathology datasets.

5 EXPERIMENTAL EVALUATION

This chapter presents the experiments performed to evaluate our approach, here, we compare our results with the results obtained from other similar methods present in the literature. Therefore, our experiments not only aim to document and explore our method numerically, but also semantically, by analyzing our outputs considering in the histopathology problem. Finally, we seek to prove the feasibility of our approach given our specific histopathology domain constraints, answer our research question and evolve the present state-of-the-art.

The Section 5.1 defines the methodology of our experiments, our datasets, baselines and metrics, together with the computational resources employed into the experiments.

The experiments we performed were divided into four sections, in Section 5.2.1 we validated our model and gathered insights about the datasets, while in Section 5.2.2 we validated our CL method and compared the results with our baselines, in the Subsection 5.2.2.1 we made an in-depth analysis of the results we obtained in the previous experiment. Additionally, in Section 5.2.3 we performed a semantic analysis of the outputs produced by the methods.

Finally, in Section 5.3 we revisited our method design, linking it with the results we obtained, and answered our Research Question. Also, we raised concerns and limitations found through our experimentation.

5.1 Methodology

The purpose of our experiments is to evaluate the capacity of our approach to actually generalize the knowledge of multiple instances of the same task over one single model. The generalization will be measured according to our metrics, and compared against the same metrics achieved by our baselines, also the outputs of our method should resemble, in a semantic meaning, to the histopathology ground truths.

To evaluate the generalization capacity over our datasets of N tasks, we designed the experiments to be independent iterations of training sessions with 10 epochs each tasks, we are using only 10 epochs because the focus of our analysis is the continual metrics, and the comparison of methods over the same configurations, we do not intend to reach the best metrics possible for the problem. Our general procedure for one experiment iteration is documented in the list bellow.

In order to produce more trustworthy experiments we performed 30 independent iterations, so the statistical results documented in our work were integrated over these iterations. The only exception to our training methodology is the **full** dataset baseline, being just a joint training of all datasets, it does not implement any kind of generalization technique other than what the model itself could learn during the training.

Iteration Steps

1. Randomly initialize the model M.
2. Train the model M for 10 epochs over the task T_i with regard of the **generalization algorithm**.
3. Evaluate the model M **metrics** over the task T_i.
4. Repeat Step 2 and 3 for the next task T_{i+1}, for all tasks T.
5. Evaluate the model M **metrics** over all tasks T.

All experiments were conducted on a dedicated machine with a 3.0GHz CPU, a 16GB GPU and 64GB of RAM. For the sake of clearness, all of our experiment parameters were documented in Table 4.

Table 4: Experiment parameters

Parameter	**Value**
Iterations	30
Datasets	19
Optimizer	Adam
Learning Rate	0.0001
Loss	Dice
Metrics	IoU, Dice
	BWT, FWT
Batch size	32
Epochs	10

5.1.1 Datasets

For our experiments, we made use of the Pannuke (GAMPER et al., 2020) dataset, it is a semi automatically generated histopathology dataset for cell nuclei segmentation and classification. Pannuke has more than 200,000 labeled nucleis in more than 7,000 patches ($256x256px$) distributed between 19 different tissue types, sampled from more than 20,000 WSI of different magnifications and from multiple data sources. This dataset is exhaustively annotated with 6 masks for clinically important classes: neoplastic cells (tumor), inflammatory tissue, connective/soft tissue cells, dead cells, epithelial cells and background area.

The Pannuke dataset was chosen given the high variability it has on its data, while maintaining a separability by tissue types, this is exactly the scenario we need for performance evaluation. Also, this dataset has the images in patches, not huge WSI files, so there is no need to pre-process the images beforehand.

To create the multiple tasks for our training routine we collected some samples separated the dataset by tissue type, creating 19 datasets, Table 5 summarizes the datasets and their instance sizes for training and testing. For our ground truth we used only the **neoplastic cells** mask, although Pannuke had 6 annotation masks, we are going to focus on one task: tumor segmentation, so the only annotation mask been used is neoplastic cells.

Table 5: Datasets

Tissue	**Train size**	**Test size**	**Total**
Adrenal gland	305	132	437
Bile duct	294	126	420
Bladder	102	44	146
Breast	1645	706	2351
Cervix	205	88	293
Colon	1007	433	1440
Esophagus	296	128	424
Head & Neck	268	116	384
Kidney	93	41	134
Liver	156	68	224
Lung	128	56	184
Ovarian	102	44	146
Pancreatic	136	59	195
Prostate	127	55	182
Skin	130	57	187
Stomach	102	44	146
Testis	137	59	196
Thyroid	158	68	226
Uterus	130	56	186
Totals:	5521	2380	7901
Averages:	290	125	415

To quickly highlight the data variations that our dataset has, we plotted some samples that exemplify different characteristics of the labeled tissues. In Figure 12 we presented two samples of the same tissue type and their annotations, notice how, even in the same tissue type, the characteristics of tumor cells differ a lot, this brings even more complexity to the generalization methods deal with. And in Figure 13 we presented three samples from different tissues types, notice now that not only the characteristics of the labels have changed but also the texture of the background, again, these variations increase even more the difficulty to perform knowledge generalization across tissues.

5.1.2 Baselines

To evaluate our method we constructed 4 comparative baselines: full dataset, EWC, MAS and LwF, all of them were constructed on top of the same model used in our approach.

The **full** dataset baseline was created by training the model in one single training loop using

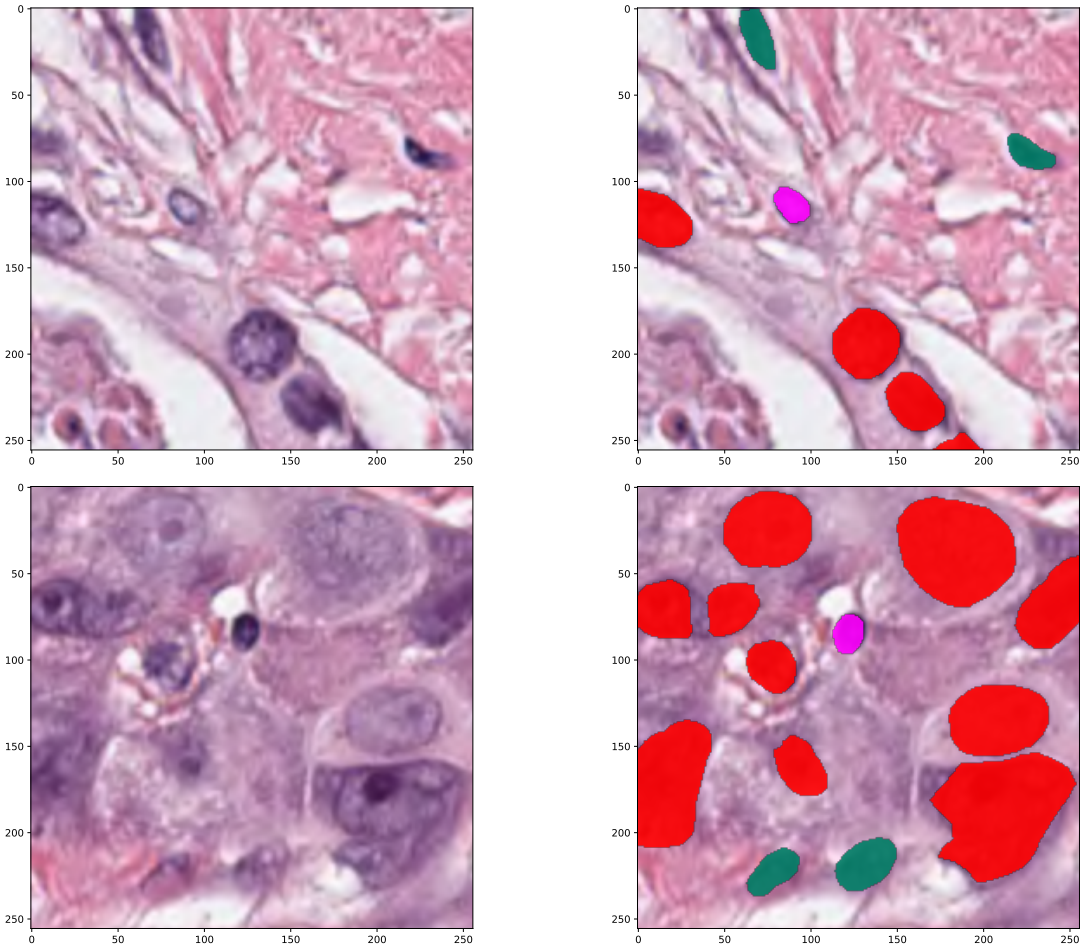

Figure 12: Breast gland tissue samples (on the left) and their respective annotations (on the right) from the Pannuke dataset. The annotations colors are red for tumors, pink for inflammatory tissue and green for connective/soft tissue.

all the datasets at once (joint training), intercalating the batches of each dataset. This baseline represent the results that the model could achieve if we had all the data available during training, which, as discussed before in Section 1.2, is not possible in the normal histopathology setting. Considering that this baseline uses all of our datasets at once, we had to keep this baseline model training for more epochs to avoid underfitting, after manual evaluation, we used **100** epochs for this baseline. This change in the training methodology was only done for this baseline, given the special case of using all of the available data in one training session.

The **EWC** and **MAS** baselines are different than our full dataset one, they follow the same training methodology of our approach, described in Section 5.1. For both of them, the regularization value was $\lambda = 100$, that is the only hyperparameter specific for these methods, controlling how much weight does the regularization term have. These two baselines were selected because they are the most documented methods in the literature, so their performance here represents the state-of-the art results that the model could achieve with off-the-shelf common methods.

The **LwF** baseline also follows our default training methodology, with the regularization

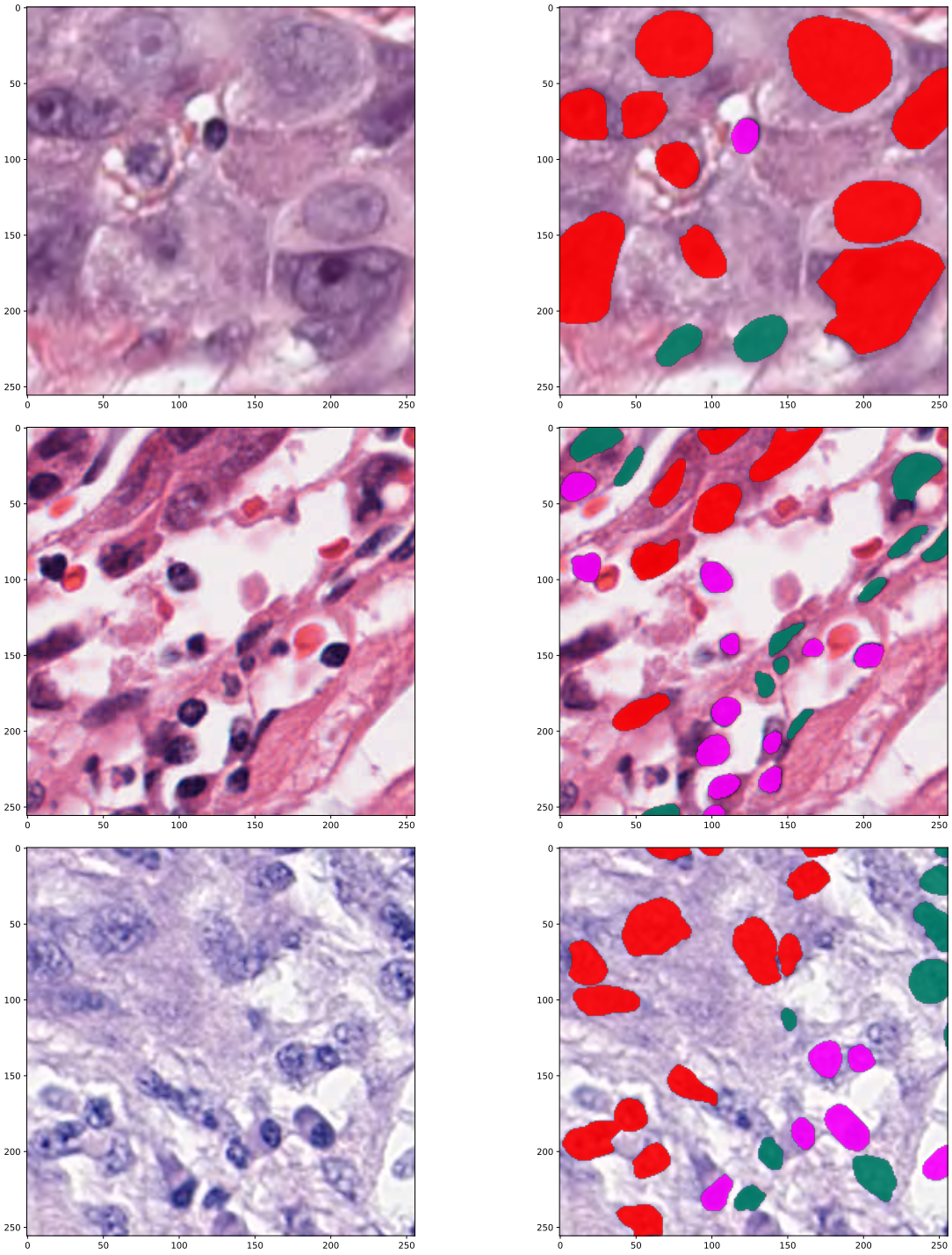

Figure 13: From top to bottom: breast, liver and head, tissue samples (on the left) and their respective annotations (on the right) from the Pannuke dataset. The annotations colors are red for tumors, pink for inflammatory tissue and green for connective/soft tissue.

term being $\lambda = 2$. Although not so common in the literature, this baseline offers a direct comparison with our method, since, in this work, we implemented an enhanced version of LwF.

5.1.3 Metrics

To measure the performance of our approach and compare it to the baselines, based on our Literature Review in Section 3, we elected four metrics into two groups: segmentation metrics and continual metrics. Given that our proposed model in this work is a binary segmentation model, we elected Dice Score and Intersection-over-Union (IoU) as our segmentation metrics to measure the model performance. Also, since our proposed method is a CL approach, we used Forward and Backward Transfer as our continual metrics, to numerically evaluate the model performance over the time.

5.1.3.1 Segmentation Metrics

The IoU index, also known as Jaccard index, is a widely used metric in 2D or 3D problems involving segmentation, object detection or tracking, the recurrence in the literature is due to the simplicity and low computational cost associated of its implementation. This metric calculates the similarity between two arbitrary shapes, and can be described as the normalized measure of the intersection of two areas or volumes. For our specific use case, we consider one of the shapes as the segmentation mask outputted by the model, while the other is the ground truth mask. Equation 5.1 presents the equation of the IoU metric.

$$IoU = \frac{|X \cap Y|}{|X \cup Y|} \tag{5.1}$$

The Dice metric, or Dice-Sørensen coefficient, is very similar to the IoU metric, with the only difference being a fixed 2 weight on the intersection, this weight tends to alleviate the output value of this metric for shapes with very low similarity. This is done because the dice score can also be used as a loss function to train segmentation models, like in our case, and, in this sense, the fixed weight alleviates the effects of really bad predictions during training. Equation 5.2 presents the equation of the Dice metric.

$$Dice = \frac{2 \times |X \cap Y|}{|X \cup Y|} \tag{5.2}$$

5.1.3.2 Continual Learning Metrics

The Forward and Backward Transfer (LOPEZ-PAZ; RANZATO, 2017) are metrics designed to evaluate, respectively, the forgetting effect and the intransigence capacity of a network over a series of continually learned tasks. These metrics are widely common in the CL literature since they provide numerical results to measure the generalization ability of a method based on its retention (forgetting) and convergence (intransigence) trade-off. It is **important** to understand that both of these metrics are not actually metric functions per se, they are common equations

based on a training methodology to evaluate CL methods. So, in this work, we are computing based on the **IoU** metric.

Forward Transfer (FWT) measures the model intransigence, its inability to learn a new task, we used this metric to evaluate how well the model can generalize its present knowledge when confronted with a new task. If a task have *positive* FWT, it means that, after learning this task, the model performed better on the next tasks, *negative* FWT is otherwise (LOPEZ-PAZ; RANZATO, 2017). The Equation 5.3 is used to calculate FWT, here, t is the number of tasks, $\bar{b}$ is the baseline metric from a model with all the data available (our FULL baseline), and $R_{i,j}$ is the metric obtained from a model trained in task i and evaluated in task j. This equation can be basically understood as the average difference between the metric of task i before training on it and its baseline metric.

$$FWT = \frac{1}{t-1}\sum_{i=2}^{t} R_{i-1,i} - \bar{b}_i \tag{5.3}$$

While Backward Transfer (BWT), measures the model forgetting, its incapacity to retain the knowledge of older tasks, we used this metric to evaluate the effect that a new task have over the performance of previous learned tasks. If a task have a *positive* BWT, it means that, after learning this task, the model performed better on the previous tasks, on the other hand, *negative* BWT is otherwise, a large negative BWT can be understood as **catastrophic forgetting** (LOPEZ-PAZ; RANZATO, 2017). The Equation 5.4 describes the calculation of BWT following the same mathematical notation of FWT. It can be understood as the average difference between the final metric of task i and its metric when the model was fit to it.

$$BWT = \frac{1}{t-1}\sum_{i=1}^{t-1} R_{t,i} - R_{i,i} \tag{5.4}$$

5.2 Results

In these sections we are going to present the partial and final results and analysis obtained from our experiments. First, we validated the model without any generalization technique, to establish the base performance of our segmentation model in each task. Later, we conducted the experiments for CL evaluation, measuring the capacity of the baseline algorithms against our method, in generalizing the models knowledge abroad the different tasks. Finally, we documented the outputs produced by each generalized algorithm, analyzing and comparing them considering the consistency and semantic meaning of the predicted masks.

5.2.1 Model

Our first experiment focus on our model performance assessment, as a way to measure how well does our model learns the segmentation of the different datasets. Considering that the model was trained in each dataset independently for 10 epochs, Figure 14 displays, for each separated task, the training and validation loss progression during the training epochs. Analyzing this figure, we can draw the conclusion that the model weights are converging with only 10 epochs, without overfitting, for all datasets. By looking at those curves we can also visualize problematics involving the performance of some tasks, it requires a more in-depth analysis of these cases. But, in the context of this initial analysis, we **successfully** validated our model, proving that it is suitable to perform the cell segmentation in any of our histopathology datasets.

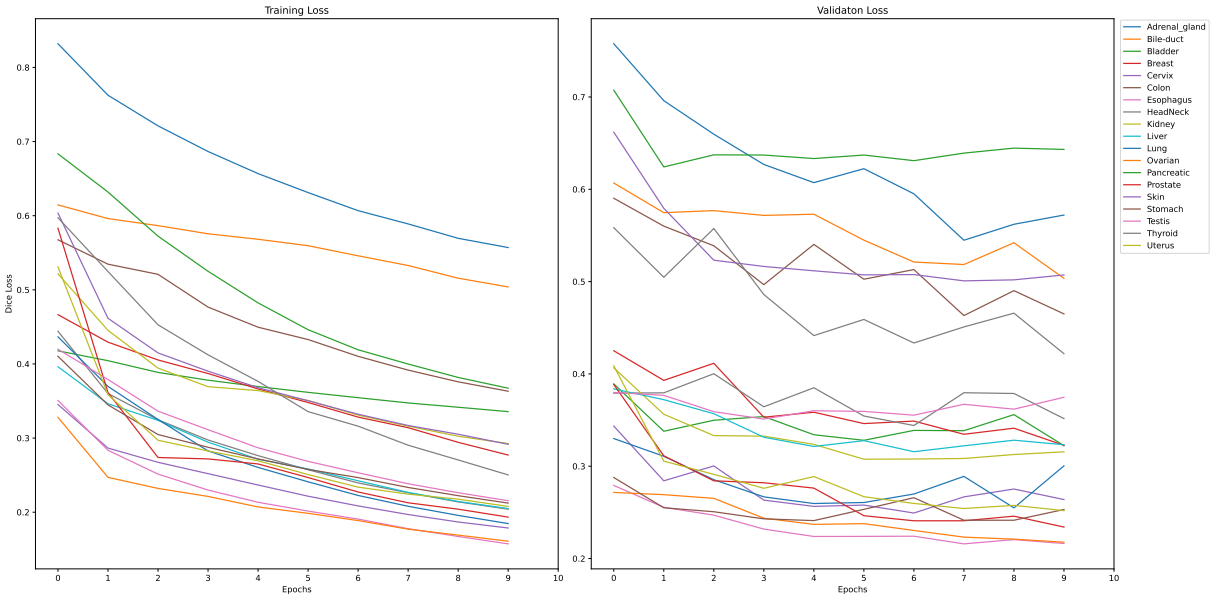

Figure 14: Training and validation model loss across the epochs for all tasks independently trained.

Now, about the found problems in the datasets, its noticeable that some validation loss curves are unstable, not smooth, and less steeper than their respective training loss curves, meaning that, for these datasets, the model is underfitted, and we may not have a good knowledge representation learned in those cases. As a result of this turbulence, not all datasets could achieve higher metrics with the same 10 epochs. Table 6 presents our segmentation metrics for each task, to better understand the performance of the tasks we established some poor performance thresholds based on the average metrics to classify the tasks. We highlighted in light red the datasets with a poor performance, where $metrics \leq (Average - Std)$. And we highlighted in light blue the datasets with a performance that is close to poor performance, where $metrics \leq (Average - 0.6 * Std)$.

All the highlighted task wold be considered as **hard** segmentation datasets of our problem, their complexity will pose hardships to the generalization algorithms, we would keep track of

these datasets, so we could investigate effects on their performance during the CL experiments.

Table 6: Model Segmentation Metrics

Task	**IoU**	$\pm Std$	**Dice**	$\pm Std$
Adrenal	0.20	0.02	0.33	0.03
Bile	0.39	0.02	0.55	0.02
Bladder	0.59	0.05	0.74	0.04
Breast	0.28	0.05	0.43	0.07
Cervix	0.53	0.01	0.69	0.01
Colon	0.22	0.03	0.36	0.04
Esophagus	0.51	0.03	0.68	0.03
HeadNeck	0.39	0.03	0.55	0.03
Kidney	0.13	0.06	0.22	0.09
Liver	0.39	0.01	0.56	0.01
Lung	0.44	0.04	0.61	0.04
Ovarian	0.60	0.01	0.75	0.01
Pancreatic	0.25	0.02	0.40	0.02
Prostate	0.35	0.04	0.52	0.04
Skin	0.23	0.02	0.37	0.03
Stomach	0.16	0.07	0.28	0.11
Testis	0.42	0.02	0.59	0.02
Thyroid	0.25	0.01	0.40	0.01
Uterus	0.61	0.01	0.76	0.01
Average	0.37±0.15		0.52±0.16	

To better illustrate this complexity discrepancy between the datasets we elaborated the Figure 15, notice the difference for each task metric from each average metric value. Also, the same three hard datasets highlighted before: Adrenal, Kidney and Stomach, are clearly bellow the average metric values. We can also pinpoint other potentially hard datasets, that are very close, slightly under or slightly over, the average lines, and when considering the complications of the CL setting, these datasets could have a even worst performance afterall. On the flip side, the presence of these hard datasets among the other ones will stress the generalization methods, really testing their performance on a difficult benchmark.

With this first experiment we validated the model suitability in performing the segmentation of histopathology images, independent of the tissue type. Also, we found differences on the difficulty of each tissue learning task, this new complexity characteristic places a new stress vector over the generalization algorithms, and must be considered during the continual learning experiments analysis.

5.2.2 Continual Learning

The goal of the next experiments is to validate our CL method and evaluate its generalization capacity against the baseline algorithms. For simplicity, we are going to document only the **IoU** metric of the models, avoiding long analysis in multiple metrics. Also, in these following

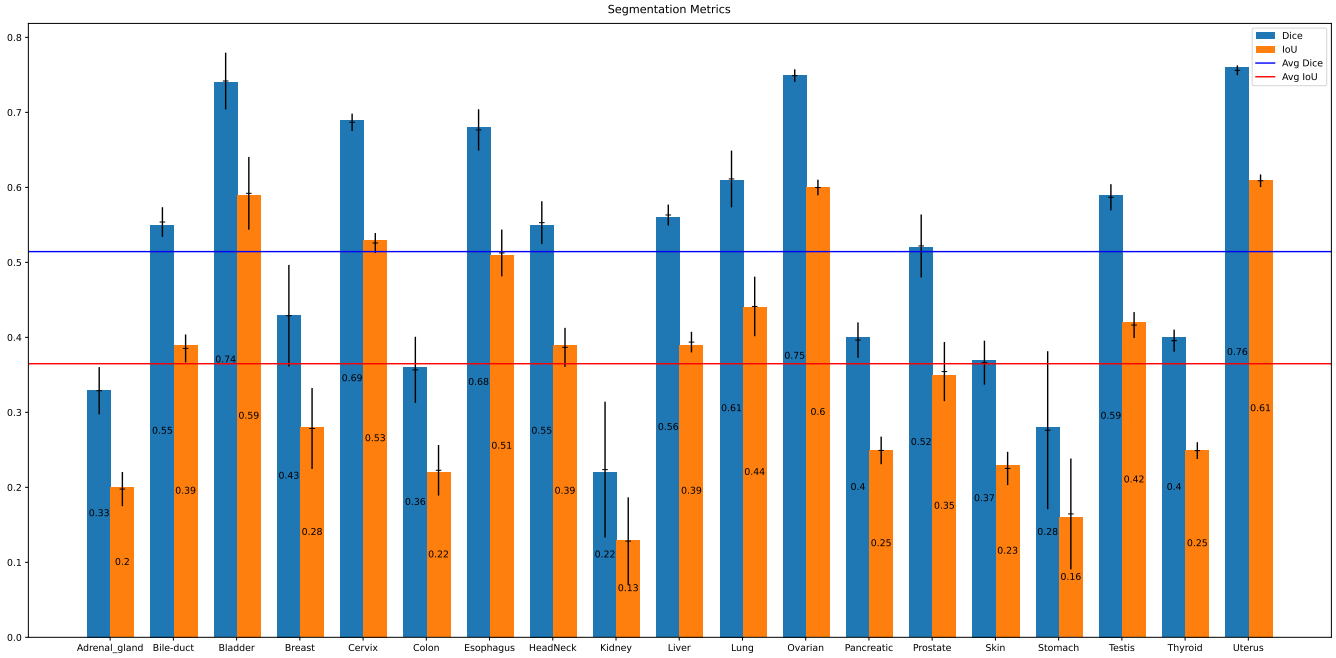

Figure 15: Segmentation metrics of all tasks independently trained. The horizontal lines represent the average metric values.

experiments we are carrying over the complexity information of the tasks, discovered in the previous experiment, on the form of a color encoding.

In the following experiment, the model was trained using the generalization algorithm consecutively in all datasets. The final model results, after training in the last dataset of the list, are documented in the Table 7. For each CL method, we documented the average of the final model metrics per task. Again, we highlighted in light red the datasets with a poor performance, where $IoU \leq (AvgIoU - StdIoU)$, and in light blue the datasets with a performance that is close to poor performance, where $IoU \leq (AvgIoU - 0.6 * StdIoU)$. The highlights in column **Dataset** were directly mapped from the assumptions of dataset complexity in our previous segmentation-only experiment, in Table 6. The last column **FULL** was drifted to the far right because this baseline has no generalization method, so it does not serve us for comparison in the matter of generalization algorithms, it serves more as an endline performance indicator. Finally, we highlighted in **bold** the algorithm with the best performance for each task.

First, we may start our analysis with the complexity/poor performance highlights. Notice how some of the datasets that had assumptions of their high complexity (colors of the Datasets column) achieved actually good metrics, for instance the *Adrenal*, *Kidney* and *Stomach* datasets. This is an indicator that the CL algorithms were able to improve their performance using the knowledge obtained from the other datasets. The same thing happened to *Breast* and *HeadNeck* when we compare the CL methods performance to the jointly-trained baseline (FULL).

On the other hand, there are datasets that we assumed were difficult to learn, and they proved us right, for instance *Colon* and *Thyroid* had poor performances in all baselines. There is also the case of the *Testis* dataset, that proved to be more difficult than we expected, having a worst

Table 7: Continual Learning: IoU Results

	IoU $\pm Std$									
Dataset	EWC		MAS		LWF		**Ours**		FULL	
Adrenal	**0.38**	0.02	0.31	0.03	0.29	0.02	0.36	0.02	0.26	0.02
Bile	0.28	0.02	**0.36**	0.02	0.27	0.02	0.34	0.01	0.30	0.02
Bladder	0.46	0.03	**0.56**	0.01	0.48	0.03	0.52	0.02	0.42	0.03
Breast	0.26	0.02	0.26	0.03	0.33	0.02	**0.37**	0.01	0.18	0.02
Cervix	0.38	0.02	0.43	0.01	0.41	0.03	**0.54**	0.02	0.47	0.02
Colon	0.04	0.00	0.06	0.02	0.15	0.01	**0.17**	0.01	0.16	0.02
Esophagus	0.31	0.04	0.38	0.05	0.50	0.02	**0.51**	0.02	0.44	0.02
HeadNeck	0.20	0.03	0.26	0.02	0.31	0.02	**0.35**	0.01	0.14	0.01
Kidney	0.36	0.02	0.32	0.04	0.40	0.01	**0.50**	0.02	0.39	0.03
Liver	0.35	0.02	**0.39**	0.02	0.32	0.03	0.38	0.02	0.38	0.01
Lung	0.36	0.01	0.38	0.02	0.36	0.01	**0.45**	0.01	0.30	0.03
Ovarian	0.32	0.03	0.45	0.01	0.42	0.02	**0.54**	0.01	0.53	0.02
Pancreatic	0.16	0.02	**0.25**	0.02	0.23	0.01	0.15	0.01	0.18	0.02
Prostate	0.40	0.02	0.40	0.03	0.49	0.02	**0.52**	0.02	0.39	0.02
Skin	0.32	0.02	0.35	0.02	0.27	0.02	**0.40**	0.02	0.22	0.02
Stomach	0.29	0.02	0.28	0.05	**0.46**	0.01	0.41	0.01	0.27	0.05
Testis	0.22	0.02	0.25	0.02	**0.27**	0.01	**0.27**	0.01	0.36	0.02
Thyroid	0.15	0.01	0.15	0.01	**0.18**	0.01	0.16	0.01	0.18	0.01
Uterus	0.35	0.03	0.44	0.01	0.44	0.02	**0.52**	0.01	0.52	0.02
Average	0.29±0.10		0.33±0.11		0.35±0.10		**0.39**±0.13		0.32±0.12	

performance with the CL methods, rather than without them. These are indicators that these datasets are still posing an issue to the generalization capability of the CL methods, and maybe even degrading the overall performance in all tasks (catastrophic forgetting), we need further experiments to access this hypothesis.

Also, we would like to highlight the cases of datasets *Adrenal*, *Bile*, *Breast*, *HeadhNeck* and *Skin*, where our model integrated these datasets **successfully**, while other CL methods had some difficulty dealing with them.

Now, numerically analyzing these results, we can highlight that the average performance of the CL methods is, in majority, higher than the FULL baseline metric, with the exception of EWC, the CL methods could leverage more of the knowledge obtained over the training of the multiple tasks.

In this sense, our approach performed, on average, **15,66%** better than all others CL methods, being our final average metric an **IoU 0.39**, that surpasses even the overall performance obtained when training in each task separately, an IoU 0.37 documented in Table 6. Finally, considering the absolute numbers per task, our approach had better results in **63%** (12) of the tasks. To visually display the performance of each method for all datasets we elaborated the Figure 16, notice our approach average line, in red, significantly over the other CL method averages.

The results obtained from these last experiments validate our approach suitability as a CL method, even further, we documented significant improvements in the generalization capabi-

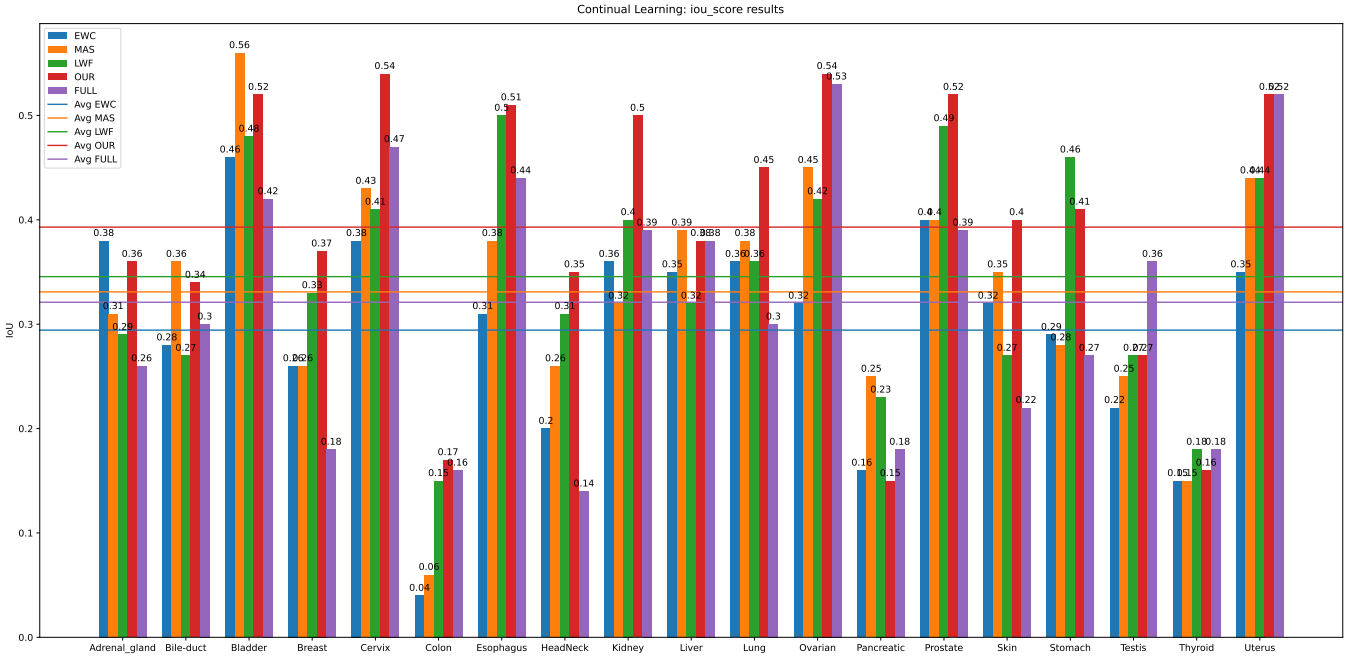

Figure 16: Final IoU for all datasets, segmented by CL method. The horizontal lines represent the average IoU values.

lity of the segmentation model when trained using our method, in comparison with other CL methods of the literature. During the experiments we noticed a poor performance in some datasets, that could be associated with the effects of catastrophic forgetting, requiring a more profound analysis.

5.2.2.1 Catastrophic Forgetting

In this sub section we are conducting experiments to measure the in-depth effects of catastrophic forgetting in each CL algorithm. The next experiment was performed, again, by training the model consecutively in all datasets making use of each CL algorithm. Just like before, our main experimentation metric is the **IoU**, but this time, this metric is used to calculate the CL metrics: Forward and Backward Transfer. Table 8 presents the FWT and BWT results per task. For simplicity, one could understand FWT and BWT as the average influence that the present dataset training had, respectively, on the performance of the next and previous datasets.

In this table, the colored highlights are following a different scheme. Considering that FWT measures the performance on future tasks, that will eventually see during the training loop, our main preoccupation should be the catastrophic forgetting effect, that happens on tasks that the had model already seen. Because of this, we are going to weight heavily on the performances in the BWT metrics, while not so much in the FWT. Therefore, we highlighted in light red the datasets with a poor performance, where $BWT \leq -0.05 \vee FWT \leq -1.0$, and in light green the datasets with a good performance, where $BWT \geq 0.05 \vee FWT \geq 1.0$. The colored complexity highlights in column **Dataset**, and the **Best** and **Worst** Methods, were directly mapped from

the findings of our previous CL experiment, in Table 7. Finally, we highlighted in **bold** the best BWT results per task.

Table 8: Continual Learning: FWT and BWT Results (based on IoU)

	EWC		MAS		LWF		**Ours**		Method	
Dataset	FWT	BWT	FWT	BWT	FWT	BWT	FWT	BWT	Best	Worst
Adrenal	0.00	**0.02**	0.00	-0.02	0.00	-0.08	0.00	0.00	EWC	LWF
Bile	-0.04	0.03	-0.07	**0.09**	-0.05	0.03	-0.07	**0.09**	MAS	EWC
Bladder	0.00	0.05	0.04	0.07	-0.02	0.08	-0.04	**0.11**	MAS	EWC
Breast	0.07	0.03	0.07	-0.07	0.10	0.03	0.12	**0.04**	Our	EWC/MAS
Cervix	-0.14	0.03	-0.13	0.00	-0.10	0.01	-0.03	**0.08**	Our	EWC
Colon	-0.13	-0.03	-0.11	-0.09	-0.06	**0.01**	-0.06	0.00	Our	EWC
Esophagus	-0.14	-0.05	-0.11	-0.04	0.04	0.00	0.00	**0.05**	Our	EWC
HeadNeck	0.07	0.02	0.07	**0.03**	0.17	-0.02	0.16	0.02	Our	EWC
Kidney	-0.11	**0.06**	-0.05	-0.06	0.00	0.00	0.07	0.02	Our	MAS
Liver	-0.09	0.03	-0.07	**0.05**	-0.05	-0.01	-0.04	0.04	MAS	LWF
Lung	0.01	**0.04**	0.04	0.00	0.06	-0.02	0.11	0.01	Our	EWC/LWF
Ovarian	-0.21	0.00	-0.13	0.00	-0.11	-0.01	-0.03	**0.01**	Our	EWC
Pancreatic	-0.01	0.02	0.04	**0.06**	0.05	-0.01	-0.05	0.01	MAS	Our
Prostate	0.00	**0.01**	0.01	-0.09	0.08	**0.01**	0.08	-0.01	Our	EWC/MAS
Skin	0.07	0.03	0.06	0.01	0.06	-0.01	0.13	**0.05**	Our	LWF
Stomach	0.02	**0.05**	-0.03	-0.04	0.15	0.01	0.09	0.01	LWF	MAS
Testis	-0.19	**0.04**	-0.10	-0.01	-0.08	-0.01	-0.10	0.01	LWF/Our	EWC
Thyroid	-0.04	**0.01**	-0.01	-0.01	0.00	0.00	-0.02	0.00	LWF	EWC/MAS
Uterus	-0.17	0.00	-0.18	0.00	-0.08	0.00	-0.01	0.00	Our	EWC
Average	-0.05	0.02	-0.03	-0.01	0.01	0.00	**0.02**	**0.03**	Our	EWC

Following our previous experiment, first, we began to analyze the complexity colored highlights. Considering the poor performance datasets *Colon*, *Pancreatic* and *Thyroid*, we can clearly see the reason for their lack of performance, all Cl methods had low BWT values for them, sometimes even slightly below zero (catastrophic forgetting), this explains their results in our previous experiment and finally establish them as the really hard datasets. Something similar happens in dataset *Testis*, that achieved close to poor performance in our previous experiment, this is now explained by the fact that most of its BWTs are close do zero.

That is the main difference in comparison to other datasets that performed relatively good, for instance *Adrenal* or *Kidney*, even though sometimes there are negative BWTs, other methods were able to preserve their knowledge, so they are not so much harder to generalize. In contrast, its noticeable that the easiest datasets are the ones pilling up positive BWTs, for instance *Bladder* and *Cervix*, therefore, their good performances in our previous experiment are explainable.

Going even more in depth, we can majorly explain the final performance of each method in each dataset by analyzing their FWT and BWT metric. For instance, taking a look at Figure 16, from our previous experiment, and at our current Table 8, we would like to highlight the datasets *Esophagus* and *Kidney*, notice how the poor performance of the methods EWC and MAS, and good performance of LWF and Ours, on these datasets, are mirrored by their FWT and BWT

values on our present experiment table. Also, for the special case of dataset *Kidney*, the EWC metric for BWT was quite high, even though the final IoU was bellow average, which indicates that there exists a sort of balance between the FWT and BWT metric.

We can confirm the existence of this balance mechanism by also analyzing the datasets *Ovarian*, *Prostate* and *Uterus*, where, even though they have not so bad BWT metrics, their final performance in the methods with bad FWT was poor. This mechanic allows us to better explain the behavior of each method for each dataset, where a method, to actually perform good at the finally trained model, should have both metrics, FWT and BWT, with generally good values over the training. Methods that perform better at one metric, than the other, will also suffer from catastrophic forgetting at the end. This can be visually verified by looking at the color distribution in Table 8, where the presence of red values in the methods is proportional to their final performance in our last experiment. Consequently, we can easily make the correlations between our last two columns, the Best and Worst methods, and the methods metrics documented on the table.

These resolutions and explanations validate our approach in contrast to the other CL methods, where our method is visibly the one with the greater amount of good performance values, and the lesser amount of the bad ones. We also have **36%** (7) of the best BWT values, tied up with EWC, but we still we have lesser negative FWTs. Finally, our performance is explained by our average **FWT 0.02** and **BWT 0.03** metrics, surpassing the other CL algorithms, and meaning that, on average, our method is the only one able to **improve** the learning of both past and future tasks.

To visually display the performance of each method for all datasets as the training goes on, we elaborated the Figure 17, notice how our approach line, for the majority of the datasets, have a positive angle, meaning that the performance keeps improving over time. Also, notice that our line is more smoother, having smaller performance fluctuations, demonstrating the stability of our approach.

With this experiment we evaluated our CL method in-depth, accessing all the improvements we achieved, and explaining our results in comparison with the baselines. The experiments in this section numerically validated our approach, and documented our statistical leverage over other CL methods, nonetheless, we still have to investigate the final model outputs, to examine and compare the predictions made by the models after trained in the CL setup.

5.2.3 Output Analysis

For our final round of experiments, we aim analyze the segmentation masks outputted by each CL method, to evaluate them semantically, and to compared them with the ground truth masks. Our goal here is to understand the effect of the differences between the CL algorithms have on their outputs quality.

In this experiment we extracted the predictions made by the models after they had been

trained following the CL setup, that is, consecutively trained in all datasets making use of each CL algorithm. Using the knowledge obtained so far with our experiments we elected 6 datasets based on their final overall performance to be representatives of the possible difficulty classes of tasks.

In Figure 18, we gathered the samples of some datasets, mounted with the predictions made by models trained with each CL method. The datasets *Ovarian* and *Bladder*, in light green, represent the easy tasks. While datasets *HeadhNeck* and *Testis*, in light blue, represent the middle-term difficulty. And finally, datasets *Thyroid* and *Colon*, in light red, represent the hard datasets. For each output, we documented the IoU metric and the overlay of the prediction, in green, over the ground truth mask, in red.

Notice in this figure, that, as the tasks get more complex, the number of tumors in the ground truth mask gets smaller, while the number of other structures that look like tumors increases. For instance, the sample of the *Colon* dataset, has many cells that would be positively classified as tumors in the context of other datasets, given its characteristics, like color and shape, but are not tumor, therefore, for all CL methods, the model outputted many false positives. And it gets even worse, given that the actual tumorous cells resembles the wealthy tissue of the other datasets, we also had many **false negatives** in all CL methods, which is a big concern for medical applications.

In the other hand, the easiest samples have very distinguishable structures for cells and tumors, even better, common tumor characteristics could be seen between different datasets. For example, the samples of *Bladder* and *Testis* have their tumors with similar tones of dark purple in a more clear background, making the prediction much easier to perform, even that all CL methods outputted good segmentation masks.

Nevertheless, when we analyze the predicted masks semantically, its noticeable that EWC and MAS produced precise segmentations, with masks that circle the tumor shapes, and, because of this preciousness, many times the masks miss the entirety of the tumor cells. In contrast, LWF produced sparse segmentations, with masks that cover the whole the tumor area, but, in the process, goes beyond the tumor edges and also marks the wealthy tissue nearby. Meanwhile, our approach sits on a comfortable sweet spot between both strategies, we produced masks that cover more tumor area than the EWC and MAS methods, without going completely rogue on the nearby wealthy tissue like the LWF outputs. In conclusion, over a semantic point of view, our approach produced segmentation masks with a visibly significant **improvement** in comparison to the other baselines.

This improvement can also be numerically verified, when analyzing the IoU per sample, our approach **outperforms** all the other CL methods by **29,9%** on average. Finally, we can also notice a general degradation of **24,15%** of the IoU metric for our baselines as the datasets get more complex, again, our approach was able to perform **better**, with only **13,48%** of IoU degradation.

With this final experiment we successfully validated our CL approach over a semantic pers-

pective. We concluded that our approach is not only suitable for generation segmentation masks for multiple types of tissues, but also significantly better than our baselines at doing so. Finally, we documented concerning results regarding the number of **false negatives** in the context of hard datasets.

5.3 Discussion

In this chapter we exhaustively experimented our method, numerically evaluating our performance in comparison to the baselines, accessing the nuances of each method results to explain their metrics, their pros and cons, and finally, analyzing the segmentation masks over the semantic point of view of the histopathology problem.

With these experiments we validated that our Unet model is suitable to perform the histopathologic segmentation, that our CL framework is capable of retaining information while learning new patterns, and that our model outputs are aligned to the necessities of our problem.

Performance-wise, our method **surpassed** all baselines metrics, even in comparison to the approach of one model trained to each dataset, or one model with all datasets at once. Unfortunately, we detected a concerning number of **false negatives** on our semantic analysis, specially on the hard datasets.

Finally, our experiments could be used to answer our research question on Section 1.3. Our method was the **only** technique able to leverage the generalization of a histopathology segmentation model, in such a way that, it fulfills the requirements of this problem domain. The problem requirements, in Section 1.2, are answered in the list bellow:

- **New tasks**: the method should allow the model to incorporate new knowledge and achieve a good performance in new tasks.

  Answer: Our approach can transfer its knowledge forward on a average of **0.02** IoU, the greatest positive transfer of all baselines.

- **Old tasks**: the method should not only preserve the learned knowledge, but also increase the performance of older tasks.

  Answer: Our approach can transfer its knowledge backward on a average of **0.03** IoU, the greatest positive transfer of all baselines.

- **Old data**: the method should not take for granted the access to data from previous tasks.

  Answer: Our approach does **not** make any use of old tasks data, and surpasses the baseline that does have access.

- **New data**: the method should also generalize the knowledge to different distributions of new data for the same old task.

  Answer: Our approach can generalize its overall knowledge, for old and new tasks, on average of **15,66%** better than all baselines.

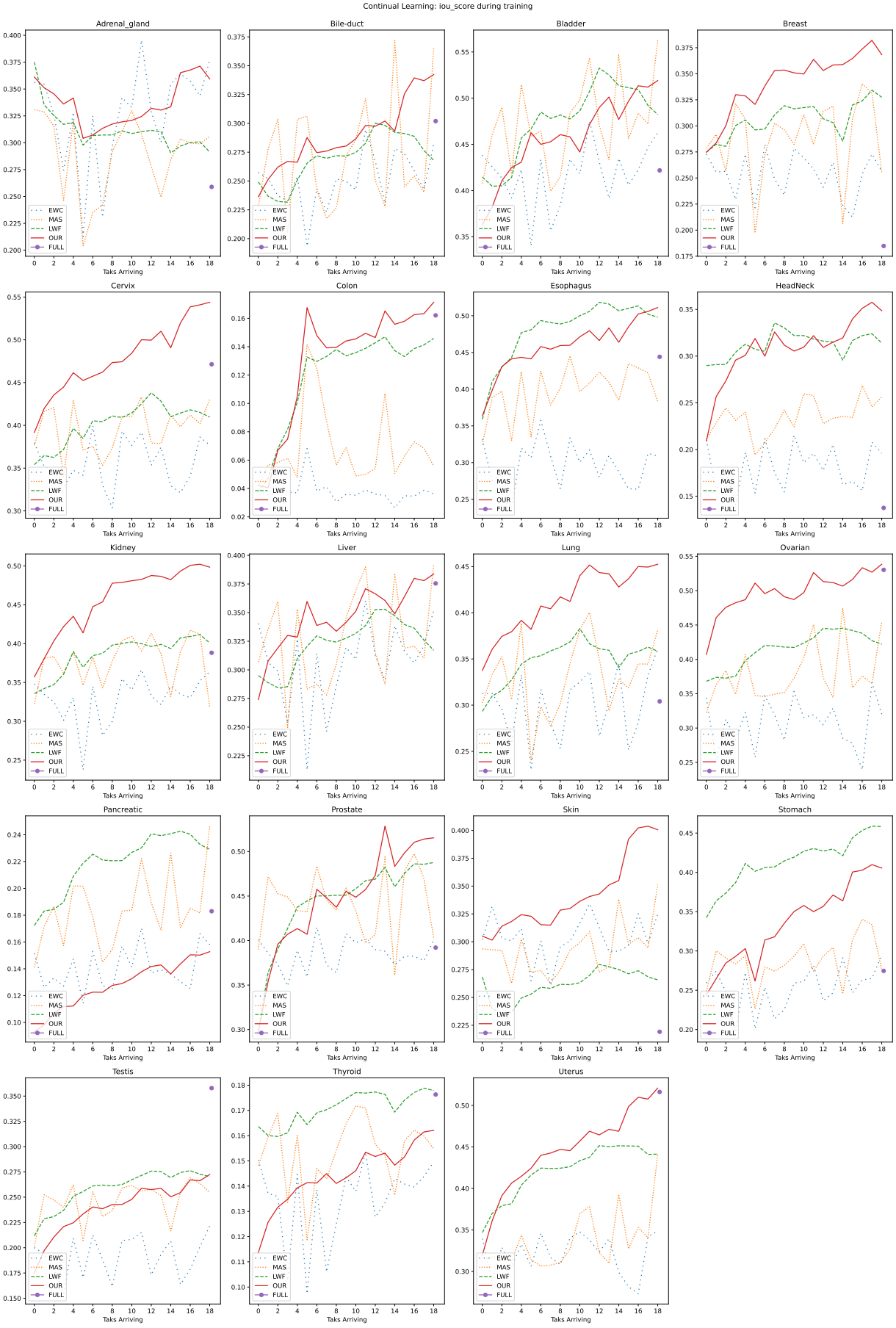

Figure 17: Continual IoU metric during the training of each dataset, the lines represent each a CL baseline, the final dot represents the FULL baseline (trained at once).

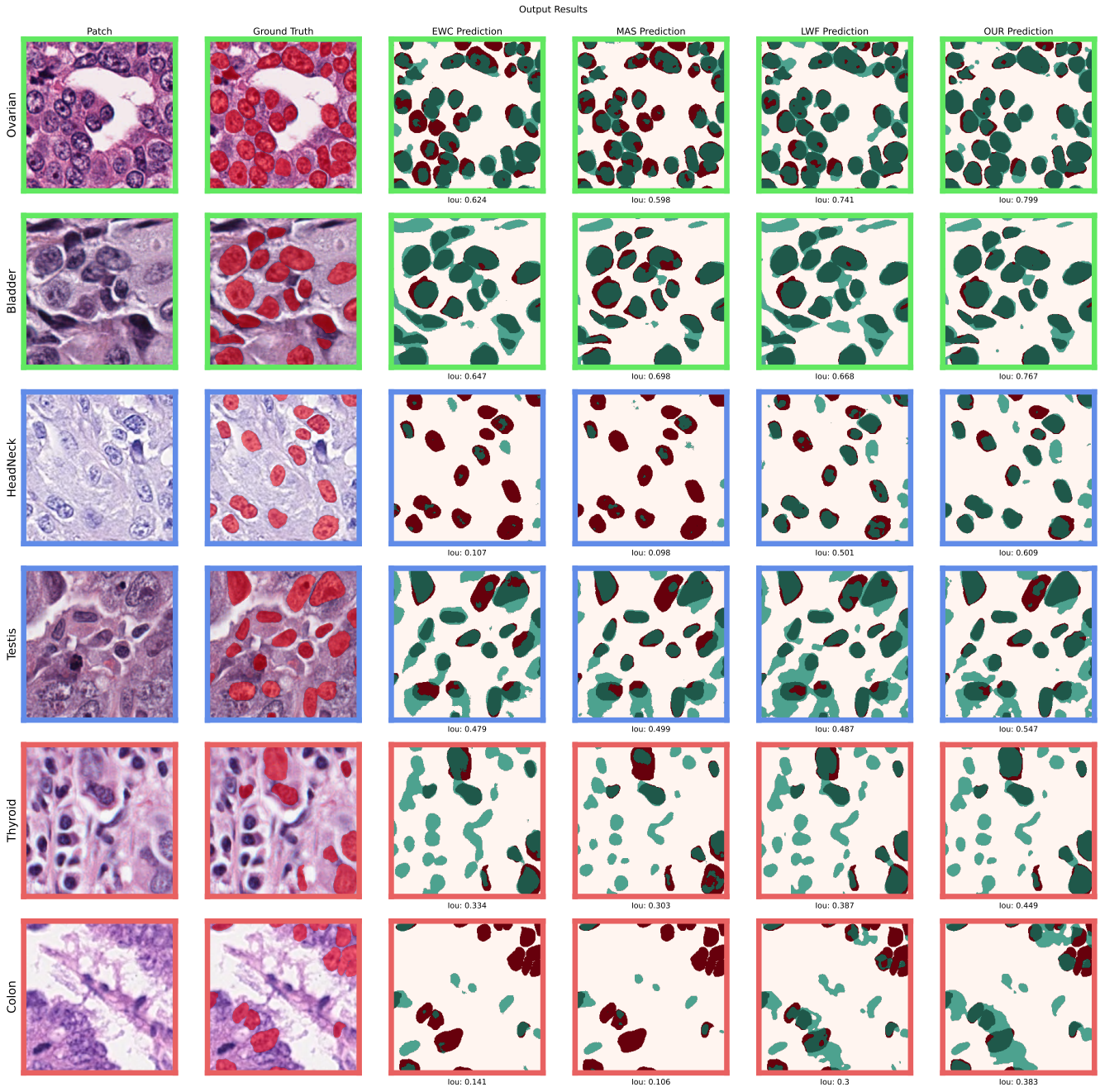

Figure 18: Samples of the selected datasets, segmented by models produced by each CL method. Datasets are top-bottom ordered by their difficulty, being the top the easiest.

6 CONCLUSION

The implementation of AI applications in the day-to-day clinic process is a topic of great interest, the possibility of using AI to enhance the performance of clinicians is hardly limited by the AI inherent limitations. The variety of clinics, processes, protocols, and patients create a plethora of sub-domains for the same task/problem. Current AI models are trained into one or few sub-domains, thus limited to a wider range of situations and applications, this non-generalization ability makes even the more accurate AI solutions lose significant performance in a real world scenario.

In this work we implemented a solution for the generalization problem into a specific clinic task: Histopathology. We discussed the aspects involving the clinic and computational problems, reviewed the current stat-of-the-art approaches specifically into the field of generalization in CPATH, designed and experimented an enhanced version of the Learning without Forgetting method to solve the continual segmentation problem over an UNet model.

Our method optimizes the common knowledge learned in previous tasks while integrates new patterns, without forgetting the previous ones. This is possible due to the replacement the hyperparameter λ by a dynamically calculated regularization term, this additional term **improves** the generalization of the UNet model by helping the stochastic optimizer to focus on the common knowledge between tasks, **smoothing** the trade-off between the new and old task.

Through exhaustive experimentation using a number of datasets greater than the literature, we evaluated our method both numerically and semantically, our results **overcame**, on average of **15,66%**, the metrics of the CL baselines, a model joint-trained in all datasets, and a scenario with a list of models, one specifically for each dataset. Additionally, we performed a deep analysis on how each CL method responds when induced to forgetting, proving that our approach was the **only** method to counter the catastrophic forgetting effects with statistically positive transfer indexes, FWT **0.02** and BWT **0.03**. Moreover, when evaluating the output segmentation masks, our results also stand out semantically, with an average improvement of **29,9%**.

Furthermore, our method relies on stochastic optimization to integrate the knowledge efficiently, so the *suggestion* mechanism we implemented with the new regularization term (generalization term) depends of the impact that each term has on the final loss. Meaning that there are possible scenarios where the optimization fails to reach a good balancing factor, or fails to follow the suggestion term, leading the catastrophic forgetting. Also, the order of the tasks can influence in the overall performance of the trained method, depending of the previous knowledge acquired and the difficulty of the incoming tasks. Finally, more experiments could be performed to establish this work as a strong baseline, such as using even more datasets, using more epochs during training, applying a normalization filter over the patches, and messing with the order of the datasets.

Finally, in this work we answered the research question stated in Section 1.3: our method was the only approach to **successfully** leverage the segmentation model considering the requi-

rements and the data-variability of the histopathology field.

In conclusion, this work presents a novel and enhanced continual learning method for histopathology segmentation, this method was statistically validated, produced outputs with improved performance in comparison to the literature, and fulfilled the continual learning requirements of a generalist CPATH application, turning it into a strong candidate to be implemented in a real world pathology clinic scenario.

REFERENCES

AL-THELAYA, K.; GILAL, N.; ALZUBAIDI, M.; MAJEED, F.; AGUS, M.; SCHNEIDER, J.; HOUSEH, M. Applications of discriminative and deep learning feature extraction methods for whole slide image analysis: a survey. **Journal of Pathology Informatics**, [S.l.], v. 14, p. 100335, 09 2023.

ALJUNDI, R.; BABILONI, F.; ELHOSEINY, M.; ROHRBACH, M.; TUYTELAARS, T. Memory aware synapses: learning what (not) to forget. In: EUROPEAN CONFERENCE ON COMPUTER VISION (ECCV), 2018. **Proceedings...** [S.l.: s.n.], 2018. p. 139–154.

ANDRYCHOWICZ, M.; DENIL, M.; GOMEZ, S.; HOFFMAN, M. W.; PFAU, D.; SCHAUL, T.; SHILLINGFORD, B.; DE FREITAS, N. Learning to learn by gradient descent by gradient descent. **Advances in neural information processing systems**, [S.l.], v. 29, 2016.

BAWEJA, C.; GLOCKER, B.; KAMNITSAS, K. Towards continual learning in medical imaging. **arXiv preprint arXiv:1811.02496**, [S.l.], 2018.

BAWEJA, C.; GLOCKER, B.; KAMNITSAS, K. Towards continual learning in medical imaging. **arXiv preprint arXiv:1811.02496**, [S.l.], 2018.

BOTTOU, L. eon. Online learning and stochastic approximations. **Online learning in neural networks**, [S.l.], v. 17, n. 9, p. 142, 1998.

BROWNLEE, J. **Master Machine Learning Algorithms**. [S.l.]: Machine Learning Mastery, 2016.

BáNDI, P.; BALKENHOL, M.; van Dijk, M.; KOK, M.; van Ginneken, B.; van der Laak, J.; LITJENS, G. Continual learning strategies for cancer-independent detection of lymph node metastases. **Medical Image Analysis**, [S.l.], v. 85, p. 102755, 2023.

CHATTERJEE, S. Artefacts in histopathology. **Journal of oral and maxillofacial pathology: JOMFP**, [S.l.], v. 18, n. Suppl 1, p. S111, 2014.

CHAUDHRY, A.; ROHRBACH, M.; ELHOSEINY, M.; AJANTHAN, T.; DOKANIA, P.; TORR, P.; RANZATO, M. Continual learning with tiny episodic memories. In: WORKSHOP ON MULTI-TASK AND LIFELONG REINFORCEMENT LEARNING, 2019. **Anais...** [S.l.: s.n.], 2019.

CHEN, B.; THANDIACKAL, K.; PATI, P.; GOKSEL, O. Generative appearance replay for continual unsupervised domain adaptation. **arXiv preprint arXiv:2301.01211**, [S.l.], 2023.

CRAWFORD, K. **The atlas of AI**: power, politics, and the planetary costs of artificial intelligence. [S.l.]: Yale University Press, 2021.

FENG, T.; YUAN, H.; WANG, M.; HUANG, Z.; BIAN, A.; ZHANG, J. Progressive learning without forgetting. **arXiv preprint arXiv:2211.15215**, [S.l.], 2022.

FINN, C.; XU, K.; LEVINE, S. Probabilistic model-agnostic meta-learning. **Advances in neural information processing systems**, [S.l.], v. 31, 2018.

FRENCH, R. Catastrophic forgetting in connectionist networks. **Trends in cognitive sciences**, [S.l.], v. 3, p. 128–135, 05 1999.

GAMPER, J.; KOOHBANANI, N. A.; BENES, K.; GRAHAM, S.; JAHANIFAR, M.; KHURRAM, S. A.; AZAM, A.; HEWITT, K.; RAJPOOT, N. Pannuke dataset extension, insights and baselines. **arXiv preprint arXiv:2003.10778**, [S.l.], 2020.

GARDEREN, K. van; VOORT, S. van der; INCEKARA, F.; SMITS, M.; KLEIN, S. Towards continuous learning for glioma segmentation with elastic weight consolidation. **arXiv preprint arXiv:1909.11479**, [S.l.], 2019.

GONZALEZ, C.; SAKAS, G.; MUKHOPADHYAY, A. **What is Wrong with Continual Learning in Medical Image Segmentation?** 2020.

GONZALEZ, C.; SAKAS, G.; MUKHOPADHYAY, A. **What is Wrong with Continual Learning in Medical Image Segmentation?** 2020.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. [S.l.]: MIT press, 2016.

HADSELL, R.; RAO, D.; RUSU, A.; PASCANU, R. Embracing Change: continual learning in deep neural networks. **Trends in Cognitive Sciences**, [S.l.], v. 24, p. 1028–1040, 12 2020.

HAYKIN, S. **Neural networks and learning machines, 3/E**. [S.l.]: Pearson Education India, 2009.

HINTON, G.; VINYALS, O.; DEAN, J. Distilling the knowledge in a neural network. **arXiv preprint arXiv:1503.02531**, [S.l.], 2015.

HINTON, G.; VINYALS, O.; DEAN, J. Distilling the knowledge in a neural network. **arXiv preprint arXiv:1503.02531**, [S.l.], 2015.

JAHANIFAR, M.; RAZA, M.; XU, K.; VUONG, T.; JEWSBURY, R.; SHEPHARD, A.; ZAMANITAJEDDIN, N.; KWAK, J. T.; RAZA, S. E. A.; MINHAS, F. et al. Domain Generalization in Computational Pathology: survey and guidelines. **arXiv preprint arXiv:2310.19656**, [S.l.], 2023.

JAHANIFAR, M.; RAZA, M.; XU, K.; VUONG, T.; JEWSBURY, R.; SHEPHARD, A.; ZAMANITAJEDDIN, N.; KWAK, J. T.; RAZA, S. E. A.; MINHAS, F. et al. Domain Generalization in Computational Pathology: survey and guidelines. **arXiv preprint arXiv:2310.19656**, [S.l.], 2023.

JANOWCZYK, A.; MADABHUSHI, A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. **Journal of Pathology Informatics**, [S.l.], v. 7, p. 29, 07 2016.

KAUSTABAN, V.; BA, Q.; BHATTACHARYA, I.; SOBH, N.; MUKHERJEE, S.; MARTIN, J.; MIRI, M.; GUETTER, C.; CHATURVEDI, A. **Continual Learning for Tumor Classification in Histopathology Images**. 2022.

KIRKPATRICK, J.; PASCANU, R.; RABINOWITZ, N.; VENESS, J.; DESJARDINS, G.; RUSU, A. A.; MILAN, K.; QUAN, J.; RAMALHO, T.; GRABSKA-BARWINSKA, A. et al. Overcoming catastrophic forgetting in neural networks. **Proceedings of the national academy of sciences**, [S.l.], v. 114, n. 13, p. 3521–3526, 2017.

KOTHARI, S.; PHAN, J.; STOKES, T.; WANG, M. Pathology imaging informatics for quantitative analysis of whole-slide images. **Journal of the American Medical Informatics Association : JAMIA**, [S.l.], v. 20, 08 2013.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. **Advances in neural information processing systems**, [S.l.], v. 25, 2012.

KUMAR, N.; GUPTA, R.; GUPTA, S. Whole Slide Imaging (WSI) in Pathology: current perspectives and future directions. **Journal of Digital Imaging**, [S.l.], v. 33, 05 2020.

LAAK, J. van der; LITJENS, G.; CIOMPI, F. Deep learning in histopathology: the path to the clinic. **Nature Medicine**, [S.l.], v. 27, p. 775–784, 05 2021.

LENGA, M.; SCHULZ, H.; SAALBACH, A. Continual learning for domain adaptation in chest x-ray classification. In: MEDICAL IMAGING WITH DEEP LEARNING, 2020. **Anais...** [S.l.: s.n.], 2020. p. 413–423.

LI, Z.; HOIEM, D. Learning without forgetting. **IEEE transactions on pattern analysis and machine intelligence**, [S.l.], v. 40, n. 12, p. 2935–2947, 2017.

LI, Z.; HOIEM, D. Learning without forgetting. **IEEE transactions on pattern analysis and machine intelligence**, [S.l.], v. 40, n. 12, p. 2935–2947, 2017.

LOPEZ-PAZ, D.; RANZATO, M. Gradient episodic memory for continual learning. **Advances in neural information processing systems**, [S.l.], v. 30, 2017.

MCCLURE, P.; ZHENG, C. Y.; KACZMARZYK, J.; ROGERS-LEE, J.; GHOSH, S.; NIELSON, D.; BANDETTINI, P. A.; PEREIRA, F. Distributed weight consolidation: a brain segmentation case study. **Advances in neural information processing systems**, [S.l.], v. 31, 2018.

MEMMEL, M.; GONZALEZ, C.; MUKHOPADHYAY, A. Adversarial continual learning for multi-domain hippocampal segmentation. In: DOMAIN ADAPTATION AND REPRESENTATION TRANSFER, AND AFFORDABLE HEALTHCARE AND AI FOR RESOURCE DIVERSE GLOBAL HEALTH: THIRD MICCAI WORKSHOP, DART 2021, AND FIRST MICCAI WORKSHOP, FAIR 2021, HELD IN CONJUNCTION WITH MICCAI 2021, STRASBOURG, FRANCE, SEPTEMBER 27 AND OCTOBER 1, 2021, PROCEEDINGS 3, 2021. **Anais...** [S.l.: s.n.], 2021. p. 35–45.

MESCHER, A. L. **Junqueira's basic histology**: text and atlas. [S.l.]: New York: McGraw Hill, 2018.

MUSUMECI, G. Past, present and future: overview on histology and histopathology. **J Histol Histopathol**, [S.l.], v. 1, n. 5, p. 1–3, 2014.

OLAF, R.; PHILIPP, F.; THOMAS, B. U-Net: convolutional networks for biomedical image segmentation. **CoRR**, [S.l.], v. abs/1505.04597, 2015.

ÖZGÜN, S.; RICKMANN, A.-M.; ROY, A. G.; WACHINGER, C. Importance driven continual learning for segmentation across domains. In: MACHINE LEARNING IN MEDICAL IMAGING: 11TH INTERNATIONAL WORKSHOP, MLMI 2020, HELD IN CONJUNCTION WITH MICCAI 2020, LIMA, PERU, OCTOBER 4, 2020, PROCEEDINGS 11, 2020. **Anais...** [S.l.: s.n.], 2020. p. 423–433.

RANEM, A.; GONZÁLEZ, C.; MUKHOPADHYAY, A. Continual hippocampus segmentation with transformers. In: IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2022. **Proceedings...** [S.l.: s.n.], 2022. p. 3711–3720.

REBUFFI, S.-A.; KOLESNIKOV, A.; SPERL, G.; LAMPERT, C. H. icarl: incremental classifier and representation learning. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2017. **Proceedings...** [S.l.: s.n.], 2017. p. 2001–2010.

ROBINS, A. Catastrophic forgetting, rehearsal and pseudorehearsal. **Connection Science**, [S.l.], v. 7, n. 2, p. 123–146, 1995.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological review**, [S.l.], v. 65, n. 6, p. 386, 1958.

RUSSELL, S. J.; NORVIG, P. **Artificial intelligence**: a modern approach. [S.l.]: Pearson, 2016.

SALVI, M.; ACHARYA, U. R.; MOLINARI, F.; MEIBURGER, K. M. The impact of pre-and post-image processing techniques on deep learning frameworks: a comprehensive review for digital pathology image analysis. **Computers in Biology and Medicine**, [S.l.], v. 128, p. 104129, 2021.

SHIN, H.; LEE, J. K.; KIM, J.; KIM, J. Continual learning with deep generative replay. **Advances in neural information processing systems**, [S.l.], v. 30, 2017.

THANDIACKAL, K.; PICCINELLI, L.; PATI, P.; GOKSEL, O. Multi-scale Feature Alignment for Continual Learning of Unlabeled Domains. **arXiv preprint arXiv:2302.01287**, [S.l.], 2023.

THRUN, S. Lifelong learning algorithms. In: **Learning to learn**. [S.l.]: Springer, 1998. p. 181–209.

VEN, G. M. van de; TUYTELAARS, T.; TOLIAS, A. S. Three types of incremental learning. **Nature Machine Intelligence**, [S.l.], v. 4, n. 12, p. 1185–1197, 2022.

WANG, Z.; ZHANG, Z.; LEE, C.-Y.; ZHANG, H.; SUN, R.; REN, X.; SU, G.; PEROT, V.; DY, J.; PFISTER, T. Learning to prompt for continual learning. In: IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2022. **Proceedings...** [S.l.: s.n.], 2022. p. 139–149.

WOHLIN, C. Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering. In: INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING, 18., 2014, New York, NY, USA. **Proceedings...** Association for Computing Machinery, 2014. (EASE '14).

XING, F.; XIE, Y.; SU, H.; LIU, F.; YANG, L. Deep Learning in Microscopy Image Analysis: a survey. **IEEE Transactions on Neural Networks and Learning Systems**, [S.l.], v. PP, p. 1–19, 11 2017.

YANG, H.; HUANG, W.; LIU, J.; LI, C.; WANG, S. Few-shot Class-incremental Learning for Cross-domain Disease Classification. **arXiv preprint arXiv:2304.05734**, [S.l.], 2023.

YOON, J.; YANG, E.; LEE, J.; HWANG, S. J. Lifelong learning with dynamically expandable networks. **arXiv preprint arXiv:1708.01547**, [S.l.], 2017.

ZHOU, S. K.; GREENSPAN, H.; DAVATZIKOS, C.; DUNCAN, J.; GINNEKEN, B.; MADABHUSHI, A.; PRINCE, J.; RUECKERT, D.; SUMMERS, R. A Review of Deep Learning in Medical Imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. **Proceedings of the IEEE**, [S.l.], v. PP, p. 1–19, 02 2021.