

EÖTVÖS LORÁND UNIVERSITY

FACULTY OF SCIENCE

INSTITUTE OF MATHEMATICS

Thesis

Bias and Transparency Analysis in Brain Tumor Segmentation



Supervisor:
Bence Bakos

Author:
Gábor Kerényi
Mathematics Expert in Data Analytics
and Machine Learning
Postgraduate Specialization Programme

Budapest, 2026

Abstract

After introducing the basics of image segmentation from a clinical, regulatory, and technical perspective, we define the Attention U-Net model and a corresponding training loop. We define DICE, Hausdorff, and GradCAM-based quality metrics to evaluate the performance of the model, as well as their differentiable variants used for training. We analyze how different metrics change when the model is post-trained with new images, while the initial training set is not available. We also analyze whether smaller or bigger tumors are impacted more during the post-training phase.

Contents

1	Introduction	4
1.1	Notations	6
1.2	Structure of the Thesis	6
2	Clinical background	7
3	Regulatory overview	9
3.1	Medical Devices	9
3.1.1	European Union	10
3.1.2	United States of America	10
3.2	Artificial Intelligence	10
3.2.1	European Union	11
3.2.2	United States of America	11
3.3	Medical Device with AI	12
4	Objectives	13
5	Dataset	14
5.1	Augmentation and transformations	14
6	Model	16
6.1	Attention U-Net	16
6.2	GradCAM	19
7	Metrics	20
7.1	DICE	20
7.2	Hausdorff distance	20
7.3	Saliency score	21
8	Training	22
8.1	Loss function	22
8.1.1	Soft DICE	22
8.1.2	Soft Hausdorff	22
8.1.3	Combined	23
8.2	Training loop	23
8.3	Results	23

9	Post-training	25
9.1	Measurement	25
9.2	Elastic Weight Consolidation	25
9.3	Post-training results	31
9.3.1	EWC with BCE loss	32
9.3.2	EWC with DICE and Hausdorff loss	32
9.3.3	No EWC	33
9.3.4	Comparison	34
9.4	Saliency score analysis	34
10	Conclusion	37
	References	38
	List of Figures	41
	List of Tables	41
	List of Algorithms	41

1 Introduction

In our modern society, cancer is the second leading cause of death worldwide[27]. A trend can be seen that the number of cancer cases is rising. There are several factors that contribute to this fact: medical imaging and diagnosis have been improved, resulting in more cases that have been missed in the past; due to a better healthcare system and available cures, people live longer, therefore there is an increased chance of developing cancer; the lifestyle has changed, such as tobacco consumption and highly processed nutrition.

Cancer has no general cure; it can only be treated with different methods. Survival rate highly depends on the time of detection, the affected tissue, and the type of cancer. There are three basic treatment types available: chemotherapy, surgical removal, and radiotherapy. The applied method(s) depend on several factors. Both diagnosis and treatment require a great amount of medical human resources and equipment. These resources are limited, and depending on how a country is developed, one or both resources may be an issue in diagnosis and delivering the treatment.

Medical image segmentation is an important step in clinical diagnosis, treatment planning, and treatment verification. It needs highly skilled medical professionals and high capital equipment. To improve the efficiency of the healthcare system, besides increasing the available resources, the precision of the treatment can be improved, and the time needed to treat a patient can be decreased.

Artificial intelligence offers an opportunity to improve both the precision and decrease the amount of time spent per patient. Image segmentation was a fully manual task in the past, while later image processing algorithms were developed to aid the segmentation. With the improved computational capacity, neural networks can also be trained to perform this task, often outperforming the classical algorithms.

One of the differences between classical segmentation algorithms and neural network-based algorithms is that, while for the former class, it can be traced how the segmentation was found, in the latter case, it is not evident. In the clinical field, precision is the most important part as it directly impacts the effectiveness of the diagnosis or treatment and the safety of the patient. As a result, currently, it is mandatory to have a human in the loop who finally reviews and approves any outcome. Medical device manufacturers are responsible for the safety and effectiveness of the medical devices, and their activities are highly regulated by laws.

There is a demand to remove the human in the loop in some cases. The medical professional shortage is extreme in developing countries. Their biggest bottleneck is not the high capital equipment, but the human resource shortage. While the equipment requires

an investment usually from the public healthcare system, and then the recurring required maintenance, medical device companies are interested in bringing their devices to these countries, and they offer special financial programs to support it. On the other hand, handling this equipment requires special know-how and experience. Higher educational systems in these countries are either not capable of producing these professionals, or they cannot meet the demand. The other problem - in contrast to the equipment - is that the medical professional might move abroad to a more developed country.

In these countries, removing the human from the loop could be a future option to provide better healthcare to the population, as it could help with the medical professional shortage. In developed countries, the focus is on delivering better treatments with more precise information. The bottleneck for this is often the medical device manufacturer, as they are not capable of delivering frequent updates to the field, incorporating all the feedback and user needs. One option for the end user would be to adjust the system based on their feedback. To achieve these, the first step is towards the XAI - eXplainable, transparent, and trustworthy AI.

In this work, I will describe a deep learning model for segmentation, along with a combined loss function. I will analyze the impact of post-training the model with new images using different metrics and evaluate if the accuracy of the small or large tumors shifts more. The aim is to evaluate whether post-training could be feasible in the field without having the initial training set.

1.1 Notations

Throughout this work, the following mathematical notations are sustained:

$\circ$	Hadamard symbol. Element-wise multiplication of matrices
$ A $	the number of elements in set A
$x \in A$	x is an element of set A
$A \setminus B$	set A without the elements of B
$A \cup B$	union of set A and B
$A \cap B$	intersection of set A and B
$\sum$	Summation of the elements
X, Y	Represent binary image
$x \in X$	a pixel coordinate in binary image X
Ω	Set of pixels
θ	the parameters of the neural network
$\log x$	Natural logarithm of x
$\ \mathbf{x}\ _2$	Euclidean ($p = 2$) norm
$\min_{x \in X} f(x)$	minimum value of function f over X
$\max_{x \in X} f(x)$	maximum value of function f over X
$\max(x, y)$	maximum of x and y
$p(A)$	probability of event A
$p(A B)$	conditional probability of event A given event B
$\text{diag}(A)$	Diagonal elements of matrix $\mathbf{A}$
$\mathcal{X}$	Sample space
$\int_{\mathcal{X}}$	Integral over $\mathcal{X}$

1.2 Structure of the Thesis

The remainder of this thesis is organized as follows: Section 2 explores the clinical background of cancer, the clinical workflow, and the different demands and challenges worldwide. We briefly describe the international regulatory framework governing medical devices and artificial intelligence in Section 3. We show an implementation of a medical segmentation algorithm in Section 6. We use a Brain Tumor MRI dataset for training and validation, described in Section 5. To measure the quality of the segmentation, we define several metrics (Section 7): DICE, Hausdorff, and GradCAM-based scores. Finally, sections 9 and 10 present experimental results regarding post-training performance shift and outline a conclusion.

2 Clinical background

Why do we need segmentation? There are several cases when we would like to identify a specific area of the human body, either based on biological function or characteristic of the cells in the target area or the surrounding tissues. Besides the identification, volume measurement can contribute to diagnosis and treatment.

In cancer treatment, the location and the size may influence the applied treatment type. Based on the volume (size of the tumor), the physician(s) may decide to remove the tumor surgically or use radiation therapy. Tumors are considered to be defective cells in the body that tend to grow and replace functional surrounding cells with tumor cells. Cancer cells appear when a normal cell mutates its DNA for some reason. It is thought that these mutations suppress normal biological processes which are protecting the cells from becoming a tumor. There are eight hallmarks identified that enable the growth of the tumor [10]:

1. Sustaining proliferative signaling
2. Evading growth suppressors
3. Avoiding immune destruction
4. Enabling replicative immortality
5. Activating invasion and metastasis
6. Inducing angiogenesis
7. Resisting cell death
8. Deregulating cellular energetics

Since the immune system is not capable of killing these ill cells, external intervention is needed. One of the primary treatment methods is surgery, where the tumor is removed, and the other one is radiation therapy (internal or external), where high-energy ionizing radiation is applied to the tumor to further damage its DNA, so it eventually dies [4]. Segmentation for external radiotherapy is crucial as a precise radiation plan needs to be developed, which irradiates the tumor as much as possible. Tumors are surrounded by healthy tissues, sometimes critical organs (called Organs At Risk - OAR), where we would like to minimize the radiation dose. We cannot avoid irradiating healthy tissues, but with optimization, it can be minimized. Irradiation is usually done in 3D space, so the tumor gets the dose from multiple angles.

Precisely localizing the tumor and the organs at risk is crucial from a safety and effectiveness point of view. As the number of cancer patients is constantly climbing, the healthcare system needs to catch up with the increased demand. One way is to improve the throughput by delivering the treatment faster, which can be narrowed down to the following stages:

- **Treatment Planning:** Delineating target volumes and OARs on planning CT/MRI scans.
- **Patient Positioning:** Aligning the patient physically to mirror the spatial coordinates of the simulation scan, adapting the plan if needed.
- **Treatment Delivery:** Administering the planned ionizing radiation dose profile.
- **Treatment Verification:** Utilizing on-board imaging (e.g., Cone Beam CT) to monitor anatomical changes over fractions.

Treatment planning uses segmentation techniques the most. Classical segmentation algorithms are industry standards to use. These algorithms produce a draft segmentation, where a healthcare professional reviews and corrects the output. The neural network-based segmentation methods usually provide a better output. As a result, the healthcare professional needs less time to finalize the segmentation. Ideally, no further correction would be needed.

Medical device manufacturers, besides measuring the quality of the segmentation with predefined metrics (DICE, Intersection Over Union, etc) - which are important from a clinical evaluation aspect - also validate it with the target user group. The user shall be able to realize time savings when using the new method over the old one.

3 Regulatory overview

Most of the countries regulate the medical devices market and also regulate the use of AI-driven products or products using AI. Manufacturers have to comply with these regulations in the target country, otherwise their products are banned. High scrutiny facilitates safe products and serves as a basic quality control as well. Medical Device regulations are comparable in all countries and have a long history, while the regulation of AI is quite new, and there are substantial differences.

3.1 Medical Devices

Medical device regulation dates back to 1976, when the Food and Drug Administration, a federal agency, released the Medical Device Amendments to the FD&C Act in the U.S. The goal of the legislation is to provide assurance that a medical device is safe and effective: it does not harm people, and it can fulfill its intended use effectively. The regulation was written in an agnostic way, as compared to other types of devices where there are safety concerns; the devices vary largely in terms of used technology, duration of use, professional vs layperson use, and being implantable or active devices. The definition is thus quite broad, built around identifying the intended function rather than any specific usage or design. The following definition is from the EU Medical Device Regulation (745/2017), but other countries have very similar ones:

”’medical device’ means any instrument, apparatus, appliance, software, implant, reagent, material or other article intended by the manufacturer to be used, alone or in combination, for human beings for one or more of the following specific medical purposes:

- diagnosis, prevention, monitoring, prediction, prognosis, treatment or alleviation of disease,
- diagnosis, monitoring, treatment, alleviation of, or compensation for, an injury or disability,
- investigation, replacement or modification of the anatomy or of a physiological or pathological process or state,
- providing information by means of in vitro examination of specimens derived from the human body, including organ, blood and tissue donations, and which does not achieve its principal intended action by pharmacological, immunological or metabolic means, in or on the human body, but which may be assisted in its function by such means.” [7]

Medical manufacturers shall provide objective evidence that they meet the legislation before putting the device to market. Depending on the risk class of the device, the legislator may require approval by the authority. Activities of manufacturing include the design and development of the devices, not only the manufacturing steps.

3.1.1 European Union

In the European Union, the 745/2017 (EU Medical Device Regulation, MDR) regulates medical device manufacturing. The regulation shall be followed as soon as it is determined by the manufacturer to meet the definition of a medical device. The regulation stipulates all activities during the lifecycle of the product, including the design and development phase. While the regulation does not contain any specific information about an Artificial Intelligence-based algorithm, the requirements are sufficiently general to enforce proper AI algorithm implementation, verification, and validation. However, the awareness of the risks of AI algorithms was not prominent even for notified bodies, and manufacturers did not use AI in their products due to the low availability of computational capacity.

3.1.2 United States of America

In the United States, the manufacturing of medical devices is regulated by the 21 Code For Regulation part 820. The U.S. was the first country which recognized that software in medical devices carries an increased risk, due to their complexity and the inability to do 100% verification on the software. The required design activities are even identical or very similar to those of the European Union. The regulator is the Food and Drug Administration (FDA), which controls the market. Since in the U.S., there is no separate regulation for AI devices, the FDA issues guidance documents to provide topic-wise interpretation of the regulation.

3.2 Artificial Intelligence

The use of Artificial Intelligence has rapidly spread in recent years due to the increase in computational capacity. In parallel, the increase in the amount of processed data also followed, which enabled having an impact on a diversity of technologies. Despite the new opportunities to boost efficiency in business or automate tasks that were not possible before, there is an increased risk of using this new technology. As AI can be applied to many fields, a wide variety of hazards shall be considered.

Countries start to recognize that leaving it being applied without control can cause severe damage to the economy, property, society, or human life. New regulations are emerging to control one way or another. Some prioritize ultimate safety and human rights, while others leave room for more creative use of AI.

3.2.1 European Union

The European Union was the first to have a very detailed regulation for AI-enabled products. It covers a wide range of products, services, and activities and is applicable to all commercial goods (it excludes research). It identifies four main product groups [16]:

- Unacceptable risk: these devices cannot be deployed
- High-risk devices: these devices require meeting harmonized standards and require a notified body¹ involvement to place them on the market.
- Limited risk devices: device is subject to certain transparency rules
- Low or no-risk devices

The medical devices fall into the high-risk category according to the regulation. The development of high-risk systems is controlled by the regulation, there shall be adequate safety, transparency NS explainability for a product. One other requirement is that a natural person shall be able to oversee its functioning, including whether it can fulfill its intended use during its lifetime. Before placing the device on the market, appropriate human oversight measures shall be defined [8].

3.2.2 United States of America

The United States of America does not have strict regulations on AI product development and usage. They have a National Policy Framework for Artificial Intelligence [23] which sets out some goals, including child protection and labor, but generally it supports the use of artificial intelligence. It also gives a guideline that the detailed regulation will be state responsibility, but the framework shall prevent too much deviation among them.

The federal Food and Drug Administration (FDA) issues guidance documents on the AI usage in medical devices, which clarifies the federal agency's standpoint on the Quality Medical System Regulation when it comes to AI-enabled devices.

¹"A notified body is an organisation designated by an EU country to assess the conformity of certain products before being placed on the market. These bodies carry out tasks related to conformity assessment procedures set out in the applicable legislation, when a third party is required." [6]

3.3 Medical Device with AI

Medical devices with an AI component or being an AI medical device product shall comply with both the EU AI Act and the EU Medical Device Regulation. A common misunderstanding persists that the regulation forbids developing an AI-enabled medical device that can be fine-tuned by the end user in the field. It is generally allowed, but it shall be ensured that the safety and performance of the device do not degrade during the fine-tuning. Both the Medical Device Regulation and the EU AI Act enforce it. It is crucial that the amount of change should not be significant, as medical device regulations require new regulatory action (in the EU, it requires a new conformity assessment, in the U.S., a new submission) from the manufacturer in this case. The non-significant changes shall be carried out according to predetermined changes set by the manufacturer.

In the U.S., the FDA has developed a guidance document in 2025 specifically on how changes shall be controlled for AI-enabled medical devices [24]. It provides specific examples of when a modification to the devices in the field was permissible or not. The Predetermined Change Control Plan is authored by the manufacturer, and it shall list what kind of changes are permitted in the field. It may contain complete re-training or fine-tuning for the model. It also guarantees, through verification and validation, that safety and effectiveness are ensured for the device after the change is implemented.

4 Objectives

In this thesis, we are looking for answers to the following:

- Can a saliency score be used to measure accuracy?
- Is there a difference in accuracy drift for small and large tumors when post-training is used to train new images from the same type of tumors?
- How does the accuracy drift change with different loss functions used to calculate the Fisher Information Matrix?

5 Dataset

We use the Brain Tumor MRI Dataset from the Science Database [5]. It consists of three types of tumors with the corresponding binary masks: Glioma, Meningioma, and Pituitary tumor. We will use the Glioma subset only, which contains 554 grayscale MRI scans, with an image size of 512×512 . Binary masks have the same size.

We plan to have two training cycles: a base training, where the model is trained to perform on the validation dataset with at least a DICE score 0.7; and a second - post-training - cycle, which will attempt to train additional Glioma images without having catastrophic forgetting. Therefore, the training dataset is further split with 95% - 5%, having a base dataset with 527 images and 27 images for the second training cycle.

We divide the base dataset into a train and a validation dataset using the 80% - 20% rule, which gives us 421 training images and 106 validation images. The tumor size distribution is shown in Figure 1. The validation dataset is further categorized into small or large tumors.

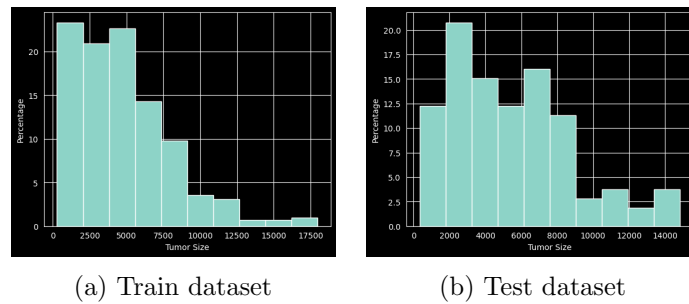


Figure 1: Size histogram

5.1 Augmentation and transformations

There are three different places where image alteration has been implemented:

- During initial image loading from storage
- Custom Dataset image access
- Dataset transformation on data access

Image loading. When images are loaded from external storage, it applies normalization to the MRI scans and transforms the mask image to $\{0, 1\}$. It also calculates a distance transform of the mask, which will be used later in the loss function. Given a binary image (in our case, it is a mask), the distance transform is a map where each pixel value represents the distance to the first background pixel. We would like to create a map where each pixel

represents the distance to the contour of the mask: if a pixel is outside the mask, it gets a positive value; a pixel inside the mask gets a positive value too. It will show a 0 value on the exact contour and borderline.

We calculate it in three steps:

1. We calculate a distance transform map, where value 0 is the background value, and value 1 is the mask value. In this map, everything outside the mask will be 0, and within the mask, from its center point, it will gradually fade out.
2. We invert the image, so the mask will get a value 0, and everything else will get a value 1. We repeat the same procedure. In the map, every pixel inside the mask will be 0, while outside, the pixel values gradually increase as the distance gets bigger from the contour of the mask.
3. The two maps are added, and as a result, only the contour will have 0 values.

This combined map is stored for each mask.

Image Access. When the custom dataset receives a random access request, it randomly performs the following operations on the image, mask, and the distance-transformed:

- It rotates the image, mask, and the distance-transformed mask by either 0, 90, 180, or 270 degrees.
- It flips the image, mask, and the distance-transformed mask horizontally.

Both random samplings are from a uniform distribution. Rotating by multiples of 90 degrees and flipping are operations that can be applied to the distance transform without needing to recalculate it.

Dataset transformation on data access. There is a limitation on what kind of augmentations can be applied: only those that do not change the mask or the distance-transformed mask. Therefore, we need to have shape-retaining transformations. As a result, scaling, rotation, and clipping cannot be used. We apply only a Gaussian Noise to the MRI scan.

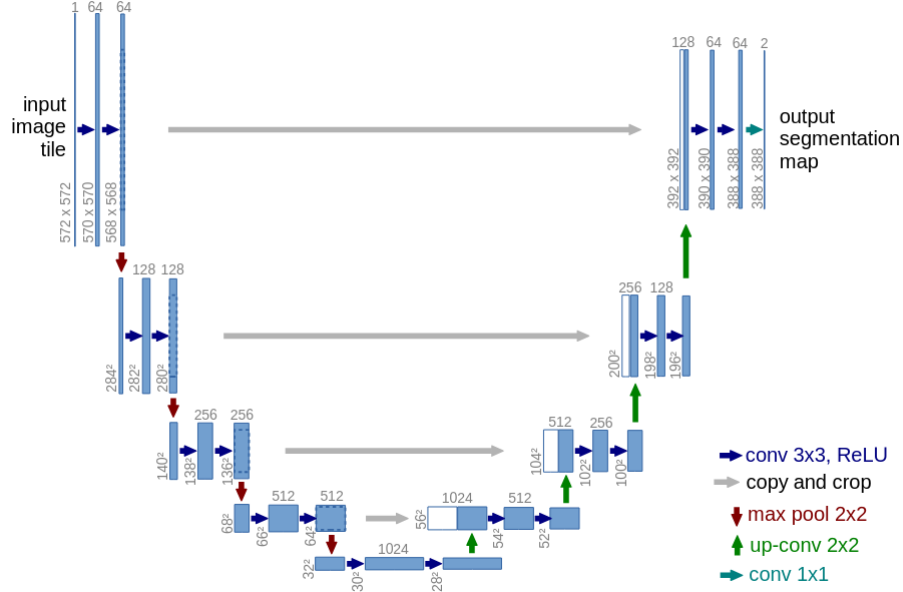


Figure 2: U-Net architecture

6 Model

For medical image segmentation, the U-Net architecture is one of the most widely used solutions. It provides good accuracy and also memory efficiency [2]. Many variants and improvements to the original U-Net architecture exist. The standard architecture is a U-shaped model, where the left side provides the encoder functionality (contracting path) and the right side provides the decoder (expansive path), and horizontally, there is the bottleneck layer, containing the richest feature representation. The number of channels on the left and right increases with depth, while the spatial resolution has the opposite attribute. There is a skip connection (concatenation) on the decoder side from the encoder on the same level, allowing the model to utilize information directly from the encoder [20]. The basic U-Net architecture is depicted in Figure 2

6.1 Attention U-Net

Natural language processing performance was boosted when the attention mechanism was introduced [25]. Based on this idea, a similar concept was proposed for the U-Net architecture by using attention gates [19]. The main idea is that the skip connections receive a

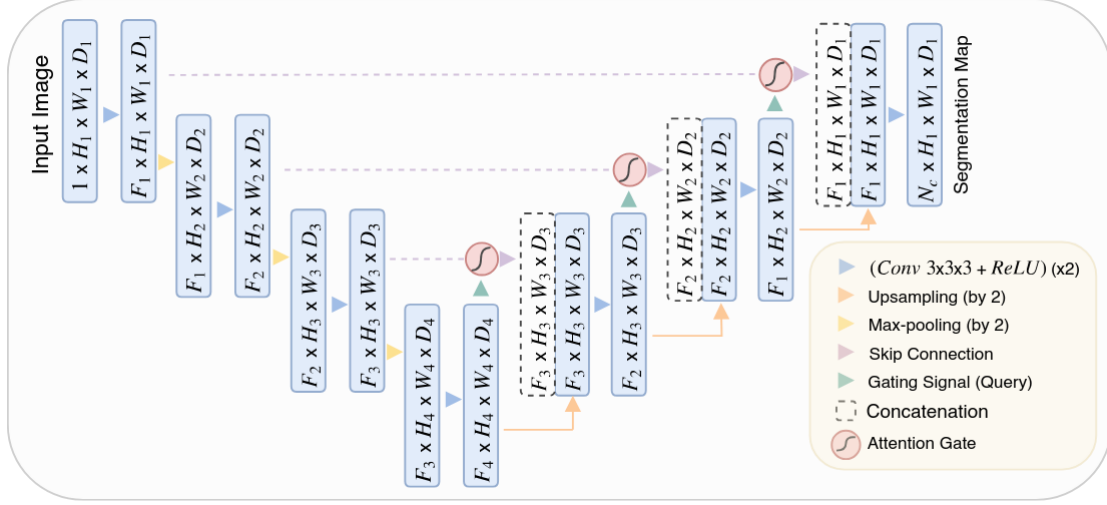


Figure 3: Attention U-Net architecture [19]

learnable attention gate, which can filter the information received from the encoder side, based on a signal from a lower decoder layer.

The attention gates (depicted in Figure 4) receive two inputs: one signal from the lower decoder (g) layer and one from the encoder (x^l). The W_x is a $1 \times 1 \times 1$ convolution with a stride of 2, so it downsamples the input to match the resolution with g . At the end of the chain, before the multiplication, the result is up-scaled to match the resolution of x^l .

The W_x convolution with a stride of 2 simply drops every second pixel, therefore we lose information. To overcome this, we apply a slight modification to the attention gate. Instead of downsampling and matching the x^l to g , we leave the spatial resolution of x^l as it is and upscale g instead. Therefore, the re-sampler is moving from the end of the gating block, right after the W_g convolution. Usually, the fastest up-scaling method is used, which is by default in `PyTorch`, is the nearest-neighbor. It produces a pixelized output, which may influence the accuracy. We choose the bi-linear interpolation techniques, which produce better results compared to nearest-neighbor, but still faster than bi-cubic interpolation[9]. The modified gating block is depicted in Figure 5.

The Attention U-Net model we use throughout this work has the following parameters:

- Base channel size is 32
- depth is 5 $\Rightarrow$ which means there are 4 down convolution blocks, 1 bottleneck block. 4 up convolution blocks

Since the dataset contains only 554 samples, and besides the augmentation used, we

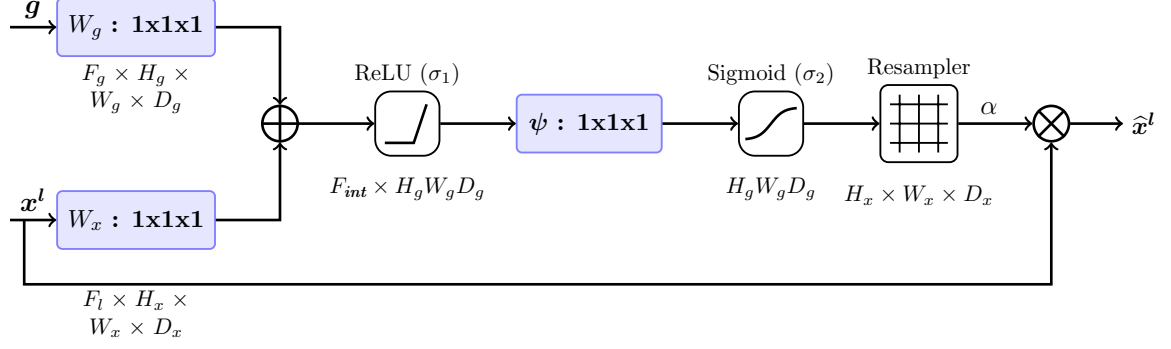


Figure 4: U-Net attention gate as in the original paper [25]

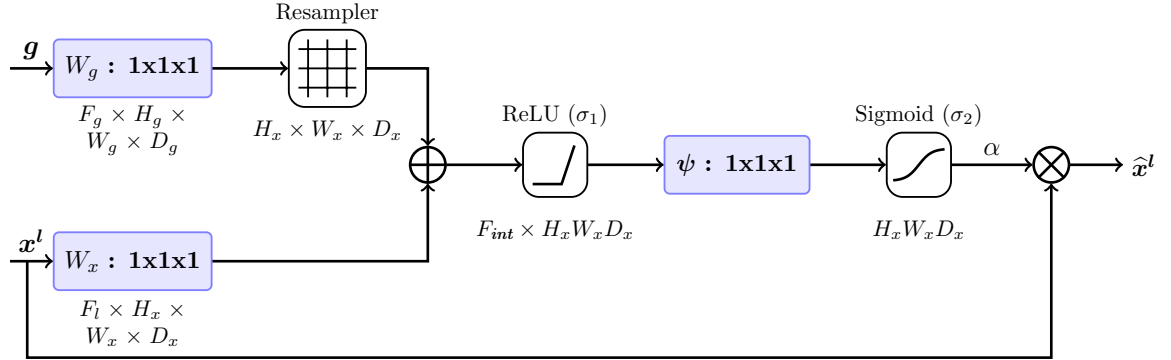


Figure 5: The modified attention gate with up-scaling

would like the model to be able to generalize better[21], therefore, a dropout layer is used in each layer, except the first one. The `Dropout2d` function is used, level-wise probability is applied (Equation (1)): 0, 0.15, 0.1, 0.15, 0.2. Dropout is not used in the gating blocks.

$$p(\text{dropout}) = 0.2 \frac{\text{level}}{4}, \text{level} \in \{0, \dots, 4\} \quad (1)$$

6.2 GradCAM

To understand which information was used in the picture by the model to generate the output mask, we can take a look at the activations in the model. A method, Classification Activation Map [28], was proposed, where areas of the image could be identified that drive the specific classification of the network. There are reports on using saliency maps for medical image classifications, gaining insights into which region influenced the model decisions the most. GradCAM was reported to underperform compared to newer methods, like ScoreCAM[3]. The efficiency and usefulness of GradCAM were evaluated in an article[1] in the ophthalmology area, where it was reported that small features were not detected by GradCAM. It also concludes that GradCAM (and similar techniques) has limitations in explainability but may better perform in debugging the model.

While it was used for classification tasks, the method can be applied generally. For segmentation tasks with Attention U-Net, the map will show which area contributes to the final mask.

7 Metrics

In any application and industry, when a computed or generated result is used for decision-making, it is important to know the extent to which we can expect an accurate result. In medical imaging, it means we would like to know, on average, how accurate the segmentation can be [18], whether it aids diagnosis or treatment. Different metrics have been developed, each of which measures the accuracy from a different aspect.

Manufacturers tend to report high accuracy for the end users, but it does not necessarily mean that the result always meets the expectation. AI-based methods outperform classical algorithms, and there are already several organs or tissue which can be segmented comparable to human performance. It raises the risk that there can be overconfidence in the algorithm, and while it is always required to validate the result, the false confidence can lead to overlooking an error. Therefore, measuring the AI performance is crucial not only in the development phase, but also after the AI solution is deployed.

We will present three metrics.

7.1 DICE

The DICE score – also known as DICE Similarity Coefficient – is the most used metric for validating the segmentation [22]. It measures the amount of overlap between two samples and gives a score between 0 and 1, where 0 means no similarity and 1 means perfect overlap. The score is calculated using the number of True Positive(TP), False Positive(FP), False Negative(FN) pixels (Equation (2)).

$$\text{DSC}(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \quad (2)$$

In medical image segmentation, the mask is a connected set of pixels. The disadvantage of using the DICE score is that when we compare our predicted area to the ground truth, and there are non-overlapping pixels far from the reference object, it may produce the same DICE score, while we see that the result is far from the expectation. The DICE score does not consider distances. Our goal is to get a segmentation that is as close as possible to the ground-truth.

7.2 Hausdorff distance

Hausdorff distance is a distance-based metric; it measures how far apart the boundaries of the two sets are. The one-sided distance measured between X and Y is defined in

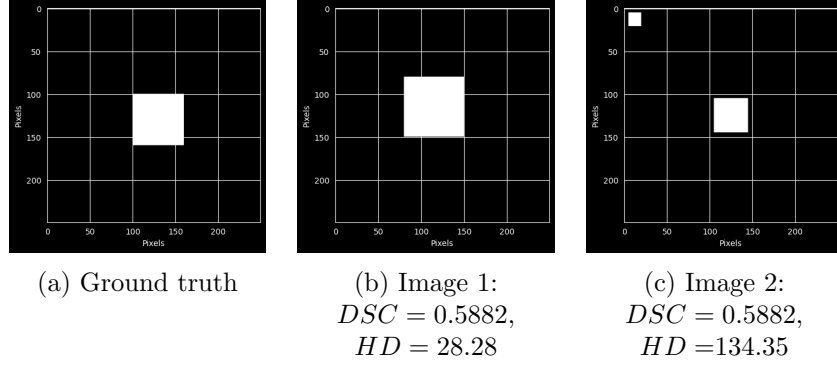


Figure 6: DICE scores: ground truth and two images with identical DICE scores

Equation (3), while the distance between Y and X is defined in Equation (4).

$$hd(X, Y) = \max_{x \in X} \min_{y \in Y} \|x - y\|_2 \quad (3)$$

$$hd(Y, X) = \max_{y \in Y} \min_{x \in X} \|x - y\|_2 \quad (4)$$

To get the bidirectional distance, we take the maximum of the one-sided distances (Equation (5)). In the medical image segmentation, we measure the distance of the boundaries of the segmented area. We use L_2 Euclidean distance, but other metrics could also be used.

$$HD(X, Y) = \max(hd(X, Y), hd(Y, X)) \quad (5)$$

7.3 Saliency score

As we saw in Section 6.2, the Grad-CAM image shows where the model is focusing. When we know the ground truth segmentation mask, then we can calculate a score indicating how much attention falls in the desired target area compared to all attentions.

$$Sc(X, Y) = \frac{\sum (G\text{Cam}(X) \circ Y)}{\sum G\text{Cam}(X)} \quad (6)$$

Where X is the predicted mask, and Y is the ground-truth mask. Bigger scores mean the model is focusing on the area of the tumor mask.

8 Training

The training is done in batches of 8 samples; additionally, gradient accumulation is applied with a value of 4. It gives us an effective batch size of 32. Training loop implements gradient clipping to avoid gradient overflows.

8.1 Loss function

The standard loss function would be to use DICE, but using the DICE score alone could still potentially mark segments far from the original mask. To overcome this, we use a combination of DICE score and Hausdorff distance as suggested by Davood Karimi and Septimiu E. Salcudean [11]. The DICE scores help maximize the overlapping area, while the Hausdorff distance minimizes pixels outside of the desired area.

Neither of the defined metrics in Equation (2) nor Equation (5) is differentiable, thus cannot be used as a loss function. To overcome this limitation, a "soft" version of them will be utilized, which are approximations.

8.1.1 Soft DICE

We follow the DICE score calculation for the loss function in [11]:

$$\text{DSC}_{\text{soft}}(p, q) = \frac{2 \sum_{\Omega} (p \circ q)}{\sum_{\Omega} (p^2 + q^2)} \quad (7)$$

where p is the ground-truth mask, and q is the predicted mask.

8.1.2 Soft Hausdorff

Hausdorff distance is non-differentiable; thus, we use an approximation, which is differentiable, the one-sided Hausdorff distance function [11]

$$\text{HD}_{\text{OS}}(p, q) = \frac{1}{\Omega} \sum_{\Omega} ((p - q)^2 \circ d_p^{\alpha}) \quad (8)$$

where q is the predicted mask and p is the ground-truth mask, and d_p is the distance transform of the ground-truth. d_p is a distance transform of the ground-truth mask p . The distance transform is a map where each pixel value represents the distance to the boundary of an area. α is defined as 2.

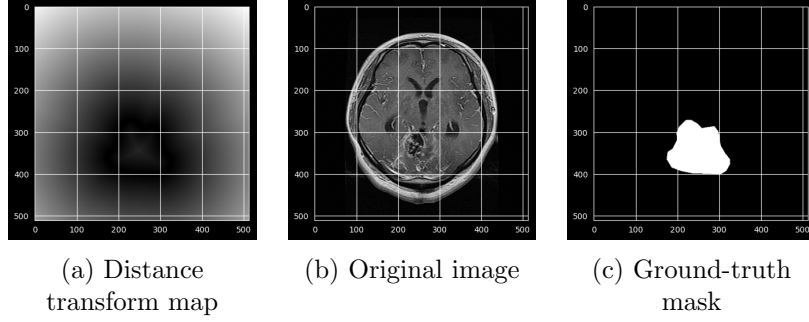


Figure 7: Distance transform

8.1.3 Combined

We use a combined loss function for training with a DICE and a Hausdorff component. Since the DICE component falls in the $[0, 1]$ range but the Hausdorff distance $HD(X, Y) \geq 0$, they are not on the same scale. In the combined loss function, we multiply the DICE score by a λ value to fall in the same magnitude (Equation (9)) [11].

$$\text{Loss}(p, q) = \text{HD}_{\text{OS}}(p, q) + \lambda(1 - \text{DSC}_{\text{soft}}(p, q)) \quad (9)$$

The λ is updated after each epoch (Equation (10)), thus the ratio in the next epoch depends on the total DICE and Hausdorff loss of the training data in the previous epoch [11]. We choose $\lambda = 100$ as the initial value.

$$\lambda = \frac{\text{HD}_{\text{OS}}(p, q)}{\text{DSC}_{\text{soft}}(p, q)} \quad (10)$$

8.2 Training loop

Due to the Hausdorff loss function, the distance transform of the ground truth mask shall be passed to the combined loss function. The distance transform is constant and pre-computed during image loading (Section 5). See Algorithm 1 for details.

8.3 Results

After 200 epochs, on the training dataset we could reach 0.761 DICE score, and on the validation dataset we could get 0.766 DICE score and Hausdorff distance of 114. During the training, the validation DICE score showed significant variance in the first 150 epochs, and in the last 50 epochs, it flattened. The Hausdorff distance shows a logarithmic decay.

Algorithm 1 Training

```

 $\lambda \leftarrow 100$ 
for  $ep \in [0, \text{epochs} - 1]$  do
  Model.train()
   $i \leftarrow 0$ 
  for  $\forall b \in \text{Batches}$  do
     $L \leftarrow \text{Loss}(b['\text{mask}'], \text{model}(b['\text{image}']), \text{mask\_dt} = b['\text{mask\_dt}'], \text{ratio} = \lambda)$ 
    L.backward()
    if  $i + 1 \bmod \text{accum\_scale} = 0 \vee i + 1 = |\text{samples}|$  then
      ClipGradNorm(model.parameters(), max_norm=1)
      optimizer.step()
      optimizer.zero_grad()
    end if
     $i \leftarrow i + 1$ 
  end for
  scheduler.step()
   $\lambda \leftarrow L.\text{total\_hd} / L.\text{total\_dice}$ 
end for

```

The loss function also shows an exponential decay behavior. The factor λ follows the same pattern; it is decreasing as DICE scores increase and Hausdorff distance decrease.

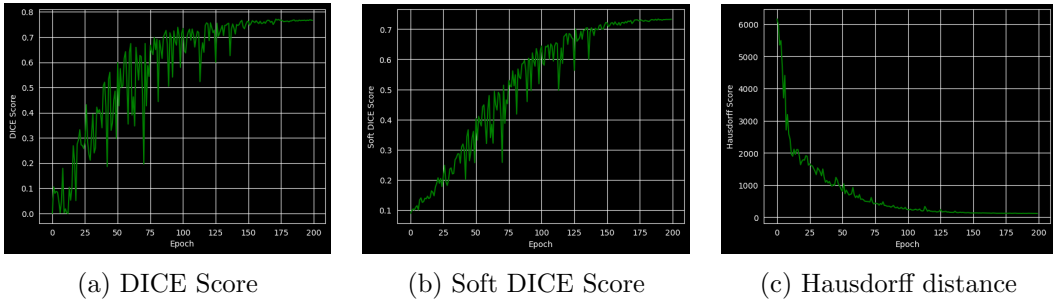
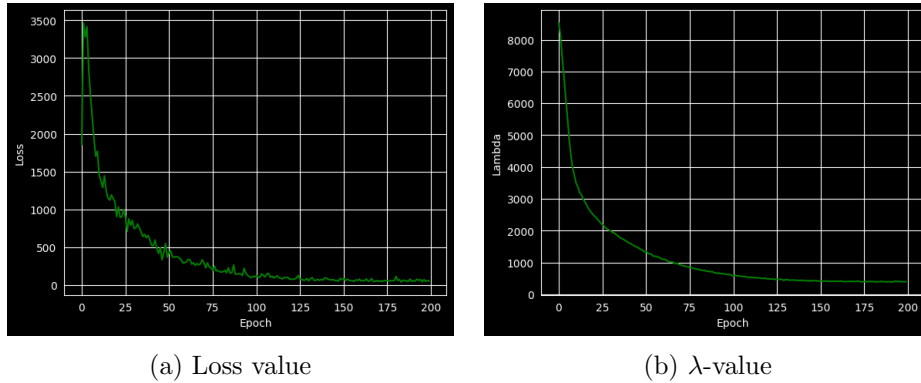


Figure 8: Base training validation scores

Figure 9: Base training loss and λ -value

9 Post-training

We simulate an on-field post-training activity by an end-user. We train an additional 24 images without the initial training set. During the training, we record the validation scores of the initial validation set, as well as the scores of the trained images.

9.1 Measurement

During the training, several parameters are measured per epoch on the training dataset and also on the initial validation dataset. For the validation set, each parameter is tracked separately for small tumors, large tumors, and all tumors. The following parameters are tracked:

- Hausdorff Distance
- DICE score
- Soft Dice score
- Saliency score

Measurements are done after each epoch.

9.2 Elastic Weight Consolidation

If the same training loop and loss function were used as in the first training, the weights would change fast to learn the new small dataset, while losing the know-how of the old one. The model would over-fit to the smaller dataset and lose generalization capability; as a consequence, the accuracy on the validation dataset would decay. To overcome this

problem, we use the Elastic Weight Consolidation[13] method, which prevents those weights from being changed that contribute to the already trained image segmentation. This method could be a mitigation of the requirement set out in AI Act Article 15:

”High-risk AI systems shall be designed and developed in such a way that they achieve an appropriate level of accuracy, robustness, and cybersecurity, and that they perform consistently in those respects throughout their lifecycle.”[8]

Let us have a neural network trained on Task A, with parameters $\theta(t)$. When we train the same neural network with samples from Task B, we would like to ensure that the new parameters $\theta(t + 1)$ can still provide an acceptable accuracy on Task A and also on Task B. In the solution space, there are local minima where Task A accuracy is small, while accuracy for Task B is high, and also when Task A accuracy is only slightly decreased while the neural network learns Task B as well.

We would like to find the most probable θ given data D . Using the Bayesian rule:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \quad (11)$$

$$\log p(\theta|D) = \log p(D|\theta) + \log p(\theta) - \log p(D) \quad (12)$$

The $\log p(\theta|D)$ is actually the negative of the loss function, which is the negative log-likelihood [13, 14]. We have two tasks: A and B. For task A, the θ is known, and let D_B be the data from task B to train. D covers the whole dataset. The equation Equation (12) can be rearranged to[13]:

$$\log p(\theta|D) = \log p(D_B|\theta) + \log p(\theta|D_A) - \log p(D_B) \quad (13)$$

When we have a discrete probability distribution, we can calculate the entropy for a probability variable using Equation (14) and for continuous distribution case we can use Equation (15)[12]:

$$H(X) = - \sum_{x \in X} p(x) \log p(x) \quad (14)$$

$$h(X) = - \int p(x) \log p(x) dx \quad (15)$$

It measures the expected uncertainty of a random experiment represented by the probability variable. It can be said, that observing a high entropy value leads to higher information gain. When $\log_2$ is used in Equation (14), the unit of information is bit.

Instead of measuring the amount of information a probability variable carries, we would like to measure the average sensitivity of parameter θ on the distribution [26]. Using the Fisher Information, we can estimate the change in the probability distribution ($\mathbf{X}$) with respect to the change in the parameter. The Fisher Information Matrix is the negative of the expectation of the Hessian of the the log-likelihood function[15].

$$J_{\mathbf{X}} = -E \left[\frac{\partial^2 \log f(\mathbf{X})}{\partial \mathbf{X} \partial \mathbf{X}^T} \right] \quad (16)$$

In our case the probability distribution is defined by the parameter space θ , and Equation (16) can be written in the form:

$$J_{\theta} = - \int_{\mathcal{X}} f(\mathbf{x}|\theta) \frac{\partial^2 \log f(\mathbf{x}|\theta)}{\partial \theta \partial \theta^T} d\mathbf{x} \quad (17)$$

Where $\mathcal{X}$ denotes the sample space of possible input images.

The Fisher Information Matrix tells us on average, how sensitive the distribution depends on parameter θ . Higher values means, that a small shift in θ would cause large shift in the distribution. The posterior probability $p(\theta|D_A)$ contains all information about which parameters in θ are important to accomplish the task. During the training we are not only interested to get to a point in the parameter space where Task B performs well, but we are also interested in the trajectory. If we can accomplish to move over those points where Fisher Information Matrix indicates low information loss, then we are preserving the knowledge for Task A. The goal is to find a solution in the parameter space among these points where Task B performance is maximal. We approximate the posterior with the diagonals of the Fisher information matrix F , which leads us to the combined loss function [13]:

$$L(\theta) = L_B(\theta) + \frac{\lambda}{2} \sum_i F_i(\theta_i - \theta_{A,i}^*)^2 \quad (18)$$

Where λ is a factor, showing how important it is to keep the ability to perform task A, θ^* represents the parameters without any modification due to learning Task B.

We can approximate the Fisher information matrix by only its diagonals, but we still need a way to calculate them. Empirical Fisher Information uses only gradients, and as presented in [14], it can be expressed as:

$$\text{diag}(F_{\text{std}}^{\text{emp}}(\theta)) = \sum_{n=1}^N \mathbf{g}_n^2 \quad (19)$$

Where $\mathbf{g}_n$ is the gradient and $\mathbf{g}_n^2$ is the element-wise square. The sum is over the the per-sample gradients.

It is common in EWC to use binary cross-entropy as a loss function when the task is binary classification or segmentation. We run a similar loop as for the training, except there will be no optimizer call within the loop. We predict each training sample with the model and calculate the loss, then we calculate the gradients. The squares of the gradients will be summed. The Fisher Information Matrix is normalized to have a maximum value of 10^3 . With this approach, we can compare outcomes with different loss functions used in the Fisher matrix calculation. With this step, the weight for the EWC loss component can be the same for the experiments.

Algorithm 2 Fisher matrix computation

```

fisher ← {}
samples=0
model.eval()
i ← 0
for  $\forall b \in \text{Batches}$  do
    samples ← samples + |b|
    model.zero_grad()
     $L \leftarrow \text{Loss}(b['\text{mask}'], \text{model}(b['\text{image}']))$ 
    L.backward()
    for name, params  $\in \text{model.named\_parameters}()$  do
        if param.grad  $\neq \text{None}$  then
            fisher[name] ← fisher[name] + (param.grad.data)2
        end if
    end for
end for
for  $\forall \text{name} \in \text{fisher}$  do
    fisher[name] ← fisher[name] / samples
    f_max ← fisher[name].max().item()
    if f_max > abs_max then
        abs_max ← f_max
    end if
end for
for  $\forall \text{name} \in \text{fisher}$  do
    fisher[name] ← fisher[name] / (abs_max / 1000)
end for
return fisher

```

The Fisher Information Matrix is classically derived by using the Binary Cross Entropy

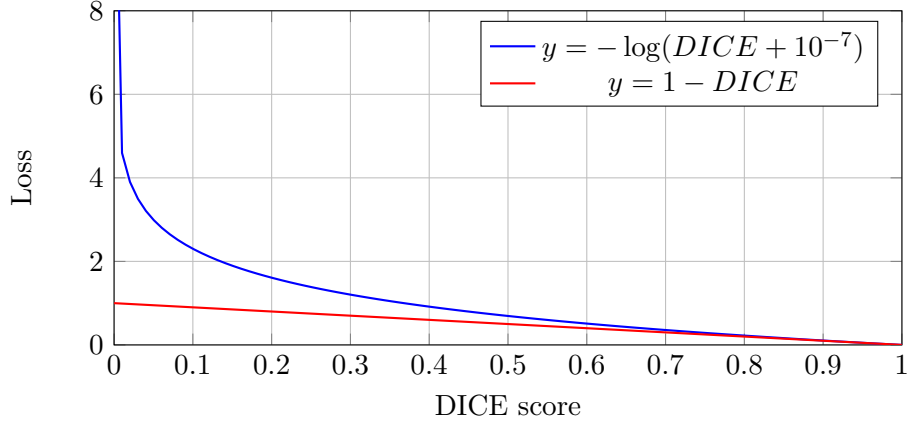


Figure 10: Negative Logarithm of DICE score vs DICE loss

loss function; our experimental loss function utilizes DICE and Hausdorff score. We would like to sustain the DICE performance of the model after task A is trained, thus we perform an experiment with a modified loss function to calculate the Fisher matrix, where the binary cross-entropy loss function is replaced with:

$$\text{Loss}_{\text{ewc}} = \frac{\text{HD}_{\text{OS}}(q, p)}{200} - \log(\text{DSC}_{\text{Soft}}(q, p) + 10^{-7}) \quad (20)$$

We applied a divisor with a value of 200 on the Hausdorff distance to match the scale of the DICE score. The value 200 was determined by performing an exploratory training and choosing a value close to the Hausdorff distance. The application of λ from Equation (9) would lead to a dynamic relation between the loss and the DICE score. We also use the negative log DICE (Figure 10). For bigger DICE scores, it behaves linearly, while for small ones, the penalty is exponential.

Where p is the ground-truth mask, and q is the predicted mask.

During the post-training phase, we use the same loss function as in the base training, but we add the Elastic Weight Consolidation penalty:

$$L(\theta) = \text{HD}_{\text{OS}} + \lambda_{\text{dice}}(1 - \text{DSC}_{\text{Soft}}(q, p)) + \frac{\lambda_{\text{ewc}}}{2} \sum_i F_i(\theta_i - \theta_{A,i}^*)^2 \quad (21)$$

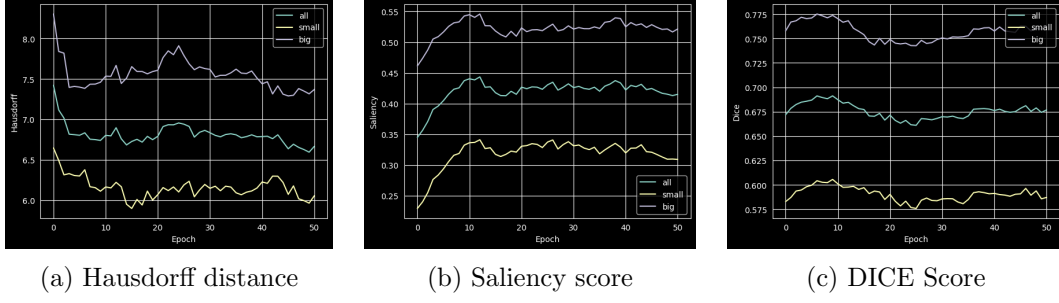


Figure 11: Post-training: initial validation set score shift

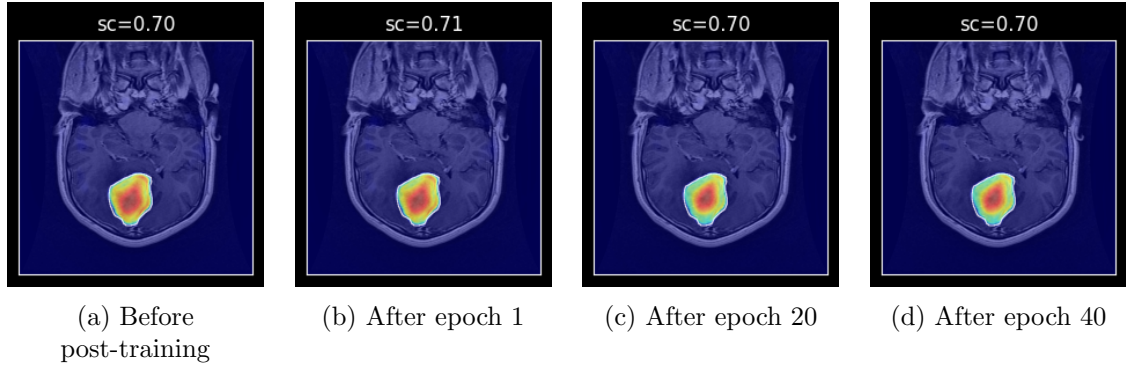


Figure 12: Post-training: Grad-CAM map on the initial validation set

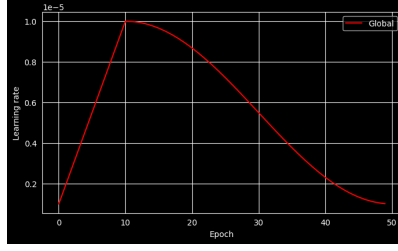
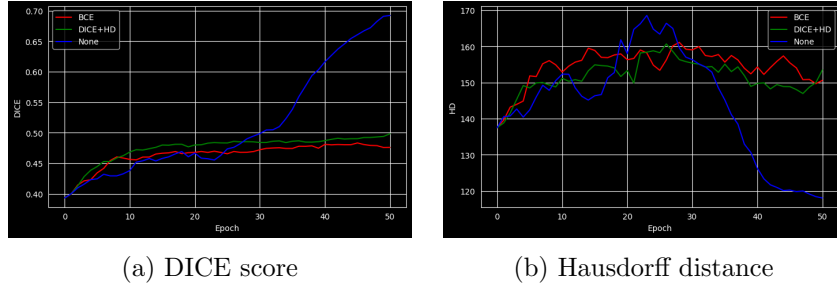


Figure 13: Learning rate during post-training



(a) DICE score

(b) Hausdorff distance

Figure 14: Post-training: DICE and Hausdorff distance of the training set

9.3 Post-training results

The model was post-trained with 24 additional Glioma images. For both experiments, the same training parameters were used:

- Linear warmup: 10 epochs
- Cosine Annealing scheduler
- Initial learning rate: 10^{-6}
- Peak learning rate: 10^{-5}
- ETA minimum: 10^{-6}
- AdamW optimizer

For the post-training phase, we do not define a validation set, as it would be too small and not representative. Instead, we measure the accuracy of the training set at the end of each epoch. The DICE accuracy starts at 0.39 (measured before the first epoch). The change of the DICE score and Hausdorff distance on the training set can be seen in Figure 14.

The learning rate is shown in Figure 13.

9.3.1 EWC with BCE loss

The Fisher Information Matrix was generated with the Binary-Cross Entropy loss function with logits output, with reduction set to `mean` (default value). The loss was multiplied by a factor of 10000. Training with the loss function in the base training loop with EWC penalty was executed for 50 epochs. λ_{ewc} was set to 10^7 .

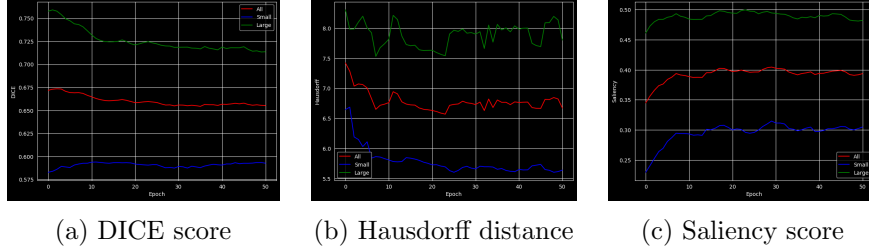


Figure 15: Validation results when Fisher Matrix was generated with BCE loss function

Small tumor scores. In general, all scores were improved for small tumors: the DICE scores changed from 0.58 to 0.59, Hausdorff distance changed from 6.64 to 5.64, Saliency score changed from 0.23 to 0.30.

Large tumor scores. The DICE score dropped from 0.75 to 0.71, Hausdorff distance dropped from 8.3 to 7.8, Saliency score improved from 0.46 to 0.48.

All tumor scores. DICE dropped from 0.67 to 0.65, Hausdorff distance improved from 7.4 to 6.6, Saliency score improved from 0.34 to 0.39.

During the training, the validation score trend did not change, no sharp change in any of the metrics. On the training set, the DICE score climbed to 0.476, the Hausdorff distance dropped to 150.

9.3.2 EWC with DICE and Hausdorff loss

The Fisher Information Matrix was generated with log Dice Loss and Hausdorff Distance. The loss was multiplied by a factor of 1000. The difference is that BCE was used with a reduction set to `sum` and therefore produces larger gradients. Training with the loss function in the base training loop with EWC penalty was executed for 50 epochs. λ_{ewc} was set to 10^7 .

Small tumor scores. The DICE scores did not change significantly, moved from 0.583 to 0.589, Hausdorff distance changed from 6.64 to 5.42, Saliency score changed from 0.23 to 0.295.

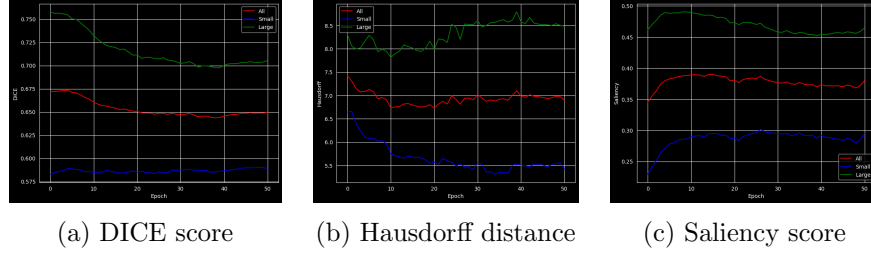


Figure 16: Validation results when Fisher Matrix was generated with $-\log \text{DICE} + \text{HD}$ loss function

Large tumor scores. The DICE score dropped from 0.758 to 0.705, Hausdorff distance increased from 8.3 to 8.47, Saliency score moved slightly from 0.461 to 0.465.

All tumor scores. DICE dropped from 0.67 to 0.65, Hausdorff distance improved from 7.4 to 6.9, Saliency score improved from 0.34 to 0.38.

During the training, the validation score trend did not change, no sharp change in any of the metrics. On the training set, the DICE score climbed to 0.498, the Hausdorff distance dropped to 153.

9.3.3 No EWC

To be able to evaluate the performance of the Elastic Weight Consolidation, the post-training was performed without the Elastic Weight Consolidation component in the loss function. The results show that there is a significant drop in the scores. The training period can be divided into three phases: steep degradation, moderate recovery, and slow recovery. There is also different behavior based on the size of the tumor. The DICE score of the large tumors tends to drop fast in the beginning of the training and goes below the score of small tumors. At epoch 9, the lines crossed each other, and also at epoch 38.

The Hausdorff distance for small tumors is improving over the training, and the final distance is below the starting distance.

Small tumor scores. The DICE score dropped from 0.58 to 0.57, Hausdorff distance changed from 6.64 to 5.63, Saliency score improved from 0.23 to 0.26.

Large tumor scores. The DICE score dropped from 0.758 to 0.69, Hausdorff distance increased from 8.3 to 8.56, Saliency dropped from 0.46 to 0.40.

All tumor scores. DICE dropped from 0.671 to 0.637, Hausdorff distance improved from 7.4 to 7.0, Saliency score improved from 0.34 to 0.33.

During the training, the validation score trend changed; the first part of the training

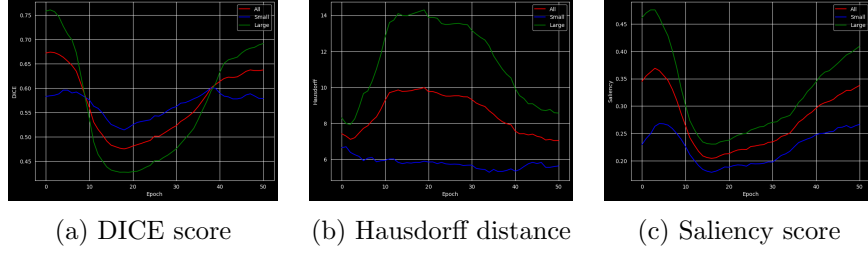


Figure 17: Validation results without EWC

Tumor size	Small			Large			All		
Metric	DICE	HD	Saliency	DICE	HD	Saliency	DICE	HD	Saliency
EWC: BCE	+	+	+	-	+	+	-	+	+
EWC: DICE +HD	+	+	+	-	-	+	-	+	+
EWC: None	-	+	+	-	-	-	-	+	+
Total + / -	9 / 1			3 / 6			6 / 3		

Table 1: Improvement / Degradation per tumor size

(epoch 1-20) showed worsening scores, and the second part of the training (21-50) showed improving scores. Except for the small tumors, all metrics declined. On the training set, the DICE score climbed to 0.69, the Hausdorff distance dropped to 118.

9.3.4 Comparison

In all EWC options (BCE, DICE+HD, None), it can be observed that small tumor segmentation quality improved, while for large tumors, the quality declined. For EWC with loss function BCE and DICE+HD, for small tumors, all metrics measured improvement when comparing epoch 1 and epoch 50. For the "large" and "all" tumor datasets, the DICE score declined for all three EWC options. The no EWC option showed the biggest improvement in accuracy on the post-training train dataset.

9.4 Saliency score analysis

To be able to evaluate if the saliency score could be used as a metric in a similar way as DICE or Hausdorff distance, we examine two properties:

- If the score is related to the target area
- If it shows agreement with another metric (DICE or Hausdorff)

The first item is a direct consequence of Equation (6), as in the nominator, the saliency map is element-wise multiplied by the ground truth mask. Since the mask pixel values can

	DICE	HD	Saliency
DICE	1	0.15	0.07
HD	0.15	1	0.61
Saliency	0.07	0.61	1

Table 2: Cohen’s Kappa for pairwise metrics

be 0 or 1, the Saliency score falls in $[0, 1]$. For the second property, we take the $+$ and $-$ values per metric and create a flat vector of them where $+$ is represented by 1 and $-$ is represented by 0:

$\text{dice} = [1, 1, 0, 0, 0, 0, 0, 0, 0]$
 $\text{hd} = [1, 1, 1, 1, 0, 0, 1, 1, 1]$
 $\text{saliency} = [1, 1, 1, 1, 1, 0, 1, 1, 1]$

We treat each metric as a rater and compare them. We measure Cohen’s Kappa pairwise, which gives information about how much agreement there is between two raters. In the medical domain, the Cohen’s Kappa statistic is used to measure the agreement between two raters on the same items. Cohen’s Kappa result falls into $[-1, 1]$, where < 0 means disagreement and 1 means complete agreement. Cohen suggested the following qualitative values [17]:

- 0.01 – 0.20 none to slight,
- 0.21 – 0.40 fair,
- 0.41 – 0.60 moderate,
- 0.61 – 0.80 substantial,
- 0.81 – 1.00 almost perfect agreement.

Cohen’s Kappa values show that the saliency score has substantial agreement with the Hausdorff distance, while ‘none to slight’ agreement with the DICE score. Hausdorff distance has also only ‘none to slight’ agreement with the DICE score.

As it could be shown that there is a relation with the target area, and also there is a substantial agreement with the Hausdorff distance, we accept that the saliency score can be used as a metric.

Based on the results, the saliency score with this model and dataset shows another nice property: the graph drawn by the values for small and large tumors seems to be visually almost parallel. The rate of change does not depend on the tumor size. We measure the

Spearman correlation between the data points for all three EWC loss functions.

	<i>Small-Large</i>	<i>Large-All</i>	<i>Small-All</i>
EWC: BCE	0.5	0.845	0.865
EWC: DICE+HD	-0.15 ²	0.766	0.413
EWC: None	0.92	0.99	0.95

Table 3: Spearman correlation of the Saliency points

When no EWC is applied, the statement holds: the rate of change of the saliency score does not depend on the size of the tumor. For EWC with BCE loss, the large tumor-all tumor pair and the small tumor-all tumor pair show high correlation. The small-large tumor pair correlation dropped to 0.5. The EWC with DICE+HD loss, despite the chart, does not confirm the saliency score correlation between tumor sizes.

²Statistically not significant: $p = 0.26$

10 Conclusion

This thesis aims to look at the bias and transparency aspects in the medical field on a brain tumor MRI segmentation task. The focus of the work is to see if post-training the model would affect the accuracy of the segmentation based on the tumor size, to see if the saliency score could be used as a metric, and to analyze how post-training behavior changes when utilizing EWC with Fisher Information Matrix calculated with different loss functions.

As a first step, for the clinical segmentation task, we introduced the regulatory landscape for medical devices and discussed the limitations and expectations for AI-enabled devices in the European Union and in the United States, and discussed the difference, that while the EU has strict regulations on AI with a main focus on safety, in the U.S., currently, it is less regulated and focuses on innovation.

The Brain Tumor MRI dataset was briefly described, and the glioma brain tumor images were split into small and large tumor sizes. Performance-friendly augmentation and transformations were defined to prevent over-training the model. A modified Attention U-Net model was presented, preserving the spatial resolution better in the attention gates. We introduced and described the basics of the GradCAM-based saliency map. We introduced DICE and Hausdorff metrics, as well as a saliency map-based score, and presented a loss function using the differentiable version of DICE and Hausdorff distance.

A hypothetical scenario was created where an already deployed model would be post-trained with new images. As it was presented from regulation perspective, it is a challenging task. We examined three different post-training variations: Elastic Weight Consolidation with the BCE loss, Elastic Weight Consolidation with DICE and Hausdorff distance loss, and no Elastic Weight Consolidation.

Our results show that small tumors with all EWC variants were less affected than large tumors. The best EWC option is when the BCE loss function is used to calculate the Fisher Matrix. During the post-training, the accuracy on the initial validation set declined at the lowest rate with the BCE loss function. We also showed that the saliency score could be used as a metric, as according to Cohen's Kappa statistic, it shows agreement with the Hausdorff distance and has a relation to the target area.

References

- [1] Asia-Pacific Journal of Ophthalmology Authors. Recent advances in ophthalmic imaging and artificial intelligence [placeholder title]. *Asia-Pacific Journal of Ophthalmology*, 13(x):xx–xx, 2024.
- [2] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *arXiv preprint arXiv:2211.14830*, 2022.
- [3] Yusuf Brima and Marcellin Atemkeng. Saliency-driven explainable deep learning in medical imaging: bridging visual explainability and statistical quantitative analysis. *BioData Mining*, 17(1):18, 2024.
- [4] Jim Cassidy, Donald Bissett, Roy A. J. Spence, and Miranda Payne. *Oxford Handbook of Oncology*. Oxford University Press, Oxford, United Kingdom, 4th edition, 2015.
- [5] Jun Cheng. Brain tumor dataset, 2022. V1; Data includes 3064 slices of meningioma, glioma, and pituitary tumors.
- [6] European Commission. Notified bodies: The role of third-party conformity assessment, 2024. Accessed: 2026-05-24.
- [7] European Union. Regulation (eu) 2017/745 of the european parliament and of the council of 5 april 2017 on medical devices, 2017. Accessed: 2026-05-24.
- [8] European Union. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act), 2024. Entered into force: August 2024.
- [9] Koustav Guha and Soumyajit Chakraborty. Study of commonly used image interpolation techniques for single frame super resolution methods. *Journal of Emerging Technologies and Innovative Research (JETIR)*, 7(9):319–323, 2020.
- [10] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, 2011.
- [11] Davood Karimi and Septimiu E. Salcudean. Reducing the Hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Transactions on Medical Imaging*, 39(2):499–513, 2020.

-
- [12] King Saud University Faculty. Fisher information and cramér-rao bound. Lecture Notes, Department of Statistics and Operations Research, 2026. Accessed: 2026-05-29.
 - [13] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences (PNAS)*, 114(13):3521–3526, 2017.
 - [14] YuXin Li, Felix Dangel, Derek Tam, and Colin Raffel. Fishers for free? approximating the fisher information matrix by recycling the squared gradient accumulator. *arXiv preprint arXiv:2507.18807*, 2025.
 - [15] Bruce G. Lindsay and Weixin Yao. Fisher information matrix: A tool for dimension reduction, projection pursuit, independent component analysis, and more. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 40(4):712–730, 2012.
 - [16] Tambiama Madiaga. Regulating facial recognition in the eu. Technical Report PE 698.792, European Parliamentary Research Service (EPRS), September 2021. Accessed: 2026-05-24.
 - [17] Mary L. McHugh. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(2):143–148, 2012.
 - [18] Dominik Müller, Iñaki Soto-Rey, and Frank Kramer. Towards a guideline for evaluation metrics in medical image segmentation. *arXiv preprint arXiv:2202.05273*, 2022.
 - [19] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.
 - [20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer International Publishing, 2015.
 - [21] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

- [22] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15(1):29, Aug 2015.
- [23] The White House. National policy framework for artificial intelligence: Legislative recommendations. Technical report, Executive Office of the President, March 2026. Accessed: 2026-05-24.
- [24] U.S. Food and Drug Administration. *Marketing Submission Recommendations for a Predetermined Change Control Plan (PCCP) for Artificial Intelligence/Machine Learning (AI/ML)-Enabled Device Software Functions*. Center for Devices and Radiological Health (CDRH), August 2025. Draft Guidance for Industry and Food and Drug Administration Staff.
- [25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS 2017)*, pages 5998–6008, 2017.
- [26] Maximilian Weimar. Fisher information flow in neural networks. Diploma thesis, Technische Universität Wien, 2022. Organisational Unit: Institut für Theoretische Physik.
- [27] World Health Organization. Cancer: Key facts, 2024. Accessed: 2026-05-24.
- [28] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *arXiv*, abs/1512.04150, 2015.

List of Figures

1	Size histogram	14
2	U-Net architecture	16
3	Attention U-Net architecture	17
4	U-Net attention gate	18
5	The modified attention gate with up-scaling	18
6	DICE scores: ground truth and two images with identical DICE scores . . .	21
7	Distance transform	23
8	Base training validation scores	24
9	Base training loss and λ -value	25
10	Negative Logarithm of DICE score vs DICE loss	29
11	Post-training: initial validation set score shift	30
12	Post-training: Grad-CAM map on the initial validation set	30
13	Learning rate during post-training	31
14	Post-training: DICE and Hausdorff distance of the training set	31
15	Validation results when Fisher Matrix was generated with BCE loss function	32
16	Validation results when Fisher Matrix was generated with $-\log \text{DICE} + \text{HD}$ loss function	33
17	Validation results without EWC	34

List of Tables

1	Improvement / Degradation per tumor size	34
2	Cohen's Kappa for pairwise metrics	35
3	Spearman correlation of the Saliency points	36

List of Algorithms

1	Training	24
2	Fisher matrix computation	28

Declaration

I, the undersigned *Gábor Kerényi*, declare that in preparing my thesis work, I used AI tools(s) as follows.

Subtask	AI tool used	Place of use	Notes (Optional)
Memory leak debugging	Gemini 3.5 Flash	Source code	–
Converting figure to L ^A T _E X	Gemini 3.5 Flash	Figure 4	Converting into TikZ figure
L ^A T _E X commands	Gemini 3.5 Flash	Entire paper	–
Code generation	Gemini 3.5 Flash	Figure 6	Only the draft was generated
Spell checker	Grammarly	Entire paper	Only grammar check, not style
BibTeX entries	Gemini 3.5 Flash	References	Create BibTeX entries from source

Other than the tasks listed in the itemized list, I did not use any AI tool for any other task when preparing my thesis work.

I declare that I have familiarized myself with the content of Rector’s Instruction No. 4/2025 (28 Oct.)³ concerning the use of AI, and that, when preparing my thesis work, I complied with the provisions of the Rector’s Instruction.

Solymár, 1 June 2026



signature

³https://www.elte.hu/dstore/document/208867/4-2025_%28X.28.%29_rektori_utas%C3%ADt%C3%A1s_MI_ai.pdf