BRNO UNIVERSITY OF TECHNOLOGY

Faculty of Electrical Engineering
and Communication

MASTER'S THESIS

Brno, 2021 Bc. Dominik Sasko

BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
FAKULTA ELEKTROTECHNIKY
A KOMUNIKAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF TELECOMMUNICATIONS
ÚSTAV TELEKOMUNIKACÍ

SEGMENTATION OF MULTIPLE SCLEROSIS LESIONS USING DEEP NEURAL NETWORKS
SEGMENTACE LÉZÍ ROZTROUŠENÉ SKLERÓZY POMOCÍ HLUBOKÝCH NEURONOVÝCH SÍTÍ

MASTER'S THESIS
DIPLOMOVÁ PRÁCE

AUTHOR
AUTOR PRÁCE
Bc. Dominik Sasko

SUPERVISOR
VEDOUCÍ PRÁCE
Ing. Martin Kolařík

BRNO 2021

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

Master's Thesis

Master's study program **Communications and Informatics**

Department of Telecommunications

Student: Bc. Dominik Sasko ***ID:*** 187518

Year of study: 2 ***Academic year:*** 2020/21

TITLE OF THESIS:

Segmentation of multiple sclerosis lesions using deep neural networks

INSTRUCTION:

Study current image segmentation methods that use deep learning. As a part of the master thesis, write a review of these methods and their properties with a focus on their use for the segmentation of multiple sclerosis lesions. In the practical part of the thesis, process the raw data provided by your supervisor. Nextly, select the best method from the examined segmentation procedures and apply it to the processed data. Justify the selection of the segmentation method and test the method on the provided and processed data. Optimize the selected method for the tested dataset. Present the test results appropriately.

RECOMMENDED LITERATURE:

[1] RONNEBERGER, Olaf; FISCHER, Philipp; BROX, Thomas. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015. p. 234-241.

[2] BROSCH, Tom, et al. "Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation." IEEE transactions on medical imaging 35.5 (2016): 1229-1239.

Date of project specification: 1.2.2021 ***Deadline for submission:*** 24.5.2021

Supervisor: Ing. Martin Kolařík

prof. Ing. Jiří Mišurec, CSc.
Chair of study program board

ABSTRACT

This master thesis focused on automatic segmentation of Multiple Sclerosis (MS) lesions on MRI images. We tested the latest methods of segmentation using Deep Neural Networks and compared the approaches of weight initialization by transfer learning and self-supervised learning. The automatic segmentation of MS lesions is a very challenging task, primarily due to the high imbalance of the dataset (brain scans usually contain only a small amount of damaged tissue). Another challenge is a manual annotation of these lesions, as two different doctors can mark other parts of the brain as damaged and the Dice Coefficient of these annotations is approximately 0.86, which further underlines the complexity of this task. The possibility of simplifying the annotation process by automatization could improve the lesion load determination and might lead to better diagnostic of each individual patient. Our goal was to propose two techniques that use transfer learning to pre-train weights to later improve the performance of existing segmentation models. The theoretical part describes the division of artificial intelligence, machine learning and deep neural networks and their use in image segmentation. Afterwards, the work provides a description of Multiple Sclerosis, its types, symptoms, diagnosis and treatment. The practical part begins with data preprocessing. Firstly, brain scans were adjusted to the same resolution with the same voxel size. This was needed due to the usage of three different datasets, in which the scans had been created by devices from different manufacturers. One dataset also included the skull, therefore it was necessary to remove it by an FSL tool, leaving only the patient's brain in the scan. The preprocessed data were 3D scans (FLAIR, T1 and T2 modalities), which were cut into individual 2D slices and used as an input for the neural network with encoder-decoder architecture. The whole dataset consisted a total of 6,720 slices with a resolution of 192 x 192 pixels for training (after removing slices where the mask was empty). Loss function was Combo loss (combination of Dice Loss with modified Cross-Entropy). The first technique was to use the pre-trained weights from the ImageNet dataset on encoder in U-Net network, with and without locked encoder weights, respectively, and compare the results with random weight initialization. In this case, we used only the FLAIR modality. Transfer learning has proven to increase the metrics from approximately 0.4 to 0.6. The difference between encoder with and without locked weights was about 0.02. The second proposed technique was to use a self-supervised context encoder with Generative Adversarial Networks (GAN) to pre-train the weights. This network used all three modalities also with the empty slices (23,040 slices in total). The purpose of GAN was to recreate the brain image, which was covered by a checkerboard. Weights learned during this training were later loaded for the encoder to apply to our segmentation problem. The following experiment did not show any improvement, with a DSC value of 0.29 and 0.09, with and without a locked encoder, respectively. Such a decrease in performance might have been caused by the use of weights pre-trained on two distant problems (segmentation and self-supervised context encoder) or by difficulty of the task considering the hugely unbalanced dataset.

KEYWORDS

Context Encoder, Data Augmentation, Deep Learning, Generative Adversarial Networks, Image Segmentation, Medical Images Processing, Multiple Sclerosis, Self-Supervised Learning, Tensorflow, Transfer Learning.

ABSTRAKT

Hlavným zámerom tejto diplomovej práce bola automatická segmentácia lézií sklerózy multiplex na snímkoch MRI. V rámci práce boli otestované najnovšie metódy segmentácie s využitím hlbokých neurónových sietí a porovnané prístupy inicializácie váh sietí pomocou preneseného učenia (transfer learning) a samoriadeného učenia (self-supervised learning). Samotný problém automatickej segmentácie lézií sklerózy multiplex je veľmi náročný, a to primárne kvôli vysokej nevyváženosti datasetu (skeny mozgov zvyčajne obsahujú len malé množstvo poškodeného tkaniva). Ďalšou výzvou je manuálna anotácia týchto lézií, nakoľko dvaja rozdielni doktori môžu označiť iné časti mozgu ako poškodené a hodnota Dice Coefficient týchto anotácií je približne 0,86. Možnosť zjednodušenia procesu anotovania lézií automatizáciou by mohlo zlepšiť výpočet množstva lézií, čo by mohlo viesť k zlepšeniu diagnostiky individuálnych pacientov. Našim cieľom bolo navrhnutie dvoch techník využívajúcich transfer learning na predtrénovanie váh, ktoré by neskôr mohli zlepšiť výsledky terajších segmentačných modelov. Teoretická časť opisuje rozdelenie umelej inteligencie, strojového učenia a hlbokých neurónových sietí a ich využitie pri segmentácii obrazu. Následne je popísaná skleróza multiplex, jej typy, symptómy, diagnostika a liečba. Praktická časť začína predspracovaním dát. Najprv boli skeny mozgu upravené na rovnaké rozlíšenie s rovnakou veľkosťou voxelu. Dôvodom tejto úpravy bolo využitie troch odlišných datasetov, v ktorých boli skeny vytvárané rozličnými prístrojmi od rôznych výrobcov. Jeden dataset taktiež obsahoval lebku, a tak bolo nutné jej odstránenie pomocou nástroju FSL pre ponechanie samotného mozgu pacienta. Využívali sme 3D skeny (FLAIR, T1 a T2 modality), ktoré boli postupne rozdelené na individuálne 2D rezy a použité na vstup neurónovej siete s enkodér-dekodér architektúrou. Dataset na trénovanie obsahoval 6720 rezov s rozlíšením 192 x 192 pixelov (po odstránení rezov, ktorých maska neobsahovala žiadnu hodnotu). Využitá loss funkcia bola Combo loss (kombinácia Dice Loss s upravenou Cross-Entropy). Prvá metóda sa zameriavala na využitie predtrénovaných váh z ImageNet datasetu na enkodér U-Net architektúry so zamknutými váhami enkodéra, resp. bez zamknutia a následného porovnania s náhodnou inicializáciou váh. V tomto prípade sme použili len FLAIR modalitu. Transfer learning dokázalo zvýšiť sledovanú metriku z hodnoty približne 0,4 na 0,6. Rozdiel medzi zamknutými a nezamknutými váhami enkodéru sa pohyboval okolo 0,02. Druhá navrhnutá technika používala self-supervised kontext enkodér s Generative Adversarial Networks (GAN) na predtrénovanie váh. Táto sieť využívala všetky tri spomenuté modality aj s prázdnymi rezmi masiek (spolu 23040 obrázkov). Úlohou GAN siete bolo dotvoriť sken mozgu, ktorý bol prekrytý čiernou maskou v tvare šachovnice. Takto naučené váhy boli následne načítané do enkodéru na aplikáciu na náš segmentačný problém. Tento experiment nevykazoval lepšie výsledky, s hodnotou DSC 0,29 a 0,09 (nezamknuté a zamknuté váhy enkodéru). Prudké zníženie metriky mohlo byť spôsobené použitím predtrénovaných váh na vzdialených problémoch (segmentácia a self-supervised kontext enkodér), ako aj zložitosť úlohy kvôli nevyváženému datasetu.

KĽÚČOVÉ SLOVÁ

Augmentácia dát, hlboké učenie, konfrontačná generatívna sieť, kontext enkodér, prenesené učenie, samoriadené učenie, segmentácia obrazu, skleróza multiplex, spracovanie medicínskych obrázkov, Tensorflow.

SASKO, Dominik. *Segmentation of Multiple Sclerosis Lesions Using Deep Neural Networks*. Brno, 2021, 62 p. Master's Thesis. Brno University of Technology, Faculty of Electrical Engineering and Communication, Department of Telecommunications. Advised by Ing. Martin Kolařík

Typeset by the `thesis` package, version 4.00; `http://latex.feec.vutbr.cz`

ROZŠÍRENÝ ABSTRAKT

Cieľom tejto diplomovej práce bolo navrhnúť metódu využívajúcu hlboké neurónové siete, ktorá by zlepšila vyhľadávanie lézií sklerózy multiplex v MRI skenoch mozgov pacientov. Problém s anotáciou dostatočného množstva dát na trénovanie neurónovej siete je všeobecne známy. V oblasti medicíny je táto výzva o niečo zložitejšia, nakoľko je nutné aby túto prácu vykonali skúsení profesionáli, a nie len široká verejnosť, ako to môže byť napríklad pri anotácií výskytu zvierat na obrázku. Hlavný problém ale nastáva pri presnosti a časovej náročnosti tohto procesu. Dvaja rôzni doktori môžu označiť iné časti mozgu ako poškodené, a s odstupom času by zmenili svoje rozhodnutie. Ako riešenie je možno použiť rôzne techniky hlbokých neurónových sietí, ktoré sa vedia naučiť segmentovať lézie aj z mála označených dát. V tejto práci prezentujeme dve metódy využívajúce transfer learning techniku na možné vylepšenie výkonu súčasných modelov hlbokého učenia.

Teoretická časť diplomovej práce začína popisom umelej inteligencie, jej delenia a aplikácie na obrazové dáta. Ďalej popisuje segmentáciu obrazu a jednotlivé techniky ktoré ju implementujú, ako Generative Adversarial Networks (GAN), enkodér-dekodér architektúry alebo Attention-Based modely. Taktiež zahŕňa augmentáciu dát a rôzne metriky vyskytujúce sa v oblasti dátovej vedy. Pri segmentácii lézií bolo nutné využiť Dice Similarity Coefficient (DSC), nakoľko počítanie pixelovej presnosti by pri nevyváženom datasete vykazovalo dobré výsledky aj pri nepresných predikciách (ak lézie obsahujú 1 % celého mozgu a ich predikcia bude 0 %, pixelová presnosť by v tomto prípade dosiahla 99 %). Koniec teoretickej časti je venovaný skleróze multiplex – čo toto ochorenia spôsobuje, kedy sa vyskytuje a aké sú možnosti diagnostiky a liečby.

Na začiatku praktickej časti je ukázané predspracovanie dát. Boli vybrané tri datasety obsahujúce MRI skeny mozgov a k nim priradených anotovaných masiek lézií. Využité boli modality FLAIR, T1 a T2. Pri prvotnej analýze sa vyskytli 3 problémy s konzistenciou dát. Jeden z datasetov mal dáta v RAW formáte, čo bolo nutné prekonvertovať do NifTI formátu na následné načítanie v jazyku Python pomocou NiBabel balíčku. Ďalšia nutnosť bola zmena rozlíšenia skenov. Nekonzistencie sa ukázali nielen medzi datasetmi, ale aj medzi súbormi v jednom datasete. Toto bolo spôsobené nasnímaním snímkov pomocou zariadení od iných výrobcov. Takisto museli mať skeny izotropné voxeli – vo všetkých troch dimenziách rovnakú veľkosť. Jeden z datasetov obsahoval lebku, a tak bolo nutné jej odstránenie pomocou nástroju FSL, pre ponechanie samotného mozgu pacienta. Tento nástroj vytvoril binárnu masku, ktorá bola následne aplikovaná na sken, aby sa lebka odstránila. Na konci predspracovania dát sme mali 3D skeny s dĺžkou jednotlivých strán 192 pixelov. Tieto skeny boli ďalej načítané do Numpy polí ako 2D rezy po z-ose, ktoré mohli

vstupovať do zvolených modelov neurónových sietí. Nakoniec bolo vybraných 23040 obrázkov (3 modality * 40 pacientov * 192 rezov), z ktorých bolo 87 % použitých na trénovanie a zvyšné boli využité na testovanie a porovnávanie predikcií. Na dáta bola taktiež aplikovaná augmentácia, ktorá obsahovala vertikálne otočenie, náhodnú rotáciu, skreslenie súradnicovej mriežky a kombináciu "shift-scale-rotate".

Prvý zvolený prístup bol transfer learning, ktorý využíval váhy predtrénované segmentáciou na ImageNet datasete. Tie boli ďalej načítané do enkodéru U-Net architektúry a následne dotrénované na našom datasete. V tomto prípade sme používali iba FLAIR modalitu, keďže trénovanie so všetkými troma modalitami vykazovalo horšie výsledky. Taktiež boli niektoré rezy vyfiltrované. Dôvodom bol malý počet lézií na mozgu pacientov, čo spôsobilo nulové anotácie na maske. Po odstránení týchto rezov ostalo 6720 obrázkov v trénovacej množine. Testovacie snímky ostali nevyfiltrované, aby sa simulovalo nasadenie na reálnom prípade. Vyskúšali sme tri možnosti načítania váh enkodéru – náhodná inicializácia a predtrénované váhy enkodéru s, resp. bez možnosti trénovania jeho váh. Transfer learning technika dosiahla lepšie výsledky (hotnota DSC približne 0,6) v porovnaní s náhodnou inicializáciou váh (hotnota DSC približne 0,4). Rozdiel medzi zamknutými a nezamknutými váhami enkodéru sa pohyboval okolo 0,02 (uzamknutie váh sa ukázalo o niečo efektívnejšie).

Ako druhý experiment bola použitá technika self-supervised learning, konkrétne context encoder pomocou GAN architektúry. Úlohou tohto modelu bolo dokreslenie skenu mozgu, ktorý bol prekrytý čiernou maskou v tvare šachovnice. Pri predtrénovaní sa využíval pôvodný počet obrázkov so všetkými troma modalitami. Po natrénovaní sme využili generátor s natrénovanými váhami na segmentáciu lézií. Maximálna hodnota DSC na testovacích dátach vystúpila na úroveň 0,29 pri neuzamknutom enkodéri, a 0,09 pri uzamknutom. V porovnaní s prvým experimentom ide o veľké zníženie sledovanej metriky. Príčinou by mohlo byť viac faktorov. Predtrénovanie váh na veľmi odlišnom probléme, ktoré nenávratne nasmeruje nasledujúce trénovanie, rozličné porovnávané architektúry (U-Net model načítaný knižnicou Segmentation Models a generátor z GAN, ktorý obsahoval batch normalization), alebo veľmi nevyvážený dataset.

DECLARATION

I declare that I have written the Master's Thesis titled "Segmentation of Multiple Sclerosis Lesions Using Deep Neural Networks" independently, under the guidance of the advisor and using exclusively the technical references and other sources of information cited in the thesis and listed in the comprehensive bibliography at the end of the thesis.

As the author I furthermore declare that, with respect to the creation of this Master's Thesis, I have not infringed any copyright or violated anyone's personal and/or ownership rights. In this context, I am fully aware of the consequences of breaking Regulation § 11 of the Copyright Act No. 121/2000 Coll. of the Czech Republic, as amended, and of any breach of rights related to intellectual property or introduced within amendments to relevant Acts such as the Intellectual Property Act or the Criminal Code, Act No. 40/2009 Coll., Section 2, Head VI, Part 4.

Brno . . . . . .

. . . . . .

author's signature

POĎAKOVANIE

V prvom rade by som veľmi rád poďakoval vedúcemu mojej diplomovej práce, pánovi inžinierovi Martinovi Kolaříkovi. Bolo mi veľkou oporou mať za sebou človeka, ktorý mi vždy ochotne pomohol a posúval ma vpred svojimi radami.
Taktiež ďakujem Vysokému učenímu technickému v Brně a predovšetkým ľuďom v organizačnej štruktúre, za poskytovanie vecnej a finančnej pomoci popri štúdiu, ako aj za možnosť absolvovať mnohé zahraničné stáže, ktoré ma obohatili nielen v profesionálnej, ale aj osobnej stránke.
V poslednom rade patrí najväčšia vďaka mojim rodičom. Za celé roky môjho štúdia ma motivovali a podporovali. Bez nich by bola táto cesta omnoho ťažšia.

Contents

List of Figures

Introduction

A current trend in computer science is to use all the data we can gather to improve people's lives. The data are being collected everywhere, *e.g.*, when using social media, buying groceries with a credit card, or just visiting a doctor for a routine checkup. However, what do data scientists use these data for? The main reason is trying to find patterns in them.
Data from the medical environment comes from doctor notes, lab results, and medical images every day. Therefore, it is essential to evaluate them without spending too much time manually examining every record. That is where data science comes to help to increase the efficiency and accuracy of diagnostics [1].
A specific example of such usage is shown in this thesis, where data science, more precisely *Deep Neural Networks* in computer vision, are used to detect damaged tissue on brain scans caused by Multiple Sclerosis (MS) better. Computer vision has many techniques, such as object detection or image reconstruction. The image segmentation technique will be used for our challenge since it classifies each pixel of an image and assigns it to the corresponding class. This can be applied to see the boundaries of a tumor, or in the case of MS, to see a lesion in the brain's gray or white matter.
Early approaches for image segmentation include techniques as thresholding, histogram-based bundling, or k-means clustering [2]. As a result of the growth of popularity in deep learning, neural networks started a new generation of image segmentation models, which remarkably improved the performance.
Nowadays, finding the lesions in brain scans is a formidable challenge even for skilled professionals [3]. This work aims to simplify the detection of lesions in brain scans and make it more precise. Even though we still depend on doctors, who must annotate the lesions, we want to train a powerful network for segmentation by augmenting the input data. The proposed solution to address this issue is a combination of self-supervised learning with transfer learning. The first model trained was a U-Net [4] model with weights pre-trained on the ImageNet dataset, which does not contain medical images. The second was GAN [5] model pre-trained with self-supervised learning on brain scans containing MS lesions.

This thesis is structured into three chapters:
The first chapter introduces the theoretical part of the problem. Mainly, it is devoted to Deep Learning and shows that it is a subset of Machine Learning and Artificial Intelligence. Furthermore, it describes the Image Segmentation technique and its usage in more depth and explains the U-Net and GAN architectures used in this work. In addition, it is explained which metrics are relevant to be used to evaluate each of the approaches, since primary ones, such as accuracy and precision, are not

applicable for this type of challenge.
The chapter also contains an introduction to Multiple Sclerosis – its basic description, symptoms, diagnosis, and related works in this domain.

The second chapter contains a description of data gathering – three chosen datasets that were pre-processed to be later used for training. Then follows an explanation of how the data were loaded and passed as an input to neural networks. Moreover, the chapter demonstrates a practical application of the chosen algorithms, where each model was trained with different settings of weights. The final part explains how the predictions were processed in order to show graphical results.

The last chapter summarizes the results for both techniques used during this research. Moreover, each of the approaches used is compared by its metrics and graphical predictions.

1 Theoretical Part

This chapter starts by explaining the basics of *Deep Learning* (DL). It continues with a specific application of DL – Image Segmentation, which is used in the practical part of this work, together with U-Net architecture of the neural network and other techniques applicable in image processing. It continues with an introduction of Multiple Sclerosis and, in the end, summarizes related works in this area.

1.1 Deep Learning

1.1.1 Introduction to Deep Learning

Deep Learning is a subset of *Machine Learning* (ML), which is an application of *Artificial Intelligence* (AI). The relationship of all three fields is shown in Figure 1.1.

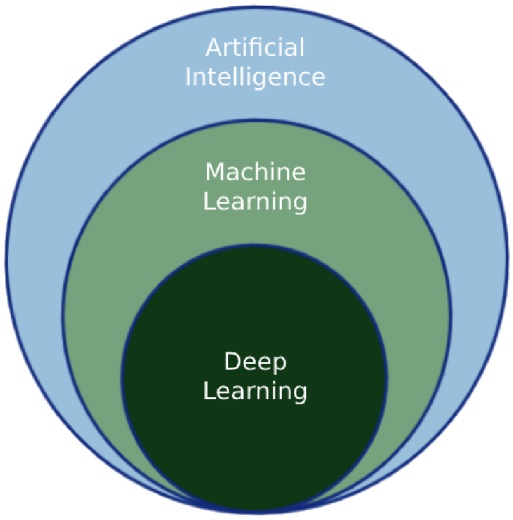

Fig. 1.1: Relationship between AI, ML and DL.

AI is supposed to simulate how a human would respond to stimulation based on human judgements and intentions.
Machine Learning enhances the AI system with the ability to learn and improve itself in order to obtain better results and predictions. These algorithms search for patterns in the data by the usage of statistical techniques to later apply actions on them.
Deep Learning is an implementation of ML, which simulates the human brain with *Artificial Neural Networks* (ANN). These networks are made of neurons (or perceptrons) connected to create a layered structure. Because of more layers organized in a row behind each other, the name *Deep Neural Networks* (DNN) is being used.

These layers can progressively extract more specific features from raw input, and therefore find patterns which ML would not be able to see.
DNNs uses many techniques, such as clustering, classification, regression or segmentation, which can be used for self-driving cars, weather prediction or medical diagnosis purposes [6, 7].

1.1.2 Image Segmentation

Detecting a dog in a picture is not a problem for human, and with modern computer vision techniques, not even for a computer. On the contrary, detecting a lesion in a 3D scan of a brain is challenging even for doctors. This is supported by the fact, that two individual professionals can mark different white spots as lesions, and additionally, the same doctor can annotate the same sample differently after some period of time [3]. This is when an image segmentation can be used. It uses various techniques to divide a picture (divide a space in 3D) into smaller parts with in common. These parts are called segments.

Image segmentation uses two techniques to segment an image: *semantic segmentation* and its successor *instance segmentation*. Semantic segmentation groups pixels into different classes, in the case in Figure 1.2 (left), the background class is gray, and the class containing all the lesions is yellow. Instance segmentation also divides pixels into corresponding classes, but each object of the same class is separated, which is shown by different colors of the same class for every individual lesion.

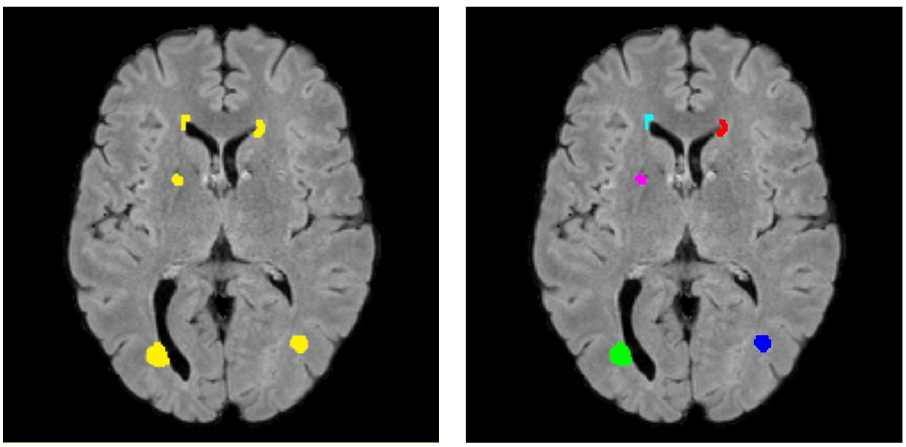

Fig. 1.2: Semantic segmentation (left) and instance segmentation (right).

To segment an image, various algorithms and approaches have been developed. In [2] Chowdhary *et al.* proposed a division by time usage of those algorithms. The oldest were developed in the late 1990s and a majority of them are not used

anymore. These include Multiresolution Method or Geodesic Minimal Path. Next group are techniques, which are still being used, although their usage is decreasing due to replacement by modern approaches. Thresholding, Edge Detection or Markov Random Field Approach belong to this group. Lastly, recent segmentation techniques were developed to deal with inaccuracies in the segmentation of medical images. They include the usage of Artificial Neural Networks [2].
Among the many techniques for segmentation, four are described in the following sections. Namely, U-Net, Generative Adversarial Networks (GANs), Transfer Learning and Attention-Based Models. First three mentioned techniques are used in the practical part of this work (Chapter 2).

1.1.2.1 U-Net Architecture

One of the specific implementations of image segmentation is the usage of U-Net. It is a fully convolutional neural network created by Ronnenberger *et al.* [4] in 2015, for the purpose of biomedical images segmentation. The network can be seen in Figure 1.3, where the left part is a *contracting path* (also called encoder) and on the right is an *expanding path* (decoder).

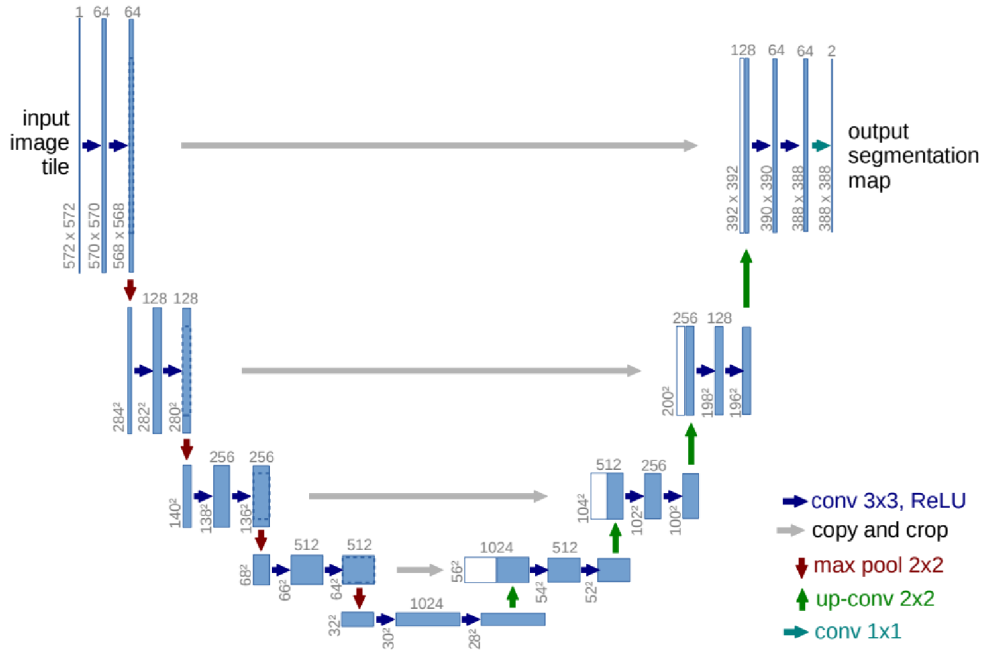

Fig. 1.3: U-Net architecture by Ronnenberger *et al.* [4]

The encoder path of U-Net consists of multiple convolutional layers, after which are max pooling layers. The decoder uses up-convolutional layers instead of pooling

layers, in order to reconstruct the compressed image. These layers and their purpose are explained below.

Convolutional Layers

Convolutional layers are used to capture a specific context of an image. As an input, they take an image, which is being processed by a kernel, also called a feature extractor. This kernel is simply a filter smaller than the input image (usually 3×3 pixels) [8], which is being applied across the input. This convolution process does linear operations that decrease the resolution and eventually create a *feature map*. The result of convolution operation is then passed through ReLU – a nonlinear activation function to map the output to a specific value.
After each downsampling step, the number of feature masks doubles. On the opposite side of U-Net, the up-convolution will halve the number of feature masks.

Many convolutional networks are used for classification tasks, which means they intend is to find an object in the picture. However, the purpose in the medical field is often to localize where the object is – label each pixel. The expanding path enables a precise localization of an object by upsampling the image. It is combined with the spatial information from the contracting path by skip connections. At the end of U-Net, a 1×1 convolution kernel is used to map a vector of 64 feature masks to a desired number of classes [9].

Max Pooling Layers

After the convolution layer, the feature masks are created. Those masks are sensitive to the location of the features. To address this issue, a reduction of dimensionality by a pooling layer is applied.
The most popular technique is Max Pooling. It takes a patch from an input image, searches through it to find the maximum value, which is then copied to an output array of numbers. In practice, filter with 2×2 dimension is often used to go through whole input image [9].

Despite the relatively older age of this architecture (given that there has been significant progress in deep learning for computer vision in past years), U-Net is still being used in current research with various modifications, *e.g.* 3D U-Net for spatial data [10] or an attention-aware U-Net [11].

1.1.2.2 Generative Adversarial Networks

Generative Adversarial Networks offer a different approach for segmenting image data. In [5], authors propose a generative model with an adversarial process. Their model consists of two neural networks, which are simultaneously trained and compete against each other (Figure 1.4). One network is a generator (G), and the second is a discriminator (D). Their purpose and functioning are described below. The main

focus of this technique is to learn how to generate new content, which resembles an actual image.

Generator

The generator uses a random vector with a fixed length from vector space as an input seed to create a new image. This vector space, referred to as *Latent Space* in Figure 1.4, contains latent variables (variables that cannot be observed directly). The goal is to set the weights of G in a way that will transform an input vector to a state that represents a targeted distribution. With the training process, G tries to maximize the discriminator network's error by creating more realistic images [12, 13].

Discriminator

The critical process happening in the discriminator is a correct classification, if the image which it received as an input is an actual image, or if the generator randomly generated it. Therefore D is trying to minimize its classification error (detect fake images) and maximizing the loss function. After the training process is finished, the discriminator is discarded since only the generator is valuable for later predictions [13].

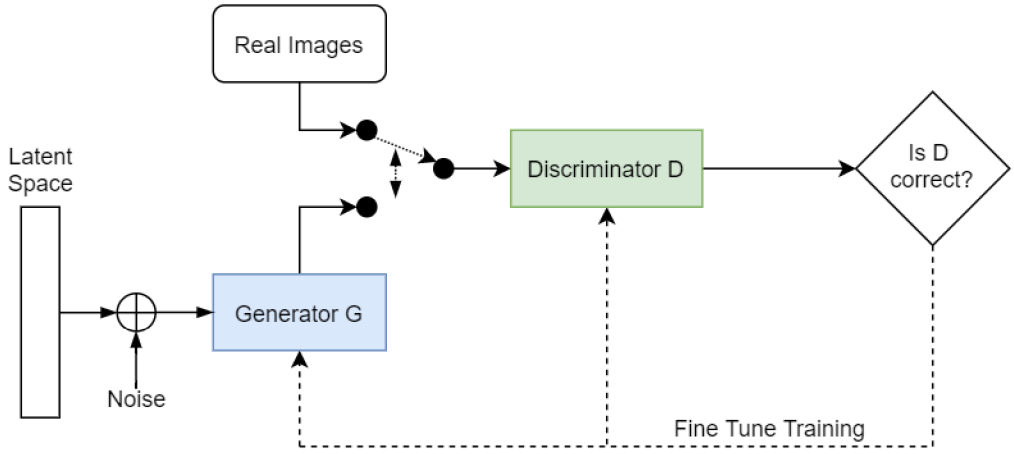

Fig. 1.4: Generative Adversarial Network architecture.

Both of these networks fine-tune their parameters by backpropagation. A unique solution is when D cannot decide if the input image is actual or generated, so the probability for mistake is 50% [5]. That indicates the generator was successfully trained to create images hardly indistinguishable from the real ones.

1.1.2.3 Transfer Learning

Since training of a neural network can be a very time-consuming process, which in some cases might not lead to desirable results, Transfer Learning is used as a shortcut

to address this issue. Instead of creating a model with randomly initialized weights that are later trained, part of the model can be initialized with already pre-trained weights on big datasets, such as *ImageNet Large Scale Visual Recognition Challenge (ILSVRC)* [14]. This dataset is part of an ImageNet project and contains 150,000 photographs. Its function is to detect the object and classify it into one of 1,000 different categories. A pre-trained model like this can be reused as a whole or just use specific layers that may be applied to a different task (*e.g.* image segmentation).

There are a few challenges to this approach. One of them is applying the Transfer Learning technique trained on the ImageNet on medical data due to significant differences in input images. ImageNet contains pictures of real-world objects (cars, animals, *etc.*), yet most of the medical images are scans of patient's bodies. Another challenge is the number of classes. Medical tasks like segmentation of MS lesion need only two classes (damaged/healthy brain tissue); therefore, the model learned to classify into 1,000 categories is redundant [15].

Raghu *et al.* [15] explored the application of Transfer Learning in medical imaging and found slightly better results in performance, despite all the above-stated challenges. The details of this work are explained in section 1.3.4

1.1.2.4 Attention-Based Models

Attention-based models were developed to train the network on suppressing irrelevant regions in an input image while highlighting other features essential for a specific task. It has been largely used for natural language processing tasks as translation or speech recognition but can be applied to better visual identification of objects in image processing as well.
Chen *et al.* [16] used attention for semantic segmentation to improve the estimation of features with different scale or features at various positions. That means the network assigns small weights for big objects and big weights for smaller objects in the picture. By differencing the weight for multi-scale features, the model showed a slight improvement in its performance.

Another use of Attention-Based Models is shown in [17]. Here the authors integrated attention into the U-Net model architecture described in section 1.1.2.1. The contracting path stayed unchanged, whilst the expanding path had *attention gate* integrated with each upsampling step.
This change resulted in increased sensitivity and prediction accuracy, with minimal computational overhead. This approach demonstrated improved tissue and organ identification and localisation for the pancreas segmentation task.

1.1.3 Metrics

Metrics are used to evaluate the results, which were predicted by the neural network. There is not one suitable metric for all the challenges, due to differences in unbalanced datasets or individuality of each problem [18]. Below are described different types of metrics for image segmentation, such as *Pixel Accuracy*, *Jaccard Index*, *Dice Similarity Coefficient* or *Hausdorff Distance*.

1.1.3.1 Pixel Accuracy

Pixel Accuracy is an early technique, which calculates how many pixels were classified correctly [18]. The equation is as follows:

$$Pixel\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1.1}$$

where *TP*, *TN*, *FP*, and *FN* are the true positive, true negative, false positive, and false negative rates, respectively.
In the case of Multiple Sclerosis and segmentation of lesions in the human brain, these metrics would not evaluate the results the way we wanted. The reason is the class imbalance. Given the ratio of brain and lesions which are present in it, if the neural network classifies all the pixels as "not lesion", the Pixel Accuracy might be even 95%, since the lesions would cover only 5% of the brain.

1.1.3.2 Jaccard Index

To address the problem of unbalanced classes, metrics as Jaccard Index or Dice Coefficient were implemented. Jaccard Index, or Intersection-Over-Union (IoU) is defined by

$$Jaccard = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}, \tag{1.2}$$

where A is a class that represents the ground truth, and B represents the predicted class [18, 19].

1.1.3.3 Dice Similarity Coefficient

Dice Similarity Coefficient (DSC) is the most used metric in validating medical volume segmentation [20]. It is calculated as two times the overlapping area divided by a total number of pixels in both images. The equation is as follows:

$$DSC = \frac{2 * |A \cap B|}{|A| + |B|} = \frac{2 * TP}{2 * TP + FP + FN}, \tag{1.3}$$

where A and B represents the same as in equation 1.2 [19, 20].
DSC and Jaccard Index are positively correlated – when one metrics evaluates one

prediction better than another, the second metrics will evaluate it likewise. DSC is always smaller than Jaccard, except at the extrema {0, 1}, where those two metrics are equal. Since they both measure the same aspects and provide the same system ranking, it is unnecessary to use both metrics [20].

1.1.3.4 Hausdorff Distance

Hausdorff Distance (HD) is the maximum of all distances from a point in one set to the closest point in the other set. In other words it represents the maximum nearest neighbor Euclidean distance between contours. It is defined as

$$h(A, B) = \max_{a \in A}\{\min_{b \in B}\{d(a, b)\}\}, \tag{1.4}$$

where A and B represents the sets, a and b are points in these sets, and $d(a, b)$ is any metric between these points.
The advantage is the possibility of quantifying the similarity of two sets without establishing one to one correspondence of points in them. That is helpful since, in many cases, the number of points in two 3D sets is not identical. The drawback is the calculation of distances between all pairs of voxels (considering 3D space). When applied to large images, this technique becomes computationally very intensive; therefore it is restricted to applications for large scale data [21].

1.1.4 Image Augmentation

To obtain satisfactory high metrics, it is necessary to train a model effectively. That can be achieved not only by a well-chosen model but also by using datasets with enough training samples. Unfortunately, labeled datasets in the medical field are often limited – therefore, overfitting might occur during the training process. To address this issue, various image augmentation techniques are applied to input data. Augmented samples then artificially enlarge training images.
There are numerous algorithms for this task. Nalepa *et al.* [22] divide data augmentation, specifically for brain tumour segmentation, into two groups:

1. **Transformation of the original data** – algorithms such as random rotation and cropping, or applying a pixel-level transformation (Gaussian noise, gamma correction, *etc.*). The advantage is a simple implementation, although a drawback can be the production of correlated images.
2. **Generation of artificial data** – these methods use Generative Adversarial Networks described in section 1.1.2.2. They can synthesize realistic examples, but the major disadvantage is a high time complexity [22].

Some of the above-mentioned methods can be combined to increase the strength of augmentation. Authors in [23] developed an algorithm to find an optimal augmentation technique, which can be applied to 3D medical images. Their approach outperformed the state-of-the-art models that used hand-made and simple augmentation methods.

An augmentation technique proposed by Eaton-Rosen *et al.* [24] try to address the problem of an imbalanced dataset for semantic segmentation. This paper introduced a 'mixmatch' technique, which applies a linear combination on training images, as well as a linear combination of training labels with ground-truth labels. Authors were inspired by 'mixup' technique [25], with a change in combining images with the highest foreground amounts with the lowest and doing this process randomly. Results of both methods were similar, although an increase in DSC compared to training without augmentation was achieved.

Augmentation might not be helpful in all cases. If the given diagnose is based on a specific location of irregularity (*e.g.* damaged brain matter), rotation or other change of the input data could discard a piece of important information. For example, cropping of an image might alter with a label or remove it entirely from the training image [26].

1.2 Multiple Sclerosis

Multiple Sclerosis is an autoimmune inflammatory disease. This means that the immune system attacks components of the body as if they are foreign – in this case, it attacks the brain or spinal cord of the nervous system [27].
MS mostly affects people between ages 20 to 50 and is three times more common in women than in men. The researches show, that MS is not contagious or directly inherited, but there are certain factors in the distribution around the world, that might help determine what causes the disease. The evidence is growing, that lack of vitamin D, smoking, and obesity play an important role in MS, and they have been identified as risk factors [28].

1.2.1 Pathophysiology

As mentioned above, the immune system attacks the layer which surrounds and protects the nerves called the *myelin sheath*, a fatty substance that insulates nerve fibers. The result may be multiple areas of scarring (sclerosis), which can eventually lead to the blockage of nerve signals. These scars are also called plaques or lesions and can be visible on brain scans as white spots [27, 29]. Figure 1.2.1 shows, how the signal traveling between neurons becomes disrupted by MS.

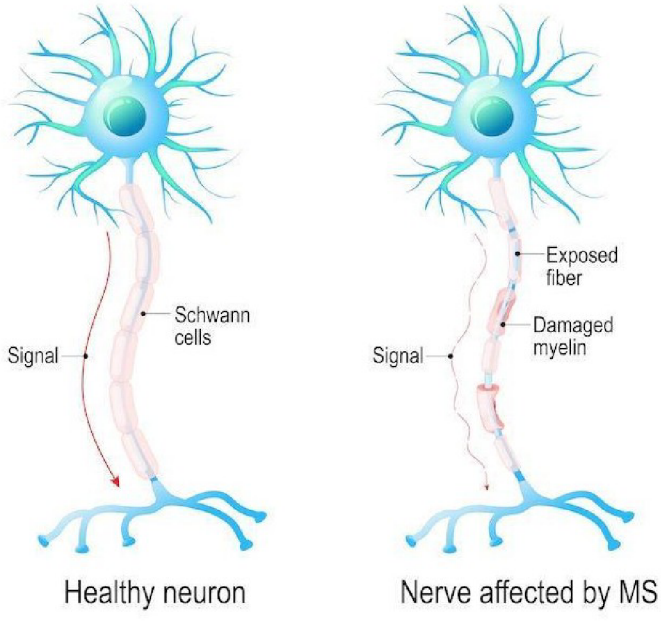

Fig. 1.5: Effects of Multiple Sclerosis on nerve fibers [29].

The cure has not yet been found, but treatments can help speed up the recovery from symptoms. However, if the formation of lesions continues, MS often leads to

physical disability. This process can take up to 20-25 years and occurs in more than 30% of patients [27][30].

1.2.2 Types of MS

Even though it is uncertain how the disease will progress, there are four primary MS disease courses (or types) defined. They concentrate on the level of disability caused over time.

Clinically Isolated Syndrome (CIS)

CIS is the first stage of neurological symptoms caused by inflammation in the central nervous system. People who experience CIS might not develop MS, because the lesions as in Figure 1.6 might not be present. Early treatment has been shown to delay the onset of MS.

Relapsing-Remitting MS (RRMS)

RRMS is the most common stage. It is characterized by attacks of increasing neurological symptoms. These attacks, which are also called relapses, happen months or years apart, and cause an increase level of disability. They are followed by the stage of partial or complete recovery (remissions). During remissions, patient's condition does not worsen. Approximately 85% of people with MS are initially diagnosed with RRMS.

Primary Progressive MS (PPMS)

PPMS is characterized by worsening of the neurological functions from the beginning, without early relapses or remissions. However, short periods of stability can occur. It occurs in approximately 15% of patients with MS.

Secondary Progressive MS (SPMS)

SPMS is very similar to RRMS. It occurs, after RRMS transitions into progressive form. In this stage, the patient's syndromes will not worsen with each relapse, but will be progressively worsening over time [35, 33].

1.2.3 Symptoms

Multiple Sclerosis is unpredictable and varies in severity. No two people have exactly the same symptoms. In some cases, it is a mild illness, but it can lead to permanent disability in others. The reason is the blockage of nerve signals, mentioned in part 1.2.1, which controls muscle coordination, strength, sensation, and vision [28]. The symptoms can occur on various parts of body.

If symptoms affect movement, MS can cause:

- Numbness or weakness in one or more limbs.
- Electric-shock sensations occurring during neck movement.

- Tremors or lack of coordination.

Common are problems with vision, such as:

- Partial or complete loss of vision.
- Double vision or blurry vision.

Multiple Sclerosis can also cause:

- Emotional changes and depression.
- Dizziness and bowel problems.
- Problem with sexual functions [28].

These symptoms can lead to psychological and social problems, due to limitations in everyday life. Although most of the them can be managed effectively with medications or rehabilitation, after correct diagnosis [30].

1.2.4 Diagnosis

Since the symptoms of the disease can be confused with other illnesses, it is difficult to determine if the patient has Multiple Sclerosis. There are various strategies for ruling out other conditions. They include magnetic resonance imaging, spinal fluid analysis and blood tests.

The process of exclusion of other diagnoses may be quick for some, however, it can also take longer periods, when repeated testing is necessary [31].
The patient is given *contrast agents*, which improve the visibility of brain structure during the scan. In this case *Gadolinium*, so the lesions in Figure 1.6 are visible.

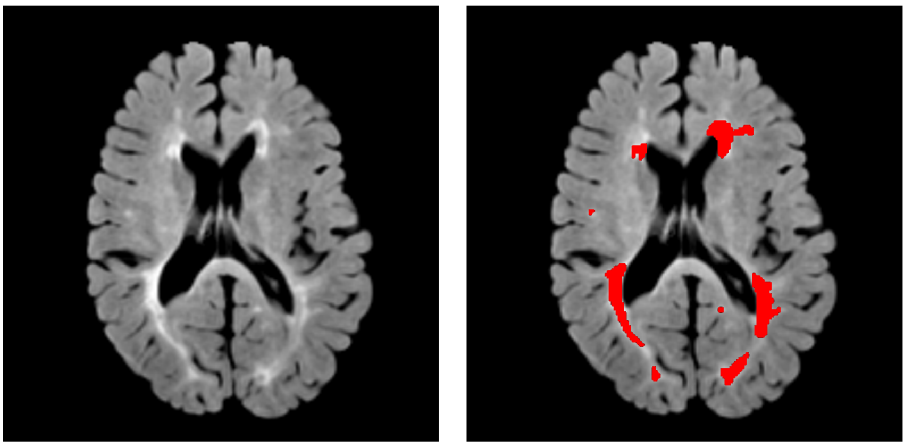

Fig. 1.6: MRI of a patient with Multiple Sclerosis (left), with corresponding annotated white matter lesions (right).

Diagnostic Criteria for MS

1. Evidence of damage in at least two separate areas of the central nervous system (brain, spinal cord and optic nerves).

2. Evidence that the damage occurred at different points in time. A true relapse, which causes the damage, must last at least 24 hours and must happen at least 30 days after the previous one.
3. Rule out other possible diagnoses [31].

The arrival of MRI has revolutionized the diagnosis and monitoring of MS. MRI demonstrates a high rate of abnormal findings used for accurate diagnoses. In a study by Paty *et al.* [32], MRI findings were abnormal in 124 of 133 patients with clinically definite MS.

Similar to MRI, computed tomography (CT) scanning has been used in the diagnosis of MS. CT scans might exclude other neurological diagnoses, but they have a low positive predictive value in the diagnosis of MS – the false-negative rate is high.

It can take few years to accurately diagnose Multiple Sclerosis since the symptoms can worsen slowly. By regular examinations with brain scans, the medical experts can determine if the patient has MS, and can find out which type it is.

1.2.5 Treatment

Once the MS is diagnosed, the patient starts with a lifetime treatment. Although there is no cure currently available, there are multiple options for treatment from the early stage of the disease. The goal is to manage symptoms, control relapses and slow down the progression of Multiple Sclerosis. This can be done by medications, rehabilitation and often patients incorporate alternative medicine as well [34].

1.2.5.1 Medications

A lot of progress has been made in developing new medications to treat MS. Since the cause is inflammation in the central nervous system, medications called *Disease Modifying Drugs* are prescribed to work with the immune system to reduce it, and hence decrease the progression of symptoms.

The treatment of attacks is realized by corticosteroids, which are given after the relapse. This decreases the recovery time but does not prevent future relapses. Many of the patients observed side effects like diabetes, osteoporosis or weight gain. Due to this, corticosteroids should not be used more than 3 times a year [37, 36].

1.2.5.2 Rehabilitation

Another possible treatment is rehabilitation. It is important to maintain an active lifestyle even after being diagnosed with MS. Rehabilitation can address the symptoms which affect mobility – functioning at home or work, personal care or

free-time activities. This can prevent complications with the weakening of muscles and de-conditioning. Physical therapy might include exercises including mobility aids (crutches, wheelchairs or poles), and training which focuses on the pelvic floor, to help with bladder issues [37].
For patients with speech and swallowing problems, a speech-language therapist can help enhance the clarity of communication as well as focus on safe swallowing.

1.2.5.3 Alternative Medicines

Complementary and alternative medicine varies from the usage of supplements and diets to meditation and Tai Chi. Those methods usually take place in combination with MS treatments prescribed by professional. Since there is not enough scientific research on these techniques, is not proven if they can be effective, or in some cases might harm the patient [34]. The most important factor in treating MS is the compliance of the patient and regularity of two previously mentioned treatments.

1.3 Related Works

This part introduces related works for Multiple Sclerosis lesion detection and shows multiple state-of-the-art techniques used for image segmentation in the medical field. Each section describes individual work, therefore to increase the readability, the quote is only at the beginning of a section.

1.3.1 Deep Learning Segmentation of Gadolinium Enhancing Lesions in Multiple Sclerosis

In paper [38] Coronado *et al.* proposed a multi-class three-dimensional convolutional neural network (CNN) to automate the detection and delineation of gadolinium (Gd) enhanced lesions. Authors focused on this type of lesions, due to the fact that not all the lesions which are shown on MRI scan are active. Patients with MS are given gadolinium contrast agents during the MRI scan to see if the size and number of lesions increase. This increase correlates with the patient's occurrence of relapses. Even though the segmentation of Gd enhancement is considered the simplest to implement, automation could make the process faster and more effective by minimizing human error.

The dataset consisted of 1006 scans of patients with relapsing-remitting MS. Preprocessing included – skull stripping, bias field correction, co-registration, intensity normalization and anisotropic diffusion filtering. The authors used 3 models with different inputs, namely *U5* (inputs images: proton density-weighted, T2w, pre- and post-contrast T1w, FLAIR), *U2* (inputs images: pre- and post-contrast T1w) and *U1* (input images: post-contrast T1w). Next, the images were used as an input to 3D U-Net, which took image patches of size $128 \times 128 \times 8$ not to exceed computational memory limits. The scans were divided by ratio 6:2:2 (training, validation and test set), and the loss function was multi-class weighted Dice. The results can be seen in Table 1.1 below. It contains the Dice Similarity Coefficient values for lesions segmentation and brain tissues – grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF), which this model additionally segmented.

Tab. 1.1: DSC values of 3D U-Net model by Coronado *et al.*

Model	**Gd-Enhancing Lesion**	**GM**	**WM**	**CSF**
U5	0.77	0.95	0.94	0.98
U2	0.72	0.80	0.82	0.83
U1	0.72	0.78	0.78	0.80

This work showed good results for automated MS lesions segmentation with the usage of a very large dataset. It also showed that usage of more modalities may be beneficial, compared to our work where it decreased the overall performance.

1.3.2 Brain Tumor Segmentation and Survival Prediction Using 3D Attention U-Net

Authors in [39] developed an attention convolutional neural network, which segments brain tumours. Another part of the work was a prediction of survival rate, but since it was done by machine learning techniques, it will not be described.
The used architecture is a 3D U-net with integrated channel and spatial attention. This 3D attention module was integrated with the decoder blocks and can be seen in Figure 1.7 below.

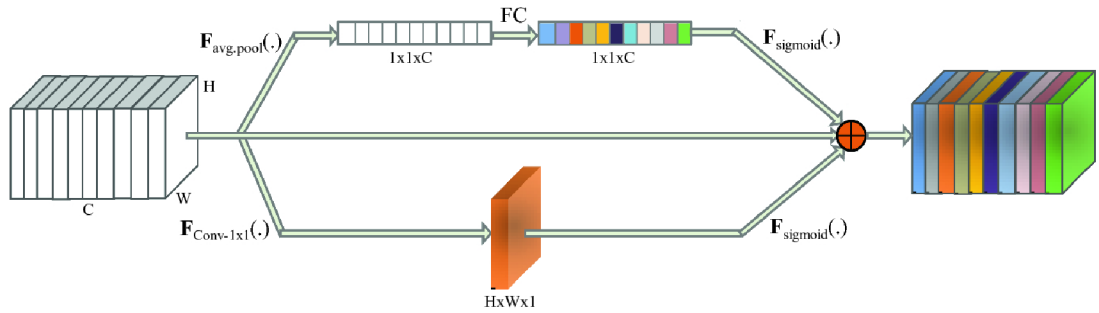

Fig. 1.7: 3D spatial and channel attention module with skip connection [39].

Firstly, to cluster all spatial feature correlations into $H \times W \times 1$ dimension, a $1 \times 1 \times C$ convolution was performed on the 3D scan. In parallel with this process, an average pooling was done and fed to the neural network to obtain the $1 \times 1 \times C$ channel correlation. Authors also integrated skip connection to make the learning more generic, and hence improve the segmentation.

The data used for this work are from the BraTS 2019 dataset, which consists of 626 scans – each scan with 4 modalities (T1, post-contrast T1-weighted, T2-weighted and T2 Fluid Attenuated Inversion Recovery), and voxel size randomly reduced to $128 \times 192 \times 192$ from original $240 \times 240 \times 155$. Data were split into 335, 125 and 166 patients (train, validation and test). Results can be seen in Table 1.2 below.

Tab. 1.2: DSC values of 3D U-Net model with attention by Islam *et al.*

Model	Enhancing Tumor	Whole Tumor	Tumor Core
3D U-Net	0.68	0.89	0.75
3D Attention U-Net	0.70	0.90	0.79

It is visible that a model with integrated attention achieved slightly higher DSC in segmenting different parts of tumors, compared to a model without the attention module, and that U-Net network used in this work might benefit by the attention module as well.

1.3.3 A Dense U-Net Architecture for Multiple Sclerosis Lesion Segmentation

A proposal in [40] contains a modified version of U-Net architecture with dense net. The formerly mentioned network has dense block added before every max-pooling and up-sampling layer. This can be seen in Figure 1.8. By the usage of dense connection, the model strengthens the feature propagation and reduces overfitting.

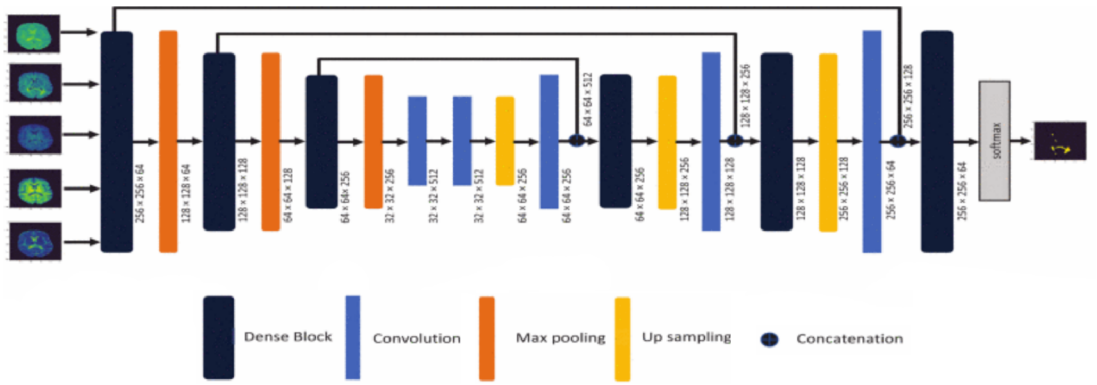

Fig. 1.8: U-Net architecture with dense blocks [40].

There were 5991 scans in total, divided at a ratio of 8:2 for training and testing data. Data were taken from MICCAI 2016 dataset, with MPRAGE, FLAIR, T1-w, T1-w gadolinium-enhanced and T2-w/DP contrast-enhanced modalities. Preprocessing steps included slicing 3D scan into 2D slices, and removing slices without any lesion in mask, normalization, centering and augmentation as rotate and flip. The results were promising, with a DSC value of 0.866 with augmentation and 0.803 without it (with the usage of softmax binary cross-entropy loss function). These results are comparable to the ratings of medical professionals.

1.3.4 Transfusion: Understanding Transfer Learning for Medical Imaging

A transfer learning technique was used by Raghu *et al.* [15] to compare its usability and performance in the medical field, in contrast to the usual usage of this method on real-world images. Authors focused on evaluating *Diabetic Retinopathy*, which is an eye disease classified into five groups of severity. They used three different architectures: ResNet, Inception-v3 and a custom smaller network consisting of 2D convolution, normalization and ReLU activation, with various numbers of channels and layers. The custom network was used to compare the performance between smaller and bigger architectures, which is not relevant for this work. The key asset

of the paper for this work was training on a small dataset, with a focus on the difference between a randomly initialized network and a pre-trained one. This dataset consisted of 5000 datapoints from *Retina* dataset, with a resolution of 587×587 used for training. The preprocessing included contrast and hue augmentation, as well as random vertical and horizontal flips.
Results showed that random initialization of the Resnet50 model performed worse by 2% of AUC (Area Under the Curve) compared to a pre-trained model. This proved a slight benefit of transfer learning usage for classification in medical imaging and motivated the idea to implement this technique also for the MS lesion segmentation.

1.3.5 Image-to-Image Translation with Conditional Adversarial Networks

Research in [41] aimed to unify different approaches for image-to-image translation tasks. Since problems as colorizing images, converting daylight picture to night-light or reconstructing photos from edge maps can be seen as a similar concept – translating an input image to output image – authors wanted to create a common framework to cover all of them. This supports the possibility of using GAN for recreating medical images, specifically brain scans used in this work.
They obtained it by *Conditional GANs (cGANs)*, which are Generative Adversarial Networks that do not use only a random vector as an input, but use a vector of features with a random vector to influence the result (Figure 1.9). This way, both the generator, as well as the discriminator see the input edge map.

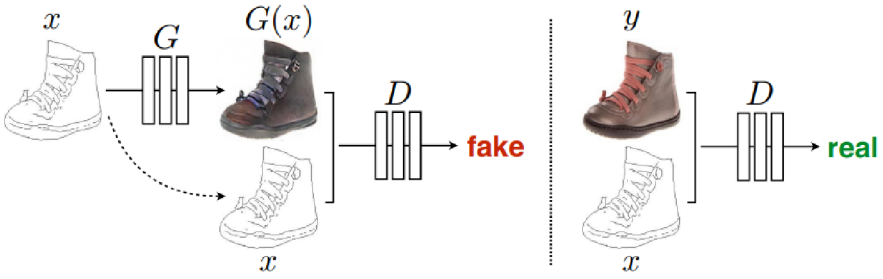

Fig. 1.9: Usage of Conditional GAN for mapping edges to photo [41].

Among many demonstrated techniques, the authors applied cGAN also for semantic segmentation. The focus was to segment objects in cityscape photos. The dataset contained 2975 training images of size 256×256. All metrics (per-pixel accuracy, per-class accuracy and class IoU) showed worse results with the usage of cGAN loss compared to L1 loss or their combination. Even though cGAN results were sharp, it produced many made-up objects, which were not included in the input picture.

2 Methodology

The methodology part begins with an explanation of preprocessing the datasets which were used for later work. It shows work with the FSL software tool, which removes the patient's skull and extracts the brain from MRI scans. This chapter also explains how the data were loaded before passing them as an input and finishes with creating and training the neural networks with generating predictions.

2.1 Dataset

2.1.1 Data Gathering

MRI scans for this work were taken from three public datasets. Together they consisted of 55 scans with manually annotated lesions by medical experts and 31 scans without annotations, used for testing. Used datasets:

- MSSEG 2016 Grand Challenge [42],
- IACL 2015 Longitudinal MS Segmentation [43],
- MICCAI 2008 MS Segmentation Challenge [44].

2.1.2 Data Preprocessing

The first stage of preprocessing the datasets was exploratory data analysis (EDA). The aim of EDA was to obtain parameters of scans. The information for individual datasets can be seen in Table 2.1 below.

Tab. 2.1: Exploratory data analysis of used datasets.

Dataset	**Format**	**Scan Resolution [mm]**	**Sequences**
MSSEG 2016	NifTI	Siemens 3T Verio: 144 x 512 x 512 Siemens Aera 1.5T: 128 x 224 x 256 Philips Ingenia 3T: 261 x 336 x 336	FLAIR T1 T2 DP T1 Gd
IACL 2015	NifTI	T1: 181 x 217 x 181 T2: 181 x 217 x 181 FLAIR: 181 x 217 x 181	T1 T2 PD FLAIR
MICCAI 2008	RAW	FLAIR, T1, T2: 512 x 512 x 512	T1 T2 FLAIR

There were three challenges with those datasets. Firstly, they were not in the same format. MICCAI 2008 data were in RAW format, whereas IACL 2015 and MSSEG 2016 were in NifTI. It was necessary to unify them to NifTI format, which could be later loaded with the *NiBabel* package in Python.
Another challenge that occurred was a various resolution of scans. In some cases, even different resolutions within the same dataset (this was caused by using divergent MRI scanners). The resampling to uniform resolution was made by the *NiBabel* package.
It was also important to have *isotropic voxels* in each scan. These voxels must have had all three dimensions with identical size. The last challenge was choosing the proper sequences for later use. Since all three datasets had *T1*, *T2*, and *FLAIR* sequences available, they were chosen. Those sequences differentiated in the settings of pulse sequences and gradients, which resulted in images with a particular appearance [45]. Below are the details for each sequence:

- **T1** – gadolinium enhanced, fat suppressed,
- **T2** – fat suppressed, fluid attenuated, susceptibility sensitive,
- **FLAIR** – special inversion recovery sequence made to null the signal for certain tissues [45].

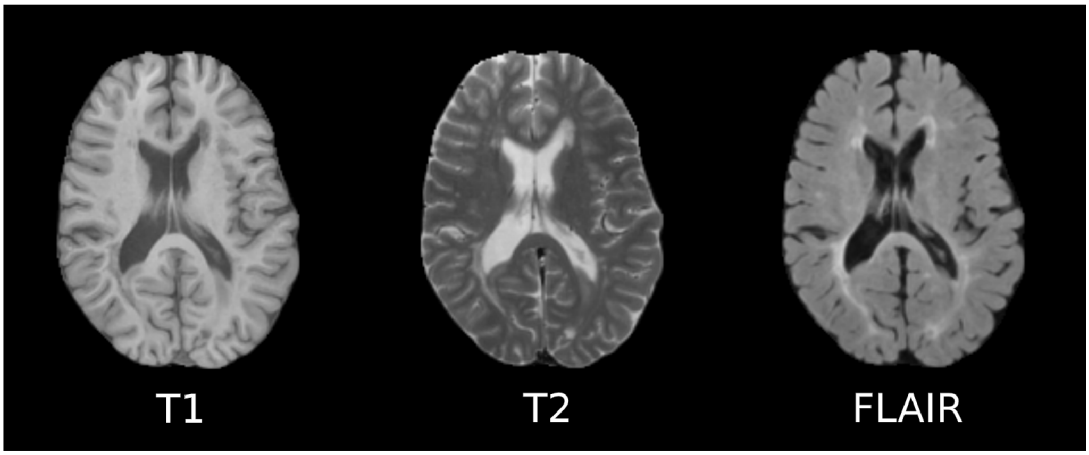

Fig. 2.1: Difference between MRI sequences on MICCAI 2008 patient.

Skull Stripping

After adjusting the datasets to the same file format and resampling to the same dimensions, MICCAI 2008 scans had to be skull stripped since these scans included not only the brain part but also the skull. For this procedure, *FSL* library was used. This library consists of special tools for the analysis of fMRI, MRI, and DTI brain image data. It can be operated and automatized through a command line, or the user can control it by GUI [46].

A *Brain Extraction Tool* (BET) helped with skull stripping and getting only the brain out of the scans. The GUI can be seen in Figure 2.2.

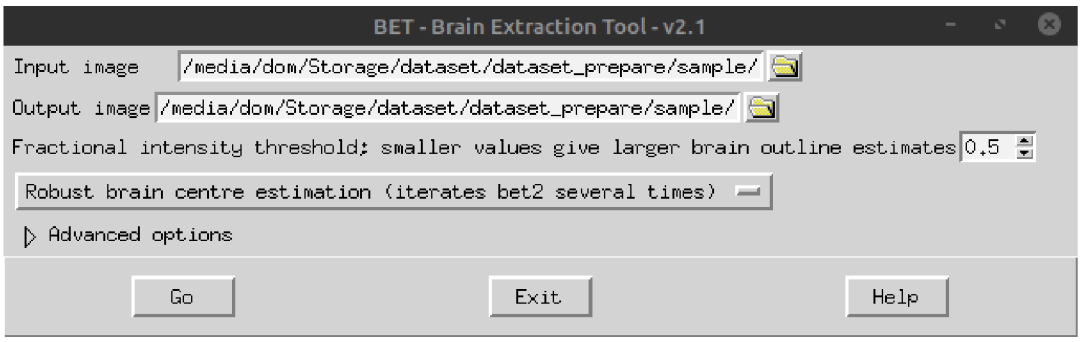

Fig. 2.2: Graphical user interface of Brain Extraction Tool.

The preprocessed image of the MICCAI 2008 scan was chosen as an *Input image*. The *Fractional intensity threshold* had to be manually set for each scan (usually around the value of 0.7 with a maximal deviation of 0.2) to achieve the required accuracy in removing the skull. In the dropdown menu, the *Robust brain centre estimation* showed the best results since it repeated the process several times.
The last setting was under the advanced options – *Create a binary brain mask image as an output*. This mask was then applied to all three sequences.

Because all three sequences were done on the same scan, the position of the brain remained unchanged. Therefore only one binary mask was needed to be created. It was essential to choose the best sequence for brain extraction to be applied to. Figure 2.3 shows the comparison of BET application on each sequence.

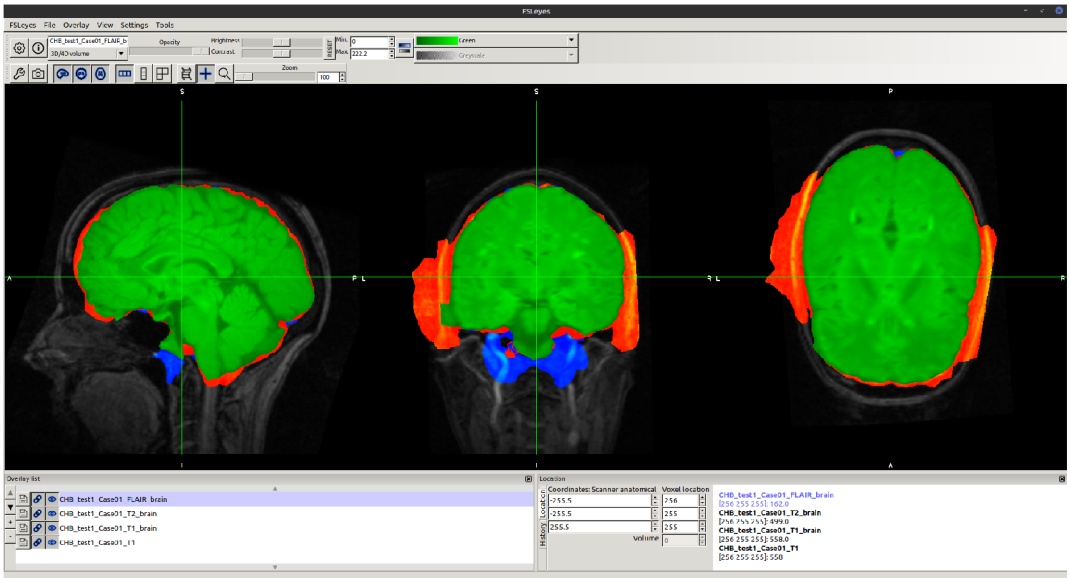

Fig. 2.3: Difference between results in all sequences after brain extraction.

The three colored brains were obtained by applying brain extractions on different sequences from patient 1 of MICCAI 2008 dataset. The colors were assigned as follows:

- Red – T2,
- Green – FLAIR,
- Blue – T1.

T2 sequence showed the worst results since the brain was extracted with many irregularities and included many parts of the skull. T1 and FLAIR sequences were similar, so the binary mask was chosen among them by individually comparing the results. The scan before the brain extraction and after can be seen in Figure 2.4.

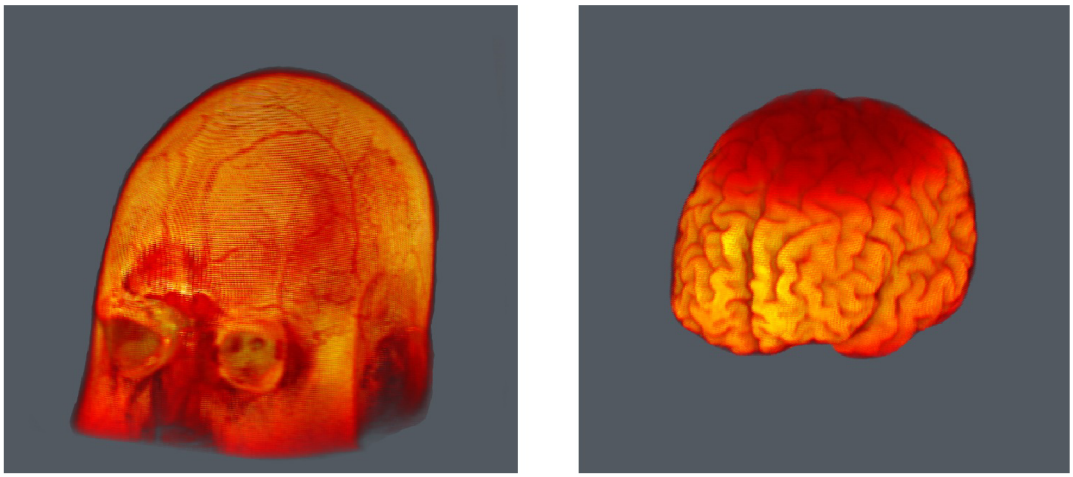

Fig. 2.4: Patient 1 of MICCAI 2008 dataset before and after brain extraction.

This was the last step of preprocessing and unifying all three datasets. Several scans out of all 55 with annotated lesions and 31 without annotations were discarded due to bad quality. The final number of scans was 40 and 15, with annotated masks and without, respectively.
After this, the data had the same format, isotropic voxel size, identical resolution and all scans contained only a brain without the skull. They were ready to be used for training the neural network.

2.2 Preparing Data for Models

After the preprocessing part was finished, the last step that needed to be done was loading and passing the input data into the neural network. All the scans had been skullstripped and had the same resolution. They were divided into two separate groups – train data and test data. The first mentioned group contained scans of

the brain and corresponding masks of lesions. The second group had only the brain scans.[1]

The NifTI files were loaded with *NiBabel* package. 2D axial cuts were made along the z-axis of the brain and saved as a list (*see* Figure 2.5). Each scan was saved after the previous one. This process was done for both groups (train and test) and for both files in the train group separately (brain and corresponding mask). Therefore we had three separate lists. After this step, the lists were converted to a Numpy array.

Since the pixel values might have varied due to different scanners, the data in arrays were normalized. This was done by dividing each value of an array by the maximum value of the array, which resulted in every value being in a range [0, 1].

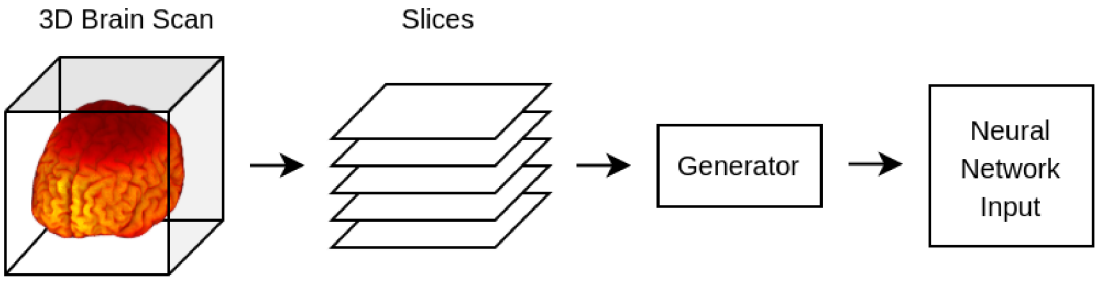

Fig. 2.5: Process of loading data for input before training.

2.2.1 Generator

The generator was another part of the loading process prior to input of the neural network. It was incorporated with "on the fly" augmentation, so the modified scans were not saved but flowed directly to the chosen model. The generator was based on *ImageDataAugmentor* [47], which supports the *Albumentations* [48] Python library. Augmentation applied on input data (with a probability of occurrence in percentage) are listed below:

- Vertical Flip (50%),
- Random Rotate 90° (50%),
- Shift Scale Rotate (50%),
- Grid Distortion (20%).

These techniques were randomly applied to the input data with a batch size of 32 slices. It was crucial that the library modified the brain and corresponding mask the

[1] In fact, even for testing data, we were using scans, which had a corresponding annotated mask, but the neural network would not see the masks. The reason for it is to show the comparison of ground truth and predicted lesions in the chapter 3 Results.

same way, in order to match annotations with damaged brain tissue. In Figure 2.6 can be seen how the brain looks before and after the augmentation.

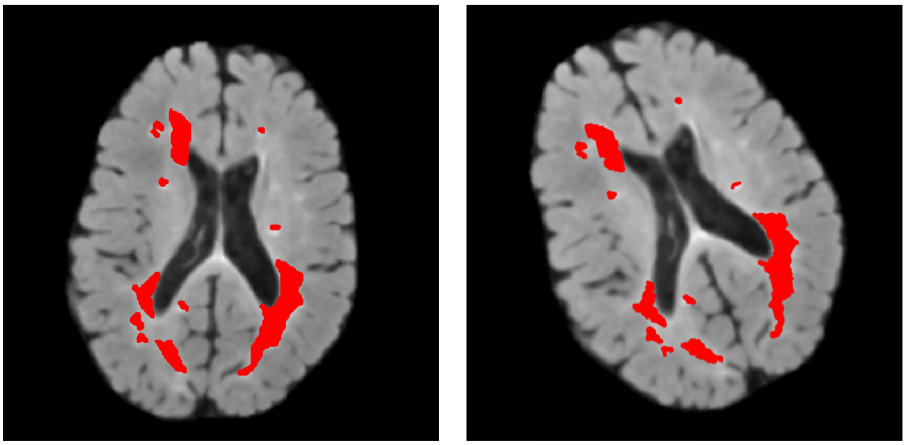

Fig. 2.6: Original brain scan with annotated lesions (left), and corresponding scan after augmentation (right).

Even though this might seem as an insignificant change for a human observer, the network considers it a new patient, which increases the model's performance and decreases overfitting. Training with and without augmentation lead to improvement of Dice Similarity Coefficient from values ≈ 0.3 to ≈ 0.5.

2.2.2 Data for U-Net Model

Due to an unbalanced distribution of MS lesions, multiple brain slices from Figure 2.5 did not contain any values in the mask slices. Training of the U-Net model with empty masks, as well as with the usage of all three modalities, showed worse results. To address this issue and narrow the focus of training, only the FLAIR modality slices containing a nonzero mask were saved to a Numpy array.
In the end, the filtering left 6,720 slices, compared to 23,040 slices of previously saved data with zero slices (120 patients $\times$ 192 slices).

2.2.3 Data for GAN Model

As for the GAN model, training was done on two different data. We did not use the ImageNet dataset to pre-train weights for transfer learning; instead, we only used the initial, unfiltered brain scans without corresponding masks with all three modalities from our prepared datasets.
After the weights of the generator were trained with GAN, the second round of training was done with the same data as for U-Net, described in section 2.2.2 above.

2.3 U-Net Training

The first architecture which we used was a 2D U-Net build by Segmentation Models [49]. It is a library based on Keras and Tensorflow that provides a high-level API to create models for segmentation tasks. The encoder weights could be either randomly generated before training or loaded as pre-trained on the ImageNet dataset. We tested the U-Net model with a ResNet-34 backbone, with three different settings of encoder weights:

- randomly initialized weights,
- pre-trained weights,
- pre-trained weights, which are locked during the training.

2.3.1 U-Net Hyperparameters

Since the U-Net model was already predefined, all the layers and activation functions were as in Figure 1.3. Other parameters were as follows:

- **Optimizer:** Adam
- **Learning Rate:** $2e^{-5}$
- **Loss Function:** Binary Crossentropy
- **Metrics:** Combo Loss[2]
- **Steps Per Epoch:** 128
- **Validation Steps:** 24
- **Batch Size:** 32
- **Epochs:** 300

The results of this testing with predictions are shown in section 3.1.

2.4 GAN Training

As mentioned in 2.2.3, GAN training was divided into two stages for the purpose of transfer learning. The first stage aimed to pre-train weights of the GAN generator on the whole dataset. The second stage used a filtered dataset with only FLAIR modality and loaded pre-trained weights from the first stage to encoder to later adapt them for MS segmentation challenge.

The pre-training code was based on Pix2Pix [41]. We implemented self-supervised learning, specifically context encoder, which learned how to reconstruct brain scans from incomplete input. The input image was combined with a "checkerboard"; therefore, it had missing pieces in it (*see* Figure 2.7) that served as an input seed.

[2]Loss proposed in [50], which combines Dice Loss and modified Cross-Entropy.

The training pipeline went as follows:

1. Discriminator updated its weights after seeing a fake (generated) image.
2. Discriminator updated its weights after seeing a real image.
3. Generator updated its weights.

Only the weights of the generator were saved (after 8,000 epochs). At this point, the generator was able to create almost identical brain scans as ground truth.

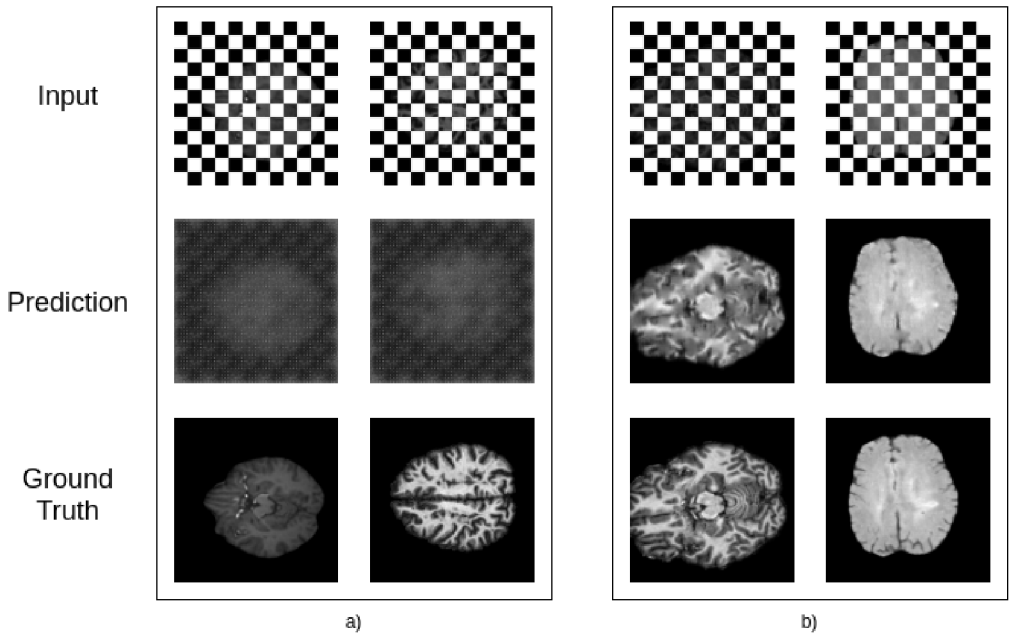

Fig. 2.7: GAN pre-training: a) epoch number 10, b) epoch number 8,000.

The second stage of GAN training used the generator, along with pre-trained weights from the first stage, to perform transfer learning. Results of randomly generated weights were used from U-Net training as a reference point. Another two settings were the same as in section 2.3.

2.4.1 GAN Hyperparameters

The generator part had encoder-decoder architecture and was used for both pre-training and later transfer learning. The only difference was changing the output activation function in the second stage to sigmoid. Other hyperparameters stayed as in the first experiment (2.3.1). The results of this testing with predictions are shown in section 3.2.

2.5 Generating Predictions

After the training of both models was done, we could create predictions of lesions. As Figure 2.8 shows, we inserted a brain scan from the testing group created in section 2.2 into the model. This model loaded the weights which had been trained on the training dataset and outputted predicted masks.

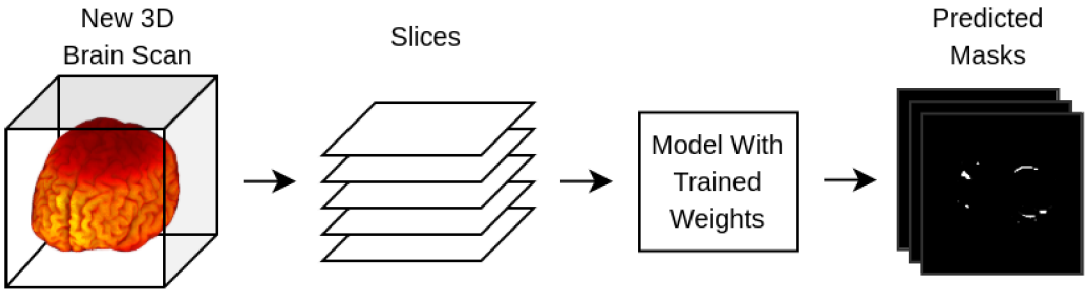

Fig. 2.8: Process of mask predicting after the training.

The grey border around predicted masks was used only for better visualisation and was not included in actual predictions. As well as rotation of the output, which corresponded with input slices (axial cuts along the z-axis).
The predicted masks were further processed with thresholding into two values – 0 (no lesion detected) or 1 (lesion detected). The thresholding ranges were individually chosen after each training to obtain the highest DSC on test data. These ranges, as well as how the results improved, are in Table 3.1. The results could be later compared with the ground truth since we also had the annotated lesions of these test scans.

Figure 2.9 indicates the difference among original mask, predicted mask and predicted mask with thresholding. After thresholding (with a thresholding point of 0.7), the image had most of the incorrectly annotated lesions removed and provided a more accurate prediction.

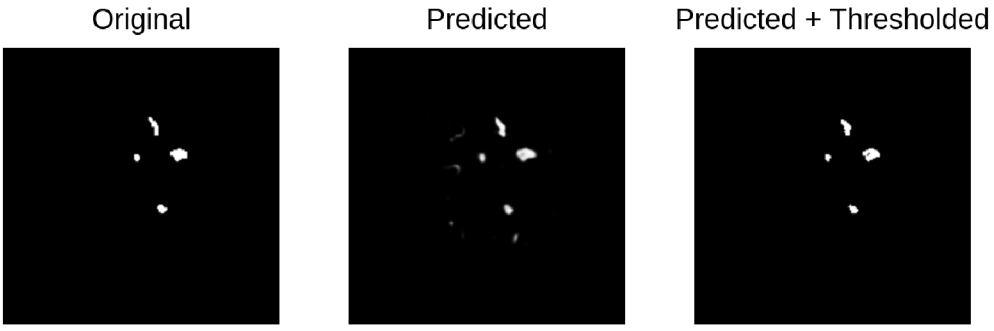

Fig. 2.9: Comparison of mask prediction results with and without thresholding.

2.6 Implementation Details

This section aims to describe the hardware, as well as the software (language, libraries and their versions) used for this work.
All the experiments were conducted on a remote Linux server at the Brno University of Technology. The machine was equipped with 66 GB of RAM and a GeForce GTX 1080 Ti graphics card, which used CUDA 11.2 to speed up the training process. The size of the whole training dataset was 1.6 GB, with one brain scan varying from 10 to 40 MB.
The language used for work was Python 3.7.8, with all libraries isolated in one Conda environment. A brief description of the main ones can be seen below.

Tensorflow 1.14.0 – open-source deep learning library.
Keras 2.3.1 – API which runs on the top of Tensorflow.
Numpy 1.19.2 – library to run mathematical functions on arrays.
NiBabel 3.2.1 – package which provides read and write access to neuroimaging files (*e.g.* MRI scans in .nii format).
Numpy 1.19.2 – library to run mathematical functions on arrays.
Segmentation Models 1.0.1 – high level API used to create various models for segmentation with possibility of loading pre-trained weights on ImageNet dataset.
Albumentations 0.5.2 – computer vision library that augments data before passing them into the input of network. In our work it was integrated with custom image data generator **ImageDataAugmentor 0.2.8**.

3 Results

This chapter presents the results of the two proposed methods. Firstly for U-Net architecture and all three settings of encoder weights, and secondly results of the GAN model.

3.1 U-Net Results

As Figure 3.1 shows, after 300 epochs of training, the Dice Similarity Coefficient reached the value ≈ 0.5 for all three testings.

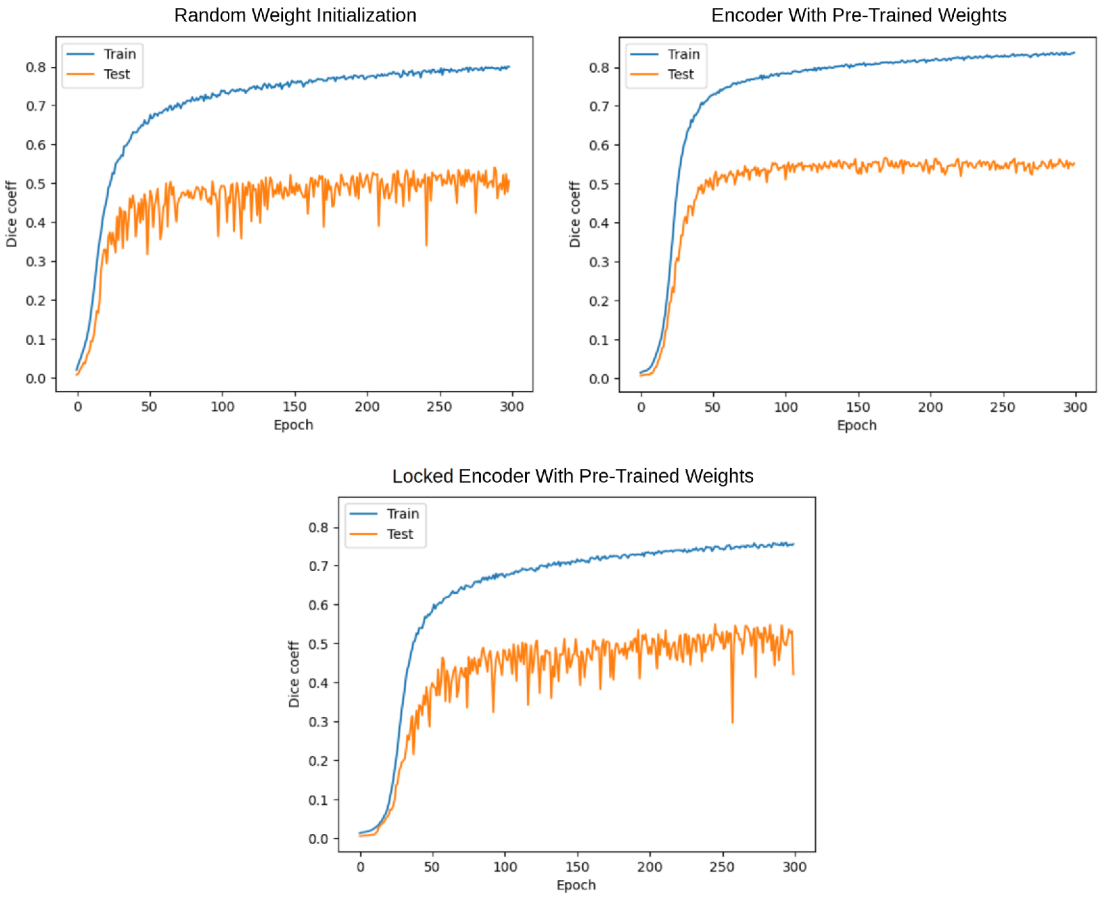

Fig. 3.1: Results of training with U-Net model.

Randomly initialized weights had proven to be moderately less effective compared to weights initialized with transfer learning. In all cases, the difference between DSC of train data and test data significant, which indicates overfitting in later epochs. The experiments with transfer learning without locked weights had faster convergence and performed best out of all three tests – reaching value ≈ 0.53 DSC on validation data.

Figure 3.2 displays the difference of predictions for each type of training and the ground truth. It is visible that more significant lesions were always correctly predicted, whereas smaller ones were in some cases omitted.

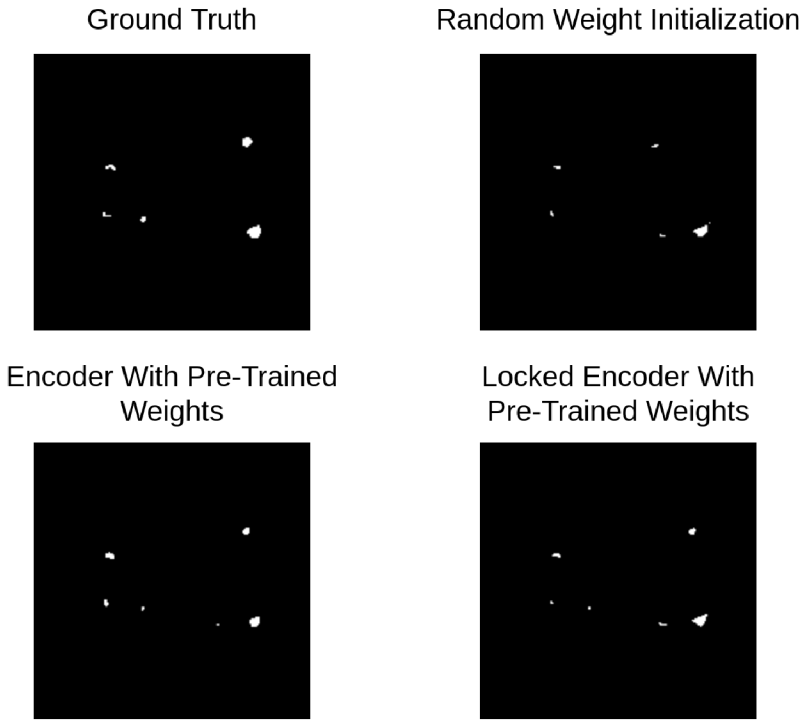

Fig. 3.2: Comparison of lesion prediction for each U-Net training method.

Predictions of the method with the highest Dice Similarity Coefficient (transfer learning with encoder without locked weights) can be seen in Figure 3.3. Results on test data, as well as a comparison of this approach with the second experiment is visible in Table 3.1.

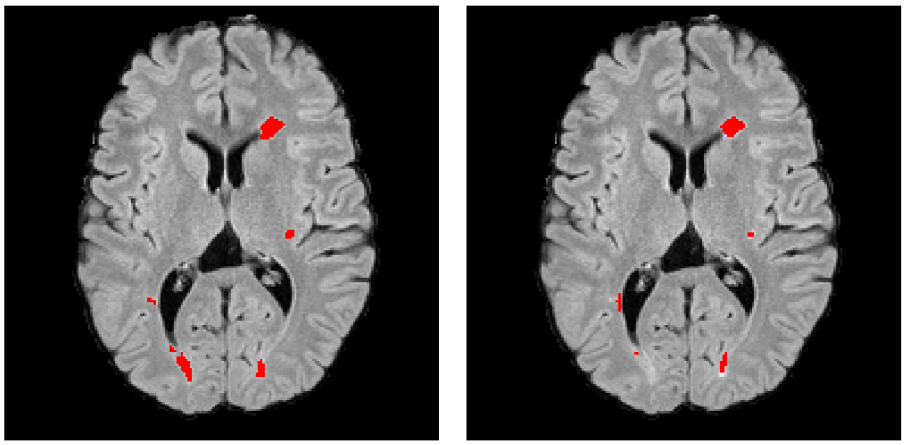

Fig. 3.3: Ground truth lesions (left) and lesions predicted with U-Net transfer learning with locked encoder (right).

3.2 GAN Results

The values of DSC after 300 epochs can be seen in Figure 3.4. The graph of training without initialization of weights is excluded (for reference value *see* Figure 3.1). The proposed method has shown a worse outcome by ≈ 0.1 DSC during training, compared to pre-trained weights with segmentation on the ImageNet dataset.

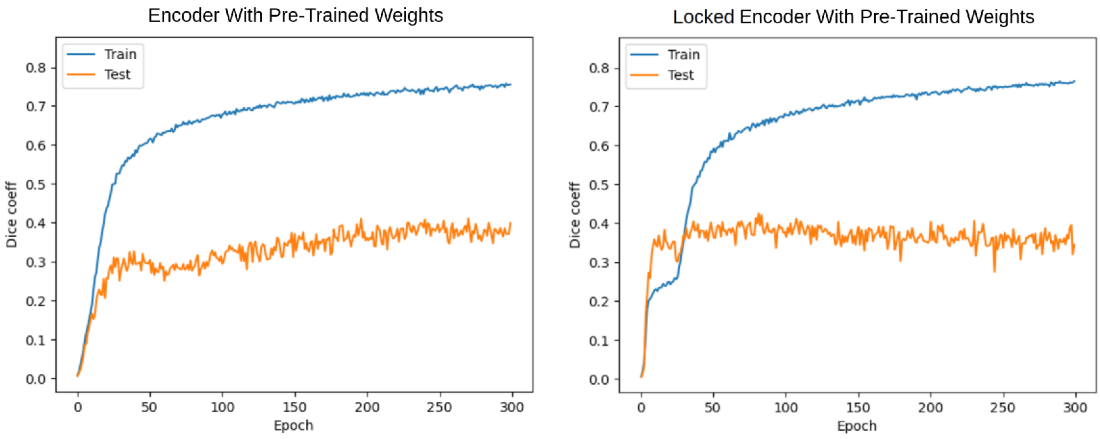

Fig. 3.4: Results of training after pre-training weights with GAN model.

Locked encoder proved to have faster convergence, although, in later training, the trend is decreasing. Worse results might have been caused by batch normalization, which must have been used in the encoder block of GAN architecture during pre-training of weights. Even though it improves the network stability, for a challenge with hugely imbalanced data, it might lower the performance.

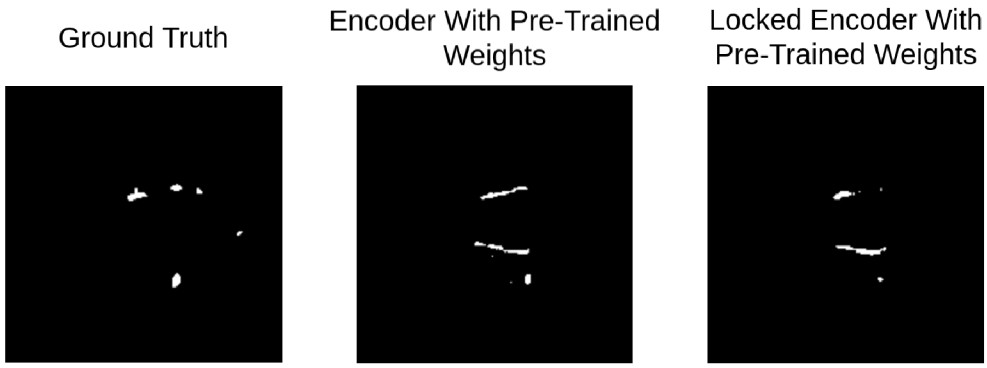

Fig. 3.5: Comparison of lesion prediction for both GAN training method.

Figure 3.5 indicates that the predicted masks are far from the ground truth. Even after thresholding, the network could not predict lesions correctly, which resulted in many false positive pixels.

How the network performed with regard to brain scan can be seen in Figure 3.6, where the experiment with best results (encoder with pre-trained weights) was able to segment lesions at the correct location but missed out on much of damaged tissue.

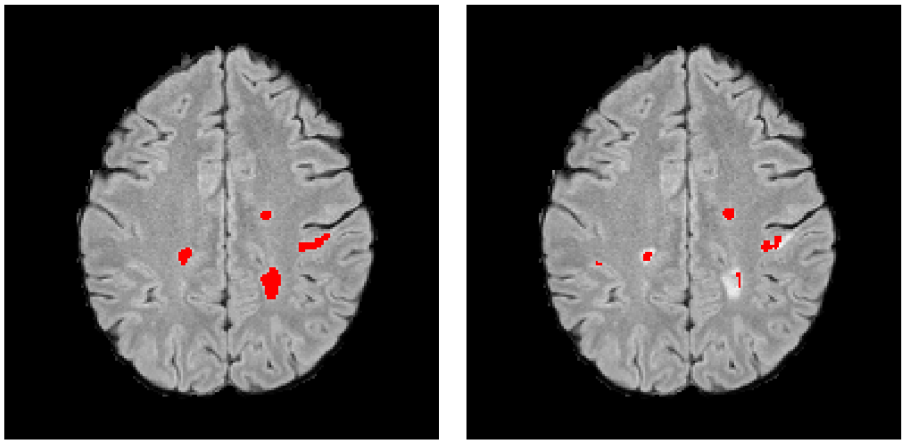

Fig. 3.6: Ground truth lesions (left) and lesions predicted with GAN transfer learning without locked encoder (right).

3.3 Results Comparison

The summary of all the results for individual methods is in Table 3.1. It is clear that pre-training weights with self-supervised learning by the GAN model did not improve the used metrics, and even after thresholding showed worse values of DSC.

Tab. 3.1: Comparison of Dice Similarity Coefficient values on training data for all proposed techniques.

Method	**Thresholding Value**	**Initial DSC**	**DSC After Thresholding**
U-Net Random Weights	0.99	0.363	0.402
U-Net Pre-Trained Weights	0.97	0.594	0.614
U-Net Locked Encoder	0.96	0.604	**0.632**
GAN Pre-Trained Weights	0.99	0.236	**0.289**
GAN Locked Encoder	0.46	0.09	0.094

There might be at least three possibilities of such a decrease in performance compared to the first and second approach. One of them is batch normalization, which was already mentioned in section 3.2.
Another one could be the difference in the implementation of our two architectures. Since we needed the first experiment to use already pre-trained weights on the

ImageNet dataset (to avoid the tedious process of pre-training them by ourselves), the usage of the Segmentation Models library was necessary. On the contrary, the GAN generator had to use a slightly modified version of U-Net, which might have caused dissimilarity in the results. Lastly, the problem might have been the difficulty of the MS segmentation task. The datasets were significantly imbalanced; hence the neural networks were not able to train properly.
In terms of the time difference, none of the experiments showed significant differences in training time (approximately 3 hours) and prediction time (1 second for a patient).

Conclusion

In this work we have tested and compared approaches to Multiple Sclerosis lesion segmentation with emphasis on transfer learning methods.
The theoretical part of the master thesis explained image segmentation and its various techniques, which are currently being used in many fields, including healthcare. It also described what Multiple Sclerosis is, as well as what causes this disease and how it can be diagnosed and treated.

The practical part started with exploratory data analysis of three datasets, which were chosen for this work. All datasets needed to be preprocessed since they had different resolutions, voxel size and file format. For a MICCAI 2008 dataset, a skull stripping had to be done to remove the skull from the scan. FSL library with the Brain Extraction Tool helped with this task. We had to check each scan manually, so the threshold for brain extraction was set correctly. After this process, all datasets were connected and loaded as a Numpy array to be further used for training.
The training included two different approaches. Both made use of transfer learning with pre-trained weights, which were obtained by two techniques. First was a U-Net architecture with pre-trained weights on the ImageNet dataset. The second experiment acquired weights from a self-supervised context encoder created with the GAN model. We compared three settings of encoder weights (randomly initialized, loaded pre-trained weights and loaded pre-trained weights with locked encoder).
The Dice Similarity Coefficient reached a value of approximately 0.63 for testing data with the best technique of U-Net (pre-trained weights with locked encoder) and for GAN around 0.29 (pre-trained weights without locked encoder).

The presented results showed a vast difference in the performance of the proposed techniques. Compared to transfer learning with U-Net, the pre-trained GAN generator reported poor results when used for this type of problem. The cause of this difference might have been caused in multiple areas. The first one is the usage of batch normalization, which was needed in the initial training of GAN and needed to remain in the generator in later training. Another challenge could have been a slight variance in encoder-decoder architectures (U-Net model was implemented with robust library Segmentation Models, whereas GAN generator had modified U-Net for the early pre-training). Lastly, the incompatibility of approaches (segmentation and self-supervised context encoder) and the imbalanced distribution of lesions could have caused worsen results. Further research could be made based on these results, with a possibility of improvement with a better architecture of the GAN generator.

This work showed a chance of performance improvement in MS lesion segmentation with transfer learning. The assignment of work has been fulfilled. The codes are available at `https://github.com/DomSas/master-thesis` under MIT license.

References

[1] When Data Science met Medicine! The Research Nest [Internet]. Medium; 2018 [cited 2020 Oct 24]. Available from: https://medium.com/the-research-nest/when-data-science-met-medicine-8d3971a0ade9

[2] Chowdhary L.C., Acharjya D.P. Segmentation and Feature Extraction in Medical Imaging: A Systematic Review. Procedia Computer Science. 2020;167:26-36. Available from: https://doi.org/10.1016/j.procs.2020.03.179

[3] Kooi T. Why Skin Lesions are Peanuts and Brain Tumors Harder Nuts [Internet]. Seoul: The Gradient; 2020 [cited 2020 Nov 20]. Available from: https://thegradient.pub/why-skin-lesions-are-peanuts-and-brain-tumors-harder-nuts/

[4] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Springer International Publishing. 2015;234-41. Available from: https://arxiv.org/pdf/1505.04597.pdf

[5] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. Advances in neural information processing systems. 2014;2672–80. Available from: https://arxiv.org/pdf/1406.2661.pdf

[6] Oppermann A. What is Deep Learning and How does it work? [Internet]. towards data science; 2019 [cited 2020 Nov 19]. Available from: https://towardsdatascience.com/what-is-deep-learning-and-how-does-it-work-2ce44bb692ac

[7] Nguyen G, Dlugolinsky S, Bobák M, Tran V, Lopéz García Á, Heredia I, Malík P, Hluchý L. Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. Artif Intell. 2019;52(1):77–124. Available from: https://doi.org/10.1007/s10462-018-09679-z

[8] Öztürk Ş, Özkaya U, Akdemir B, Seyfi L. Convolution Kernel Size Effect on Convolutional Neural Network in Histopathological Image Processing Applications. Bucharest: 2018 International Symposium on Fundamentals of Electrical Engineering (ISFEE). 2018;1-5. Available from: https://doi.org/10.1109/ISFEE.2018.8742484

[9] Yamashita R, Nishio M, Do R.K.G., Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging. 2018;9:611–29. Available from: https://doi.org/10.1007/s13244-018-0639-9

[10] Qamar S, Jin H, Zheng R, Ahmad P, Usama M. A variant form of 3D-UNet for infant brain segmentation. Future Generation Computer Systems. 2020;108:613-23. Available from: https://doi.org/10.1016/j.future.2019.11.021

[11] Jin Q, Meng Z, Sun C, Cui H, Su R. RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans. Frontiers in Bioengineering and Biotechnology. 2020;8:1471. Available from: https://doi.org/10.3389/fbioe.2020.605132

[12] Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N, Terzopoulos D. Image segmentation using deep learning: A survey. arXiv preprint arXiv:2001.05566. 2020 Jan. Available from: https://arxiv.org/pdf/2001.05566.pdf

[13] Rocca J. Understanding Generative Adversarial Networks (GANs) [Internet]. towards data science; 2019 Jan [cited 2020 Mar 4]. Available from: https://towardsdatascience.com/understanding-generative-adversarial-networks-gans-cd6e4651a29

[14] Russakovsky O, Deng J, Su H., Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg A, Fei-Fei L. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis. 2015;115, 211–252. Available from: https://doi.org/10.1007/s11263-015-0816-y

[15] Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: Understanding transfer learning for medical imaging. arXiv preprint arXiv:1902.07208. 2019. Available from: https://arxiv.org/pdf/1902.07208.pdf

[16] Chen L.-C., Yang Y, Wang J, Xu W, Yuille A. L. Attention to scale: Scale-aware semantic image segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016;3640–9. Available from: https://arxiv.org/pdf/1511.03339.pdf

[17] Oktay O, Schlemper J, Folgoc L.L., Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla N.Y., Kainz B, Glocker B. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999. 2018. Available from: https://arxiv.org/pdf/1804.03999.pdf

[18] Caesar H. Restoring the balance between stuff and things in scene understanding. 2018. Available from: https://era.ed.ac.uk/handle/1842/31352

[19] Hashemi S. R., Mohseni Salehi S. S., Erdogmus D, Prabhu S. P., Warfield S. K., Gholipour A. Asymmetric Loss Functions and Deep Densely-Connected Networks for Highly-Imbalanced Medical Image Segmentation: Application to

Multiple Sclerosis Lesion Detection. IEEE Access. 2019;7:1721-35. Available from: https://doi.org/10.1109/ACCESS.2018.2886371

[20] Taha A.A., Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging. 2015;15:29. Available from: https://doi.org/10.1186/s12880-015-0068-x

[21] Zhang D, He F, Han S, Zou L. An efficient approach to directly compute the exact Hausdorff distance for 3D point sets Integrated Computer Aided Engineering. 2017:24(3);261–77. Available from: https://doi.org/10.3233/ICA-170544

[22] Nalepa J., Marcinkiewicz M., Kawulok M. Data Augmentation for Brain-Tumor Segmentation: A Review. Front. Comput. Neurosci. 2019 Dec;13:83. Available from: https://doi.org/10.3389/fncom.2019.00083

[23] Xu J., Li M, Zhu Z. Automatic data augmentation for 3d medical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer. 2020 Oct;378-87. Available from: https://arxiv.org/pdf/2010.11695.pdf

[24] Eaton-Rosen Z, Bragman F.J., Ourselin S, Cardoso M.J. Improving Data Augmentation for Medical Image Segmentation. 2018. Available from: https://openreview.net/pdf/09274d29f227a2c6cce8e32f25df8df5dbf18df1.pdf

[25] Zhang H, Cisse M, Dauphin Y.N., Lopez-Paz D. Beyond empirical risk minimization. arXiv preprint. 2017 Oct. Available from: https://arxiv.org/pdf/1710.09412.pdf

[26] Shorten, C, Khoshgoftaar T.M. A survey on Image Data Augmentation for Deep Learning. J Big Data. 2019 July. Available from: https://doi.org/10.1186/s40537-019-0197-0

[27] WILSON J. A. Brain Imaging in Multiple Sclerosis [Internet]. Medscape; 2019 [cited 2020 Sep 24] Available from: https://emedicine.medscape.com/article/342254-overview

[28] Symptoms & Diagnosis [Internet]. National Multiple Sclerosis Society [cited 2020 Sep 25]. Available from: https://www.nationalmssociety.org/Symptoms-Diagnosis/MS-Symptoms

[29] What Is Multiple Sclerosis? [Internet]. ZiMS Foundation [cited 2020 Sep 24]. Available from: https://zimsfoundation.org/multiple-sclerosis/

[30] Multiple sclerosis [Internet]. Mayo Clinic [cited 2020 Sep 25]. Available from: https://www.mayoclinic.org/diseases-conditions/multiple-sclerosis/symptoms-causes/syc-20350269

[31] Multiple Sclerosis Diagnosis [Internet]. National Multiple Sclerosis Society [cited 2020 Sep 25]. Available from: https://www.nationalmssociety.org/Symptoms-Diagnosis/Diagnosing-MS

[32] Paty D. W., Oger J. J. F., Kastrukoff L. F., HashimotoS. A., Hooge J. P., Eisen A. A., Eisen K. A., Purves S. J., Low M. D., Brandejs V., Robertson W. D., Li D. K. B. MRI in the diagnosis of MS: a prospective study with comparison of clinical evaluation, evoked potentials, oligoclonal banding, and CT. Neurology. 1988;28. Available from: https://doi.org/10.1212/WNL.38.2.180

[33] Understanding Multiple Sclerosis (MS) [Internet]. Healthline [cited 2020 Sep 26]. Available from: https://www.healthline.com/health/multiple-sclerosis#types

[34] Treating MS [Internet] National Multiple Sclerosis Society [cited 2021 Jan 19]. Available from: https://www.nationalmssociety.org/Treating-MS

[35] Multiple Sclerosis Diagnosis [Internet]. National Multiple Sclerosis Society [cited 2020 Sep 26]. Available from: https://www.nationalmssociety.org/What-is-MS/Types-of-MS

[36] Understanding Multiple Sclerosis (MS) [Internet]. Healthline [cited 2021 Jan 19]. Available from: https://www.healthline.com/health/multiple-sclerosis

[37] Treatment - Multiple sclerosis [Internet]. National Health Society [cited 2021 Jan 19]. Available from: https://www.nhs.uk/conditions/multiple-sclerosis/treatment/

[38] Coronado I, Gabr R.E., Narayana P.A. Deep learning segmentation of gadolinium-enhancing lesions in multiple sclerosis. Multiple Sclerosis J. 2020 May. Available from: https://doi.org/10.1177/1352458520921364

[39] Islam M, Vibashan V.S., Jose V.J.M., Wijethilake N, Utkarsh U, Ren H. Brain Tumor Segmentation and Survival Prediction Using 3D Attention UNet. Lecture Notes in Computer Science. Springer;11992. 2020 May. Available from: https://doi.org/10.1007/978-3-030-46640-4_25

[40] Kumar A, Murthy O.N., Ghosal P, Mukherjee A, Nandi D. A Dense U-Net Architecture for Multiple Sclerosis Lesion Segmentation. TENCON 2019-2019 IEEE Region 10 Conference. 2019 Oct;662-667. Available from: https://ieeexplore.ieee.org/document/9185760

[41] Isola P, Zhu J.Y., Zhou T, Efros A.A. Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017;1125-34. Available from: https://arxiv.org/pdf/1611.07004.pdf

[42] MS Segmentation Challenge Using a Data Management and Processing Infrastructure [Internet]. France: FLI-IAM: France Life Imaging - Information Analysis and Management; 2016 [cited 2020 Oct 25]. Available from: https://portal.fli-iam.irisa.fr/msseg-challenge/overview

[43] Lesion Challenge [Internet]. Smart-stats-tools; 2015 [cited 2020 Oct 25]. Available from: https://smart-stats-tools.org/lesion-challenge

[44] MS lesion segmentation challenge 2008 [Internet]. MS lesion segmentation challenge 2008; 2008 [cited 2020 Oct 25]. Available from: http://www.ia.unc.edu/MSseg/index.html

[45] MRI sequences (overview) [Internet]. Radiopaedia [cited 2020 Nov 5]. Available from: https://radiopaedia.org/articles/mri-sequences-overview

[46] FSL [Internet]. FSL [cit. 2020 Nov 5]. Available from: https://fsl.fmrib.ox.ac.uk/fsl/fslwiki

[47] Tukiainen M. ImageDataAugmentor [Software]. GitHub. 2019 [cit. 2021 Apr 24]. Available from: https://github.com/mjkvaak/ImageDataAugmentor

[48] Buslaev A, and Iglovikov V. I, Khvedchenya E,Parinov A, Druzhinin M, Kalinin A. A. Albumentations: Fast and Flexible Image Augmentations. Version 0.5.2 [Software]. GitHub. 2020 [cit. 2021 Apr 24]. Available from: https://github.com/albumentations-team/albumentations

[49] Yakubovskiy P. Segmentation Models. Version 1.0.1 [Software]. GitHub. 2019 [cit. 2021 Apr 25]. Available from: https://github.com/qubvel/segmentation_models

[50] Taghanaki S.A., Zheng Y, Zhou S.K., Georgescu B, Sharma P, Xu D, Comaniciu D, Hamarneh G. Combo loss: Handling input and output imbalance in multi-organ segmentation. Computerized Medical Imaging and Graphics. 2019;75;24-33. Available from: https://arxiv.org/pdf/1805.02798.pdf

List of Symbols, Quantities and Abbreviations

3D	Three-dimensional
AI	Artificial Intelligence
ANN	Artificial Neural Networks
API	Application Programming Interface
BET	Brain Extraction Tool
cGANs	Conditional Generative Adversarial Networks
CIS	Clinically Isolated Syndrome
CNN	Convolutional Neural Network
DSC	Dice Similarity Coefficient
DL	Deep Learning
DNN	Deep Neural Networks
DTI	Diffusion Tensor Imaging
EDA	Exploratory Data Analysis
FLAIR	Fluid Attenuated Inversion Recovery
GANs	Generative Adversarial Networks
Gd	Gadolinium
GUI	Graphical User Interface
IoU	Intersection-Over-Union
MRI	Magnetic Resonance Imaging
fMRI	Functional Magnetic Resonance Imaging
FN	False Negative
FP	False Positive
HD	Hausdorff Distance
ML	Machine Learning

MS	Multiple Sclerosis
NifTI	Neuroimaging Informatics Technology Initiative
PPMS	Primary Progressive MS
ReLU	Rectified Linear Unit
RRMS	Relapsing-Remitting MS
SPMS	Secondary Progressive MS
TN	True Negative
TP	True Positive