Master's Programme in Computer, Communication and Information Sciences

Deep Learning Based Analysis of Prostate Cancer from MP-MRI

Classification, detection and segmentation of prostate cancer lesions from MP-MRI with deep learning based methods

---

Pedro David Carneiro Neto

Aalto University

MASTER'S THESIS

Aalto University
MASTER'S THESIS 2020

Deep Learning Based Analysis of Prostate Cancer from MP-MRI

Classification, detection and segmentation of prostate cancer lesions from MP-MRI with deep learning based methods

Pedro David Carneiro Neto

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology.
Otaniemi, 25 May 2020

Supervisor:	professor Juho Kannala
Advisor 1:	doctor Saad Ullah Akram
Advisor 2:	doctor Harri Merisaari
Advisor 3:	medical doctor Ivan Jambor

Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

Author
Pedro David Carneiro Neto

Title
Deep Learning Based Analysis of Prostate Cancer from MP-MRI

School School of Science

Master's programme Computer, Communication and Information Sciences

Major Computer Science - Big Data and Large-Scale Computing **Code** SCI3042

Supervisor professor Juho Kannala

Advisor doctor Saad Ullah Akram; doctor Harri Merisaari; medical doctor Ivan Jambor

Level Master's thesis **Date** 25 May 2020 **Pages** 74+4 **Language** English

Abstract

The diagnosis of prostate cancer faces a problem with overdiagnosis that leads to damaging side effects due to unnecessary treatment. Research has shown that the use of multi-parametric magnetic resonance images to conduct biopsies can drastically help to mitigate the overdiagnosis, thus reducing the side effects on healthy patients. This study aims to investigate the use of deep learning techniques to explore computer-aid diagnosis based on MRI as input. Several diagnosis problems ranging from classification of lesions as being clinically significant or not to the detection and segmentation of lesions are addressed with deep learning based approaches.

This thesis tackled two main problems regarding the diagnosis of prostate cancer. Firstly, a deep neural network architecture, XmasNet, was used to conduct two large experiments on the classification of lesions. Secondly, detection and segmentation experiments were conducted, first on the prostate and afterward on the prostate cancer lesions. The former experiments explored the lesions through a two-dimensional space, while the latter explored models to work with three-dimensional inputs. For this task, the 3D models explored were the 3D U-Net and a pretrained 3D ResNet-18. A rigorous analysis of all these problems was conducted with a total of two networks, two cropping techniques, two resampling techniques, two crop sizes, five input sizes and data augmentations experimented for lesion classification. While for segmentation two models, two input sizes and data augmentations were experimented. Moreover the experiments were conducted for both sequences independently, and within the lesion classification problem, the experiments were also conducted for both sequences simultaneously. However, while the binary classification of the clinical significance of lesions and the detection and segmentation of the prostate already achieve the desired results (0.870 AUC and 0.915 dice score respectively), the classification of the PIRADS score and the segmentation of lesions still have a large margin to improve (0.664 accuracy and 0.690 dice score respectively). It was also studied how some flaws in the dataset can be addressed to improve the results of all these problems. Further research on the problem is still needed, but nonetheless, this thesis established sufficient ground for future work to be conducted.

Keywords deep learning, computer vision , prostate cancer, segmentation, classification, computer-aided diagnosis systems

Contents

1. Introduction

Prostate cancer is one of the biggest causes of cancer deaths in men around the world. It has also been consistently ranked as one of the most frequent cancers to be diagnosed in men by a myriad of organizations worldwide, with 84 different countries having it as the most diagnosed cancer in men [1]. Predictions for 2020 indicate that the number of new cases in the United States will be around 191,930, and 33,330 deaths caused by the disease [2]. The disease's risk varies with some factors like the race, medical records of the family, age and daily diet choices [1], and it only affects men since the prostate is a gland that belongs to the male reproductive system [3]. Regarding these factors, it has been shown that the disease is not only more likely to be diagnosed in black men, but it also shows higher mortality rates [4]. Nevertheless, family medical record is still rather important because having a first-degree relative with prostate cancer doubles the chances of an individual to be also affected in the future [5]. Although responsible for enormous fatalities every year, it does not have a high mortality rate, especially if it is detected in early stages. Hence, the detection in these stages can prevent death and avoid other treatments, with respective side effects and harms, especially if the disease spreads to other body parts.

It is estimated that two-thirds of the affected patients do not show any symptoms [6]. Frequently, the initial phase of the diagnosis is the screening phase that can be performed through two exam. The first example exam is the measurement of the prostate-specific antigen (PSA), that uses blood tests to detect the levels of the antigen. If they are high, there is a chance of prostate cancer, even though the values may change with other diseases. Thus, it does not guarantee the presence of cancer [7]. The second screening technique is the digital rectal examination (DRE) [8] that consists in the insertion of one finger in the rectum to directly feel anomalies in the

prostate. However, these exams are controversial among the scientific community, for example, the PSA has been shown to increase the risks of overdiagnosis that usually implies overtreatment [9] and DRE considered to be ineffective [10, 11]. Moreover, both tests are lacking evidence of benefits in reducing the mortality of the disease [12, 8]. Concerns related to the overtreatment (i.e. unnecessary biopsies) are supported by several harms that may result from the treatment (i.e. loss of sexual functions, blood in the urine) [13, 14].

Since screening is not accurate enough to detect prostate cancer, these methods are often used as a procedure to decide between doing biopsies or not. High values generally require a biopsy, a process that consists of the direct removal of prostate samples using a special needle [15]. The process is usually performed transrectally and implies some risks to the patient due to the invasiveness of the procedure, thus, since the mid-1980s, ultrasound images have been used as a guide, and by 2014 it was still the most common approach [16]. However, prostate cancer cannot be properly detected by ultrasound images since these have poor resolution while displaying soft tissue [17] leading to poor quality and rigour of biopsies. More recently, studies on multi-parametric magnetic resonance images (mp-MRI) have shown not only a better quality in identifying prostate cancer [17] by detecting cancer that was missed by traditional blind biopsies [18, 19, 20], but also decreasing the cases of overdiagnosis by 89.4% [21]. Thus, for the purpose of this thesis, mp-MRI is the medical imaging technique used and it is further detailed in Section 2.3.

Over the years, several researchers have attempted to build systems, both hybrid and fully automatic systems, that could help humans diagnosing the disease. The development of these systems required different techniques, where the most common were machine learning algorithms. Endeavouring to diagnose the disease by detecting and grading lesions, early research covered prostate segmentation, lesion detection, lesion classification, and lesion segmentation. Some of these tasks led to the creation of challenges and competitions as a way to motivate the research on those topics. ProstateX [22] for clinical significance classification of lesions and the ProstateX2 [23] for lesion detection are some of those challenges, in which automated methods achieved promising results. In recent years, exponential growth in computer power, development of new machine learning algorithms and the integration of them with computer vision led to a significant improvement of computer-aided diagnosis systems (CAD).

The whole new approach, called deep learning, required no handcrafted features and was the source of significant improvements, not only in other fields, but also in cancer screening, with some systems able to outperform radiologists [24]. Chapter 2 explores previous research on the topic with Section 2.2 focusing on details regarding deep learning and Section 2.4 diving deeper into computer-aided diagnosis systems.

This thesis presents a research study on a recent dataset (IMPROD) [25] [26] that contains samples of mp-MRI images, biomarkers, segmentation masks and lesion scores for 157 men with prostate cancer. The dataset is further explored in Chapter 3. The two main problems addressed in this thesis are the classification of lesions, including PIRADS and clinical significance, and the detection and segmentation of the prostate cancer lesions. For these problems different experiments were performed on the dataset, firstly using a convolutional neural network (CNN), named XmasNet [27], to classify lesions regarding its clinical significance and a variation of this architecture to predict the PIRADS score and and for the other ones regarding the detection and segmentation of lesions on spatial images using two neural network architectures called 3D UNet [28] and a pretrained 3D ResNet-18 [29]. The latter experiments were evaluated in two different problems in an attempt to explore their capabilities with prostate segmentation before applying it to lesions. Moreover, the architectures have shown results in a variety of different image segmentation problems. Experiments are respectively detailed in Section 5.2 and Section 5.3 after introduction of the methods to be used in Chapter 4. Finally in Chapter 6 results are discussed and the thesis concluded.

2. Background

Ever since computational methods could learn from data, prostate cancer has been studied, with more recent techniques being applied directly to medical images. The interaction of these methods with humans has been evolving, from methods that required human fine-tuning and corrections after the prediction, to algorithms that can predict with enough quality to be considered automatic.

In this chapter, the current clinical practice for diagnosing prostate cancer is described. Moreover, the advantages and disadvantages of the current methodology are explored.

Deep Learning techniques are also introduced and contextualized with the requirements of the problem, based on previous research on similar topics. Furthermore, it is explained how multi-parametric magnetic resonance images work, and why they should be used to detect, segment and classify prostate cancer lesions, due to the differences in the contrast of those regarding soft-tissue appearance against common CT images.

Finally, previous research on computer-aided diagnosis systems for prostate cancer are discussed with previous deep learning-based CADs being mentioned and some details explored.

2.1 Prostate cancer: current clinical practice

Manually diagnosing prostate cancer is a challenging process that requires a considerable amount of expertise, diligence, and potentially several test examinations. Thus, the process is rather expensive and time consuming that has a high rate of overdiagnosis. Treatments are also frequently associated with a myriad of secondary effects with distinct types of severity.

The process of diagnosing and treating prostate cancer is frequently divided into four main phases. First, the screening where an expert tries

to assess by direct contact or blood analysis if cancer is present. Medical images are captured and analyzed to guide the third phase, the biopsy, which is used to confirm the presence and severity of cancer. The latter phase is rather invasive and may lead to potential harm. Finally, the treatment phase, where previously captured or new medical images can be used to guide the surgery. In the following paragraphs, these four phases are described and explained, and the potential automatization of those will be discussed, either to reduce the required resources, the required time, or to decrease the overdiagnosis.

2.1.1 Prostate cancer screening

Prostate cancer screening is used to detect suspicions of prostate cancer, however, it is also used in individuals without symptoms, but that might have an undiagnosed tumor [8]. Screening is usually the first step in the detection of prostate cancer, and it may improve the chances of early detection. The screening tools used by doctors are the digital rectal examination (DRE) and the analysis of the prostate-specific antigen (PSA) in the blood. And despite their frequent use, some experts disagree with their value, especially when thresholded with the potential harms of the overdiagnosis leading unnecessarily the patient to the subsequent phases [30]. Considering the risks and the advantages of these techniques [31, 32], patients frequently are invited to decide jointly with the responsible doctor if they should perform the screening exams or not [33].

A DRE requires a doctor or a specialized person to introduce a gloved finger into the rectum to feel the prostate in an attempt to detect anomalies with its size. Its use is frequent in clinical practice as the first screening test, however, several studies, institutions, and experts have questioned the use of this test [11, 10], whereas others recommend this test to be used as the second line of test, one that should be used after PSA tests [34].

Testing for the presence of the prostate-specific antigen can be useful to detect anomalies in the prostate, therefore, it is used as a screening method. The PSA is produced in the prostate and can be detected with a blood test and a low quantity of this antigen in the blood indicates that the prostate is healthy. Yet, prostate cancer is not the only potential problem to raise the quantity of this antigen in the blood [35], because prostatitis and benign prostatic hyperplasia can also be possible causes [36]. It is clear that PSA is not enough to diagnose prostate cancer, and that it might led to overdiagnose that is the cause of considerable harms due to further

unneeded treatments. Thus, it is also recommended to not administer the test to young men due to the fact that it would potentially diagnose cancers that did not require immediate or future interventions [37].

2.1.2 Image-guided prostate cancer biopsy

Generally, image-guided prostate cancer biopsy is performed when there is at least a minor suspicion that prostate cancer may be present, for instance, high values in the prostate-specific antigen blood test, or anomalies found in the digital rectal examination. It is an invasive procedure that consists of the introduction of a special needle to remove samples from the patient's prostate for further tests and exams. It requires the introduction of the needle either in the urethra, the perineum, or transrectally, however, the latter is the most frequent[38].

There are two distinct types of prostate biopsies, one where ultrasound is used to guide the intervention (TRUS), and another where MRI images are used. Despite being widely used, and frequently mentioned as the most common method [39, 16], TRUS biopsies have, since 2005, lost ground to MRI-guided biopsies due to a better soft-tissue resolution, and higher potential to characterize prostate cancer [17]. In fact, the MRI-guided biopsies' potential was already shown in practice, and its ability to detect more cancers [19] is verified by studies that confirm that it improved the detection of clinically significant prostate cancers by 17.7% [21]. Moreover, it also reduced the overdiagnosis by 89.4% [21], and the correlation between the biopsy and the pathology is higher when using MRI-guided biopsies [40].

As a consequence of the invasiveness of the procedure, some harm might be inflicted on the patient as a side effect. The most frequent side effects include blood in the urine, in 31% of the cases, rectal pain, burning when urinating, poor erections and urinary infection potentially requiring hospitalization [41, 42]. Due to this fact, in some cases, where there were previous negative biopsies, experts may recommend only the analysis of an MRI image if PSA values continue increasing [43].

One of the areas that can benefit most of the automation is the analysis of medical images of prostate cancers. Firstly, the process of having a doctor annotate MRI images manually, either to guide a biopsy or to analyze the cancer lesions, is expensive and time consuming. To extend the implementation of this technique is crucial to reduce the costs, and here automation can assist if it succeeds to annotate the images with the

same precision of the expert. Furthermore, if the annotation quality is increased by an automated tool, it might represent fewer biopsies, and fewer treatments, by reducing overdiagnosis. The reduction of these two interventions potentially could greatly reduce the risk of side effects and harm to the patient.

2.1.3 Prostate cancer treatment

Treating prostate cancer is challenging and often it requires a combination of surgical and non-surgical interventions [44]. Cancers in a more advanced state sometimes spread to other parts of the body near the prostate, endangering more organs and creating difficulties in the treatment. Therefore, these require special treatments such as hormonal therapy and chemotherapy, with minor exceptions, for example, when there is a limited amount of metastasis, sometimes radiation treatment is used [45]. Other less advanced tumors can be treated with surgery, external beam radiation therapy or cryosurgery. However, it is not wise to combine radiotherapy after a failed surgery since it may cause other problems (e.g. different cancers) [46, 47]. Both these procedures have similar prospects for the side effects after a five-year period [48].

Despite all the available treatments, radical prostatectomy is the main treatment for prostate cancer. Its effectiveness has been growing in recent years, with robotic-assisted procedures not only being available and common [49] but also reducing the stay in the hospital [50]. Both this procedure and radiotherapy have significant side effects that can inflict harm, to the patient, that may considerably affect his life. Within the side effects, we can include stress urinary incontinence and erectile dysfunction. The latter affects nearly all the patients that undergo treatment [51].

2.2 Deep Learning

Deep learning is the usage of deep neural networks (i.e. a neural network with several hidden layers stacked) to tackle problems ranging from natural language processing to computer vision and speech recognition [52]. It has been growing throughout the years, since the publication of the paper describing an architecture of a deep convolutional neural network called AlexNet that won the ImageNet challenge in 2012 by a considerable margin against traditional methods (e.g. previous state of the art algorithms)

Figure 2.1. Examples of use cases of deep learning with pose estimation, object detection and object segmentation with the respective models KeypointRCNN, MaskR-CNN and DeepLabV3.
Image from the following blog post - https://pytorch.org/blog/torchvision03/

[53]. The performance in the challenge of these networks is now superior to the performance of humans [54].

Further research showed that deep learning was not only able to outperform previous models and techniques on classification problems, but it was also able to become the state of the art in a myriad of other problems such as the translation of a text to other languages, the detection [55, 56] and the segmentation [57] of objects in images. Figure 2.1 shows examples of some computer vision tasks and models, such as pose estimation, object detection and segmentation. These models can be used to detect a variety of objects from images, with slight variations, such as drawing a bounding box around an object or creating a segmentation mask for the object. There are plenty of different applications for these models, and they range from biomedical applications [57] to autonomous vehicles [58] and video description [59].

From these methods, the segmentation is the one with more successful applications in biomedical domains, such as segmentation of cancer lesions [60], which is one of the main problems, applied to prostate cancer, that this thesis tries to solve. This section and respective subsections focus on the necessary deep learning foundations to be able to solve the proposed problems.

2.2.1 Artificial neural networks

Some computational algorithms were built based on and inspired by a specific biological system. One of these algorithms, called Artificial Neural

Networks, was inspired by the nature neurons present in biological systems like the human brain.

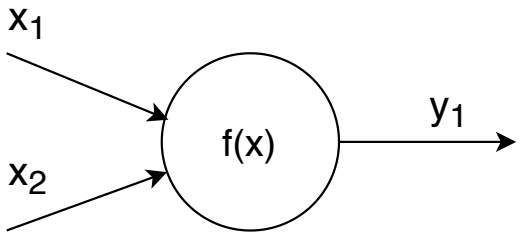


Figure 2.2. An Artificial Neuron (also known as perceptron), basic unit of neural networks, with 2 input values, and one output value subject to the function f(x).

Similarly to what it is possible to observe in nature, an artificial neural network has a basic unit, the artificial neuron. However, for historical reasons, the unit is sometimes called the perceptron and it is shown in Figure 2.2. Several of these artificial neurons can be stacked either vertically or horizontally in order to create more complex compound units able to learn how to approximate nonlinear functions.

To interact with the perceptron, it is necessary to feed it input values $\vec{x}$ to which it is applied a function $f(\vec{x})$ that originates output values $\vec{y}$. Either $\vec{x}$ and $\vec{y}$ can be represented with one or multiple values. As mentioned previously, these structures learn how to approximate nonlinear functions, however, $f(\vec{x})$, the applied function, is just a representation of a linear function $f(\vec{x}) = w^t\vec{x}+b$, where both b and w are learnt values, denominated respectively as bias and weights. Therefore as an attempt to approximate those nonlinear functions, researchers introduced nonlinearities after each function, called activation functions. Some of the most popular activations functions are the Tanh function, the Sigmoid function, Softmax function and the rectified linear unit (ReLU) [61], and while those are broadly used some other activations, such as Mish [62] and Leaky rectified linear unit (Leaky ReLU) [63], have been used recently, showing promising results in some use cases.

$$\begin{aligned} \vec{y} &= \phi(f(\vec{x})) \\ &= \phi(w^t\vec{x} + b) \end{aligned} \tag{2.1}$$

As a result of including this activation in our artificial neuron, the output

is now given by the expression shown in the equation 2.1.

$$\text{ReLU:} \quad \phi(x_i) = \begin{cases} 0, & \text{if } x_i \leq 0 \\ x_i, & \text{if } x_i > 0 \end{cases} \tag{2.2}$$

$$\text{Sigmoid:} \quad \phi(x_i) = \frac{1}{1 + e^{-x_i}} \tag{2.3}$$

$$\text{Softmax:} \quad \phi(x_i) = \frac{x_i}{\sum_j e^x_j} \tag{2.4}$$

Equations 2.2, 2.3 and 2.4 represent three of the most popular activation nonlinearities that are useful and important to the scope of the problems addressed in this thesis. Each one of these activations has particularities, either in the range of the output, the computations needed to computed and the derivative, or the circumstances where they are seen as particularly useful. ReLU, equation 2.2, has been used frequently as the activation that follows a convolutional layer (see subsection 2.2.3) since it has been shown that they usually improve deep neural networks training [64]. The computations needed to calculate this activation and its derivative are minimal, and the output ranges from $[0, \infty[$. While the ReLU range does not limit a maximum value to the output, both sigmoid and softmax functions limit the output to a value in the interval $[0, 1[$. The sigmoid activation, equation 2.3, is particularly useful and popular for binary classification problems, for instance, when classifying if an image is from a dog or not, the output, between 0 and 1, can be interpreted as a probability of being a dog. On the other hand, the softmax activation, equation 2.4, is widely used on multiclass classification problems, due to two particular characteristics. First it is its range, and secondly the fact that it is a function of all the output values, in a way that the sum of the softmax of all the output values must sum to 1. Thus enabling it to be interpreted as a probability of belonging to that class, in other words it normalizes the outputs of one node based on the value of all nodes.

In Figure 2.3 it is possible to observe the behavior of the three nonlinearities and compare them with the linear function. It is also worth noting that despite having the same scale on the x-axis in all the plots, each of

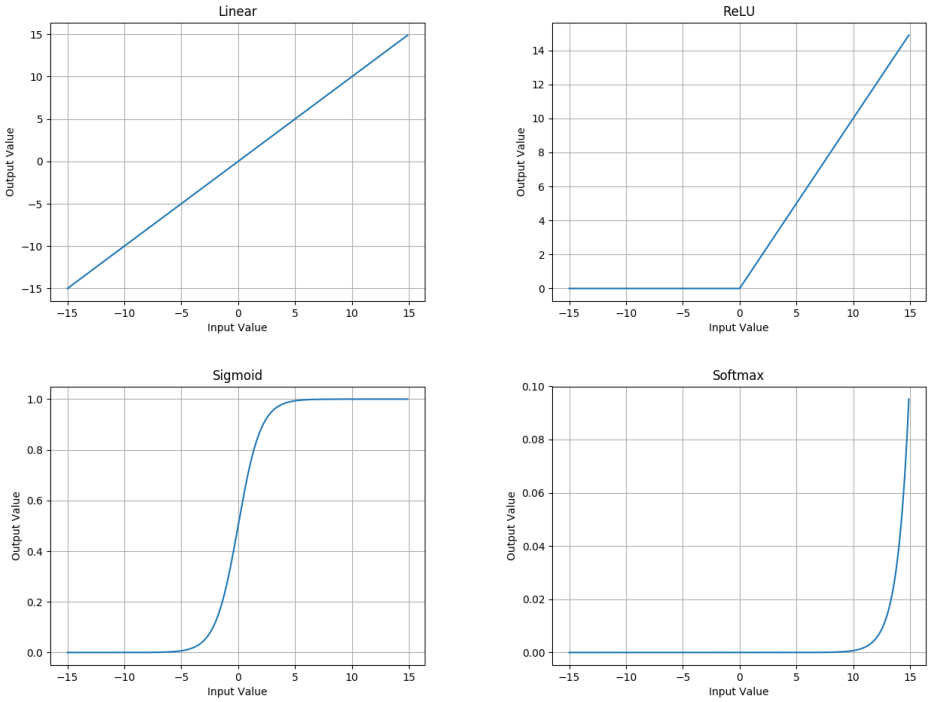


Figure 2.3. Behavior of the ReLU, Sigmod and Softmax activation functions compared with the linear function. It is worth noting the scale of the y-axis in each plot.

them has a specific scale for the y-axis. This is of particular relevance when comparing ranges of functions, their progression and the impact that it may have on training (e.g. what if all the values are 0 before being given to a ReLU nonlinearity). However, despite allowing the network to learn more complex functions, the introduction of nonlinearities transforms the optimization problem in a non-convex optimization problem, meaning that it is now more complicated to optimize when compared to a convex problem [65].

An artificial neural network is not only characterized by the number of input values, the number of hidden layers with the respective number of nodes and the number of output values, but also by how these layers and each node is stacked and connected to the others. One of the oldest and more popular neural networks is the feed-forward, and it contains several layers, the input, and output layers and also a stack of one or more hidden layers. However, each layer has also a stack of artificial neurons (or nodes), increasing the number of nodes increases the complexity of the learned function but also makes the training process more difficult. The same happens when the number of hidden layers is increased. The layers of this network are usually denominated as fully-connected layers, as a result of the fact that every node of the layer n feeds its output to all the nodes of the layer $n + 1$ as input. In other words, this means that a simple network

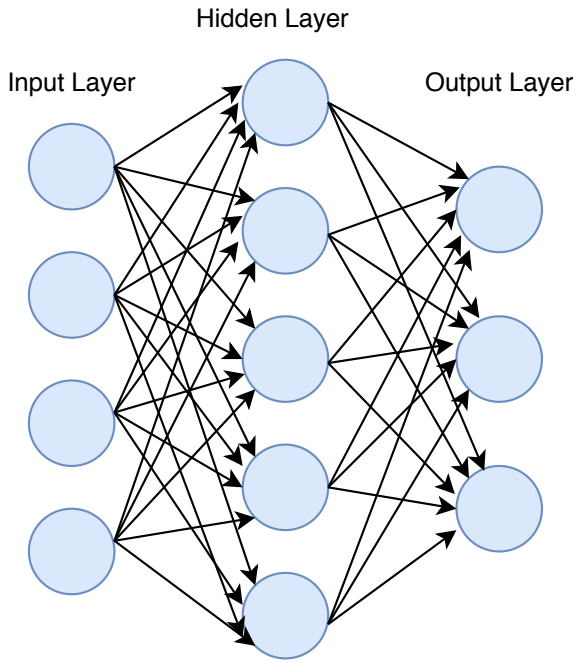


Figure 2.4. A simple fully connected (also known as feed forward) artificial neural network with 4 input values, 1 hidden layer composed of 5 neurons and 3 output values.

as seen in Figure 2.4 contains, if all the bias are ignored, 35 parameters to be learned in the optimization process.

2.2.2 Optimization of neural networks

Due to its non-convex nature, artificial neural network optimization cannot be solved by an analytical mathematical expression. Instead, an algorithm that is a special case of reverse accumulation, designated back-propagation, must be used [66, p. 217-218]. This algorithm solves the values of the parameters (weights) of a neural network in an iterative two-step process, the forward and the backward step. Despite being independent algorithms, back-propagation often takes advantage of the gradient descent algorithm to update the weights and try to find a solution [66, p. 200]. Despite the fact that configuring the hyperparameters carefully usually allows the algorithm to get a solution that has the reqired performance in practice, there is no guarantee about finding an optimal solution or even a solution that is close to the optimal. Optimization of neural networks is usually called the training of the network, and only during training, the backward step is computed.

In order to know how to update the network, and to be able to integrate gradient descent, it is necessary to give an objective to the optimization process. Therefore, during training, besides the input, $\vec{x}$, and the output, $\vec{y}$, it is necessary to have a label representing the expected value for $\vec{y}$ [68, 69].

Table 2.1. Loss functions and the type of machine learning problems where they are frequently used.

Name	Problem
Cross-Entropy loss	Classification
Logistic loss	Classification
Dice loss	Image Segmentation
Focal Loss[67]	Detection
Mean Square Error	Regression

Frequently this label is mentioned as ground truth. Some optimizations do not require labels (e.g. unsupervised learning) [70], however, for the scope of this thesis, only supervised learning is covered. In supervised learning problems, during training, the objective is to obtain an output each iteration closer to the ground truth, and the closeness of both is computed using a loss function. Loss functions generate what can be seen as a numerical error that can be used to analyze the performance of the model. A proper optimization algorithm can, through the minimization of the error, update the weights of the network in a way that will better approximate the optimal function and generate results closer to the ground truth.

There is a myriad of different loss functions, each of them having different properties, advantages, and disadvantages, being arguably more appropriate to one type of problem than to others. Some of these can be seen in Table 2.1 with a reference to the type of machine learning problems where they are frequently used.

The forward step was previously described, and it is the process used to generate an output from an input. After computing an output during training, the next step is to compare it with the ground truth through the calculation of an error, using, as mentioned before, a loss function. The backward step then takes advantage of a dynamic programming technique to compute, with the chain rule, the gradients of the input with respect to the loss function. This process goes from the last layer to the first, computing each gradient in one whole iteration, and, for instance, does not repeat redundant computations in order to compute the gradients of some layer [66, p. 200-220]. The gradients are then fed to a gradient descent algorithm that will perform a gradient step, which is used to finally update the weights [66, p. 200]. Different variations of gradient descent can be used, such as mini-batch gradient descent, which uses a small batch of the training set at each gradient step, or stochastic gradient descent, which

uses only one sample per gradient step [71, p. 303-307]. These variations frequently reduce computation cost (e.g. not all the data is loaded in the memory or in the GPU) and help to avoid local optimums, however, the convergence rate decreases [72, p. 351-368].

In the inference phase, no gradients are computed and new samples are given to the network. The network must be able to generalize and have error margins in never seen data close to the ones given in the training set, after training. When this does not happen it is said that the network overfitted the training data and strategies to avoid this problem range from reducing the complexity of the network to adding a regularization term to the loss function or increasing significantly the number of training samples [73, p. 1-12].

2.2.3 Convolutional neural networks

The application of deep learning and deep artificial neural networks to computer vision tasks led to the detection of a problem with feed-forward models. Even small images have a large number of pixels, for example, a 128x128x3 (height x width x color channels) image contains $128 * 128 * 3 = 49,152$ pixels, which represent the number of nodes in the input layer. The effect of these pixels get worse when we consider that the next layer will connect all the nodes to these input nodes, and that complex tasks require more nodes. Thus, a model that receives this image has 100 nodes in the second layer and five outputs totaling $49,152 * 250 + 250 * 5 = 12,289,250$ parameters to be optimized. Optimization of such a large number of parameters is rather complicated, despite the fact that the network shown is not even a deep neural network.

To construct feasible networks that work with images as input, researchers started replacing most of the fully-connected layers with convolutional layers. Networks with these layers are called convolutional neural networks (CNN). Inspired by biological receptive fields, the CNN architecture mimics, to some extent, this feature of the animal visual cortex [74, 75]. The aforementioned receptive field is used by animals as detectors, to detect special characteristics in an image, such as edges. The approximation can be achieved with the convolutional operator [76].

A convolutional layer, the main building block of convolutional neural networks, is typically defined by a set of filters/kernels, which are the representation of their weights. Despite going through all the input, these filters are usually considerably smaller than the input size. In order

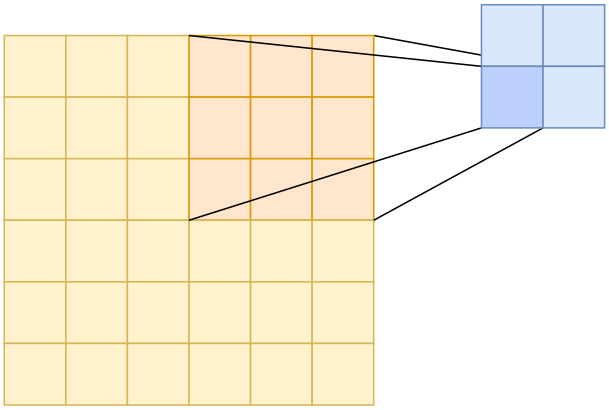


Figure 2.5. Visual representation of a convolution happening in a 2 dimensional input of size 6x6. This convolution has dilatation of 0, stride of 3, padding of 0 and a kernel size of 3. The output is of size 2x2

to detect the spatial characteristics of the input, each kernel is used to compute the dot product between the kernel and the image, throughout the entire width and height, generating an activation matrix similar to what can be seen visually represented in Figure 2.5. This process can also be mentioned as a convolution operation. Furthermore, it also includes a third dimension, the channels where the size of that dimension represents the number of stacked filters to be used in the convolution. Using this dimension is what allows one convolutional layer to be able to learn several features in the spatial representation of the input in one single convolution. These characteristics might be useful for future layers to understand and learn more complex spatial shapes.

Besides the filters dimension and the number of channels, the output shape of these layers is controlled by other hyperparameters that require careful attention and tuning to approximate. The three hyperparameters are dilation, stride, and zero-padding.

- **Stride** - Represents the movement of the filter along with the image height and width, it affects the size of the output since a larger stride will represent a smaller sized output [77]. If the filter is seen as a sliding window that moves around the input, the stride is the step size, in terms of pixels, that the sliding window should move. If the stride is smaller than the size of the kernel, it means that there will be overlapping between activations in the output, otherwise, the same input pixel will not be in more than one activation.

- **Dilation** - The filter usually convolves on a specific size of consecutive pixels, however, sometimes it is useful to utilize a specific type of

convolutions, called dilated convolutions [78]. This parameter, dilation, represented by an integer, is the spacing between each pixel to be convolved, i.e. a regular convolution has dilation of 0 since there are 0 pixels of spacing between each pixel to be convolved in the input.

- **Zero-padding** - Some use cases require that a padding of 0's is added to the input, this increases the size of the input and is usually used to control the spatial size of the output, for example, when it is required that the output size is the same of the input size.

$$Dim(Y) = \frac{Dim(X) + 2 * P - D * (K - 1) - 1}{S} + 1 \tag{2.5}$$

The output size $Dim(Y)$ can be written in function of the input size $Dim(X)$, the filter size K, the stride S, the dilation D and the padding P by the expression given in the equation 2.5 [79].

In convolutional layers, the optimization process does not try to optimize the weight of every single neuron, what it does is optimize the existent values in the sliding window. There is an assumption that if the values of the kernel are able to acquire information at some spatial position, they should be able to do the same in the others. In other words, it means that if a convolutional layer is composed of five channels, each with a filter of size 4x4 and it gets as input a 128x128x3 image, the number of parameters of the layer will be $5 * 3 * 4 * 4 = 240$. This characteristic, that is mentioned as parameter sharing, contributes to two main aspects that might improve the computational performance and the results obtained. The first one is that it does contribute to the translation invariance characteristic that is frequently associated with CNN. Secondly, they require considerably fewer parameters, easing the optimization of the network, and allowing its design to be deep (i.e. include more layers) [80]. The optimization of convolutional neural networks can be done, similarly to feed-forward networks, with backpropagation and gradient descent [81].

Besides convolutional layers, convolutional neural networks frequently have max/min/average pooling layers that try to reduce the spatial representation of the image by using a sliding window and max, min or average operations to pick the output value. These layers are particularly useful to reduce the size of the input, however recently researchers have discarded

them [82] or reduced the size of their filters [83].

To finish a convolutional network architecture, usually, one or more fully connected layers are added to do the high-level reasoning. This works similarly to what was shown before regarding artificial neural networks, including the activations. Including several of these layers might have a negative impact on the computational cost, thus usually only one to three are added after several layers of convolutions and poolings.

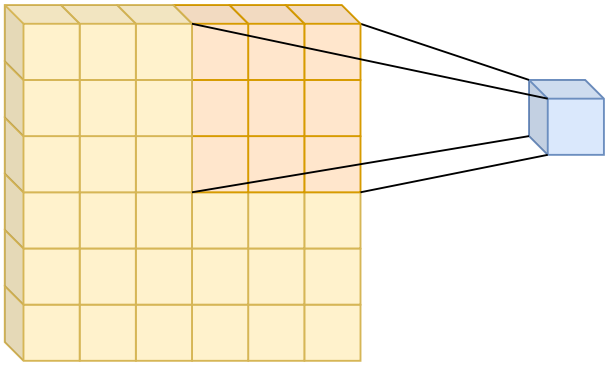


Figure 2.6. Visual representation of a convolution happening in a 3 dimensional input of size 6x6x1. This visualization shows the capture of a single activation using a kernel size of 3x3x1. The full output is not displayed.

While this subsection focused on convolutional neural networks working in two dimensions, the architecture is not limited to it. Convolutions on three dimensions are possible, and they work similarly, however instead of a sliding window (2-d kernel), it is used a sliding cube (3-d kernel). In Figure 2.6 it is possible to observe a simple case of how these convolutions work to generate one entry of the output matrix.

2.2.4 Image segmentation

Similar to other research fields, computer vision has also benefited from the rise of deep learning-based methods. Not only previous results were improved in a myriad of tasks and use cases, but also the process was simplified due to the fact that deep learning does not require manual feature extraction. One of the most frequent topics to be mentioned in research articles is image segmentation, that is other words can be described as the process of dividing an image, given as output, into segments. Some images can have their analysis simplified simply by changing their representation to something with more significance to the problem [84, 85].

Segments can vary accordingly to the problem, however, frequently the main use of image segmentation is to find segments that contain some object, and to draw its boundaries. For instance, if a problem consists of

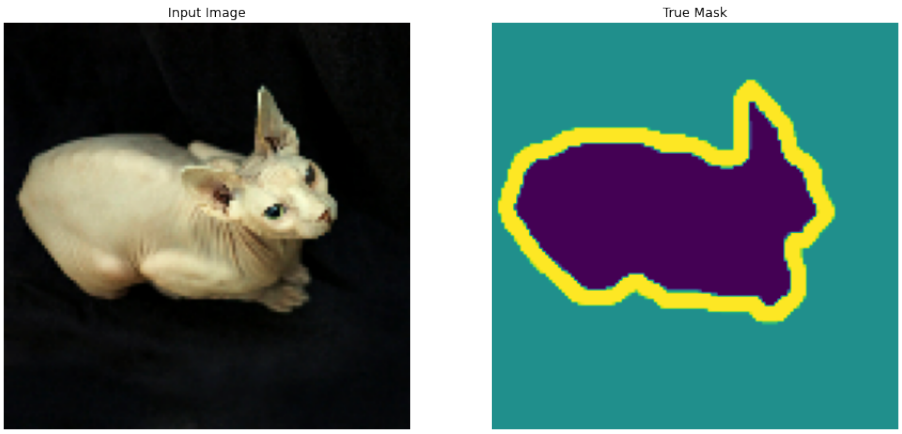


Figure 2.7. Segmentation example, where the boundaries and the animal image are segmented as separated segments. On the left it is possible to observe the original image, on the right the 3 different segments, the background (blue), the boundaries (yellow) and the cat (purple). Image from https://www.tensorflow.org/tutorials/images/segmentation

segmenting the ball from a football game picture, the idea is that the output includes the same label in each pixel where the ball is represented and another label (e.g. a background label) in the other ones. This representation not only gives the position of the ball, as it gives its boundaries too. Figure 2.7 shows an example of a segmentation problem, where the goal is not only to segment the cat, but to segment separately the boundaries between the cat and the background.

The interception of image segmentation with other research areas also led to significant advances and some even more promising future prospects. One of the areas, where it has been applied to, was medical imaging [86, 87], with several uses cases, such as detecting and segmenting tumors [88], surgery planning and studying the anatomical structure to detect, and diagnose problems [89].

2.2.5 Data augmentations for biomedical images

Training a regular artificial neural network takes a considerable amount of time and data, and some specific areas, such as deep learning, where the neural networks also extract the features from the data, it is required even larger volumes of data. This is also true to avoid problems like overfitting, which might hurt the performance of the network when it is exposed to never seen samples, due to the fact that the complex model fitted almost perfectly the testing data, decreasing the generalization capabilities. Moreover, it is important that the data has representations of a great part of the input data distribution, so it can properly approximate the real distribution, in other words, when trying to segment prostate

lesions it is important that the model trains on different tumors shapes, sizes, and aggressiveness.

Constructing a large medical dataset is an almost impossible task due to diverse factors. To start with the labeling of the samples must be precise, accurate and correct, which requires expert knowledge. Nevertheless, not only the price of these radiology experts is rather expensive, the process of manually annotating the labels in a medical scan is time-consuming, and a complicate process. Hence, the datasets provided by these clinical experts are frequently small in size. Furthermore, recently the restrictions imposed by privacy laws created extensive problems for the investigation labs, health institutions or clinical trial lab to combine their data and joint efforts to create a large meaningful dataset for deep learning. Comparing to other Computer Visions datasets with millions of available samples for training, the frequently less than one thousand samples in medical images datasets shows the lack of proper data in these problems. Moreover, medical images are frequently represented with one color channel, which can be seen as one extra challenge to the learning process.

There are, however, techniques that were created as an attempt to bypass this problematic lack of data. The technique with the most success cases is called data augmentation, or more specifically image augmentation when referring to image data. Since one of the advantages of having a big dataset is to have a high variation in the data which drives the network to focus on the important features instead of focusing on specific of each image, these techniques apply small random transformations to the data. The application of these augmentations can be done statically and applied to all the dataset, however, the best results have been attained with generators, that will apply on the fly the augmentations to an image randomly varying the magnitude and which transformations are applied. Generators usually help to further tackle the overfitting problem.

Several transformations that can be applied to the data, however, not all of them are appropriate to be used in medical images. Considering the impact that the features of medical images are important, and should be preserved, the most frequently used augmentations include the use of linear transformations. These can be divided into rigid and non-rigid transformations, with the latter not keeping the image shapes unchanged. Some of the rigid transformations are translation, rotation, flipping.

- **Translation** - Shifting of the image towards some direction, changing

the absolute position of the image elements.

- **Rotation** - Rotating the training image by a randomized number of degrees.

- **Flipping** - Due to the fact that anatomical structures are, in several cases, symmetric, this technique mirrors the image either vertically or horizontally. This also tackles the lack of variety of the dataset, in the sense that images givens could, without any particular medical reason, have the region of interest located more frequently in on side of the organ.

Examples of the non-rigid linear transformations are stretching and shearing.

- **Stretching** - Zooming in and out with different ratios stretches the image, and changes both absolute and relative features of the image. However, this transformation must be used carefully since if its magnitude is too high it can hurt the model performance instead of improving it.

- **Shearing** - While stretching works in one direction, shearing is a similar transformation that consists of moving the top and the bottom, or the left and the right side of the image in different directions. It can be seen as stretching in two directions.

Finally, in an attempt to improve the generalization of the model across different medical imaging capture machines, it might be useful to apply a transformation that will vary the intensity of the gray-scale pixels in the image. This also helps to optimize the model in a more robust way when the limited dataset had all the samples captured by the same machine.

2.3 Multi-parametric MRI

The complexity of the task of detecting, segmenting, or diagnosing disease from medical images is due to several factors. First of all, not only it requires appropriate expert knowledge so the image is properly analyzed, but the results are affected by the type and quality of the image being visualized. In other words, not all techniques to capture an image are

adequate for the visualization and accurate diagnosis of either prostate cancer lesions or the classification of the aggressiveness of those lesions.

For prostate cancer, two main imaging techniques have been used over the years. Ultrasound (US) scan was the main technique to diagnose prostate cancer, however, in recent years, Magnetic Resonance Imaging (MRI) started to have its use increased in several countries. One of the reasons for this increase is that ultrasound images have poor soft tissue resolution whereas magnetic resonance images show better resolutions [17]. Moreover, MRI has been used to help doctors assessing other details, such as the difficulty of the surgical procedure, and which direction should they follow when planning the surgery [90].

To capture an image of the organs of the body, magnetic resonance images rely on radio waves and in strong magnetic fields. Since it does not use radiation or ionization it can be seen as a better solution when compared to alternatives like Computed Tomography and X-Rays. However, MRI requires longer capture times, which can become an obstacle when this technique becomes widely used on a higher number of patients. The mapping of the organs is based on the detection of water and fat in the body through the interaction and excitation of hydrogen atoms present in those biological tissues. Small changes to the configuration can generate different types of images with a huge variety of contrasts and use cases applied to different diagnoses, these different types are frequently called sequences. When a magnetic resonance image is composed of two or more sequences, it usually is called multi-parametric magnetic resonance image (mpMRI) [91] [92]. Some of the most common sequences are T1 weighted (T1W), T2 weighted (T2W), diffusion-weighted image (DWI) and apparent diffusion coefficient (ADC). Each of these sequences captures different details, therefore they have different uses based on the problem to analyze and diagnose (e.g. DWI and T2W are frequently seen in prostate cancer studies).

2.3.1 PI-RADS

Prostate Imaging Reporting and Data System, also known as PI-RADS is a set of standards to capture and report prostate cancer images with mpMRI and assess from them the risk of clinically significant cancer being present. The first version of these standards was especially focused on the clinical significance classification [93], whereas the second version, PI-RADSv2, focused also in creating global standards for MRI [94]. Overall,

both versions aimed to improve the quality of image capture and reporting.

A plethora of studies have tried to assess the impact of these standards across the entire workflow of a prostate cancer diagnosis. Both versions have shown positive results in classifying clinically significant prostate cancer lesions [95, 96], however, some limitations regarding the classification of small ($\leq$0.5mL) significant lesions (GS$\geq$4+3) have been noticed [96]. Furthermore, PI-RADS has proved to be useful in other applications as the detection of the extension of cancer outside of the prostate [97] which is an important step in the staging of the cancer. Predicting when the active surveillance termination period should be defined, based on the aggressiveness of lesions, is another of the uses of this set of standards and evaluation criteria [98].

Table 2.2. PI-RADS scores [99]

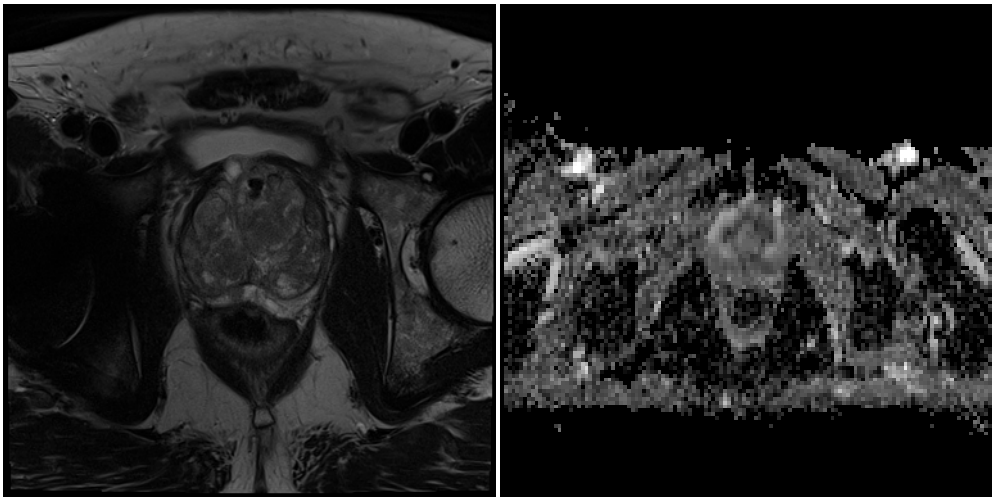
PI-RADS	Probability of clinical significance
1	Very low
2	Low
3	Intermediate
4	High
5	Very High

Table 2.2 shows the scores that compose the PI-RADS score system and their meaning in terms of risk of the cancer being clinically significant.

2.3.2 T2W

After excitation, tissues affected follow a relaxation process, on T2 weighted a specific relaxation process, by the name Spin-Spin relaxation, is used in order to generate the image. This process of capturing the images is based on the decay of the transverse component of the magnetization vector to an equilibrium state. The decay is exponential and the time necessary to reach the equilibrium given in function of a constant known as spin-spin relaxation time, or T2 [100, 101]. The constant T2 (given in the order of seconds for protons) varies, and each biological tissue has its own, for example, water-based tissues range from 40 to 200 ms whereas fat-based tissues range from 10 to 100 ms.

Variation in the capture time of different tissues generates different signal intensities, therefore it is possible to observe a contrast between some tissues and body organs based on what is in their composition. Some of the high signals in T2W images are given by water tissues [102], such as



(a) T2W sequence. ProstateX dataset [22].

(b) ADC sequence. IMPROD dataset [25, 26]

Figure 2.8. Visualization of T2W and ADC slices, from a mp-MRI of the prostate, side by side.

inflammation, edema, tumors or infections [103] whereas bone, air, protein-rich fluids, and fibrosis originate low signals [102, 103]. The representation of a high signal is a white, brighter pixel while a low signal is represented by a darker pixel. In Figure 2.8a it is shown an example of a T2W image of a prostate MRI.

2.3.3 DWI-ADC

The diffusion of water molecules in tissue is constrained and affected by the interaction of water molecules with different components of that tissue, such as fiber and proteins. Images are generated through the analysis of water diffusion patterns that are used to map the microarchitecture of biological tissues into an image with the appropriate contrast to detect and diagnose diseases and anomalies [104, 105]. Images generated with this technique are called diffusion-weighted magnetic resonance images (DWI).

The composition of a tumor, frequently highly cellular, which originates more constraints to water diffusion allows this type of pathology to originate a high signal in DWI images [106]. Hence, this sequence type is frequently used to detect, stage and monitor cancer tumors.

Standard diffusion-weighted images have an inherent relation with T2 weighted images, which in some cases has a possible negative impact on the image capture. For instance, lesions that do not restrict water diffusion are not supposed to have a high signal, however, if their T2 relaxation time is long then the pixels will be bright [107]. To reduce this effect, a special type of DWI, called apparent diffusion-coefficient (ADC), is used. The

capture of an ADC image requires several diffusion-weighted images to be captured with different weights, the rate of diffusion is then calculated from the change in the signal between weights. Despite the similarities and the relation between ADC and DWI, the second uses bright pixels to represent constraint movement, whereas ADC uses dark pixels for that representation [108]. In Figure 2.8b it is possible to observe an example of a ADC image of a prostate MRI, and it is also possible to observe that this image has frequently lower resolution when compared to T2W images, as can be seen by comparing with Figure 2.8a.

2.4 CAD to prostate cancer

Several researchers have attempted to develop solutions that would help to diagnose prostate cancer. These solutions varied greatly in their methodology, not only in terms of deep learning models, but in the way that they saw the problem. For instance, some solutions tried to classify the clinical significance of lesions from diffusion-weighted images using a dataset with 427 patients [109]. Others directed their efforts to segmentation problems from these magnetic resonance images.

Since it is an easier problem, considerable research was conducted for the prostate segmentation with varying results regarding different datasets and methodologies. For this problem, some researchers used only T2W sequences, either from public datasets such as PROMISE12 [110] or private [111]. In an effort to further increase the performance of single sequence segmentation, some other researchers co-registered both DWI/ADC images with T2W images. While in some cases this was the only approach [112], it is possible to see in others that the co-registration, in fact, increases the performance when compared to single sequence inputs [113].

Some authors decided to conduct more extensive experiments and efforts, endeavoring to detect and segment prostate cancer lesions from these images. The results reflected the hardness of the task at hand, and in some cases, neither of the independent sequences or a combination of both was enough to surpass a dice score of 0.60 [113]. However, there has been some progress with some solutions obtaining a dice score of 0.64 which is considerably closer to the reported dice score for two expert clinicians of 0.67 [112].

A myriad of other techniques was attempted ranging from traditional machine learning with features extracted and treated manually by re-

searchers, probabilistic methods, and others that required post-processing of the output by experts. Since the main focus of this thesis is the use of deep learning based methods, the focus of this section was in those methods previously developed by other researchers.

3. Improd Dataset

Recent progress in machine learning and deep learning research has been backed up by the improvement of public datasets for a myriad of domains. These datasets allow a more objective evaluation of the methods since they are all tested on the same data. However, that is not the case with datasets containing prostate magnetic resonance images, since those do not combine high image and annotations quality and different types of annotations. Therefore, for this project, the existence of a private dataset, IMPROD, reuniting all this characteristic is of greater importance, even if the objective comparison with other work might be slightly affected. Moreover, not only the quality is important, but the size and some other statistical details of the dataset are important.

The analysis of a dataset includes several questions to be answered such as "What is the source of the dataset?", "How was the dataset annotated?", "Why is this dataset valuable for the problem that is being solved?" and "How do those annotations align with the experiments necessary to solve the problem?". Furthermore, since this dataset contains medical images it is also important to discuss the expertise of the annotator and the quality of the dataset. Moreover, studying if the data available is representative of its real-world distribution that would allow generalizing to the deployment of the model.

This chapter discusses all the previously mentioned questions addressing, in particular, the IMPROD dataset and its characteristics. Further important statistics to set up, interpret and evaluate experiments using the methods detailed in the previous chapter are analyzed thoughtfully to understand the weaknesses and strengths of this particular dataset in the context of this thesis.

3.1 Source of images and annotations

The IMPROD (Improved Prostate Cancer Diagnosis - Combination of Magnetic Resonance Imaging and Biomarkers) dataset originated from a clinical trial conducted in a joint effort by the Turku University Hospital and the University of Turku. The trial included 175 patients between 40 to 85 years with suspicions of prostate cancer supported by screening results (i.e. abnormal DRE or 2,5-25ng/ml PSA in two measurements).

In order to maximize the quality, the dataset was, not only gathered using images captured with recent and high-quality magnetic resonance scanners (Magnetom Verio 3T, Erlangen, Germany), but it was also carefully annotated by experts in the field that belong to the institutions conducting the clinical trial. This maximization of the data quality due to the techniques used denotes an attempt to improve the outcome of machine learning and deep learning solutions in a myriad of prostate cancer-related problems.

From the possible image sequences captured by a magnetic resonance machine, the selected to compose the dataset are T2W and ADC sequences captured from five diffusion-weighted images with a b value varying between 0 and 500. Both of them are accompanied by the respective prostate mask, and if there are one or more lesions, by the lesion masks. Distinct lesions have different independent masks, even if the patient and the magnetic resonance image are the same. Each sequence was manually segmented to construct the mask of the lesions and the prostate under the premise that despite requiring twice the necessary work, it generates better quality annotations. Concerning the dimensions of each sequence, T2W is the largest of them with 360x360 with a varying number of slices. ADC, on the other hand, is considerably smaller with only 128x128, also with a varying number of slices.

Besides images and corresponding masks, metadata of all lesions of every patient is also gathered to facilitate development of automated methods. It includes scores interpreted from magnetic resonance images, such as the Gleason Grade Group and both the PIRADS and Likert scores. Moreover, a Gleason score given is also provided allowing for potential studies on the overdiagnosis problem too. The analysis and the classification of lesions can be done using several approaches and interpretations of this data. In this project, since, lesions labels indicating clinical significance are not given directly in the dataset, they are inferred based on the provided details

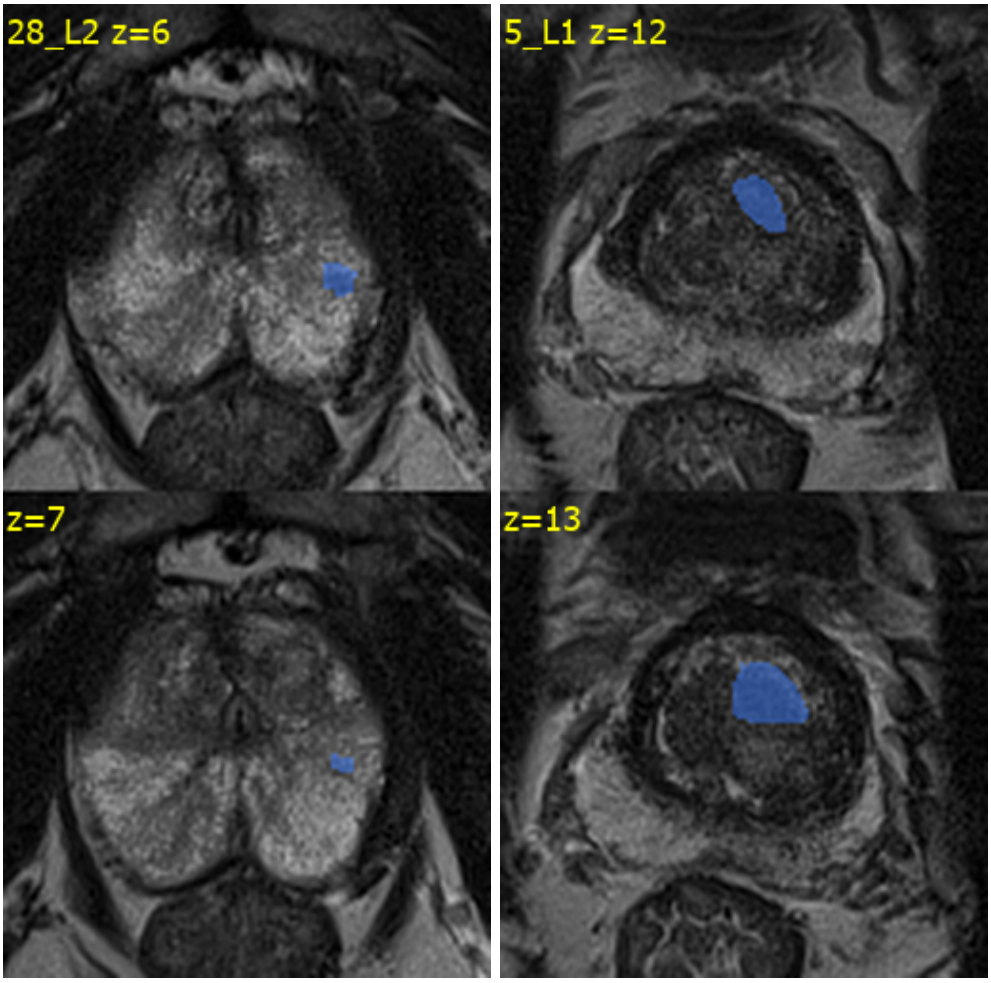


Figure 3.1. Example extracted from the IMPROD dataset of lesions masks over a T2W multi-parametric magnetic resonance image. The figure on the left is the second lesion of the patient identified as 28, with upper and lower images representing respectively the consecutive slices 6 and 7. The figure on the right is the lesion number one of the patient identified as 5, with upper and lower images representing respectively the consecutive slices 12 and 13. Note the lesion size difference between patients and between slices of the same patient.

concerning each lesion. Similar to the aforementioned mask annotations, experts from the institutions involved in the clinical trial were responsible for given the scores of each lesion.

3.2 Clinical Interpretation of the data

Interpreting the data from a clinical perspective is crucial to understand and solve the problem. Notwithstanding the fact that deep learning architectures implicitly learn feature extraction, other details must be considered in order to properly set up the experiments. For instance, what the meaning and the reason behind the distinction of a lesion as clinically significant, or the three-dimensional representation of the data, and how masks are visualized.

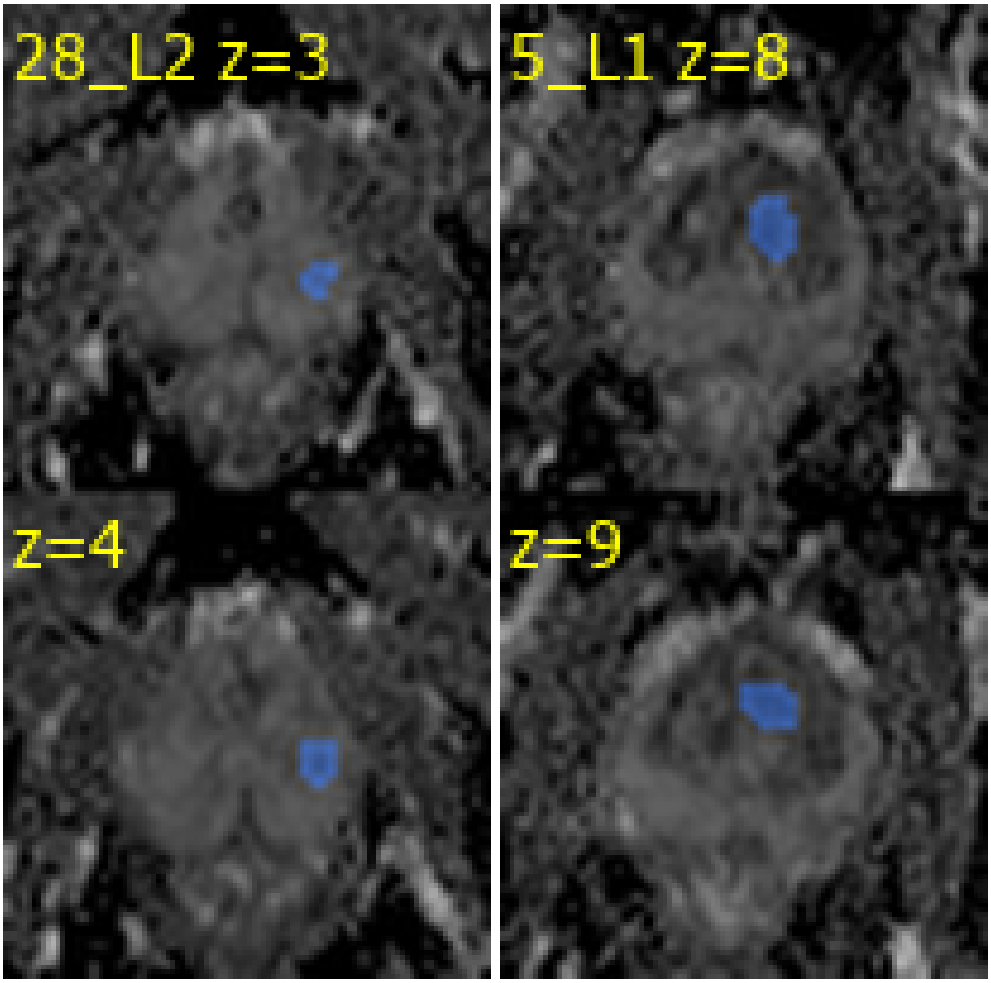


Figure 3.2. Example extracted from the IMPROD dataset of lesions masks over a ADC multi-parametric magnetic resonance image. The figure on the left is the second lesion of the patient identified as 28, with upper and lower images representing respectively the consecutive slices 3 and 4. The figure on the right is the lesion number one of the patient identified as 5, with upper and lower images representing respectively the consecutive slices 8 and 9. Note the lesion size difference between patients and between slices of the same patient.

The images are divided into three dimensions, the first two, height and width, are common to regular two-dimensional images. The third dimension that gives extra spatial information is in this case the slices. This is captured by collecting prostate images in a different position with a given spacing between them, which is relevant to the way the images are processed. Thus, even though the same institution usually captures the images with the same technique, there has been an attempt to create a standard technique to collect these images, allowing them to combine in the future several datasets.

A clinically significant lesion can be defined through several scoring systems such as the Gleason grade groups (GGG), the PIRADS and Likert scores given from analysis of multi-parametric magnetic resonance images. Some datasets use the real Gleason score verified after the surgery. For this

project, the Gleason grade group from the MRI (GS_MRI) is used. Positive labels are given to lesions that belong to group two, whereas lesions from groups one and zero are classified with a negative label of zero.

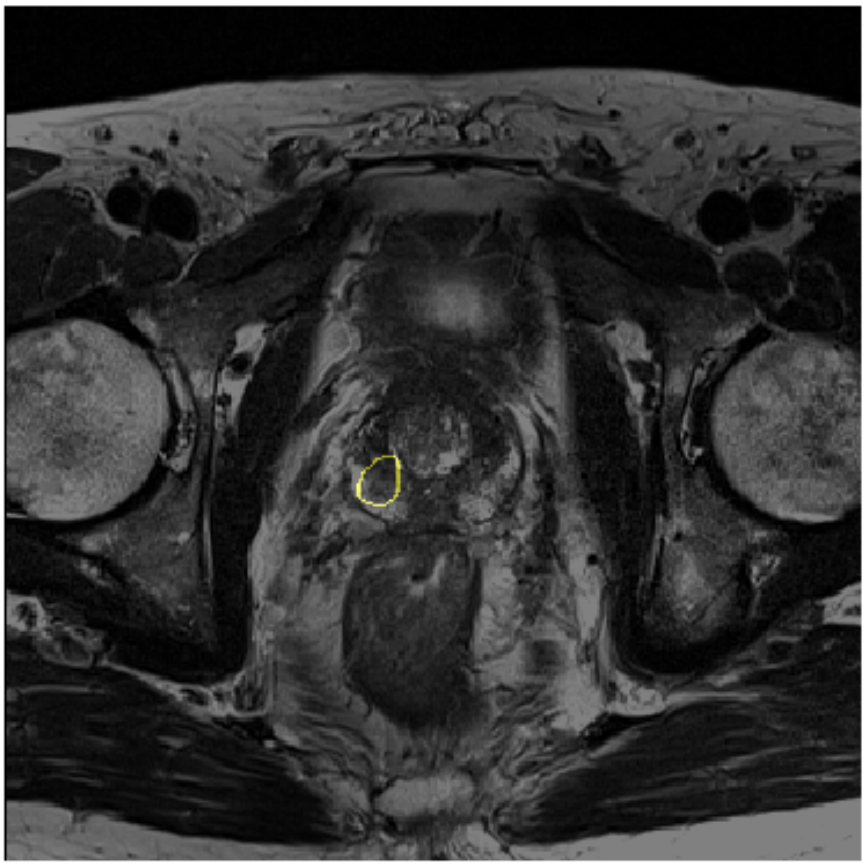


Figure 3.3. Example of a lesion contour in a T2W sequence. This contour is drawn from the segmentation mask and it is only used from visualization purposes. Slice 13 from the patient with id 10.

Lesions are much less represented in the dataset than the prostate, thus generating less positive labels and an increasingly difficult task due to a more skewed dataset. Figure 3.1 shows the visualization of lesions masks in T2W sequences. In the first analyzes of the image, it is perceptible that lesions do not have the same size. Comparing this image to the one in Figure 3.2 denotes two key details, firstly the resolution of the latter, representing ADC sequences, is inferior. Secondly, both figures include the same lesions, however, they are visible in different slices of each sequence, in other words, these sequences are not registred. Moreover, in Figure 3.3 it is possible to see the contour of the lesion drawn from the segmentation mask.

3.3 Dataset statistics

The size and the quality of the annotations are undoubtedly important to the quality of the learning. However, some other key factors might influence the ability of the model to generalize to or work properly with all the classes and potential requirements of the problem. For instance, as previously mentioned, the classification of both clinically significant lesions or the PIRADS score is one of the problems that this thesis is trying to solve, and they can be affected by skewed datasets. In other words, a dataset where one label has a much significant representation than the others, or when one label does not have proper representation (e.g. 95% of the dataset from the same class).

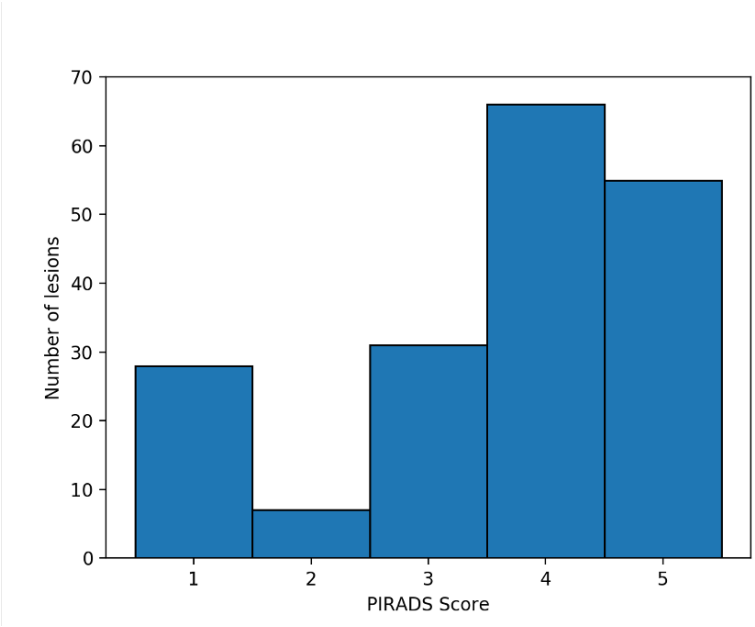


Figure 3.4. Distribution of the PIRADS score across lesions. This histogram includes all the images and lesions in the dataset, even those that do not have individual lesion masks, such as most of the lesions with a PIRADS score of 1.

All the lesions in the dataset are classified by experts according to the PIRADS system, and one of the important points to the prediction of this score is to see if the dataset has a proper representation of all classes. Thus, in Figure 3.4 a histogram with the distribution of the scores is shown. The most represented labels are those of the scores four and five, accounting for more than 60% of the labels. On the other hand, label two has a poor representation of less than 10% of the data. Despite being present in this histogram, not all the lesions that have a PIRADS score have an independent mask, so it is impossible to locate them in the magnetic resonance image. Thus, since the lesion cannot be located through the mask coordinates they are excluded from all the final classification

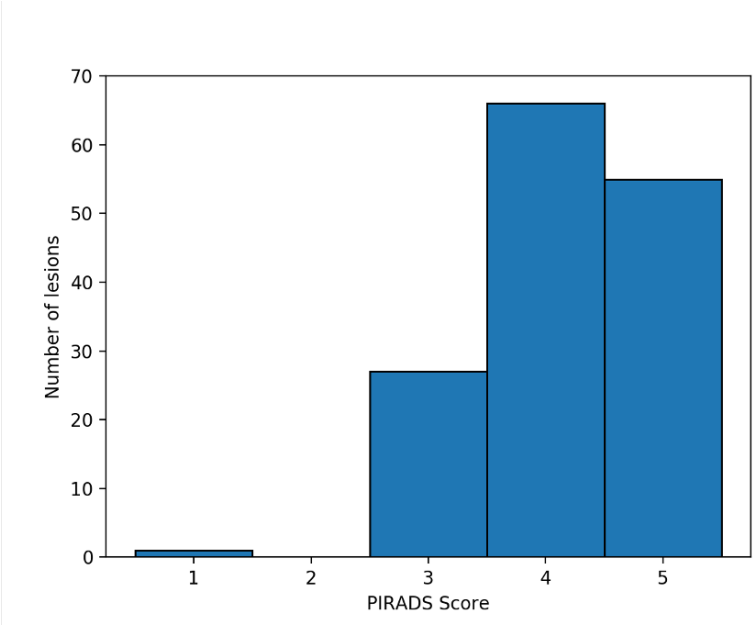


Figure 3.5. Distribution of the PIRADS score across lesions. This histogram includes all the images and lesions in the dataset, even those that do not have individual lesion masks, such as most of the lesions with a PIRADS score of 1.

problems. Furthermore, these lesions are lower grade lesions of labels one and two, which might become considerably problematic due to the lack of representation of these. The distribution of the PIRADS scores after the exclusion of lesions without mask can be seen in Figure 3.5. It is relevant to note that there is no lesion with a score two and only one with score one.

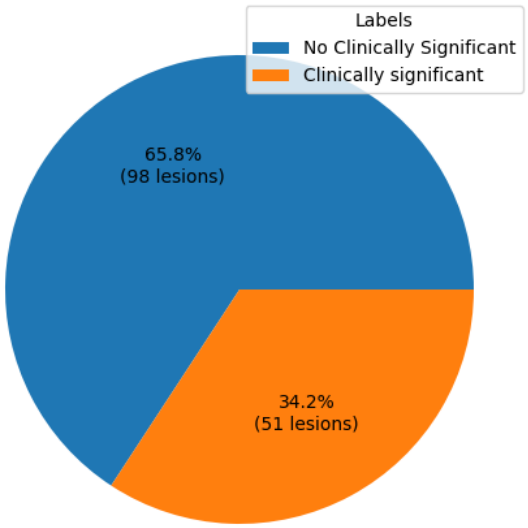


Figure 3.6. Distribution of the Clinically Significant (positive label) and No Clinically Significant Lesions (negative label). The former is represented by the orange slice, while the latter is represented by the blue slice and represents most of the lesions. This pie plot only includes the lesions which have mask representations of it and can be used in classification problems.

After removing the lesions that do not have masks and cannot be located,

it is possible to start studying the distribution of clinically significant lesions. The exclusion of the lesion leads to a considerable reduction in the number of available samples for classification problems with only 149 lesions to be used. In figure 3.6 a pie chart shows the representation of the positive and negative labels in the dataset of lesions that can be cropped and classified by a machine learning model. In contrast to what is seen in Figure 3.4, since clinical significance is not, in the context of this thesis, related to the PIRADS score, the presence of no clinically significant lesions is much higher. The majority of the dataset, 65.8%, has a negative label, where the remaining samples 34.2% are clinically significant, thus, associated with a positive label.

Lesion characteristics have an impact on how they are classified, and some of those can be studied in order to optimize the solution search space or understand the problems related to one unfitting approach. The size of the lesion is one of those elements that are relevant to decide how to approach the problem. Through this several analyses can be done, such as the height-width relationship, the area of lesions and the direct impact of those in the clinical significance of the lesions. All the images were resampled to a 224x224 height and width while keeping intact the number of slices.

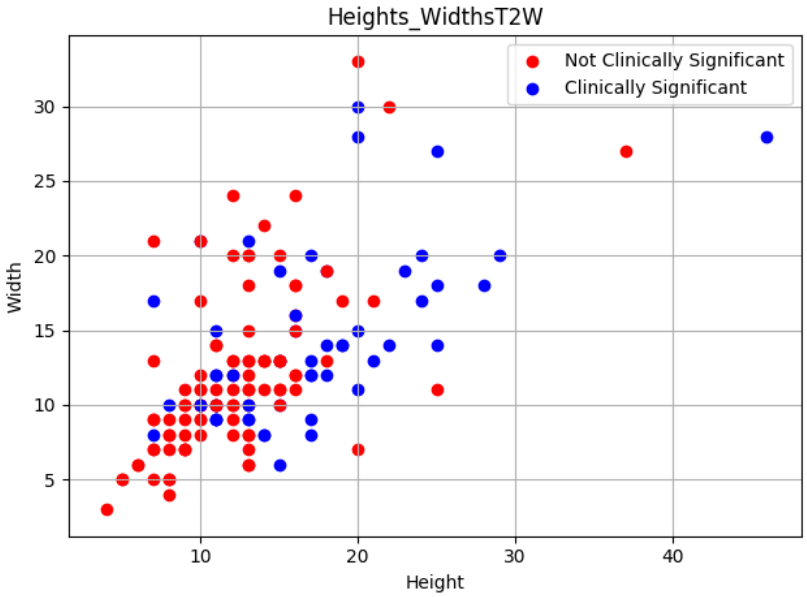


Figure 3.7. Scatter plot displaying the relationship between the weight and the height of lesions in T2W sequences in the slice that maximizes the area of the lesion. Only lesions with masks are used to generate this plot that also indicates with blue if the lesion is clinically significant or with red if it is not clinically significant. This allows to understand the influence of the size in the significance. Lesions are measured after a resampling of the images from 360x360 to 224x224 keeping the same number of slices.

A potential relationship between the height and the size of the lesions

is explored in T2-weighted, in Figure 3.7 and in apparent diffusion coefficient images, in Figure 3.8. Despite the original size difference, with the resampling both of the scatter plots show provide information to acquire similar knowledge and draw identical conclusions from them. T2W lesions width values are between 2< W_{T2W} <35 while for ADC the values are 2< W_{ADC} <25. The value fore the height, are 2< H_{T2W} <50 and 2< H_{ADC} <40 respectively. It is also worth noting, that the absolute difference between the height and the width is in the majority of the cases limited to ten, indicating that both dimensions have similar values. Moreover, a more careful analysis denotes that the size seems to have an influence on the classification of the lesions as clinically significant. It is shown that lesions with both dimensions with values below ten do not have, with a high degree of certainty, clinical significance. However, as the size increase, the impact in the classification seems to be blander, and although having a larger size denotes more likelihood of being clinically significant, it is not a sufficiently accurate indicator to draw any conclusion related to the significance. Both images show several lesions with high dimensions and negative labels. In 3.7 the resampling has an unnoticeable effect on the distribution of the sizes, but since ADC sequences were resampled to a higher resolution it is possible to notice in Figure 3.8 small consistent gaps between the sizes.

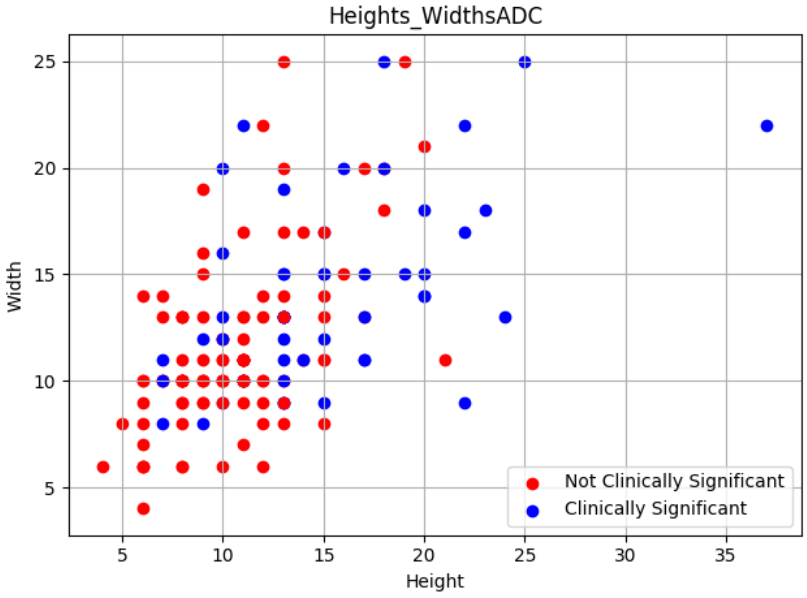


Figure 3.8. Scatter plot displaying the relationship between the weight and the height of lesions in ADC sequences in the slice that maximizes the area of the lesion. Only lesions with masks are used to generate this plot that also indicates with blue if the lesion is clinically significant or with red if it is not clinically significant. This allows to understand the influence of the size in the significance. Lesions are measured after a resampling of the images from 128x128 to 224x224 keeping the same number of slices.

Several strategies can be used in order to locate, crop and prepare a lesion to be classified. Yet, not all the strategies are adequate for this problem, and the selection of the strategy to be used is highly dependent on the range of sizes of lesions. For instance, if there is a broad range of sizes, a technique that just does a crop of a fixed size centered in the center of the lesion might be problematic. Smaller lesions crops will include an excessive amount of background, which might interfere with the learning since the region of interest is insignificant when compared with the surrounding. On the other hand, if a lesion is larger than the predefined crop size important parts for the classification might be cropped out, also hurting the performance of the model. Therefore, knowing the size of the lesions and how the cropping technique might affect the performance is valuable to understand the reasons behind the success or lack of it in further experiments.

Implications of the size of lesions are broad and affect other tasks such as the segmentation of those lesions. To start with, a larger lesion area impacts directly the number of positive labels in each mask, for example, smaller lesions have fewer pixels thus fewer labels, and this interferes with how skewed the dataset is. Secondly, a considerable variation in the area occupied by lesions might affect negatively the performance, since models sometimes deduct correlations that might make the predictions more prone to a specific size. To analyze the occupied area of lesions, the area of one lesion was considered to be approximated by the multiplication of its width and height dimensions. However, it was only considered the slice where each lesion had the biggest approximated area since it is the same slice to be used in classification problems. In Figure 3.9 and 3.10, a box plot of this area in ADC and T2W sequences is shown respectively. And even though both show similar means, and as a consequence of higher width and height maximum values, T2W sequences are more prone to outliers with larger areas. This might imply problems when cropping the lesions.

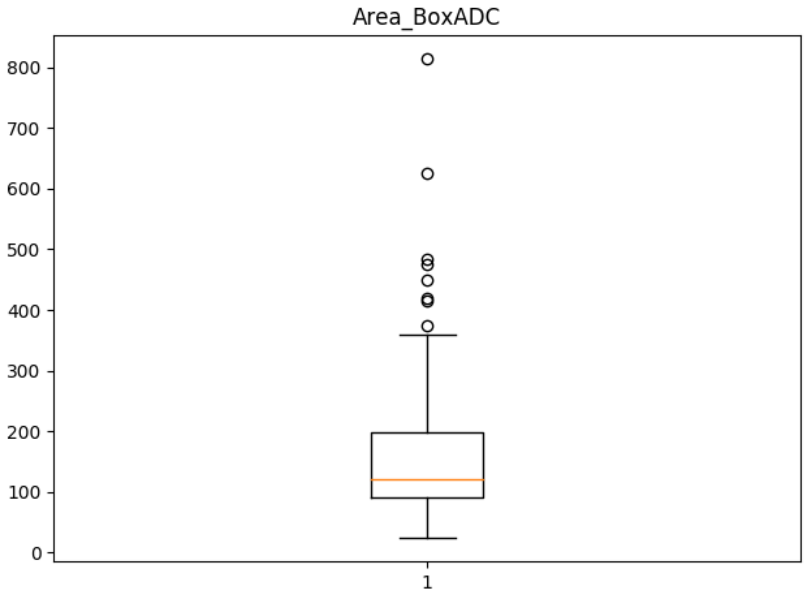


Figure 3.9. Box plot of the area occupied by ADC lesions. This takes into account only the size of the lesion in the slice that maximizes its size, and only lesions with masks are used to generate this data. Lesions are measured after a resampling of the images from 128x128 to 224x224 keeping the same number of slices.

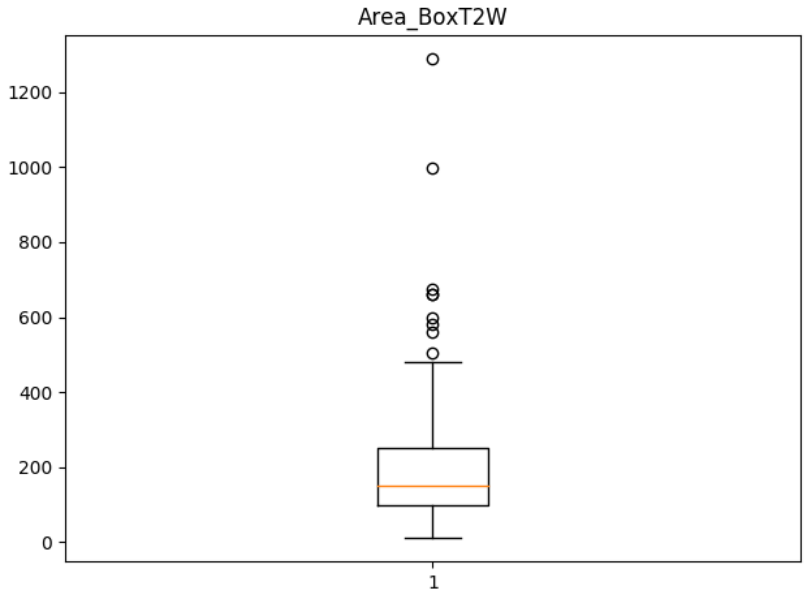


Figure 3.10. Box plot of the area occupied by T2W lesions. This takes into account only the size of the lesion in the slice that maximizes its size, and only lesions with masks are used to generate this data. Lesions are measured after a resampling of the images from 360x360 to 224x224 keeping the same number of slices.

4. Methods

In the context of this thesis, there are two central research directions. The first of them is to approach the data to work in a classification problem, more specifically, the classification of lesions. Whereas the other focuses on detection and segmentation from the multi-parametric magnetic resonance images. These problems require distinct approaches and methodologies, as well as computing power, which shall be taken into consideration while designing the model.

The classification of lesions is a broad domain due to the most variate scores scales given to the lesions, hence this problem can be divided into two smaller problems that follow more narrow guidelines. To start with, a binary classification problem based on the prediction of the clinical significance of the lesions. And, another more complex problem aiming to predict the PIRADS score, thus a multi-class classification problem. Both of these problems share common characteristics, however, dataset differences and other smaller details entitle each of them to be individually researched and explored.

Despite employing similar methods, detecting and segmenting lesions or the prostate are two distinct problems. On one hand, the same loss functions, models or data processing techniques can be used. On the other hand, the methods working in one task might perform poorly on the other one or even show strange behavior.

This Chapter's focus is to establish the research grounds used in the context of this thesis, as well as discuss other factors such as research questions, problems to be solved and implications of the chosen methods. Furthermore, it will be explained how the conducted research will proceed for the experiments and study alternative solutions and techniques.

4.1 Lesion classification

Prostate cancer lesions vary in shape, size, and intensity, even their aggressiveness changes accordingly. Being able to define and categorize the aggressiveness of a lesion is one of the roles of experts when dealing with multi-parametric magnetic resonance images of the prostate. Moreover, the classification can be performed according to the guidelines of several score metrics and scales, requiring a deep knowledge of the problem and years of study in order to classify them correctly. Yet, it is a tedious, expensive and prone to errors task, since there is a considerable quantity of bad diagnosis, especially overdiagnosis with some lesions being classified as more aggressive than what they actually are.

In an attempt to solve the classification problem, two smaller tasks were undertaken and similar methods were employed. For both problems the input is the same, so the base model to tackle them was the same, with some minor differences and some other distinctions in the optimization and evaluation process.

4.1.1 Data

The original format of the data was not appropriate for the problems at hand. Therefore, some previous processing was required to generate appropriate input and labels that could be used in these tasks. Originally the data was available as the 3D representations of magnetic resonance images of the prostate and surrounding organs. However, the expected format to solve these problems was a 2D representation of the lesions, avoiding as much as possible unnecessary body elements.

The first step to attain the desired data format was to spatially locate the lesions in the 3D space. Lesions were previously detected and marked by experts, therefore the use of the masks to locate them is possible. This also means that lesions that appear in the data but do not contain an individual mask cannot be part of it and need to be excluded. To efficiently locate the lesions, their domain area was approximated, and the initial and final points were marked independently in every slice as shown in Figure 4.1. In the figure, it is possible to see at the green starting points and the red ending points if the lesion is interpreted from top left to bottom right. These points can be used to draw a bounding box without predefined size that contains the entire lesion and potentially some background too. Alternatively, they can be used to approximate center and draw a fixed

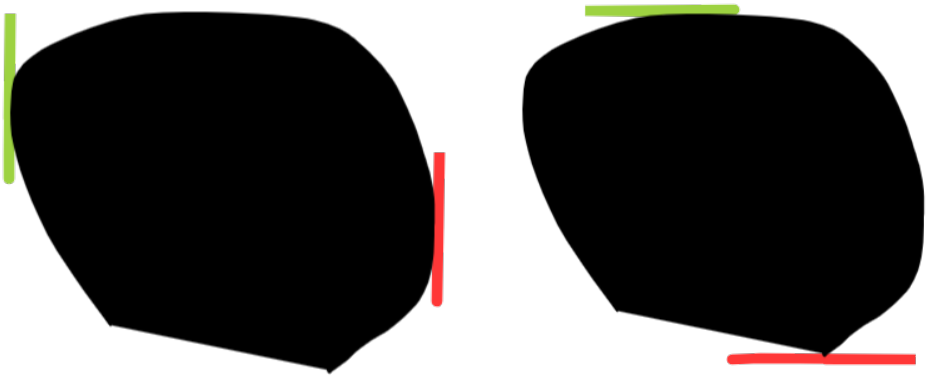


Figure 4.1. Example of how the lesions are located and their area approximated. The lesion is represented by the black area. In the image in the left, the green line represents the left most point of the lesions while the red represents the point that is more to the right. In the image in the write the green represent the point that has the biggest height and the red line represents the point with lowest height. This points can be used to approximate the lesion area to a square, or to point the center of the lesion.

size bounding box that might crop the lesion, or capture almost exclusively background.

Following the location of the lesion at each slice, the approximated area of these lesions was calculated on a slice basis. The slices were then ordered based on the largest lesions size, and the one that maximized the size was chosen. A crop to the lesion was performed in that slice and the output of the cropping was saved as an image to input to the classification model. The label of the lesion was respectively associated with the new two-dimensional image.

4.1.2 XmasNet

In all the machine learning tasks the model is as important as the data, and in this case, the same happens. Previous challenges such as ProstateX already tried to address the clinical significance classification problem. As a result, many submissions attained reasonable values at this task, therefore, one of the submissions was selected to be the base of the classification system developed in this thesis.

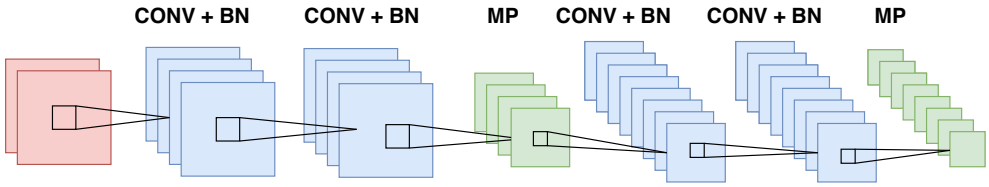


Figure 4.2. XmasNet [27] based network architecture with 2 input channels, and without fully-connected layers. Red squares represent each one channel of the input layer. Blue squares represent each 8 channels of a convolutional layer with a kernel size of 3, padding of 1 and stride of 1 plus a batch normalization layer right after. Green layers represent each max pooling layers with kernel size of 2 and stride of 2.

The selected architecture, named XmasNet achieved the second place with an AUC of 0.84 in the ProstateX challenge surpassing traditional machine learning techniques [27]. Additionally to the model architecture, the submission presented also a technique to preprocess the data and feed it to the neural network. This preprocessing is considerably different from that was used in this project, therefore the only inspiration is the model [27].

Figure 4.2 partially shows the model, more specifically, it shows the common layers to both classification problems. This part is responsible for extracting useful information from the lesions, in other words, the features of the lesions, regarding for instance, shape, size or intensity. The architecture despite being shown to have two input channels works also with one input channel, depending on how many sequences are used. The input is followed by a convolution that increases the number of channels to 32 and another convolution that keeps the number of channels. To these, a max-pooling layer follows, down-scaling the size by half, and feeding it to a convolutional layer that will double the number of channels. Another convolution and max-pooling layer follow, keeping the number of channels and halving the size again. It is also worth noting that every convolution has a kernel size of three and padding of one and they are followed by a batch-normalization layer and ReLU non-linearity. Max-pooling layers have both kernel and stride set to two.

4.1.3 Clinical significance classification

Predicting if a lesion has clinical significance or not is a typical machine learning binary classification problem. In this sort of problem, the model predicts if an input belongs to the positive or negative classes which in this case are being clinically significant or not respectively. The output of the network is one single value, between zero and one, the probability of belonging to the positive class.

In order to tackle this task, to the partial convolutional model shown in Figure 4.2, some extra layers, as seen in Figure 4.3, were appended. These layers are responsible for using the extracted features from the convolutional part and compute from them the prediction, and are specific for the binary classification problem. The output of the last max-pooling layer is flattened to one single dimension, besides the batch dimension, and is then fed as input to three fully-connected layers. The first layer receives the flat vector that has a length dependent on the size of the lesion image

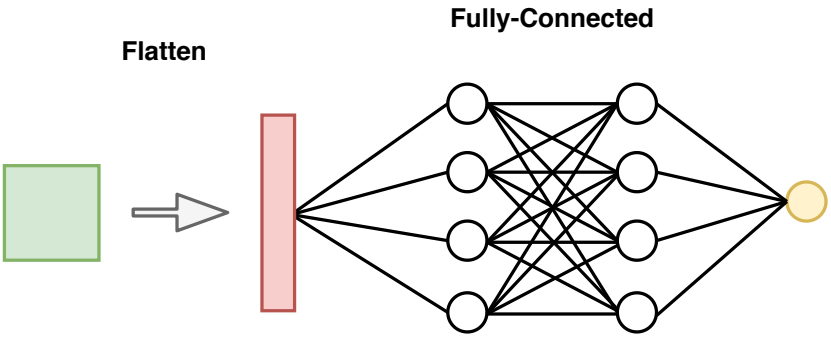


Figure 4.3. Additional layers to be added to the network for a binary classification of the clinical significance of lesions. Green square represents the last max pooling layer. Arrow to the write represents the flatten of the square data to a vector that can be fed to fully convolutional layers. Yellow circle represents the output followed by a sigmoid activation.

given to the network accordingly to the formula $\frac{height*width}{4} * 64 = length$, for example, a 32x32 lesion image will originate a vector of length of $8^2 * 64 = 4096$. This first layer connects this input vector to 1024 nodes, while the second layer connects the 1024 node to the 256 following nodes. Finally, the third and last layer generates one output from all of the 256 nodes and applies a Sigmoid function in order to convert it to a probability. The first two layers are followed by ReLU non-linearity.

$$\begin{aligned} BCE(y,t) &= -t * log(y) - (1-t) * log(1-y)) \\ &= \begin{cases} -log(y) & if \quad t = 1 \\ -log(1-y) & if \quad t = 0 \end{cases} \end{aligned} \tag{4.1}$$

To optimize the network and to approximate the output the binary cross-entropy loss was minimized. The expression for this loss function can be seen in the Equation 4.1 where t stands for the expected class (one or zero), and y is the output given by the network that ranges between zero and one. This loss function is a special case of the general cross-entropy loss function and is frequently used in binary classification problems. Moreover, the stochastic optimization was performed by the Adam optimizer with a learning rate of $7 * 10^{-4}$ and a weight decay of 10^{-4}. Each step computed the results and afterward the gradients of a mini-batch of size 8.

4.1.4 PI-RADS classification

Classifying a lesion as clinically significant or not is a useful task, however, it can frequently be seen as rather limiting and lacking detail. Therefore, in a clinical environment, the PIRADS score is a better classification system

for the lesions, since it does try to measure the likelihood of a lesion to be from a clinical significance cancer or not. In practical terms, it means that the network had to be changed in order to produce five different outputs. Moreover, not only the outputs were required to be between zero and one, but their sum had to result in the value one.

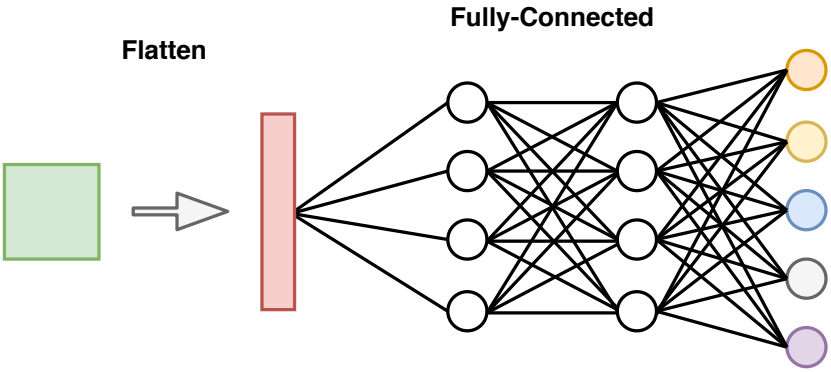


Figure 4.4. Additional layers to be added to the network for a multi-class classification of the PIRADS score of lesions with five possible outputs. Green square represents the last max pooling layer. Arrow to the write represents the flatten of the square data to a vector that can be fed to fully convolutional layers. Yellow circles represents the outputs followed by a softmax activation.

The model used in this task was a variation from the partial convolutional model displayed in Figure 4.2. To the last max-pooling layers, some other layers were added, such layers are seen in Figure 4.4. For the most part, these layers work similarly to the ones approached in the previous problem, with the flattening of the matrix, the input dependent number of nodes and the node count of posterior layers. Nevertheless, the output was in this problem represented by five different nodes, where each node represents a class or a score from the PIRADS scale. Moreover, to the output a Softmax function is applied, so that they can be used to minimize the cross-entropy loss.

$$\begin{aligned} CE(y,t) &= -log\left(\frac{e^{y_t}}{\sum_j^C e^{y_j}}\right) \\ &= -y_t + log\left(\sum_j^C e^{y_j}\right) \end{aligned} \tag{4.2}$$

The optimization of the network was similar to the one described in the previous problem, although there were small differences. To start with, the learning size was decreased to a value of $5 * 10^{-5}$ to ensure smoother convergence, while the optimizer and the weight decay were kept the same. The loss function was also slightly different since it had to handle all

the five inputs and the five classes. Cross-entropy loss was used and its equation is represented in Equation 4.2, where y represents the predictions vector and t the ground truth class. Moreover, as seen in the equation t can also be used as an index of the vector y.

4.2 Detection and segmentation

Analyzing a medical image is not an effortless simple task, not only it requires a considerable amount of training and expertise, but due to low resolution or the format of the images, some elements such as organs and their boundaries are difficult to locate. A machine learning segmentation model tries not only to detect these elements but to segment them and delineate their boundaries. In the multi-parametric magnetic resonance images, two segmentation tasks can be defined and tackled, the segmentation of the prostate and the segmentation of prostate cancer lesions. Methods for both these tasks are discussed in this section, and how the solution for each one was approached.

For the most part, the tasks are similar with the main goal being to predict a mask close to the ground truth mask. Therefore, similar methods were used in both cases, similar architectures trained and tested for both problems, and the same loss functions were also used. However, this section focus on similarities further details on how the problems differ are discussed in the following sections.

4.2.1 Loss function

There are several loss functions that could have been applied to the problem at hand, however, only dice loss was picked. This loss function comes from the similarity measure Sørensen–Dice coefficient applied to a three-dimensional space.

$$DICE = \frac{2 * TP}{2 * TP + FN + FP} \tag{4.3}$$

The formula to calculate the Dice coefficient or score is in Equation 4.3 in terms of binary data that composes the masks. The intuition behind this score is to measure the correctly predicted values divided by the sum of those values and all the values predicted wrongly. A perfect value of one in this function would require that no label of one is incorrectly predicted as

zero (a false negative) and that no label of zero is wrongly predicted as one (a false positive). Although the binary formulation works perfectly in an evaluation setup, it is not appropriate for training.

$$DICE(Y, M) = \frac{2 * Y * M}{Y * Y + M * M} \tag{4.4}$$

During the training of the model, the mask predictions are supposed to be given in probabilities, in other words, each pixel in the three-dimensional space is supposed to have a probability of being classified with a positive label. While in the final output to test a prediction a threshold can be established to attribute a label of one after that specific value, during training the loss function must be able to work with the probabilities. Therefore, in order to use the Dice coefficient effectively in these problems, and to approximate the value of the real mask, the formulation to be used is shown in the Equation 4.4. For that formulation Y represents the predicted mask while M represents the target mask.

$$DICE\_L(Y, M) = 1 - DICE(Y, M) \tag{4.5}$$

Moreover, a loss function usually measures the distance between prediction and target and not the similarity, hence, the Dice Score in Equation 4.4 can be further adapted to represent a function to be minimized. The adaptation is rather simple, and it does only require that to a value of one, the Dice Score is subtracted. The formulation for the loss is given in Equation 4.5. Despite not being used directly as loss, the Dice Score is also a great tool to measure the similarity and evaluate the performance of the model further on.

$$Comb(Y, M, n) = \frac{1.5 * \sum_i^n BCE(Y_i, M_i)}{n} + DICE\_L(Y, M) \tag{4.6}$$

Despite the wide use of Dice loss to optimize models in image segmentation tasks, there are other losses that can be explored and combinations of losses. For this thesis, a weighted combination of the dice loss with the binary cross-entropy loss was used. The formulation of this loss can be seen in Equation 4.6 and the weights of both are respectively 1 and 1.5. Uniting losses requires some thought and attention, for this particular

task, the combination of these was conceived in order to try and improve the overlapping of the predicted and ground truth mask.

4.2.2 Models

In terms of the architectures to tackle these problems, the first model to be used was a 3D U-Net, that was specifically designed for segmentation in medical imaging problems [28].

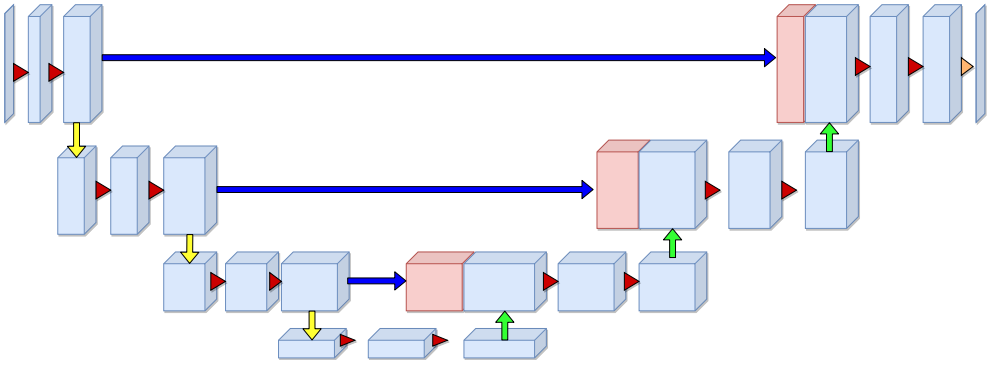


Figure 4.5. 3D U-Net [28] architecture. Blue arrows represent a concatenation operation (blue cubes in the tail of the arrow are seen as red cubes in the head of the arrow). Red arrows represent a 3D convolution followed by batch normalization and ReLU. Yellow arrows are max pooling operations while green arrows are up-convolutions. The final orange arrow represents simply a convolution to generate the output.

The 3D U-Net model architecture can be seen in Figure 4.5. It is characterized by two main features in its design. To start with the model has "deep levels" where after two convolutions operations across three dimensions, the input is downscaled to half of the size and the number of channels doubled. After reaching the desired depth, in this case after three downscaling operations, the input is upscaled until it returns to the original dimension. In spite of having, between upscaling operations, 3D convolutions that do not change the size, the first convolution is responsible for reducing the number of channels. Another particularity of this model is the shortcut connections between levels, where the output of one level in the downscaling side is concatenated to the input of the same level in the upsample side. These shortcut connections are especially important to ensure that no information is lost between all the convolutions, and to ensure that spatial information is preserved from the original image and through the entire downscale-upscale process.

The second model used in these problems was a 3D ResNet-18 [114] that is considerably more complex than 3D U-Net as can be seen in Table 4.1. However, the complexity does not guarantee better learning due to the considerable difference between architectures and different types of

Table 4.1. Comparison of the number of parameters of different models

Code	Model name	Parameters
U	3D U-Net	19,069,955 [28]
R	3D ResNet-18	32,990,000 [29]

connections. The construction of this model is based on building blocks, and in this thesis, the original model was slightly adapted to work with the dimensions of the input.

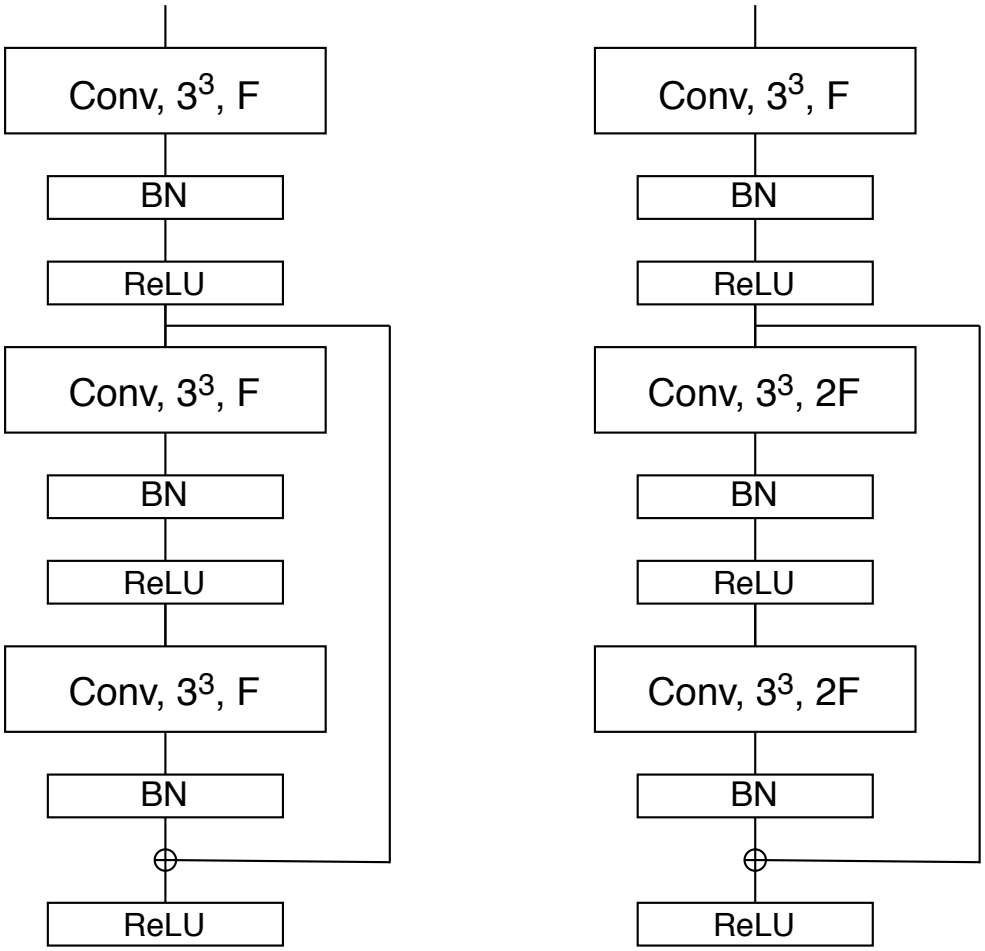


Figure 4.6. Adapted architecture of the basic block for three dimensional residual convolutional neural networks (3DResNet)[114]. This block is composed by the following components: BN as Batch Normalization; ReLU as the Rectified Linear Unit nonlinearity; and Conv, X^3, F as a convolution with a kernel of size X across the three dimensions and that maps the input to an output with F channels. Basic block on the left is the Block A while the one on the right is Block B.

The adapted building blocks, seen in Figure 4.6, differently from the original version have a stride of size one, removing the downsampling properties of these layers. These building blocks have three convolutional layers each followed by a batch normalization layer and a ReLU activation. The difference between both blocks is that the block on the left (Block

A) does not change the number of channels, whereas Block B doubles the number of channels in the first convolution. Moreover, both versions include a shortcut residual connection between the output of the first convolutional layer and the output of the third convolutional layer.

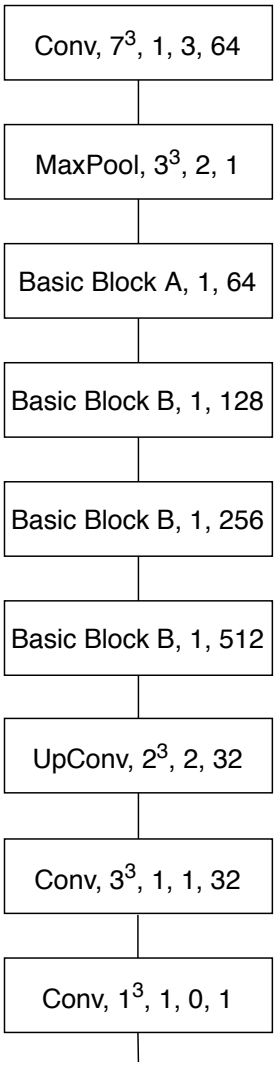


Figure 4.7. Adapted architecture of a 3DResNet-18 [115] from the MedicalNet project, which trained several three dimensional neural networks in a combined dataset of various organs [29]. The weights of the network are available and any adaptation of the network did not affect the original parameters.

The complete architecture is seen in Figure 4.7. It is composed of several instances of the basic blocks, as well as other layers. The networks start by convolving the input with a 7x7x7 kernel and mapping one channel to 64. The following max-pooling layer downsamples to half the size and then feeds it to a sequence of four basic blocks, with the first being of type A and the following of type B. The original size is then restored by an up-convolution and processed by two more convolutional layers. The final output has one channel subject to sigmoid at each pixel. The network used was already pretrained in another dataset for other organs [29] and the weights were used as initial weights of this model.

4.2.3 Data augmentations

Neither the prostate cancer lesion segmentation problem nor the prostate segmentation problem had abundant data and examples to learn from. As seen before, machine learning and especially deep learning algorithms rely considerably on the number of samples in the training set that represent potentially different situations and variability between samples. Moreover, variability in the data and a considerable number of distinct samples is a mechanism to avoid also overfitting and improve the generalization capabilities of the model at hand.

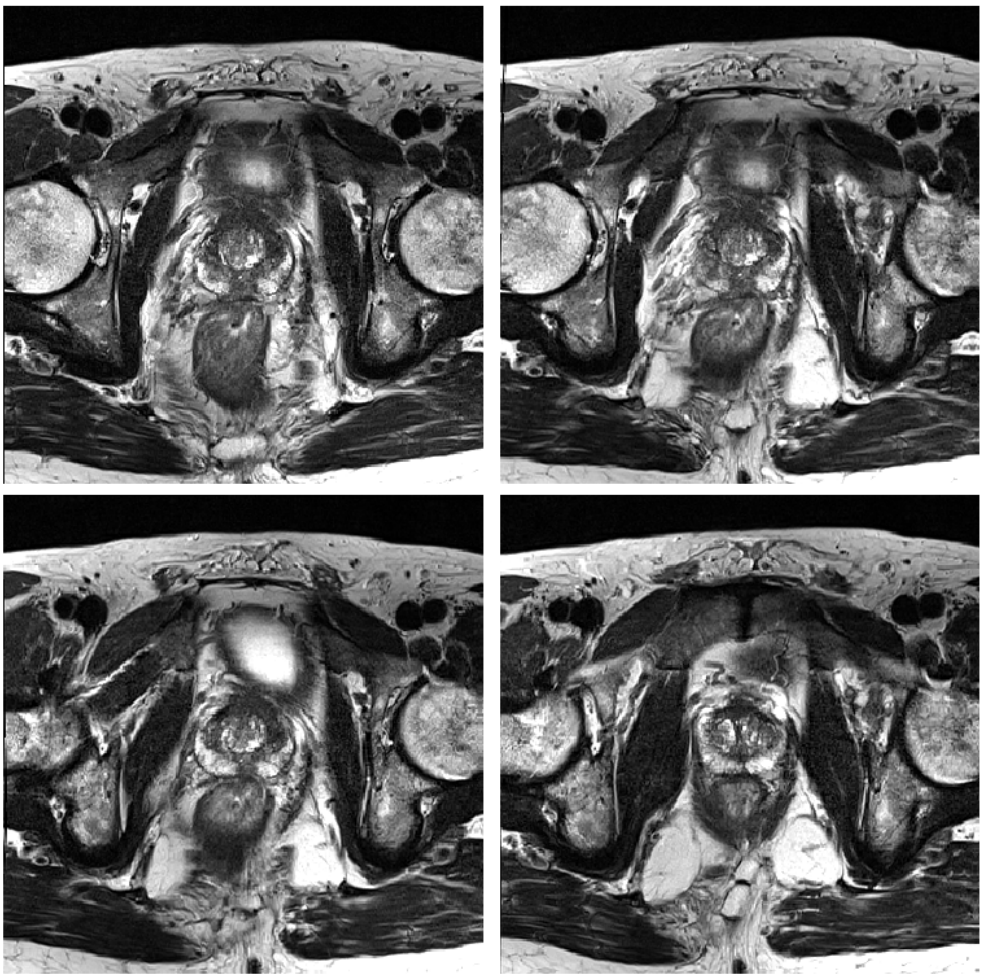


Figure 4.8. Example of a elastic deformation augmentation applied to a T2W image. The original image is represented in the top left corner and the other are generated by deformations with a spline order of 3, an alpha of 1000, and a sigma of 40. The deformations are applied to the whole 3D, however, in this image only results in slice 13 are shown. This T2W sequence is from the IMPROD dataset.

In Figure 4.8 it is shown an example of elastic deformation, that was one of the augmentations used. The application of this transformation to the image and to the mask increases the variability of the data by giving slightly different shapes and sizes to the prostate cancer lesions. Other transformations were also applied, such as random contrast, random

Poisson noise, random Gaussian noise, and random rotations.

4.2.4 Prostate segmentation

The main goal of this problem is to predict the segmentation mask of the prostate from multi-parametric magnetic resonance images. This problem can be seen from two perspectives, first, it was attempted to achieve an overlapping between the ground truth and the predicted mask. And it can also be seen as the attempt to classify individually each pixel with a positive or negative label.

Due to being present and visible in a considerable number of slices in each MRI, and also to the fact that the prostate occupies a considerable part of the slice when visible, the number of positive labels is abundant. Moreover, size and shape variations between patient prostates exist, but they are small and possible to learn by the model. The problem is, therefore, simpler than other versions, for example, lesion classification, and requires less tuning to attain the expected results.

This problem was used as a baseline for the models, a test to the performance of the model, and the entire pipeline process that led to the segmentation. The performance was evaluated in different sequences, variable resampling sizes, and data augmentation in order to gather details regarding the ability to solve this problem and the impact of these hyperparameters.

4.2.5 Lesions segmentation

Predicting a mask for prostate cancer lesion in multi-parametric magnetic resonance images is a problem considerably more difficult than the segmentation of the prostate. First, a lesion appears in fewer slices than the prostate and that combined with the smaller size the lesions results in much less positive pixels. Moreover, lesions are not mainly centered on the magnetic resonance image, and not only their position varies, but the shape and the slices where the lesion is visible vary too. Finally, the number of lesions to be segmented vary between patients, and this increases the difficulty of the problem. This is also a far more difficult problem to humans than just segmenting the prostate.

Each patient might have several lesions, thus it will also have several masks, one for each prostate cancer lesion. The model, independently of which architecture is used, do not try to identify individual lesions.

Instead, the model tries to segment all lesion voxels present in one mp-MRI. Therefore, it is necessary to create a mask composed by the union of all the individual lesion masks of a patient. Thus, to combine the masks, since the lesions do not overlap, a sum of all the masks is performed.

5. Experiments and Results

Once the methods are established, it is time to conduct several research experiments within the problem to be solved regarding the methods and the data. Experiments help to understand which techniques work and which ones do not work. Moreover, careful experimentation can improve already working models. For example, the data to be used in one experiment can be resized, the cropping can be performed differently, and the magnitude of data augmentations can be increased or decreased. Slight changes in any of these can have a considerable impact on the results, thus the experiments must be conducted thoroughly.

In spite of not being the only indicator of good research, results are crucial, and they should be a relevant part of any research. Frequently results are compared to different research projects due to the use of the same public dataset, however, since this thesis uses a private dataset, the results are compared between several different experiments, with considerable parameters changed between experiments.

In this chapter, the process used to conduct experiments is explored, the results for each experiment are given and results are carefully detailed and compared. Furthermore, an analysis of the methods is conducted, and the reasons for the different performance results are discussed with the advantages and disadvantages of the methods for each one of the problems presented initially in this thesis. Finally, some of the experiments are illustrated with visual examples of their performance, either using plots or segmentation masks side by side.

5.1 Evaluation

Deep learning problem evaluation is usually performed in unseen data, in order to also assess the capabilities of generalization from a specific model.

In other words, if the model is overfitted to the training data. Splitting the dataset to construct a test set can be done with different ratios, such as 20%/80% or 25%/75%. However, the performance of the algorithm is frequently biased by the selected split, and results can appear to be better or worse depending on the data present in each set. In this problem a cross-validation technique with five different folds is used.

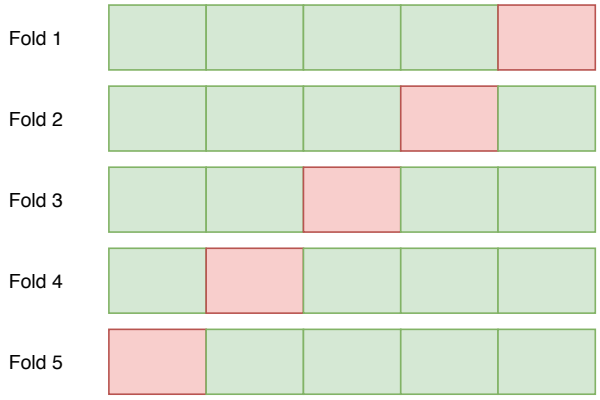


Figure 5.1. Structure of the five folds. Each rectangle represents a subset of the dataset containing 20% of the data. Green rectangles indicate that the data contained is part of the training set. On the other hand the red rectangle indicate that the data is used to test and evaluate the model.

The construction of the folds can be seen in Image 5.1, with each rectangle corresponding to 20% of the data present in the dataset. To accurately analyze the performance in a problem, a model is trained for each fold on the 80% training data (green rectangles) and tested afterward on the data present in the red rectangle of each fold. Finally, the results of all folds are averaged. It is also important to note that this technique works with different performance metrics, the data present in one rectangle is the same in all the folds and the split is performed patient-wise.

5.2 Lesion classification

Experiments on the classification of lesions are conducted independently for both problems with different losses, models, and metrics. However, some elements are common to experiments in both problems. Both work with two-dimensional images of lesions cropped from the slice where the approximated area of the lesion is bigger. The cropping of this lesion can be performed with a bounding box with a fixed size centered in the center of the lesion or an adjustable bounding box that resizes to have the exact size of the lesion that is resized afterward to the desired size. The cropping is performed at 64 and 32 pixels (with an adjusted bounding box the lesion

is resized to have this dimension), while the model receives a 56x56 and 28x28 image respectively. Moreover, the augmentations applied were the same in both experiments with the images being subject to the following random operations: crop with the size of the input of the model, horizontal flip, vertical flip, and an affine transformation with 15 degrees of rotation, a scaling between 0.50 and 1.50, a shear of 15, and translations of at most 0.07 in each direction. Whenever data augmentations are not used, the image is simply centrally cropped to the size of the input of the model.

5.2.1 Clinical significance classification

To perform binary classification of prostate cancer lesion experiments, it was necessary to define some configurations for the experiments, in other words, a specific combination of techniques that can be applied in several experiments. The experiments can vary in the type of cropping technique to be used, the sizes of the cropping and the input size also vary, as well as if data augmentations are used or not. Since all three of these parameters have two different options each it results in $2^3 = 8$ different configurations to be used and evaluated.

Table 5.1. Different configurations of the experiments. Includes details of the cropping technique used, if data augmentation techniques were used and what were the values used for cropping the lesions and the input size for the model.

Config.	Crop Size	Input Size	Crop Type	Augmentations
A	64	56	Adjusted Box	No
B	64	56	Adjusted Box	Yes
C	64	56	Fixed Box	No
D	64	56	Fixed Box	Yes
E	32	28	Adjusted Box	No
F	32	28	Adjusted Box	Yes
G	32	28	Fixed Box	No
H	32	28	Fixed Box	Yes

The impact of the parameters is further analyzed individually in order to understand not only which combinations work, and which do not, but also to understand the independent impact of the parameter. Table 5.1 shows the eight different configurations. Each configuration is associated with a letter, and it will be referred through that letter during the rest of this section. All the experiments' configurations described in the initial phase of the experiment, before retrieving the lesions, resample the MRI to 224x224 while keeping the original number of slices. An analysis of the effect of the resampling was also conducted and the results discussed.

Cropping the lesions

The impact of the cropping technique in the performance of the model was not clear, therefore, some experiments comparing both techniques described in this thesis were conducted.

Table 5.2. Comparison of the performance of the experiments using different sequences types that used distinct sizes and different cropping techniques. For this comparison all the selected experiments configurations do not include data augmentations. Bold represents the best performance configuration in that sequence type. All the values in this table represent AUC scores that were computed in a cross-validation setup for each of the five folds, and averaged across them.

Config.	ADC	T2W	ADC+T2W	Average
A	**0.863**	0.730	0.839	0.811
C	0.737	0.592	0.701	0.677
E	0.856	**0.776**	**0.844**	**0.825**
G	0.762	0.706	0.760	0.742

Table 5.2 illustrates the results of each experiment by the sequence used as input of the model. It is possible to observe that configurations, which crop the lesion with a bounding box adjusted to the lesion size and resized afterward to the crop size, show better results regardless of this size (A and E). Both these experiments demonstrated considerable performance gains when compared to the use of a bounding box with a fixed size. Moreover, despite the slightly better performance of configuration A with ADC sequences when compared to the configuration E, the latter displays substantially better results with T2W and with the combination of both sequences, averaging also in a better AUC for the three different inputs. Regarding the other configurations, G is by far superior to C in all the inputs. Thus, it is possible to infer that a crop size of 32 with an input size of 28 further improves the performance regardless of the cropping technique used.

Data augmentation

Similarly to the cropping, the effect of data augmentations in the performance of the model was also analyzed. The experiments to study its impacts were divided into two different groups in order to establish a fair comparison between configurations since, as seen before, the cropping type impacts the performance. Therefore, the cropping technique was the criterion to divide configurations into groups.

Table 5.3 displays the results for the experiments that used the best performing cropping technique. Differently from what was expected, the

Table 5.3. Analysis of the effects of data augmentation techniques in experiments that used an adjusted box as cropping technique. Bold represents the best performance configuration in that sequence type. All the values in this table represent AUC scores that were computed in a cross-validation setup for each of the five folds, and averaged across them.

Config.	ADC	T2W	ADC+T2W	Average
A	**0.863**	0.730	0.839	0.811
B	0.840	0.772	0.830	0.814
E	0.856	0.776	**0.844**	0.825
F	0.855	**0.815**	0.828	**0.833**

data augmentations do not seem to positively impact all the different sequences used as input. For instance, on ADC sequences, the impact seems to be negative, and while it can be neglected when comparing E and F, it is somewhat larger when comparing A and B. In contrast, using T2W sequences as the input seems to greatly impact the results in a positive manner and this can be observed by direct comparison of A with B and E with F which show improvements of approximately 0.04. However, when combining both sequences, the impact is clearly negative in all the experiments analyzed. In spite of having some negative impact, data augmentations have improved the average AUC of a configuration when compared to a similar one without data augmentations.

Table 5.4. Analysis of the effects of data augmentation techniques in experiments that used a fixed box as cropping technique. Bold represents the best performance configuration in that sequence type. All the values in this table represent AUC scores that were computed in a cross-validation setup for each of the five folds, and averaged across them.

Config.	ADC	T2W	ADC+T2W	Average
C	0.737	0.592	0.701	0.677
D	0.725	0.679	0.749	0.718
G	0.762	**0.706**	**0.760**	**0.742**
H	**0.818**	0.640	0.757	0.738

The impact of data augmentations on the experiments with different cropping techniques is also different and distinct as seen in Table 5.4. Furthermore, the impact also seems to be based on the cropping and input sizes used. For instance, on larger sizes (C and D) the impact on ADC seems to be negative with T2W and the combination of sequences benefiting greatly from the augmentations. Whereas on smaller crop sizes (G and H), ADC has a positive impact when using augmentations while the other two different inputs have a decreased performance. It is also worth noting that different from the results in Table 5.3, the augmentations do

not improve the average AUC of all the configurations, instead, it decreases the area under the curve in smaller input sizes.

Resampling

All the previous experiment configurations started by resampling the magnetic resonance images to a 224x224 dimension while keeping the number of slices untouched. However, this resampling undoubtedly affects the performance, either negatively or positively. In order to properly assess this impact, all the previous experiments were replicated without the initial resampling and the results compared. Since in the previous section it was shown that both data augmentations and the cropping technique impact the results, the comparisons were performed between examples that had these parameters in comparison, while the size of the crop and the input size were the only variants within a results table.

Table 5.5. Comparison of the performance of the model if no resampling for 224x224 is used, between configurations that use no augmentations and use a fixed box crop technique. Bold represents the best performance configuration in that sequence type. All the values in this table represent AUC scores that were computed in a cross-validation setup for each of the five folds, and averaged across them. WR stands for "With Resampling" whereas NR stands for "No Resampling"

Config.	ADC	T2W	ADC+T2W	Average
C - WR	0.737	0.592	0.701	0.677
C - NR	0.620	0.627	0.676	0.641
G - WR	**0.762**	**0.706**	**0.760**	**0.742**
G - NR	0.631	0.656	0.655	0.647

Table 5.6 shows the results of the configuration with a fixed box with worse performance, C, and the one with the best performance, G. Neither include data augmentations, and the size of their crop varies with G having the smaller value. In these cases, not resampling the images hurt the average AUC score, especially of ADC sequences. This, however, can be explained with the fact that the original size of ADC sequences is rather small, thereby the lesions are also small, meaning that it may be difficult to spot. On the other hand, while T2W performance decreased on configuration G, it improved on C. The reason behind this is the fact that the original size of T2W is considerably larger than the resampling size, meaning that a larger crop size might better capture the lesions while a small one might crop important parts of lesions.

Analyzing Table 5.6 where the results of two configurations with an adjusted box and no augmentations are shown, A and E, it is possible to

Table 5.6. Comparison of the performance of the model if no resampling for 224x224 is used, between configurations that use no augmentations and use a adjusted box crop technique. Bold represents the best performance configuration in that sequence type. All the values in this table represent AUC scores that were computed in a cross-validation setup for each of the five folds, and averaged across them. WR stands for "With Resampling" whereas NR stands for "No Resampling"

Config.	ADC	T2W	ADC+T2W	Average
A - WR	0.863	0.730	0.839	0.811
A - NR	0.834	0.741	0.818	0.798
E - WR	0.856	**0.776**	**0.844**	**0.825**
E - NR	**0.870**	0.750	0.818	0.813

observe similar patterns with lower average scores on experiments without resampling, and improved results on the experiment with larger crop size without resampling. However, in configuration E, the results for ADC show an improvement to 0.87, which means that this configuration is the best performing.

Table 5.7. Comparison of the performance of the model if no resampling for 224x224 is used, between configurations that have data augmentations and use a adjusted box crop technique. Bold represents the best performance configuration in that sequence type. All the values in this table represent AUC scores that were computed in a cross-validation setup for each of the five folds, and averaged across them. WR stands for "With Resampling" whereas NR stands for "No Resampling"

Config.	ADC	T2W	ADC+T2W	Average
B - WR	0.840	0.772	**0.830**	0.814
B - NR	0.839	**0.825**	**0.830**	0.831
F - WR	**0.855**	0.815	0.828	**0.833**
F - NR	0.820	0.803	0.814	0.812

Finally, Table 5.7 shows a comparison of the equivalent configurations shown in Table 5.6 but with data augmentations. One interesting observation from this table is that the experiment B without resampling attains a performance boost that allows it to outperform in the average AUC column all the other configurations besides F with resampling. This happens due to a major improvement in the performance using T2W sequences, that is in a majority of the experiments below 0.80, and achieves a higher maximum with this configuration. This also allows configuration B to be the only experiment to have a better performance without resampling than with resampling. Arguably, this can be associated with the combination of data augmentations that attenuated the degradation of the performance of ADC sequences and further improved the performance of T2W sequences.

For each fold used in the cross-validation, a Receiver operating charac-

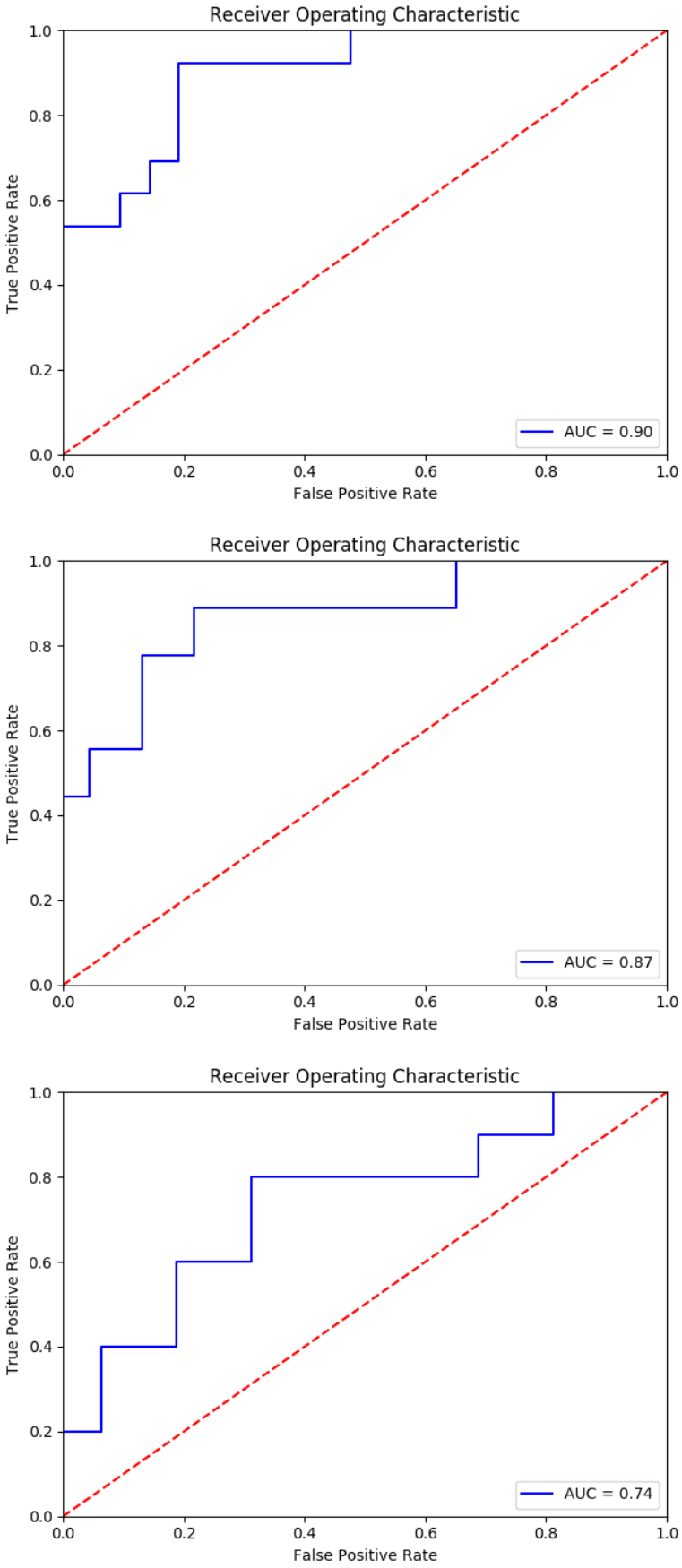


Figure 5.2. Receiver operating characteristic (ROC) curve of the best performing model for some selected folds out of the five folds present in the problem. Folds are respectively from top to bottom numerated as 1, 3 and 4.

teristic (ROC) curve was generated and the AUC for that fold calculated. In Figure 5.2, it is possible to see these curves for some of the folds of the experiment configuration E without the resampling on ADC sequences. This was the configuration to achieve the best AUC score overall when used on ADC sequences. The folds represented in the image are the folds 1, 3, and 4. These were selected because fold 1 is, alongside with following 2 and 5, the best performing fold, while 3 and 4 are the worst performing. Despite having an inferior performance, fold 3 is close to the performance of the others, whereas fold 4 is distant by 0.16. Moreover, the latter fold has performed poorly in all the configurations, in some cases being below 0.60.

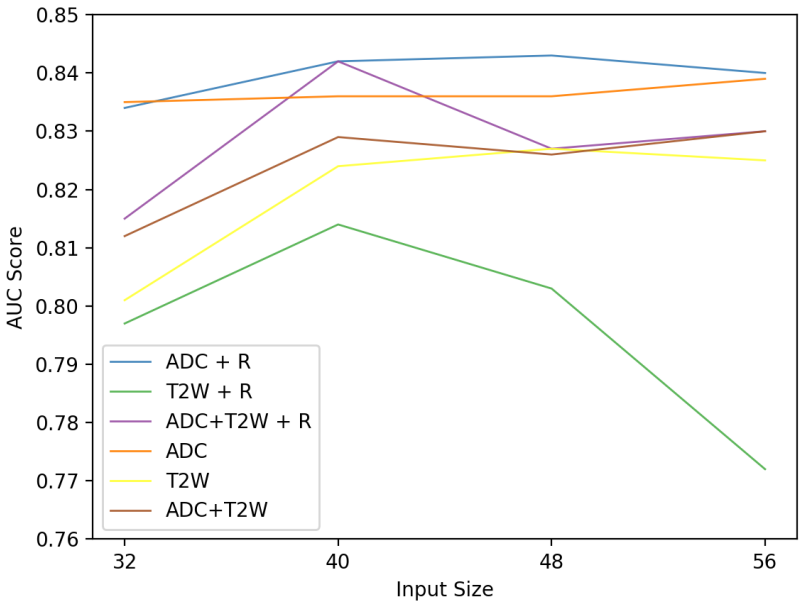


Figure 5.3. Comparison of the AUC score with the variation of the input size from a adjusted box crop with a size of 64. The results are shown for four different input sizes: 32, 40, 48 and 56. Although the selected size for performance comparison is 56, it is possible to see that some sizes show better performance, and that the selected size leads to a decreased performance of the T2W with resampling. The letter R in front of the sequence means that before cropping the lesion was resampled.

The final experiments regarding the binary classification of lesions were conducted to analyze variation in the performance of experiments with a crop size of 64 over different input sizes as seen in Figure 5.3. No data augmentations were used, and the crop was performed by adjusting the bounding box to the size of the lesion. Evaluations were performed in each of the three different input types (ADC, T2W, and ADC+T2W), with and without resampling for the following input sizes: 32, 40, 48, and 56. The results are interesting, and while it is impossible to find a common input size that maximizes the AUC in every experiment, it is possible to infer

that 32 is an ineffective size overall. Moreover, all the experiments using T2W, either alone or combined with ADC, have a peak in performance when using an input of 40. The performance does not suffer significant variations when the selected size is greater or equal to 40 in most of the experiments, however, the experiment using T2W sequences as input with resampling suffered a considerable degradation in performance with larger input sizes.

5.2.2 PI-RADS classification

For the classification of the lesions accordingly with their PIRADS score, six different experiments were performed to assess the performance of the model in this problem. Before the experiments, it was also verified that only one of the lesions that included mask had a PIRADS score below 3. It was problematic and difficult to assess if the model can perform and learn how to predict the score of lesions with low scores. This also increased the difficulty of the training and the optimization of the model for the existent lesions. Results for these experiments are given in terms of accuracy and confusion matrices. Since in this problem the classes are not independent and can be ordered in terms of proximity to each other, these matrices were constructed to indicate if the target class was missed by one or two classes. Moreover, the performance was also verified if the results aggregated both scores 4 and 5.

Table 5.8. Table with the results of the accuracy of the experiments for all the input sequences and with or without data augmentations. Bold highlights the best performance column-wise. This results were calculated for separated scores 4 and 5. All the values in this table represent AUC scores that were computed in a cross-validation setup for each of the five folds, and averaged across them.

Augmentations	ADC	T2W	ADC+T2W	Average
No	0.631	0.658	0.644	0.644
Yes	**0.644**	**0.664**	**0.651**	**0.653**

The first evaluation of the results was performed across all the five different PIRADS scores. The results are shown in Table 5.8 and it is also possible to analyze the impact of data augmentations. And despite the improvements provided by those, the results are below what was expected. This can be explained based on the skewness of the dataset that does not include any sample for lesions with a score of 2, and only one lesion has as a score of 1. Differently from what was seen in the previous classification problem, the performance of using T2W as input is superior

to the performance of using only ADC or combining both.

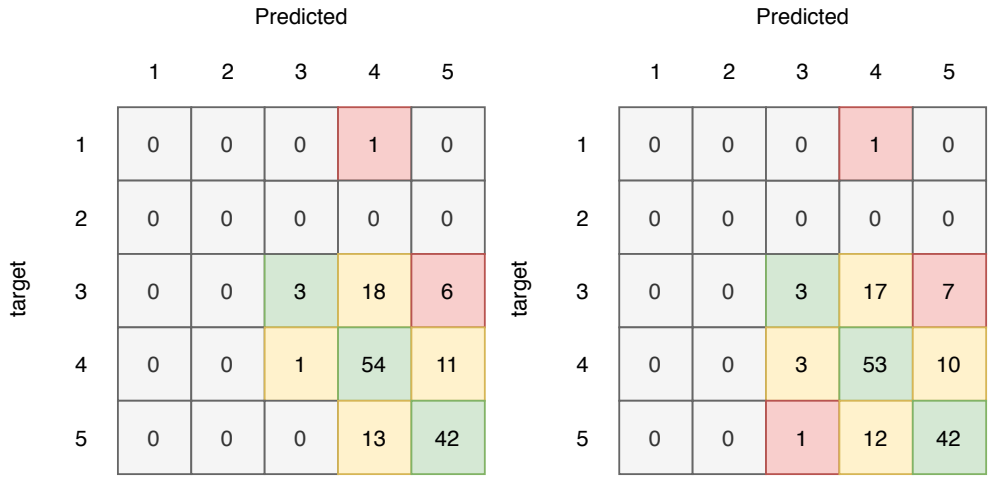


Figure 5.4. Confusion matrix for experiments to predict the PIRADS score using resampling, adjusted box cropping with a size of 32 and 28 input size on T2W sequences. Left image included data augmentation and had an accuracy of 66.4% whereas the other did not included data augmentations with an accuracy of 64.8%. Green - Correct, Yellow - Missed by 1 class , Red - Missed by more than one class

Figure 5.4 displays the results of Table 5.8 as confusion matrices. Both images represent experiments that used T2W sequences as input, however, the left image includes data augmentations, and as seen before a better performance. It is also possible to observe that there is an overestimation of lesions with a ground truth score of 3, being in a majority of the cases classified as 4 or 5. As mentioned before, the skewness of the dataset is most likely the reason for this overestimation.

Table 5.9. Table with the results of the accuracy of the experiments for all the input sequences and with or without data augmentations. Bold highlights the best performance column-wise. Results were calculated for classes 4 and 5 as a single class. All the values in this table represent AUC scores that were computed in a cross-validation setup for each of the five folds, and averaged across them.

Augmentations	ADC	T2W	ADC+T2W	Average
No	**0.852**	0.805	**0.846**	**0.834**
Yes	0.826	**0.826**	0.832	0.828

In order to explore the effects of this overestimation, scores 4 and 5 were grouped in the final result. Not only overestimation, but also underestimation is studied, however, with less impact. The results in Table 5.9 reflect the impact of using this joint group which for both ADC and combined sequences as the input changes the best performing model.

Despite the best results shown in Table 5.9, in Figure 5.5 it is possible to observe a deficiency of the model in classifying lesions with a score of 3. In the left image, 70% of lesions with label 3 are overestimated and 52% in the right image. These high values represent the effect of the distribution

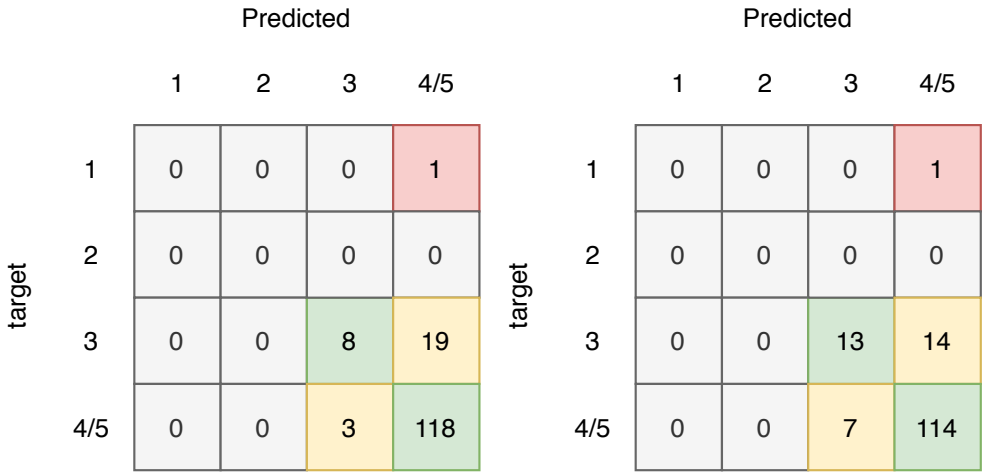


Figure 5.5. Confusion matrix for experiments to predict the PIRADS score using resampling, adjusted box cropping with a size of 32 and 28 input size without data augmentations. Left image combined both T2W and ADC and had an accuracy of 84.6% whereas the other did not included T2W sequences with an accuracy of 85.2%. Green - Correct, Yellow - Missed by 1 class , Red - Missed by more than one class

of the dataset in the results.

5.3 Detection and segmentation

The experiments to detect and segment prostate cancer lesions and to detect and segment the prostate were executed similarly. The same preprocessing techniques, the same models, and the same resampling sizes were all used in experiments for both problems. In this section the 3D U-Net and the 3D ResNet-18 models are experimented with different input sizes and data augmentations.

Table 5.10. Different configurations of the experiments. Includes the dimensions of the input to the model as well as the loss function to be minimized while optimizing the model.

Config.	Size	Loss Function
A	128x128x16	3D Dice
B	128x128x16	3D Dice+BCE
C	224x224x16	3D Dice
D	224x224x16	3D Dice+BCE

Table 5.10 lists the dimensions used and the input sizes for each configuration. The first size 128x128x16 was selected because it is the original size of ADC sequences regarding width and height. While 224x224x16 was used since the original width and height (360x360) of T2W sequences was too large to even fit one batch of size one in the GPU memory (RTX

2080ti 11 Gb). Models for these problems ran in parallel with two GPUs, which means that half of the batch was run in one GPU while the other half ran in the other. Moreover, the use of 16 slices is due to the architecture of the 3D U-Net model that reduces every dimension by half at each downsample block. Furthermore, the loss function used in each configuration varies also, with some configurations using exclusively a 3D dice loss, whereas the others use a combination of this latter loss with the binary cross-entropy loss applied pixel-wise. The evaluation of the experiments was performed with the resized ground truth mask to the size of the respective configuration.

5.3.1 Prostate segmentation

Regarding the process of segmenting the prostate, the experiments conducted had their results compared based on the resampling size used, based on the models and finally the impact of data augmentations was also analyzed and carefully discussed. The experiments using 3D U-Net ran considerably faster than the 3D ResNet-18 experiments. While the former used a learning rate of $5 * 10^{-4}$ the latter used a learning rate of $2 * 10^{-4}$. Both versions of the experiments ran for 200 epochs and used the Adam optimizer with a weight decay of 10^{-4}. Regarding the size of the batches used, configurations A and B used a batch size of four while C and D used two as batch size. This difference is due to GPU memory limitations.

Input size and loss function

Both the size of the input used to feed the model and the loss function used to calculate the error and backpropagate it in order to optimize the network are important factors in the performance of a model. Thus, before comparing the models against each other, it is important to compare and understand the effects of these parameters on the performance of each model.

Table 5.11 shows the performance of the 3D U-Net for all four configurations and the two different input sizes. It is possible to observe three different patterns. First, the performance using T2W sequences is always superior to the performance of using ADC sequences in the same configuration. Secondly, in configurations with the same loss, the performance is always superior, for all the input sequences in the configuration with a smaller input size. Finally, the performance improves when using a combination of loss functions.

Table 5.11. Experiments on the segmentation of the prostate using a 3D U-Net model in the different configurations. All the values in this table represent DICE scores that were computed in a cross-validation setup for each of the five folds, and averaged across them. The average column represents the mean of both sequences. Bold highlights the best performance column-wise.

Config.	Model	ADC	T2W	Average
A	3D U-Net	0.893	0.889	0.891
B	3D U-Net	**0.898**	**0.911**	**0.905**
C	3D U-Net	0.873	0.882	0.878
D	3D U-Net	0.893	0.908	0.901

Despite the improvements that the loss function brought to the 3D U-Net, Table 5.12 shows that the same is not reflected in the 3D ResNet-18 model. Not only is it difficult to analyze and detect any pattern, but also the use of different losses seemed to have a residual impact on the performance of the model. While having slightly more impact on ADC sequences, the input size impact can also be neglected when T2W sequences are used.

Table 5.12. Experiments on the segmentation of the prostate using a 3D ResNet-18 as model in the different configurations. All the values in this table represent DICE scores that were computed in a cross-validation setup for each of the five folds, and averaged across them. The average column represents the mean of both sequences. Bold highlights the best performance column-wise.

Config.	Model	ADC	T2W	Average
A	3D ResNet-18	**0.895**	0.886	**0.891**
B	3D ResNet-18	0.888	0.886	0.887
C	3D ResNet-18	0.876	0.888	0.882
D	3D ResNet-18	0.872	**0.891**	0.882

Models

Comparison between models is also useful in order to understand how they perform and in which conditions should a model be used. The comparison of these models was only performed on configurations that had an input size of 128x128x16, configurations A and B.

On Table 5.13 it is possible to observe that the performance of the models varies, especially when a combination of loss functions is used. While having a quite similar performance with the 3D dice loss, the 3D U-Net outperforms the other model when the binary cross-entropy is also used. These results are different from what would be expected considering the complexity difference of both models. However, the particular U-shaped architecture has shown to achieve slightly better results, and since it is also faster to train, further experiments in the segmentation of the prostate

Table 5.13. Comparison of the models 3D ResNet-18 and 3D U-Net in the configurations A and B, the ones that gave better results overall in the segmentation of the prostate. All the values in this table represent DICE scores that were computed in a cross-validation setup for each of the five folds, and averaged across them. The average column represents the mean of both sequences. Bold highlights the best performance column-wise.

Config.	Model	ADC	T2W	Average
A	3D U-Net	0.893	0.889	0.891
A	3D ResNet-18	0.895	0.886	0.891
B	3D U-Net	**0.898**	**0.911**	**0.905**
B	3D ResNet-18	0.888	0.886	0.887

use the 3D U-Net as the model.

Data Augmentations

In an attempt to further improve the performance of the experiments, a couple of variations of these were performed using data augmentation techniques. The data augmentations used did not, however, have a high magnitude, since it is important to restrain the distortions that may go beyond the distribution of the data in the real-world.

Table 5.14. Analysis of the impact of data augmentation techniques in the performance of the 3D U-Net model in the configurations A and B in segmenting the prostate. All the values in this table represent DICE scores that were computed in a cross-validation setup for each of the five folds, and averaged across them. The average column represents the mean of both sequences. Bold highlights the best performance column-wise.

Config.	Model	Augmentation	ADC	T2W	Average
A	3D U-Net	No	0.893	0.889	0.891
A	3D U-Net	Yes	0.902	0.900	0.901
B	3D U-Net	No	0.898	**0.911**	0.905
B	3D U-Net	Yes	**0.915**	0.910	**0.913**

As expected, the use of these augmentations slightly improved the performance of the model for both the loss functions used. The improvement margin was similar in both cases, however, due to the initial better performance of the combination of both losses, the use of augmentations in that experiment improves the DICE score to a value of 0.915. Another particularity is that these new experiments show better performance in ADC images when compared to T2W images, which is the opposite of previous experiments.

Final Results

Configuration B, including data augmentations and using a 3D U-Net as the model, was the best experiment of this problem. Thus, the experiment was selected to be analyzed fold-wise. Table 5.15 shows the performance of the model on the different folds, and it is possible to observe that the results do not vary significantly with fold 3 being the only one to be below a 0.910 DICE score.

Table 5.15. Performance in each fold of the model that performed the best in each sequence in the table 5.14. All the values in this table represent DICE scores . The average column represents the mean of both sequences. Bold highlights the best performance column-wise.

Fold	ADC	T2W	Average
1	0.9182	**0.9164**	**0.9173**
2	0.9136	0.9113	0.9125
3	0.9086	0.9023	0.9055
4	0.9146	0.9103	0.9125
5	**0.9188**	0.9116	0.9152

The results in this problem indicate that the segmentation is being performed with significant quality and the visualizations of the predictions are also similar to the visualizations of the ground truth. Figure 5.6 shows predictions for consecutive slices in the fold 5 of this experiment using ADC sequences as input.

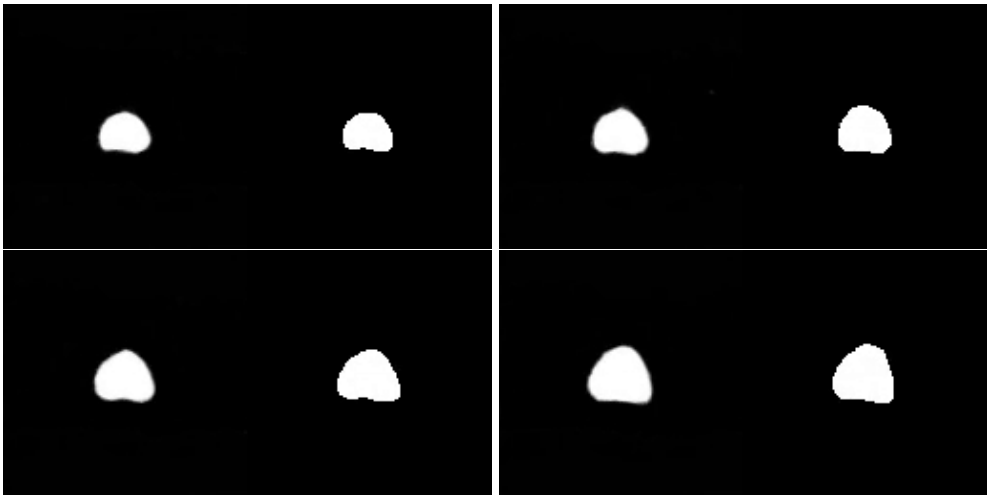


Figure 5.6. Sample predictions made by a 3D U-Net model from the fold 5 of experiment configuration B with data augmentations using ADC as input. Images represent slices with mask predictions on the same patient compared to the ground truth. Left side of each image contains the prediction whereas the right size contains the label.

Table 5.16 shows a comparison between the methods presented in this paper and other methods presented in literature. It is important to note that this is not an objective comparison of methods, since a different

dataset was used. However, these results denote that despite performing in different dataset, the presented method achieves and perhaps surpasses the expected performance.

Table 5.16. Comparison of the results presented in other papers that attempted to approach prostate segmentation with deep learning methods. Bold highlights the best performance column-wise. All the values in this table represent DICE scores.

Method	ADC	T2W	ADC+T2W	Best
Ours	**0.92**	**0.91**	-	**0.92**
[113]	0.86	0.86	**0.88**	0.88
[111]	-	0.85	-	0.85
[110]	-	**0.91**	-	0.91
[112]	-	-	0.87	0.87

5.3.2 Lesion segmentation

The segmentation of prostate cancer lesions had similar experiments to the previous problem regarding the segmentation of the prostate. Similar analyses were conducted on the input size, data augmentations, and different models. Despite these similarities that also include the same Adam optimizer, the same weight decay value of 10^{-4} and the same batch size for the different configurations (A and B have a batch size of four while C and D have a batch size of two in order to fit in the GPU) there are some differences. Due to being a slightly more complicated problem, the models trained for 250 epochs in this problem, and the learning rates were $2 * 10^{-4}$ for the 3D U-Net Model and $7 * 10^{-5}$ for the 3D ResNet-18 model. The results in the experiments regarding this problem reflect the difficulty of the task at hand.

Input size and loss function

Studying the loss function and the input size that optimizes the results of each model is important in order to understand the limitations of the model and how they can be explored.

An increase in the size of the input dramatically increases the time needed in the backward and forward pass of the training, and also the inference time. Therefore, increased training and inference can only be justified if these larger sizes represent significant performance improvements. Table 5.17 illustrates the results for the experiments using a 3D U-Net model. It is possible to observe two patterns in the table. Firstly, once more,

Table 5.17. Experiments on the segmentation of the prostate cancer lesions using a 3D U-Net as model in the different configurations. All the values in this table represent DICE scores that were computed in a cross-validation setup for each of the five folds, and averaged across them. The average column represents the mean of both sequences. Bold highlights the best performance column-wise.

Config.	Model	ADC	T2W	Average
A	3D U-Net	0.500	0.500	0.500
B	3D U-Net	**0.674**	**0.530**	**0.602**
C	3D U-Net	0.630	0.511	0.571
D	3D U-Net	0.652	0.516	0.584

when no data augmentations are being used, the combination of 3D dice loss and binary cross-entropy shows significant improvements regardless of the sequence given as input or the size of the sequence. Secondly, the performance of ADC sequences is usually superior to the performance of T2W sequences, and in most cases, there is a considerable difference between the performance of both.

Table 5.18. Experiments on the segmentation of prostate cancer lesions using a 3D ResNet-18 as model in the different configurations. All the values in this table represent DICE scores that were computed in a cross-validation setup for each of the five folds, and averaged across them. The average column represents the mean of both sequences. Bold highlights the best performance column-wise.

Config.	Model	ADC	T2W	Average
A	3D ResNet-18	0.500	0.385	0.443
B	3D ResNet-18	**0.501**	0.438	0.470
C	3D ResNet-18	0.500	0.501	0.501
D	3D ResNet-18	0.500	**0.503**	**0.502**

The results observed in Table 5.18 are rather poor. Despite the slight performance increase on T2W sequences when the size is increased, neither 3D dice nor the combined loss seem to be enough to optimize the model. Smaller sizes show a performance degradation of T2W sequences, thus an inferior performance compared to the ADC sequences.

Models

Differently from what was seen in the prostate segmentation problem, the models show a completely distinct performance when compared against each other.

Table 5.19 displays the values for these experiments, and it is possible to observe that not only the performance of the 3D ResNet-18 model does not go beyond 0.501, but it gets near that value with ADC sequences since T2W performs always poorly. The results of the best 3D U-Net model are

Table 5.19. Comparison of the models 3D ResNet-18 and 3D U-Net in the configurations A and B, the ones that gave better results overall in the segmentation of prostate cancer lesions. All the values in this table represent DICE scores that were computed in a cross-validation setup for each of the five folds, and averaged across them. The average column represents the mean of both sequences. Bold highlights the best performance column-wise.

Config.	Model	ADC	T2W	Average
A	3D U-Net	0.500	0.500	0.500
A	3D ResNet-18	0.500	0.385	0.443
B	3D U-Net	**0.674**	**0.530**	**0.602**
B	3D ResNet-18	0.501	0.438	0.470

34.5% and 21.0% higher for ADC and T2W sequences respectively when compared to the other model in the same configuration. Once more the performance of the 3D U-Net model outperforms the 3D ResNet-18, this time by a significant margin. Further experiments use the 3D U-Net as the model.

Data Augmentations

Data augmentation will once more be used to see how much further the performance of the model can be improved. The previous two best-performing experiments in the comparison of the models were used in two new experiments that included data augmentations in the Data loading stage.

Table 5.20. Analysis of the impact of data augmentation techniques in the performance of the 3D U-Net model in the configurations A and B to segment prostate cancer lesions. All the values in this table represent DICE scores that were computed in a cross-validation setup for each of the five folds, and averaged across them. The average column represents the mean of both sequences. Bold highlights the best performance column-wise.

Config.	Model	Augmentation	ADC	T2W	Average
A	3D U-Net	No	0.500	0.500	0.500
A	3D U-Net	Yes	**0.690**	0.491	0.591
B	3D U-Net	No	0.674	**0.530**	0.602
B	3D U-Net	Yes	0.681	0.526	**0.604**

The impact of the augmentations in the performance of the model did not seem to be positive when T2W sequences were given as input with results slightly worse than the ones without data augmentations. However, the performance in ADC sequences greatly benefited from the inclusion of such transformations. Observing the data presented in Table 5.20, it is possible to observe a small 1% improvement when ADC sequences were given to a model optimized with the combined loss function, yet the performance of this model was already far superior to previous models

without data augmentations. The largest performance boost is with the use of the 3D dice and ADC sequences, where the augmentations help the model to achieve a performance 38% higher when compared to the model without augmentations. This new value is the best DICE score in all the experiments.

Final Results

Using data augmentation in configuration A gave the best results for ADC inputs, while for T2W, configuration B without augmentations had the best results. Both these experiments can be seen in Table 5.21 with the results in each fold discriminated.

Table 5.21. Performance in each fold of the model that performed the best in each sequence in the table 5.20. All the values in this table represent DICE scores. The average column represents the mean of both sequences. Bold highlights the best performance column-wise.

Fold	ADC	T2W	Average
1	0.4686	0.2964	0.3825
2	0.6722	0.5214	0.5968
3	0.7003	0.5832	0.6418
4	**0.8302**	**0.6564**	**0.7433**
5	0.7896	0.6005	0.6951

It is important to note that the performance varies significantly between folds, especially between folds 1 and 4 with the latter having 77.1% and 121.5% higher performance in ADC and T2W sequences respectively when compared to the performance of the former. It is also worth noting that while T2W results are relatively poorer, the variation fold-wise is similar to the variation that occurs with ADC sequences. The best performing fold for ADC sequences displays somewhat promising results.

Figure 5.7 displays some predictions given of the model for the fold 4 of ADC sequences. The predictions have some distance from the ground truth, either by being smaller, having a somewhat more simplistic shape, or simply by appearing one or two slices before expected.

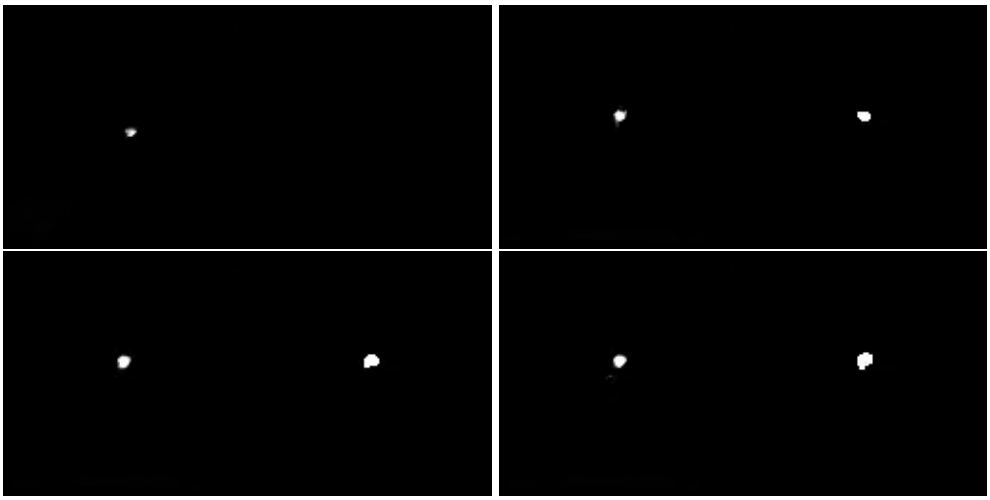


Figure 5.7. Sample predictions made by a 3D U-Net model from the fold 4 of experiment configuration A with data augmentations using ADC as input. Images represent slices with mask predictions on the same patient compared to the ground truth. Left side of each image contains the prediction whereas the right size contains the label.

6. Conclusion and Future Work

This research approached deep learning methods for prostate cancer diagnosis problems based on multi-parametric magnetic resonance images. From the information contained in these images, the purpose was to understand which problems could be solved or addressed using computer vision methods and deep learning. Some of the problems required lesions to be classified. Others required the prostate or prostate cancer lesions to be detected in the MRI and segmented accordingly to the annotations given by experts.

Deep learning approaches dwell in a world that is unreachable for some domains, the world of big data and massive volumes of information. This is, for now, the case of medical use cases. The information is very limited, not only due to the requirements of experts, but also due to privacy matters. Hence, it was a challenge that required a thorough analysis of the problem, the preprocessing techniques, the models and the hyperparameters to use. Overall, a deep analysis of the performance of each problem was conducted and the results discussed.

The image quality and the annotations carefully made by experts supported the learning process of the model. However, other characteristics were considerably less positive. First of all, the small number of samples hurt the performance. Secondly, the distribution of the PIRADS score of lesions with mask increased the difficulty of the problem. Moreover, the overall number of patients was rather small. This affected the performance in some tasks, such as the segmentation of lesions. The latter task due to its nature requires significantly more data in order to achieve better results.

Regarding the classification of lesions, the performance varied greatly between tasks. Results in the cross-validation score presented in Section 5.2.1 for the clinical significance of lesions were superior to the 0.84 AUC

obtained by XmasNet in the ProstateX challenge [27]. This while using fewer sequences as input, thus also using fewer parameters [27]. And while the challenge dataset had poor data quality, the number of images, from the 346 patients, was considerably higher. Hence, the results not only matched the expectations, but they can also be considered promising. It is expected that if the number of samples increases while keeping the quality of the data, the performance may also increase. Despite these results, the distribution of the PIRADS score from lesion with masks did not allow the performance in this task to be the expected. The performance was significantly worse than the binary classification and was not nearly enough to be adopted in a clinical situation or the problem to be considered solved.

The process of detecting those lesions and segment them is rather complicated. Thus, in order to test and validate the research methodology to be followed, one experiment was conducted before segmenting lesions. The extra prostate mask annotations given in the dataset were used to detect and segment the prostate itself. The results shown in Section 5.3.1 exceeded expectations. Not only both the models perform as required on both sizes and with both losses. But experiments were further improved with data augmentation techniques. The results were not only promising, but they also surpassed performance in previous research [110, 111, 112, 113]. Therefore, all the experiments were repeated for lesion segmentation, yet, without achieving results close to the previous problem. The results indicate, however, some promising details. Such as the performance increase in the experiments that used data augmentations. These details may indicate that increasing the data in the dataset is also crucial to improve upon the 0.69 DICE score obtained in this task. And while some folds performed poorly, others presented results that achieved a DICE score up to 0.83. Understanding the reason behind these distinct results might in the future lead to gains in the performance. And despite the results in this task being inferior to the ones in the previous segmentation task, it is important to note that the results are somewhat slightly better than the ones presented by previous researchers [112, 113].

The relevance of the work presented in this thesis goes beyond the simple analysis of the results. The thesis demonstrates that deep learning methods can be applied to solve computer vision tasks on multi-parametric magnetic resonance images regarding the prostate cancer diagnosis. Despite only being able to achieve the desired results on two of the four

tasks. The remaining tasks have shown a potential that can be achieved by addressing some of the flaws regarding the data previously mentioned.

Further research is still needed to study other plausibly relevant questions. Studies on the performance of the model in classifications tasks can be conducted by averaging the prediction of two models trained in different sequences accordingly to different configurations. It can also be analyzed if the Gleason score can be better predicted through this data compared to the PIRADS score. Regarding the segmentation, the registration of both sequences or the addition of more augmentations can be empirically studied. All these future work is based on the results given in this thesis that conducted and presented more than 112 different experiments. More experiments were conducted in the background and were not presented in the thesis but they were essential to tune the necessary hyperparameters for the final experiments. The goals of the thesis were therefore achieved, and the ground for future work on this dataset established.

Bibliography

[1] Bernard W Stewart and Christopher P Wild. *World Cancer Report 2014*. International Agency for Research on Cancer, 2014.

[2] Key Statistics for Prostate Cancer | Prostate Cancer Facts.

[3] Alfred Sherwood Romer and Thomas S. (Thomas Sturges) Parsons. *The vertebrate body*. Saunders, 1977.

[4] Richard P Gallagher and Neil Fleshner. Prostate cancer: 3. Individual risk factors. Technical Report 7, 1998.

[5] Maurice P A Zeegers, Annemarie Jellema, and Harry Ostrer. Empiric risk of prostate carcinoma for relatives of patients with prostate carcinoma. *Cancer*, 97(8):1894–1903, 2003.

[6] David C Miller, Khaled S Hafez, Andrew Stewart, James E Montie, and John T Wei. Prostate carcinoma presentation, diagnosis, and staging. *Cancer*, 98(6):1169–1178, 2003.

[7] The American Cancer Society. Can Prostate Cancer Be Found Early?, 2019.

[8] PDQ Screening and Prevention Editorial Board. *Prostate Cancer Screening (PDQ®): Health Professional Version*. 2002.

[9] Leonard G. Gomella, Xiaolong S. Liu, Edouard J. Trabulsi, Wm Kevin Kelly, Ronald Myers, Timothy Showalter, Adam Dicker, and Richard Wender. Screening for prostate cancer: The current evidence and guidelines controversy. *Canadian Journal of Urology*, 18(5):5875–5883, oct 2011.

[10] Leen Naji, Harkanwal Randhawa, Zahra Sohani, Brittany Dennis, Deanna Lautenbach, Owen Kavanagh, Monica Bawor, Laura Banfield, and Jason Profetto. Digital Rectal Examination for Prostate Cancer Screening in Primary Care: A Systematic Review and Meta-Analysis. *Annals of family medicine*, 16(2):149–154, mar 2018.

[11] David C. Grossman, Susan J. Curry, Douglas K. Owens, Kirsten Bibbins-Domingo, Aaron B. Caughey, Karina W. Davidson, Chyke A. Doubeni, Mark Ebell, John W. Epling, Alex R. Kemper, Alex H. Krist, Martha Kubik, C. Seth Landefeld, Carol M. Mangione, Michael Silverstein, Melissa A. Simon, Albert L. Siu, and Chien Wen Tseng. Screening for prostate cancer USPreventive servicestaskforcerecommendation statement. *JAMA - Journal of the American Medical Association*, 319(18):1901–1913, may 2018.

[12] Richard M Hoffman. Screening for prostate cancer, 2020.

[13] Matthew J Roberts, Harrison Y Bennett, Patrick N Harris, Michael Holmes, Jeremy Grummet, Kurt Naber, and Florian M E Wagenlehner. Prostate Biopsy Related Infection: a Systematic Review of Risk Factors, Prevention Strategies and Management Approaches. *Urology*, 2016.

[14] A. El-Shater Bosaily, C. Parker, L. C. Brown, R. Gabe, R. G. Hindley, R. Kaplan, M. Emberton, H. U. Ahmed, Mark Emberton, Hashim Ahmed, Ahmed El Shater Bosaily, Alex Kirkham, Alex Freeman, Charles Jameson, Richard Hindley, Christopher Parker, Colin Cooper, Robert Oldroyd, Richard Kaplan, Louise Brown, Rhian Gabe, Yolanda Collaco-Moraes, Cybil Adusei, Katie Ward, Sophie Stewart, Katie Thompson Claire Mulrenan, Hannah Gardner, Carlos Diaz-Montana, Chris Coyle, Mark Sculpher, Rita Faria, David Guthrie, John Chester, Richard Cowan, Michael Jewitt, H. Ahmed, J. Coe, A. El-Shater Bosaily, M. Emberton, A. Freeman, M. Hung, C. Jameson, A. Kirkham, S. Punwani, R. Scott, Richard Hindley, A. Edwards, H. El-Mahallawi, D. Peppercorn, J. Smith, A. Thrower, M. Winkler, K. Ansu, T. Barwick, S. Edwards, L. Honeyfield, N. Qazi, B. Statton, V. Stewart, E. Temple, N. Burns-Cox, P. Burn, K. Gordon, H. Routley, A. Maccormick, D. Paterson, A. Henderson, E. Bernsten, R. Casey, D. Day, S. Ghosh, J. James, P. J. McMillan, G. Russell, R. Persad, J. Ash-Miles, M. Elmahdy, S. Pandian, C. Shiridzinomwa, M. Sohail, A. Treasure, M. Ghei, V. Conteh, L. Harbin, R. Katz, J. Kumaradevan, A. Trinidade, A. Verjee, T. Dudderidge, J. Smart, D. Rosario, J. Catto, F. Selem, I. Shergill, and S. Agarwal. PROMIS - Prostate MR imaging study: A paired validating cohort study evaluating the role of multi-parametric MRI in men with clinical suspicion of prostate cancer. *Contemporary Clinical Trials*, 42:26–40, may 2015.

[15] Amit R Patel and J Stephen Jones. Optimal biopsy strategies for the diagnosis and staging of prostate cancer. *Current opinion in urology*, 19(3):232–7, may 2009.

[16] H. Y. Bennett, M. J. Roberts, S. A.R. Doi, and R. A. Gardiner. The global burden of major infectious complications following prostate biopsy, jun 2016.

[17] David Bonekamp, Michael A. Jacobs, Riham El-Khouli, Dan Stoianovici, and Katarzyna J. Macura. Advancements in MR imaging of the prostate: From diagnosis to interventions. *Radiographics*, 31(3):677–704, may 2011.

[18] Geoffrey A. Sonn, Shyam Natarajan, Daniel J.A. Margolis, Malu MacAiran, Patricia Lieu, Jiaoti Huang, Frederick J. Dorey, and Leonard S. Marks. Targeted biopsy in the detection of prostate cancer using an office based magnetic resonance ultrasound fusion device. *Journal of Urology*, 189(1):86–92, 2013.

[19] Leonard Marks, Shelena Young, and Shyam Natarajan. MRI-ultrasound fusion for guidance of targeted prostate biopsy, jan 2013.

[20] Timur H. Kuru, Matthias C. Roethke, Jonas Seidenader, Tobias Simpfendörfer, Silvan Boxler, Khalid Alammar, Philip Rieker, Valentin I. Popeneciu, Wilfried Roth, Sascha Pahernik, Heinz Peter Schlemmer, Markus Hohenfellner, and Boris A. Hadaschik. Critical evaluation of magnetic resonance imaging targeted, transrectal ultrasound guided transperineal fusion biopsy

for detection of prostate cancer. *Journal of Urology*, 190(4):1380–1386, oct 2013.

[21] Morgan R. Pokorny, Maarten De Rooij, Earl Duncan, Fritz H. Schröder, Robert Parkinson, Jelle O. Barentsz, and Leslie C. Thompson. Prospective study of diagnostic accuracy comparing prostate cancer detection by transrectal ultrasound-guided biopsy versus magnetic resonance (MR) imaging with subsequent mr-guided biopsy in men without previous prostate biopsies. *European Urology*, 66(1):22–29, 2014.

[22] PROSTATEx - Overview.

[23] PROSTATEx-2 Challenge.

[24] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg C Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona J Gilbert, Mark Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher J Kelly, Dominic King, Joseph R Ledsam, David Melnick, Hormuz Mostofi, Lily Peng, Joshua Jay Reicher, Bernardino Romera-Paredes, Richard Sidebottom, Mustafa Suleyman, Daniel Tse, Kenneth C Young, Jeffrey De Fauw, and Shravya Shetty. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

[25] Hannu J Aronen. Improved Prostate Cancer Diagnosis - Combination of Magnetic Resonance Imaging and Biomarkers - Full Text View - ClinicalTrials.gov.

[26] Harri Merisaari, Ivan Jambor, Otto Ettala, Peter J. Boström, Ileana Montoya Perez, Janne Verho, Aida Kiviniemi, Kari Syvänen, Esa Kähkönen, Lauri Eklund, Tapio Pahikkala, Paula Vainio, Jani Saunavaara, Hannu J. Aronen, and Pekka Taimen. IMPROD biparametric MRI in men with a clinical suspicion of prostate cancer (IMPROD Trial): Sensitivity for prostate cancer detection in correlation with whole-mount prostatectomy sections and implications for focal therapy. *Journal of Magnetic Resonance Imaging*, 50(5):1641–1650, nov 2019.

[27] Saifeng Liu, Huaixiu Zheng, Yesu Feng, and Wei Li. Prostate Cancer Diagnosis using Deep Learning with 3D Multiparametric MRI. mar 2017.

[28] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. jun 2016.

[29] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

[30] Ethan Basch, Thomas K Oliver, Andrew Vickers, Ian Thompson, Philip Kantoff, Howard Parnes, D Andrew Loblaw, Bruce Roth, James Williams, and Robert K Nam. Screening for prostate cancer with prostate-specific antigen testing: American Society of Clinical Oncology Provisional Clinical Opinion. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 30(24):3020–5, aug 2012.

[31] A. M. D. Wolf, R. C. Wender, R. B. Etzioni, I. M. Thompson, A. V. D'Amico, R. J. Volk, D. D. Brooks, C. Dash, I. Guessous, K. Andrews, C. DeSantis, and R. A. Smith. American Cancer Society Guideline for the Early Detection of Prostate Cancer: Update 2010. *CA: A Cancer Journal for Clinicians*, 60(2):70–98, mar 2010.

[32] Kirsten L. Greene, Peter C. Albertsen, Richard J. Babaian, H. Ballentine Carter, Peter H. Gann, Misop Han, Deborah Ann Kuban, A. Oliver Sartor, Janet L. Stanford, Anthony Zietman, and Peter Carroll. Prostate Specific Antigen Best Practice Statement: 2009 Update. *The Journal of Urology*, 2013.

[33] Prostate Cancer Guidelines: Guidelines Summary, Genetic Testing, Multiparametric Magnetic Resonance Imaging.

[34] Prostate Cancer: Early Detection Guideline - American Urological Association.

[35] William J. Catalona, Jerome P. Richie, Frederick R. Ahmann, M'Liss A. Hudson, Peter T. Scardino, Robert C. Flanigan, Jean B. Dekernion, Timothy L. Ratliff, Louis R. Kavoussi, Bruce L. Dalkin, W. Bedford Waters, Michael T. Macfarlane, and Paula C. Southwick. Comparison of Digital Rectal Examination and Serum Prostate Specific Antigen in the Early Detection of Prostate Cancer: Results of a Multicenter Clinical Trial of 6,630 Men. *Journal of Urology*, 151(5):1283–1290, may 1994.

[36] Vicki Velonas, Henry Woo, Cristobal Remedios, and Stephen Assinder. Current Status of Biomarkers for Prostate Cancer. *International Journal of Molecular Sciences*, 14(6):11034–11060, may 2013.

[37] Matthew R. Cooperberg, Peter R. Carroll, and Laurence Klotz. Active surveillance for prostate cancer: Progress and promise, sep 2011.

[38] M Ghei, S Pericleous, A Kumar, R Miller, S Nathan, and B H Maraj. Finger-guided transrectal biopsy of the prostate: a modified, safer technique. *Annals of the Royal College of Surgeons of England*, 87(5):386–7, sep 2005.

[39] Mike Stuckey. Tales from a prostate biopsy - Health - Men's health - Low Blow | NBC News.

[40] Thomas Hambrock, Caroline Hoeks, Christina Hulsbergen-Van De Kaa, Tom Scheenen, Jurgen Fütterer, Stefan Bouwense, Inge Van Oort, Fritz Schröder, Henkjan Huisman, and Jelle Barentsz. Prospective assessment of prostate cancer aggressiveness using 3-T diffusion-weighted magnetic resonance imaging-guided biopsies versus a systematic 10-core transrectal ultrasound prostate biopsy cohort. *European Urology*, 61(1):177–184, jan 2012.

[41] Neil Bell, Sarah Connor Gorber, Amanda Shane, Michel Joffres, Harminder Singh, James Dickinson, Elizabeth Shaw, Lesley Dunfield, and Marcello Tonelli. Recommendations on screening for prostate cancer with the prostate-specific antigen test. *CMAJ*, 186(16):1225–1234, nov 2014.

[42] Matthew J. Roberts, Harrison Y. Bennett, Patrick N. Harris, Michael Holmes, Jeremy Grummet, Kurt Naber, and Florian M.E. Wagenlehner. Prostate Biopsy-related Infection: A Systematic Review of Risk Factors, Prevention Strategies, and Management Approaches, jun 2017.

[43] Baris Turkbey and Peter L. Choyke. Future Perspectives and Challenges of Prostate MR Imaging, mar 2018.

[44] Oliver Sartor and Johann S. De Bono. Metastatic prostate cancer, feb 2018.

[45] Bert Dhondt, Elise De Bleser, Tom Claeys, Sarah Buelens, Nicolaas Lumen, Jo Vandesompele, Anneleen Beckers, Valerie Fonteyne, Kim Van der Eecken, Aurélie De Bruycker, Jérôme Paul, Pierre Gramme, and Piet Ost. Discovery and validation of a serum microRNA signature to characterize oligo- and polymetastatic prostate cancer: not ready for prime time. *World Journal of Urology*, 37(12):2557–2564, dec 2019.

[46] Christopher J.D. Wallis, Alyson L. Mahar, Richard Choo, Sender Herschorn, Ronald T. Kodama, Prakesh S. Shah, Cyril Danjoux, Steven A. Narod, and Robert K. Nam. Second malignancies after radiotherapy for prostate cancer: Systematic review and meta-analysis, mar 2016.

[47] V. Mouraviev, B. Evans, and T. J. Polascik. Salvage prostate cryoablation after primary interstitial brachytherapy failure: A feasible approach. *Prostate Cancer and Prostatic Diseases*, 9(1):99–101, mar 2006.

[48] Christopher J.D. Wallis, Adam Glaser, Jim C. Hu, Hartwig Huland, Nathan Lawrentschuk, Daniel Moon, Declan G. Murphy, Paul L. Nguyen, Matthew J. Resnick, and Robert K. Nam. Survival and Complications Following Surgery and Radiation for Localized Prostate Cancer: An International Collaborative Review, jan 2018.

[49] Surgery for Prostate Cancer.

[50] Dragan Ilic, Sue M. Evans, Christie Ann Allan, Jae Hung Jung, Declan Murphy, and Mark Frydenberg. Laparoscopic and robotic-assisted versus open radical prostatectomy for the treatment of localised prostate cancer, sep 2017.

[51] Erectile Dysfunction After Prostate Cancer | Johns Hopkins Medicine.

[52] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-Rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury. Deep Neural Networks for Acoustic Modeling in Speech Recognition. Technical report.

[53] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. Technical report.

[54] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. Technical report.

[55] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, jun 2017.

[56] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. apr 2018.

[57] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9351, pages 234–241. Springer Verlag, may 2015.

[58] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo R-CNN based 3D Object Detection for Autonomous Driving. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:7636–7644, feb 2019.

[59] Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. Video description: A survey of methods, datasets, and evaluation metrics, oct 2019.

[60] Parvin Yousefikamal. Breast Tumor Classification and Segmentation using Convolutional Neural Networks. may 2019.

[61] Vinod Nair and Geoffrey E Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, Madison, WI, USA, 2010. Omnipress.

[62] Diganta Misra. Mish: A self regularized non-monotonic neural activation function. *arXiv preprint arXiv:1908.08681*, 2019.

[63] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier Nonlinearities Improve Neural Network Acoustic Models. Technical report, 2013.

[64] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning, may 2015.

[65] Prateek Jain and Purushottam Kar. Non-convex Optimization for Machine Learning. *Foundations and Trends in Machine Learning*, 10(3-4):142–336, dec 2017.

[66] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[67] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, aug 2017.

[68] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*.

[69] Stuart Russel and Peter Norvig. *Artificial intelligence—a modern approach 3rd Edition*. 2012.

[70] Geoffrey E. Hinton and Terrence J. (Terrence Joseph) Sejnowski. *Unsupervised learning : foundations of neural computation*. MIT Press, 1999.

[71] Matt Taddy. *Business data science : combining machine learning and economics to optimize, automate, and accelerate business decisions*.

[72] Paul H.Sra. *Optimization for Machine Learning (Neural Information Processing Series)*. 2011.

[73] Douglas M. Hawkins. The Problem of Overfitting, jan 2004.

[74] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243, mar 1968.

[75] Kunihiko Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1(2):119–130, jan 1988.

[76] D. Marr and E. Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London - Biological Sciences*, 207(1167):187–217, 1980.

[77] CS231n Convolutional Neural Networks for Visual Recognition.

[78] Fisher Yu and Vladlen Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, nov 2015.

[79] torch.nn — PyTorch master documentation.

[80] Hamed Habibi Aghdam and Elnaz Jahani Heravi. *Guide to Convolutional Neural Networks*. Springer International Publishing, 2017.

[81] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, dec 1989.

[82] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings*. International Conference on Learning Representations, ICLR, dec 2015.

[83] Benjamin Graham. Fractional Max-Pooling. dec 2014.

[84] Linda G. Shapiro and George C. Stockman. *Computer vision*. Prentice Hall, 2001.

[85] Lawrence W Lee and San Francisco. United States US 2004OO59754A1 (12) Patent Application Publication. Technical Report 10, jul 2003.

[86] Dzung L. Pham, Chenyang Xu, and Jerry L. Prince. Current Methods in Medical Image Segmentation. *Annual Review of Biomedical Engineering*, 2(1):315–337, aug 2000.

[87] Mohamad Forouzanfar, Nosratallah Forghani, and Mohammad Teshnehlab. Parameter optimization of improved fuzzy c-means clustering algorithm for brain MR image segmentation. *Engineering Applications of Artificial Intelligence*, 23(2):160–168, mar 2010.

[88] E Ben George. MR Brain Image Segmentation using Bacteria Foraging Optimization Algorithm. Technical report.

[89] Sridharan Kamalakannan, Arunkumar Gururajan, Hamed Sari-Sarraf, Rodney Long, and Sameer Antani. Double-edge detection of radiographic lumbar vertebrae images using pressurized open DGVF snakes. *IEEE Transactions on Biomedical Engineering*, 57(6):1325–1334, jun 2010.

[90] Nelly Tan, Daniel J.A. Margolis, Timothy D. McClure, Albert Thomas, David S. Finley, Robert E. Reiter, Jiaoti Huang, and Steven S. Raman. Radical prostatectomy: Value of prostate MRI in surgical planning, aug 2012.

[91] Amirhessam Tahmassebi, Georg J Wengert, Thomas H Helbich, Zsuzsanna Bago-Horvath, Sousan Alaei, Rupert Bartsch, Peter Dubsky, Pascal Baltzer, Paola Clauser, Panagiotis Kapetas, Elizabeth A Morris, Anke Meyer-Baese, and Katja Pinker. Impact of Machine Learning With Multiparametric Magnetic Resonance Imaging of the Breast for Early Prediction of Response to Neoadjuvant Chemotherapy and Survival Outcomes in Breast Cancer Patients. 2018.

[92] Maria Adele Marino, Thomas Helbich, Pascal Baltzer, and Katja Pinker-Domenig. Multiparametric MRI of the breast: A review. *Journal of Magnetic Resonance Imaging*, 47(2):301–315, feb 2018.

[93] Jelle O. Barentsz, Jonathan Richenberg, Richard Clements, Peter Choyke, Sadhna Verma, Geert Villeirs, Olivier Rouviere, Vibeke Logager, and Jurgen J. Fütterer. ESUR prostate MR guidelines 2012. *European Radiology*, 22(4):746–757, 2012.

[94] Jeffrey C. Weinreb, Jelle O. Barentsz, Peter L. Choyke, Francois Cornud, Masoom A. Haider, Katarzyna J. Macura, Daniel Margolis, Mitchell D. Schnall, Faina Shtern, Clare M. Tempany, Harriet C. Thoeny, and Sadna Verma. PI-RADS Prostate Imaging - Reporting and Data System: 2015, Version 2. *European Urology*, 69(1):16–40, jan 2016.

[95] J. E. Thompson, P. J. Van Leeuwen, D. Moses, R. Shnier, P. Brenner, W. Delprado, M. Pulbrook, M. Böhm, A. M. Haynes, A. Hayen, and P. D. Stricker. The diagnostic performance of multiparametric magnetic resonance imaging to detect significant prostate cancer. *Journal of Urology*, 195(5):1428–1435, may 2016.

[96] H. A. Vargas, A. M. Hötker, D. A. Goldman, C. S. Moskowitz, T. Gondo, K. Matsumoto, B. Ehdaie, S. Woo, S. W. Fine, V. E. Reuter, E. Sala, and H. Hricak. Updated prostate imaging reporting and data system (PIRADS v2) recommendations for the detection of clinically significant prostate cancer using multiparametric MRI: critical evaluation using whole-mount pathology as standard of reference. *European Radiology*, 26(6):1606–1612, jun 2016.

[97] Nicola Schieda, Jeffrey S. Quon, Christopher Lim, Mohammed El-Khodary, Wael Shabana, Vivek Singh, Christopher Morash, Rodney H. Breau, Matthew D.F. McInnes, and Trevor A. Flood. Evaluation of the European Society of Urogenital Radiology (ESUR) PI-RADS scoring system for assessment of extra-prostatic extension in prostatic carcinoma. *European Journal of Radiology*, 84(10):1843–1848, oct 2015.

[98] Hamidreza Abdi, Farshad Pourmalek, Homayoun Zargar, Triona Walshe, Alison C. Harris, Silvia D. Chang, Christopher Eddy, Alan I. So, Martin E. Gleave, Lindsay Machan, S. Larry Goldenberg, and Peter C. Black. Multiparametric magnetic resonance imaging enhances detection of significant tumor in patients on active surveillance for prostate cancer. *Urology*, 85(2):423–429, feb 2015.

[99] How to Read Your Prostate MRI Report.

[100] A. Abragam. *Principles of Nuclear Magnetism (The International Series of Monographs on Physics)*. Clarendon Press, 1983.

[101] Timothy D.W. Claridge. *High-Resolution NMR Techniques in Organic Chemistry: Third Edition*. Elsevier Inc., may 2016.

[102] University of Wisconsin. Magnetic Resonance Imaging.

[103] Keith A. Johnson. basic MR imaging.

[104] Joel Dunn and Paul Kenneth Marsden. Hyperpolarised 13C NMR Spectroscopy for studying cardiac metabolism View project PET/MRI Reconstruction View project. *Physics in Medicine & Biology*, 56(13):6441, 2015.

[105] Klaus Dietmar Merboldt, Wolfgang Hanicke, and Jens Frahm. Self-diffusion NMR imaging using stimulated echoes. *Journal of Magnetic Resonance (1969)*, 64(3):479–486, oct 1985.

[106] Dow-Mu Koh and David J. Collins. Diffusion-Weighted MRI in the Body: Applications and Challenges in Oncology. *American Journal of Roentgenology*, 188(6):1622–1635, jun 2007.

[107] Mark Hammer. MRI Physics: Diffusion-Weighted Imaging - XRayPhysics.

[108] Allen D. Elster. Causes of restricted diffusion - Questions and Answers in MRI.

[109] Sunghwan Yoo, Isha Gujrathi, Masoom A. Haider, and Farzad Khalvati. Prostate Cancer Detection using Deep Convolutional Neural Networks. *Scientific Reports*, 9(1):1–10, dec 2019.

[110] Davood Karimi, Golnoosh Samei, Yanan Shao, and Septimiu Salcudean. A deep learning-based method for prostate segmentation in T2-weighted magnetic resonance imaging. jan 2019.

[111] Minh Nguyen Nhat To, Dang Quoc Vu, Baris Turkbey, Peter L. Choyke, and Jin Tae Kwak. Deep dense multi-path neural network for prostate segmentation in magnetic resonance imaging. *International Journal of Computer Assisted Radiology and Surgery*, 13(11):1687–1696, nov 2018.

[112] Aharon Feldman, Zhenzhen Dai, Eric Carver, Chang Liu, Joon Lee, Milan Pantelic, Mohamed Elshaikh, and Ning Wen. Prostate and Prostate Cancer Segmentation Using a Deep Learning-Based Object Detection Algorithm. *Clinical Research*, may 2019.

[113] Patrick Schelb, Simon Kohl, Jan Philipp Radtke, Manuel Wiesenfarth, Philipp Kickingereder, Sebastian Bickelhaupt, Tristan Anselm Kuder, Albrecht Stenzinger, Markus Hohenfellner, Heinz-Peter Schlemmer, Klaus H. Maier-Hein, and David Bonekamp. Classification of Cancer at Prostate MRI: Deep Learning versus Clinical PI-RADS Assessment. *Radiology*, 293(3):607–617, dec 2019.

[114] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? Technical report.

[115] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition. Technical report.

A. Prostate Segmentation

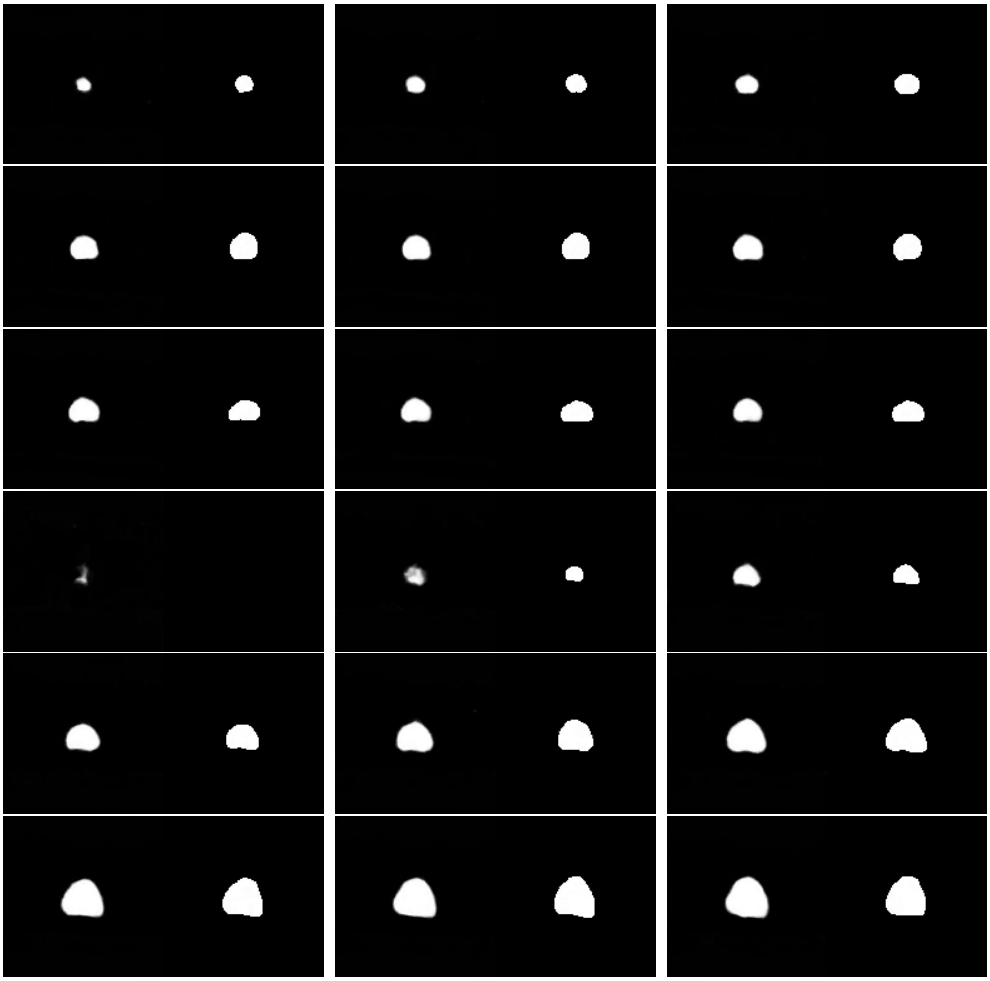


Figure 1.1. Sample predictions made by a 3D U-Net model from the fold 5 of experiment configuration B with data augmentations using ADC as input. Images represent slices with mask predictions on the same patient compared to the ground truth. Left side of each image contains the prediction whereas the right size contains the label.

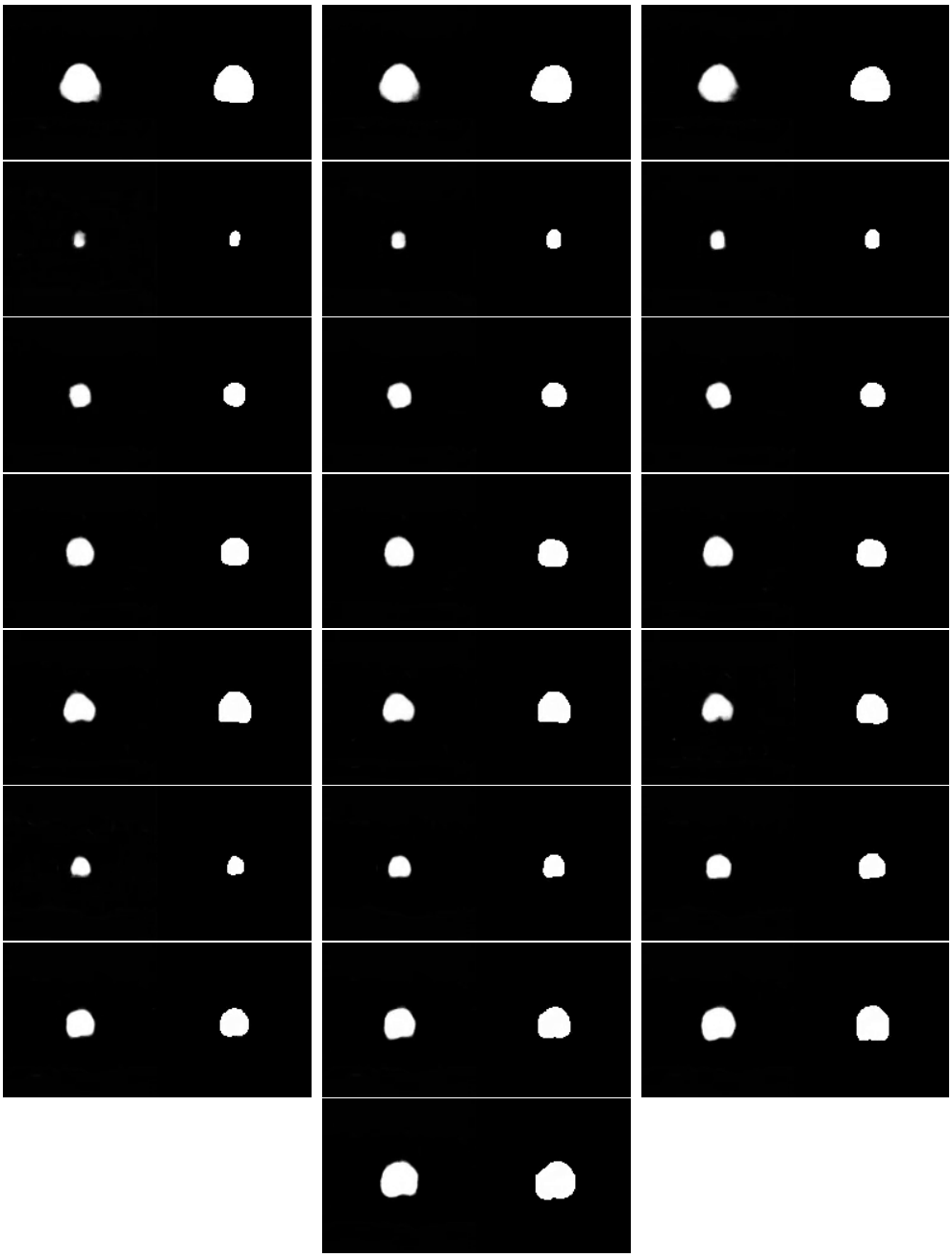


Figure 1.2. Sample predictions made by a 3D U-Net model from the fold 5 of experiment configuration B with data augmentations using ADC as input. Images represent slices with mask predictions on the same patient compared to the ground truth. Left side of each image contains the prediction whereas the right size contains the label.

B. Lesions Segmentation

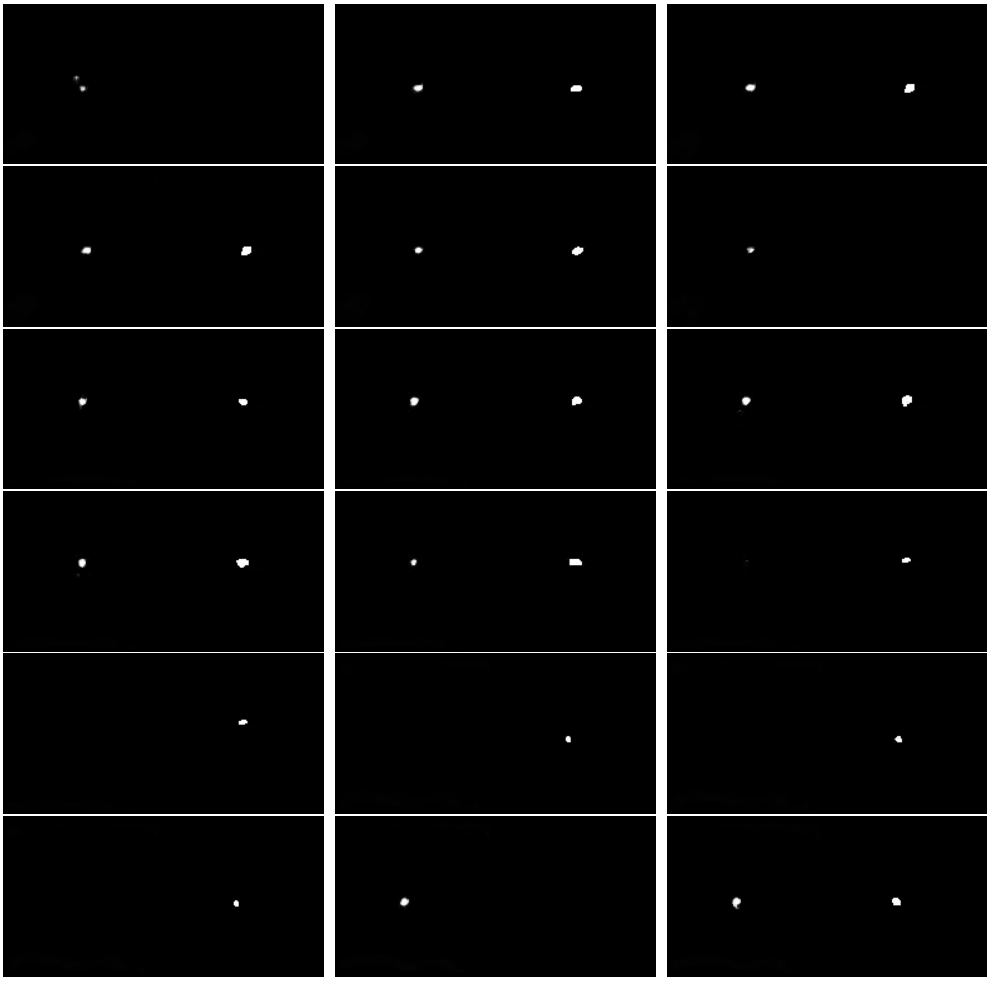


Figure 2.1. Sample predictions made by a 3D U-Net model from the fold 4 of experiment configuration A with data augmentations using ADC as input. Images represent slices with mask predictions on the same patient compared to the ground truth. Left side of each image contains the prediction whereas the right size contains the label.

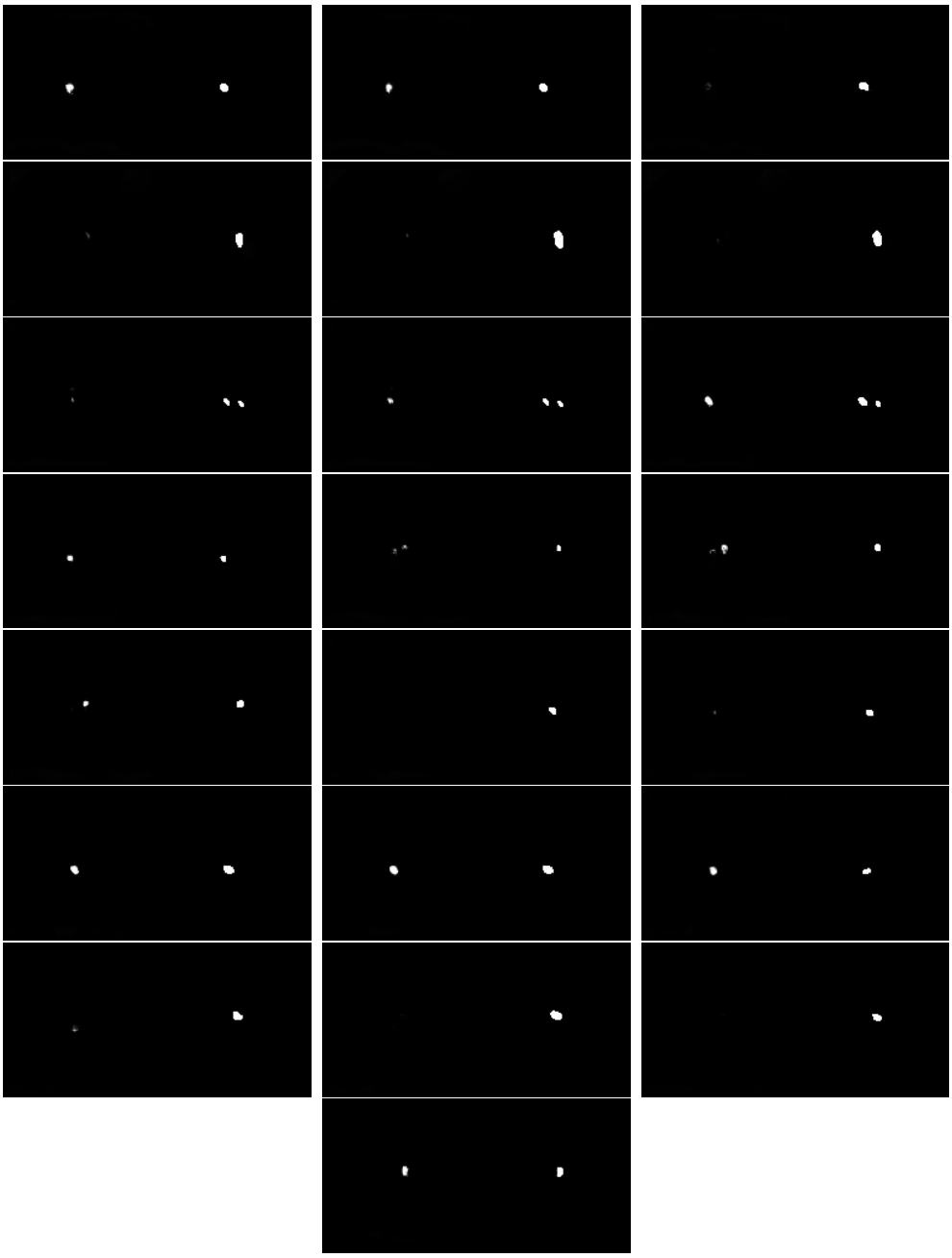


Figure 2.2. Sample predictions made by a 3D U-Net model from the fold 4 of experiment configuration A with data augmentations using ADC as input. Images represent slices with mask predictions on the same patient compared to the ground truth. Left side of each image contains the prediction whereas the right size contains the label.