POLITECNICO DI TORINO

Dipartimento di Ingegneria Meccanica e Aerospaziale

Corso di Laurea Magistrale

in Ingegneria Biomedica

Tesi di Laurea Magistrale

Colorectal cancer segmentation on MRI images using Convolutional Neural Networks

Supervisors

prof. Gabriella Balestra

dott. Valentina Giannini

Candidate

Jovana Panic

A.A. 2018/2019

Index

Index figure

Index tables

Index equations and formulas

Sommario

Lo studio ha come obiettivo quello di studiare diverse Convolutional Neural Networks (CNNs) al fine di segmentare su immagini di Risonanza Magnetica (MR) il tumore colorettale.

Le immagini fornite per lo studio sono immagini di risonanza magnetica (MR) costituite da varie sequenze: nello specifico T2 pesata (T2w), DWI (Diffusion Weighted Images) con il valore B pari a 1000, e le immagini ADC. Inoltre, sono state fornite le maschere di segmentazione manuale su tre fette per paziente eseguite da radiologi, sia per la sequenza T2w sia per la sequenza DWI.

Nella prima fase dello studio sono stati implementati 3 sistemi che utilizzano le convolutional neural networks che classificano ciascuno ROI di 3x3, 6x6 e 9x9 pixel. Ogni sistema è costituito a sua volta da tre CNNs che classificano ciascuno ROI estratte dalla sequenza T2w, dalla sequenza DWI B1000 e sequenza ADC. Tutte le sequenze vengono prima sottoposte ad un pre-processing che consente l'individuazione della regione all'interno della quale c'è il tumore e la normalizzazione delle immagini. La maschera di segmentazione finale viene quindi ottenuta sottoponendo le varie maschere ottenute dalle tre reti al majority voting.

Dai valori di Dice Coefficient, Precision e Recall è possibile notare come il sistema migliore risulti essere quello che classifica ROI di 3x3 pixels. Dalla letteratura però si può osservare che le prestazioni del sistema non sono soddisfacenti.

Nella seconda fase dello studio è stata implementata una rete neurale con un'architettura innovativa. Questa infatti è caratterizzata da due parti, la prima parte di "discesa" che riprende la struttura classica della CNN, e la seconda parte di "risalita" che risulta essere simmetrica alla prima, creando in questo modo una struttura a U, da qui il nome "U-Net". Il vantaggio principale introdotto grazie a questa particolare architettura consiste nella capacità di classificare singolarmente il pixel, invece della ROI.

Sono stati analizzati varie architetture, così da meglio comprendere il funzionamento della rete e ottimizzarla per l'obiettivo dello studio. Tutte le reti sono state addestrate utilizzando soltanto le immagini relative alla sequenza T2w.

La prima rete implementata è costituita da quattro livelli di discesa, mentre la seconda da cinque. La differenza principale tra le due è la profondità alla quale la rete estrae le informazioni utili per la creazione della maschera di segmentazione finale.

Dati i valori di Dice Coefficient, Precisione Recall, la seconda U-Net risulta avere le prestazioni migliori. Al fine di ottimizzare la rete, è stata implementata una terza rete, mantenendo lo stesso numero di livelli di discesa, ma con un numero di epoche di allenamento maggiore. Grazie all'aumento del tempo di allenamento, le prestazioni sono migliorate, rendendole così paragonabili a quelle presenti in letteratura.

Introduction

The colorectal cancer (CRC) is a malignant tumor arising from the inner wall of the colon (the longest part of the large intestine) and/or the rectum (the last part of the large intestine before the anus) (fig.01). It is the third most commonly diagnosed cancer in males, the second in females, and presents the highest incidence rate in Australia/New Zealand, Europe and Northern America (1). Moreover, the incidence of colorectal cancer is increasing in certain countries where risk has been historically low, and it is increasing among people younger than 50 years old. In contrast to incidence trends, decreasing colorectal cancer mortality have been observed in several countries worldwide and are most likely attributed to colorectal cancer screening, reduced prevalence of risk factor, and/or improved treatments (2). In the following scheme it is possible to observe the colorectal cancer incidence rates by sex and world ara (fig.01 and fig.02).

Figure 1: Areas where the colorectal cancer can occur

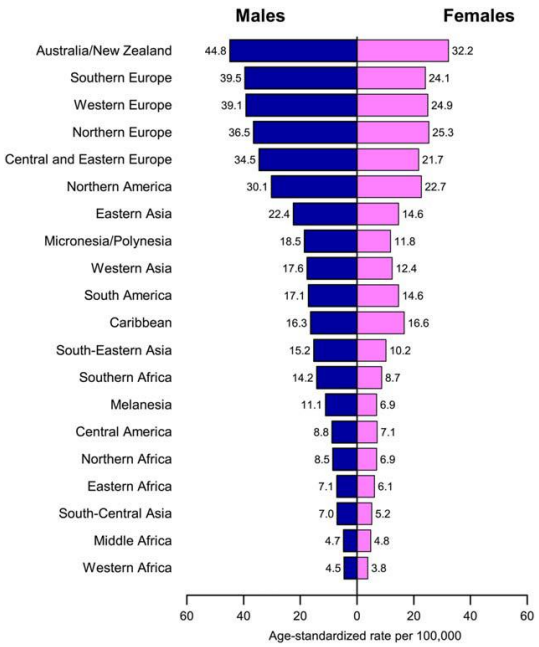


Figure 2: Colorectal Cancer Incidence Rates by Sex and World Area

Risk factors for this kind of cancer include colon polyps (from which most of the colorectal cancers develop), long-standing ulcerative colitis, genetic family history (i.e. the Lynch syndrome), and diabetes II.

Preventive measures for colorectal cancer include maintaining a healthy body weight, being physically active, minimizing consumption of red and processed meat and alcohol, and avoidance of smoking (1).

Since colon polyps and early colon cancer can have no symptoms, the regular screening can detect colorectal polyps that can be removed before they become cancerous, as well as detect cancer at an early stage when treatment is usually less extensive and more successful. If the cancer is not detected during the early stage, surgery is the most common treatment.

There are several accepted screening options:

- The guaiac-based fecal occult blood test [FOBT]
- The immune-chemical FOBT [or fecal immunochemical test]
- Flexible sigmoidoscopy
- Stool DNA test
- Computed tomography [CT] colongraphy
- Double-contrast barium enema
- Colonscopy
- Magnetic resonance [MRI]

Since the MRI shows a higher spatial resolution compared to CT, it is more and more used during the diagnosis, preoperative prediction and therapeutic effect evaluation in CRC. For these reasons, an accurate segmentation of colorectal tumors using MRI is crucial. In clinical routines, the segmentation is carried with manual or semi-manual techniques by experienced radiologists. This process is time-consuming, highly operator-dependent and tedious. As a result, several efforts have been made toward the development of valid techniques for the automated detection of colorectal cancers (3).

The existing tumor segmentations can be categorized into two groups: **generative models** and **discriminative models**. Generative models acquire prior informations through probabilistic medical image registration, where the potential deformations caused by the tumor are not taken into account. The main drawback is the fact that the only published applications are related for the brain tumor segmentation (2). Discriminative models acquire a large set of features from the medical images, in order to classify each pixel. There are several variants of these models. Day et al. (2) developed a region growing methods (CCRG), but it has not been tested with non-uniform uptake distributions, so it will fail in specific situations. Irving et al. (4) propose a method based on detection of the tumor by analysing super-voxel neighbourhood contrast characteristics of homogeneous tumors subregions, but it is a semi-automatic segmentation since it needs the manual intervention of the radiologist.

Lately different deep learning techniques are used for segmentations tasks in medical filed (5), i.e. Convolutional Neural Networks (CNNs), since they have shown good performances (3).

Deep learning (also known as deep structured learning or hierarchical learning) is part of the family of machine learning methods based on learning data representations, as opposed to task-specific algorithms. It learns hierarchical feature representation from image data. This means that it can generate high level feature directly from raw images (6). Learning can be supervised, semi-supervised or unsupervised (for clustering tasks). There are several studies which uses different types of Neural Networks. Jlan et al. (3) present a method based on a Fully Convolutional Network (FCN) for the colorectal cancer segmentation on T2-weighted magnetic resonance. The method consists on the application of a deep convolutional network where each convolutional block's outputs (which carry deep multiscale features) are fused together to determine the tumoral area on normalized images (each image is divided in ROIs 96x96 pixels only in the colorectal area). The main issue is the fact that this method allows to classify the ROIs, not the single pixel.

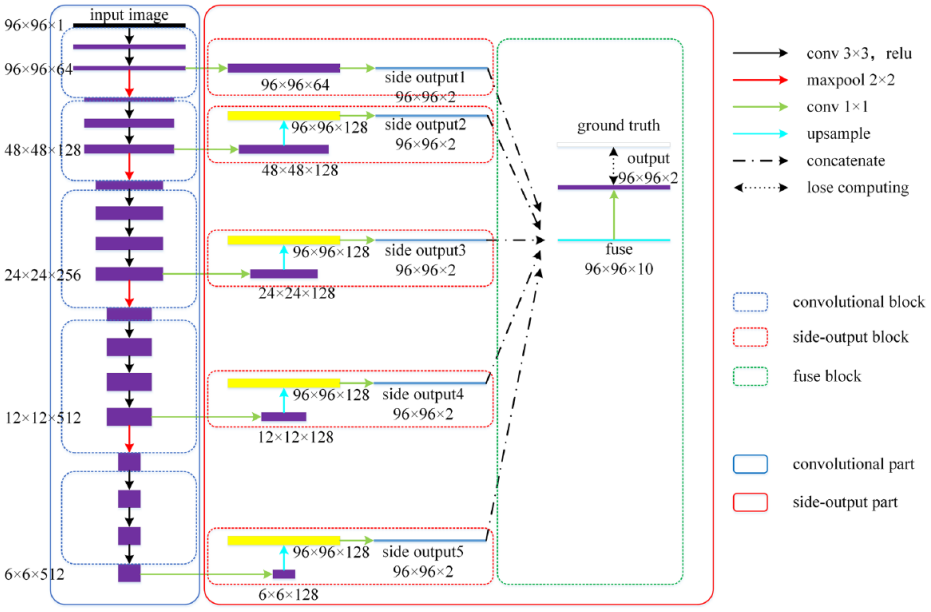


Figure 3: Architecture of the Neural Network proposed by Jlan et al. (3)

In another study conducted by Trebeschi et al. (7) a Convolutional Neural Network (CNN) has been trained on the multiparamentric MRIs to classify each voxel into tumor or non-tumor class. The standardized images have three channels: T2-weighted image, aligned DWI b-1000 image and aligned DWI b-0 image. The main drawback specified in the paper is the fact that in case of high FOV values, different healthy areas are classified as tumor.

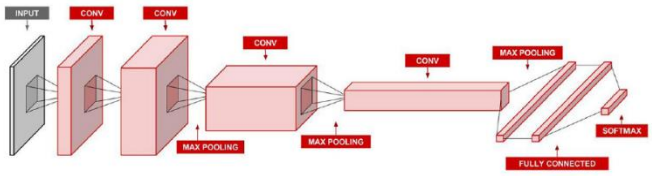


Figure 4: Architecture of the Neural Network by Trebeschi et al (7)

Huang et al. (8) proposed a 3D ROI-aware U-Net for ROI localization and intra-ROI segmentation using standardized T2-weighted images. The network consists on two

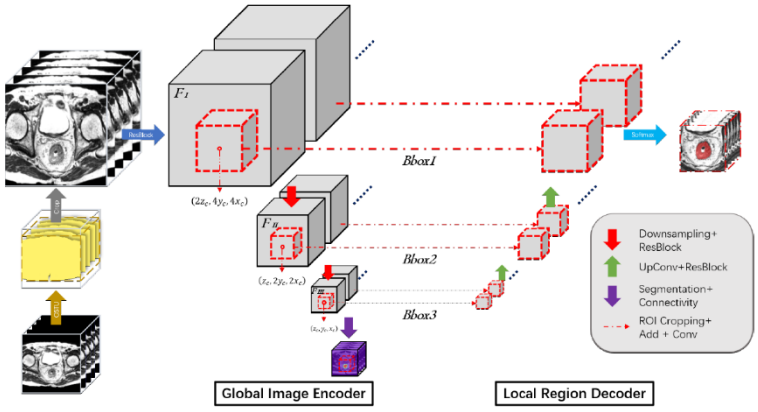


Figure 5: Architecture of the Neural Network by Huang et al. (8)

components: Global Image Encoder which extracts the features, ad the Local Region Decoder which aims to segment the colorectal tumor

Another example of study related to the colorectal cancer segmentation is proposed by Soomro et al. (9). They have presented a method which combines 3D fully convolutional neural networks and 3D level set for a fine tunning of the training phase.

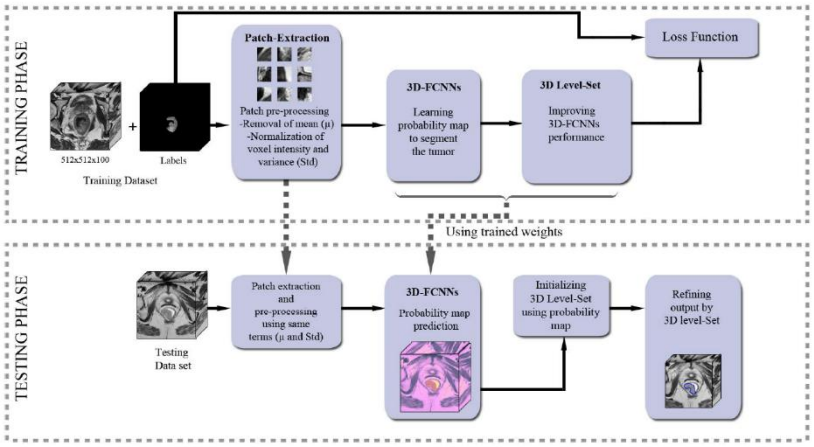


Figure 6: Work flow of the model presented by Soomro et al. (9)

The main drawback of these published method is the fact that in order to obtain better results, the deep learning algorithm is not enough and need the support of the level-set method.

Another Fully convolutional neural network is proposed by Huang et al. (10). It differs from the other models because of the use of a hybrid Dice-based loss function. This modification enables the use of unbalanced dataset (which is very common among these papers). Although, the performances are not very high. The main issue is the fact that the system needs three trained NNs for the analysis of the slices of a single patient.

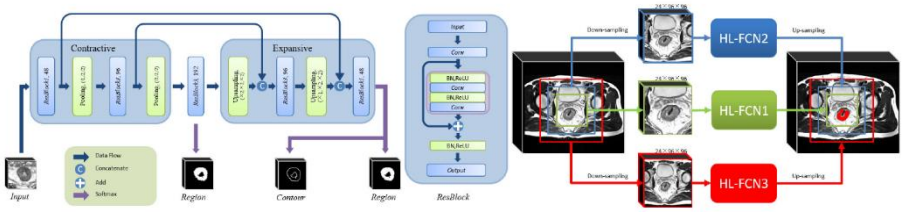


Figure 7: Architecture of the Neural Network proposed by Huang et al. (10) and work flow of the application of the NNs.

All the presented papers, given the limitation due to the lack of large medical image dataset, are characterized using data augmentation approaches, such as special filtering, adding noise, rotation, cropping to increase the size of training data.

The purpose of the master thesis is to analyse different Neural Networks for the colorectal cancer segmentation, without the use of any data augmentation approaches.

Convolutional Neural Network – CNN

Materials and Methods

Subject and Study Dataset

For this study 33 patients from the Candiolo Cancer Institute (IRCC Candiolo) with proven locally colorectal carcinoma were chosen. Among them there are 22 males and 11 females with adenocarminomas (28 cases) and mucinous carcinomas (5 cases). All patients have undergone multiparametric (mp) MRI (11), consisting of T2 weighted (with size 512x512) and diffusion weighted imaging (DWI) (with size 256x256), both axially angled. The diffusion sequence was performed using b-values B0 and B1000. Thanks to the DWI sequences it is possible to evaluate the ADC sequence applying the following formula [eq.01].

Equation 1: ADC formula

$$ADC = \frac{1}{b}\ln\frac{S(0)}{S(b)}$$

S(0): DWI sequence with b-value 0
S(b): DWI sequence with b-value

For all the T2w sequences of the patients, an initial mask was created using a k-mean algorithm, and three slices of the tumoral volume were then manually adjusted by a radiologist. The manually segmented slices were used as ground truth. Moreover, a manually segmented mask of the DWI B1000 has been made by the candidate with the support of a radiologist. The final dataset consisted of 99 slices for each sequence, used for the creation of the training set and part of the test set. In fact, the test set consists of all the slices of the patients, including the not manually modified ones.

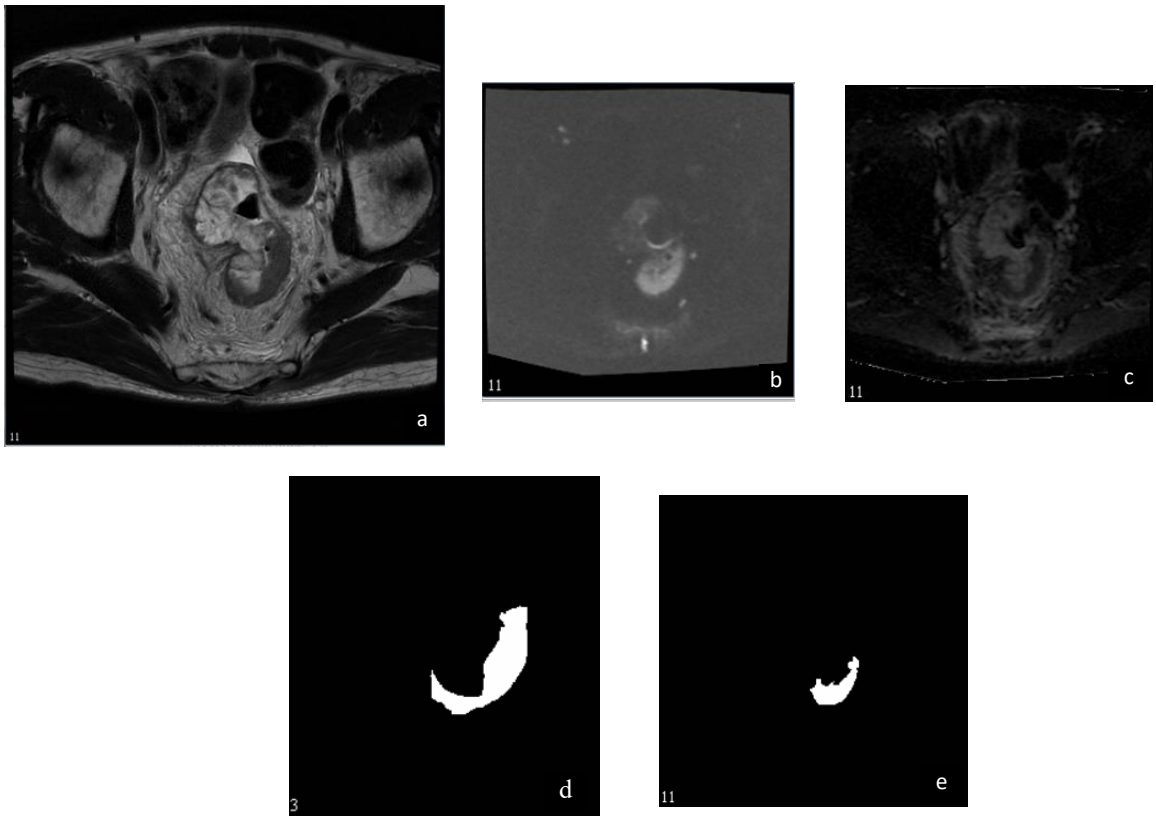


Figure 8: Example of T2w(a), DWI B1000(b), ADC(c), segmentation mask of the T2w sequence(d) and the manual mask of the DWI sequence (e) of patient 107

Pre-processing

The pre-processing consists of two main phases: the cropping phase, and the extraction of the ROIs. The cropping phase aims to automatically identify the region where the tumor is.

The cropping phase algorithm consists on different steps (they are summarized in the fig.09):

- Application of the Fuzzy c-mean clustering on all the slices of the DWI B1000 sequence. For each slice the method identifies 4 clusters and centroids;
- Extraction of the mask related to the clusters with centroid between the 50^{th}-percentile and 85^{th}-percentile on each slice. This step aims to identify the pixels which are probably belonging to the tumor. In fact, in this sequence the cacer is characterized with high intensity values. Moreover, since the image is noisy, there are some artefacts (12) which are characterized with very high intensity values, so the upper threshold aims to minimize their effects on the identification of the area of interest;
- Evaluation of the area of interest summing up all the masks obtained in the previous step, thus to identify the pixels which are more often identified as tumor;

- Removal of the pixels belonging to the border and the upper half of the image, since the cancer is known to be in the colorectal area;
- Binarization of the image;
- Extraction of the box crop which include the tumoral tissue.

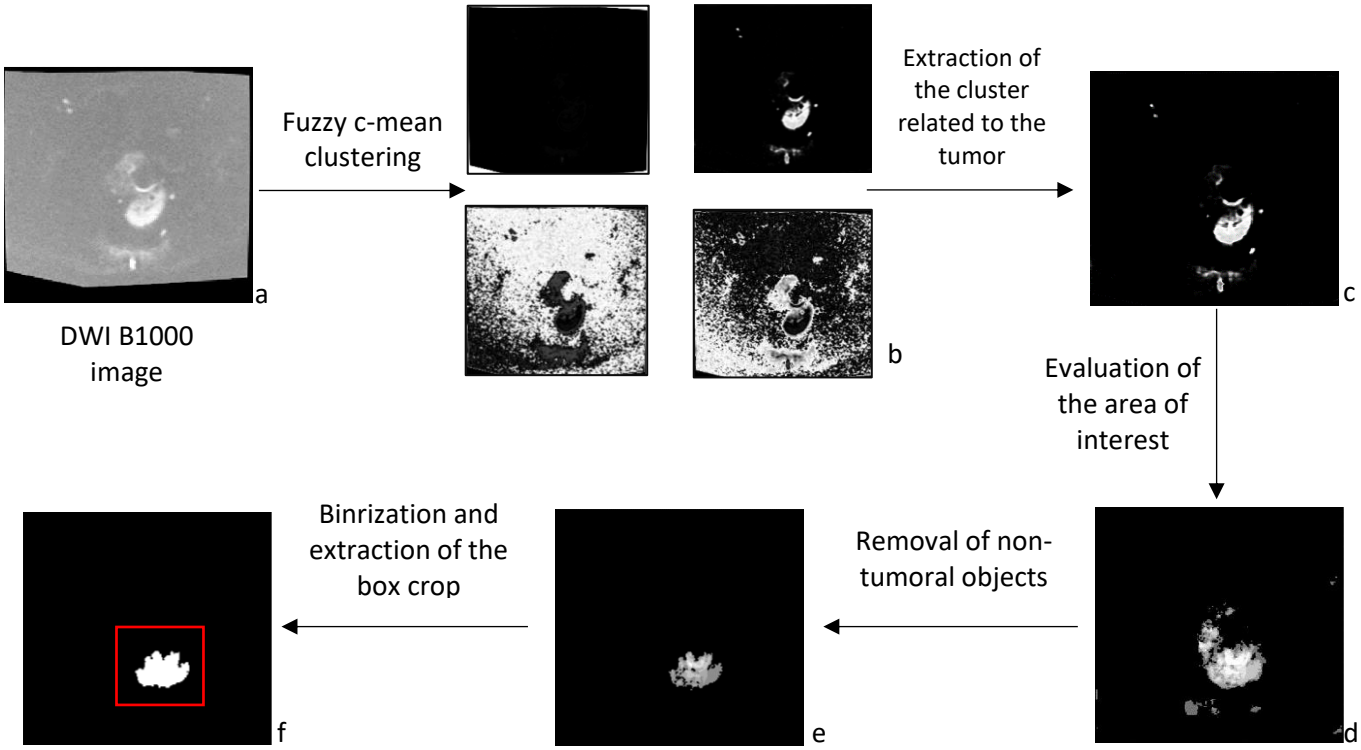
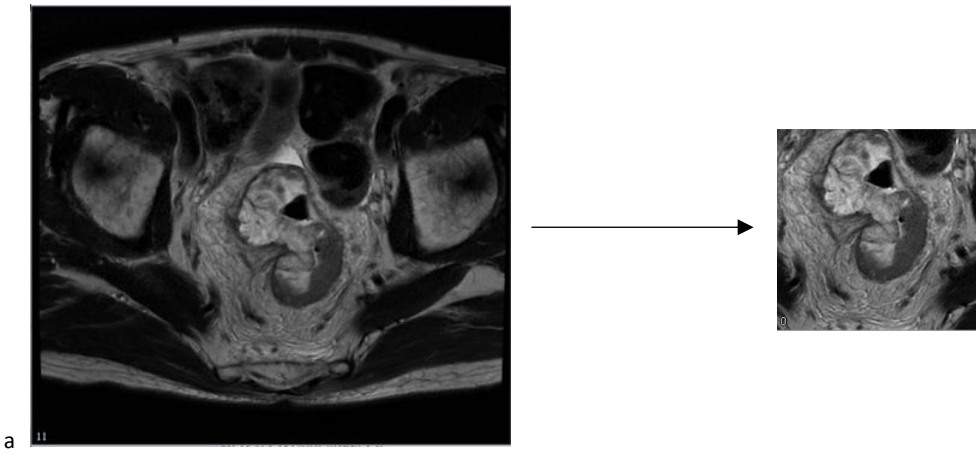


Figure 9: Pre-processing steps. From the B1000 (a), 4 cluster are identified by the Fuzzy c-mean clusterign (b) and the one that satisfies the condition is selected for each slice (c). Then all the masks related to the considered cluster are summed in order to define the initial area of interest (d). The identified objects which are close to the borders and in the upper half of the image are removed (e). Thanks to the created binary mask (f) the box crop is defined (red rectangle in f).

From the obtained mask, which contains the tumor, the region of interest is extracted and applied on the T2w, DWI B1000 and ADC sequences (fig.10).



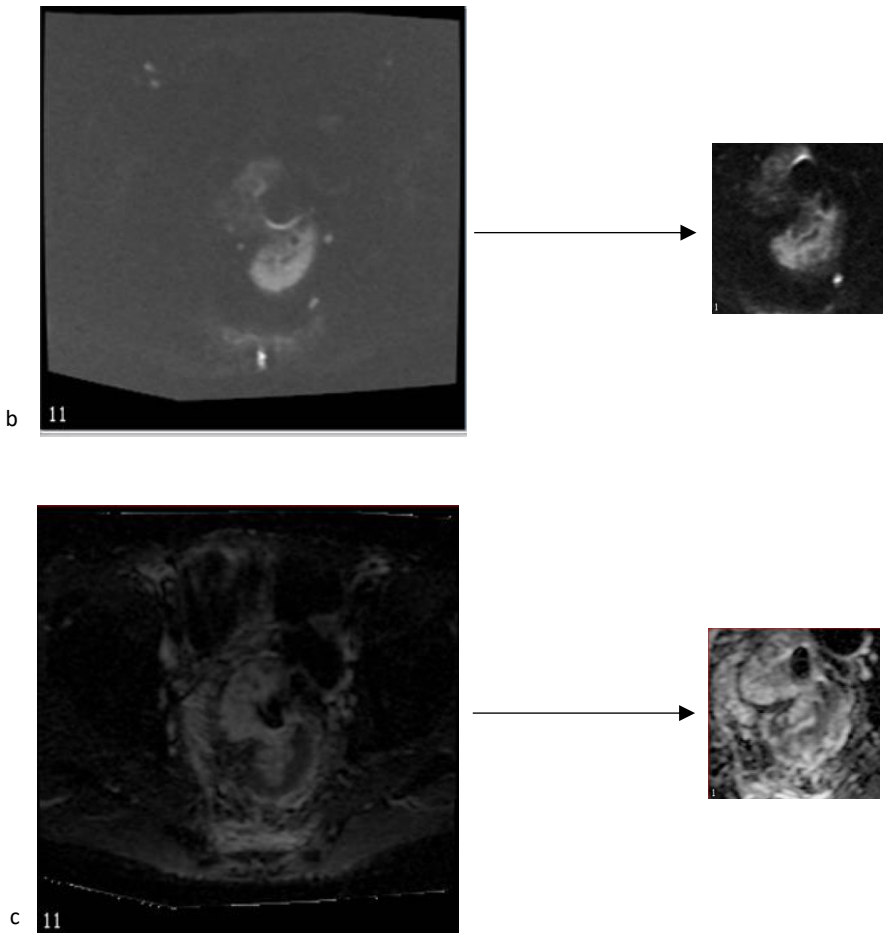


Figure 10: Example of cropped images – T2w sequence (a), DWI B000 sequence (b) and ADC sequence (c)

All the cropped sequences are then subject to the normalization through the min-max scaling [eq.02], thus to have the intensities of the images between 0 and 1.

Equation 2: Min-max scaling formula

$$img_{normalized} = \frac{img - min}{\max - min}$$

After the normalization there is the extraction ROIs phase. This process consists on dividing the cropped images in regions of dimensions 3x3, 6x6 and 9x9 without overlap.

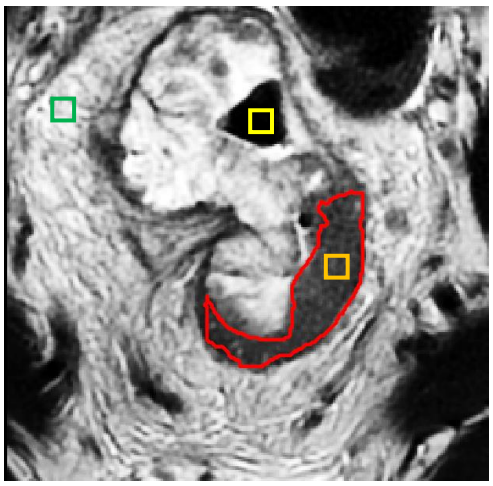


Figure 11: Example of ROIs label on the cropped T2w image. The green ROI represents the 0 class, the yellow ROI the 1 class and the orange ROI the 2 class. The red line is the manual segmentation of the tumor.

The ROIs are labeled in three different classes:

- ***Class 0 / Bright non-tumoral ROI***: ROIs fully belonging to the background which contains the 85% of the pixels considered higher than the median intensity of the cropped image (green square in fig.11);
- ***Class 1 / Dark non-tumoral ROI***: ROIs fully belonging to the background which contains the 85% of the pixels considered lower than the median intensity of the cropped image (yellow square in fig. 11);
- ***Class 2 / Tumoral ROI***: ROI fully belonging to the tumoral area (orange square in fig. 11).

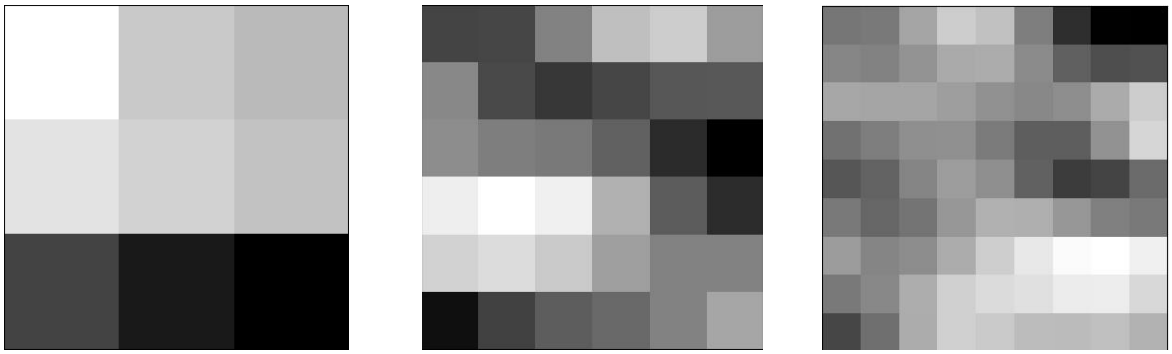


Figure 12: Examples of ROI 3x3, 6x6 and 9x9

For a balanced dataset all the ROIs related to the class 2 are considered, while the ROIs of class 0 and 1 are randomly selected thus to have equal number of tumoral and non-tumoral samples. The dataset is then divided in two groups, the training dataset (70%) and the testing dataset (30%). This process is the same for all the sequences, T2w, DWI (B1000) and ADC.

Method

Brief introduction to the Convolutional Neural Networks

The Convolutional Neural Network (CNN) belongs to the class of Deep Neural Network and it uses a variation of the multilayer perceptrons. It is a feedforward neural network, where a signal flows through the network without forming cycles or loops. This process can be expressed as in the following equation [eq.03] (2):

Equation 3: General expression of the signal through the network

$$F(x) = f_N(f_{N-1}\left(\dots\left(f_1(x)\right)\right))$$

where N denotes the number of hidden layers, and f_i represents the function in the corresponding layer i.
The main functional layers include Convolutional layer, Activation layer, Pooling layer, Batch Normalization layer, Fully Connected layer, and Predictive layer.
In the Convolutional layer, *F(x)* is composed of multiple convolution kernels ($g^1 \dots g^{k-1}$, g^k). Each g^k represents a linear function in the k-th kernel, which can be represented as follow [eq.04] (2):

Equation 4: Linear function in the kth kernel of the Convolutional layer

$$g^k(x,y) = \sum_{u=-m}^{m} \sum_{v=-n}^{n} \sum_{w=-d}^{w} W_k(u,v,w) I(x-u, y-v, z-w)$$

where (x, y, z) denotes the position of the pixel in input I, W_k denotes the weight of the k-th kernel, m, n, and w denote the height, width, and depth of the filter. The result is the so-called *Feature Maps*, which consist on k maps where each pixel identifies the value of the specific feature obtained with the previous formula [eq.04].
In the Activation layer, *F(x)* is a pixel-wise non-linear function, i.e. Rectified Linear Unit (ReLU).
The Pooling layer aims to reduce the dimensions of the feature maps; in other words, combines the outputs of neuron clusters at one layer into a single neuron in the next layer. Usually the pooling layer considers the maximum value of the clusters.
The Batch Normalization layer "normalizes" the values obtained from the previous layers (all the channels of the feature maps) by subtracting the batch mean and dividing by the batch standard deviation [eq. 05], thus to increase the stability of the network. Thanks to this layer each layer of a network is able to learn by itself a little bit more independently of other layers.

Equation 5: Equation of the Batch Normalization process

$$\check{x} = \frac{x_i - \mu_B}{\sigma_B + \varepsilon}$$

Where μ_B is the batch mean, σ_B the batch standard deviation and ε the bias.
The Fully Connected layer connects every neuron in the previous layer to the neuron in the next layer. This aims to collect all the relevant deep features for the classification.

The last layer, the Predictive layer, provides the result of the classification of the network giving as output a vector containing the belonging probability score of the object to all the classes. The most commonly used functions are *Sigmoid* and *Softmax* [eq.06].

Equation 6: Equation of Sigmoid and of the Softmax functions

$$\varphi(z) = \frac{1}{1 + e^{-z}} \qquad \varphi(z) = \frac{e^{z}}{\sum_{k=1}^{K} e^{x}}$$

Other important components which define the structure of the network and its training process are: the optimizer, the learning rate, and the loss function. The optimizer evaluates the gradient error which is used for the modification of the weights during the backpropagation process, thus to minimize the error between the predicted output and the desired one. The most commonly used is the *Adam* (*Adaptive moment estimation*), which is an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments, and *SGD* (S*tochastic Gradient Descent)*. The learning rate defines how the weights are modified, affecting the computational time of the training process. The loss function affects the training phase, by evaluating the error rate. This function must be minimized. Among the most commonly functions used there are the *binary cross-entropy* (used for binary classification) and the *categorical cross-entropy*.

Equation 7: Equation of the Binary Cross-entropy and Categorical Cross-entropy loss functions

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log\big(p(y_i)\big) + (1 - y_i) \cdot \log\big(1 - p(y_i)\big) \qquad H_p(q) = \sum_{i=1}^{N} q(y_i) \cdot \log(p(y_i))$$

Where y_i is the probability of the object q to belong to class p.

Thanks to these characteristics the CNN automatically learns a hierarchy of increasing complex features directly from data (3), thus reducing the need of feature engineering, which is one of the most complicated and time-consuming parts in machine learning, especially in processing redundant image data.

In the last years the CNN has shown good performances in image recognition problems (13), thus it has been more and more used for medical image segmentation (2). However, the input of CNNs is limited to relatively small images, due to the Fully Connected layers, thus it is not used directly on large images. For this reason, in this study different Convolutional Neural Networks systems have been implemented as ROIs classifiers, one for the 3x3 ROIs, one for 6x6x ROIs, and the last one for the 9x9 ROIs, thus to analyse how much the resolution affects the performances of the system.
In order to improve the accuracy of the net two other networks have been trained, one with the DWI B1000 and ADC images, as proposed in one of the papers (7).
Thus, each system consists of three CNNs: one for the classification of the ROIs belonging to the T2w cropped images, one for the ROIs belonging to the DWI B1000 cropped sequence and the one for the ROIs belonging to the ADC cropped sequence. The probability scores to all classes are obtained from each CNN and using the majority voting system it is possible to evaluate the class for the ROI and creating the segmentation mask.

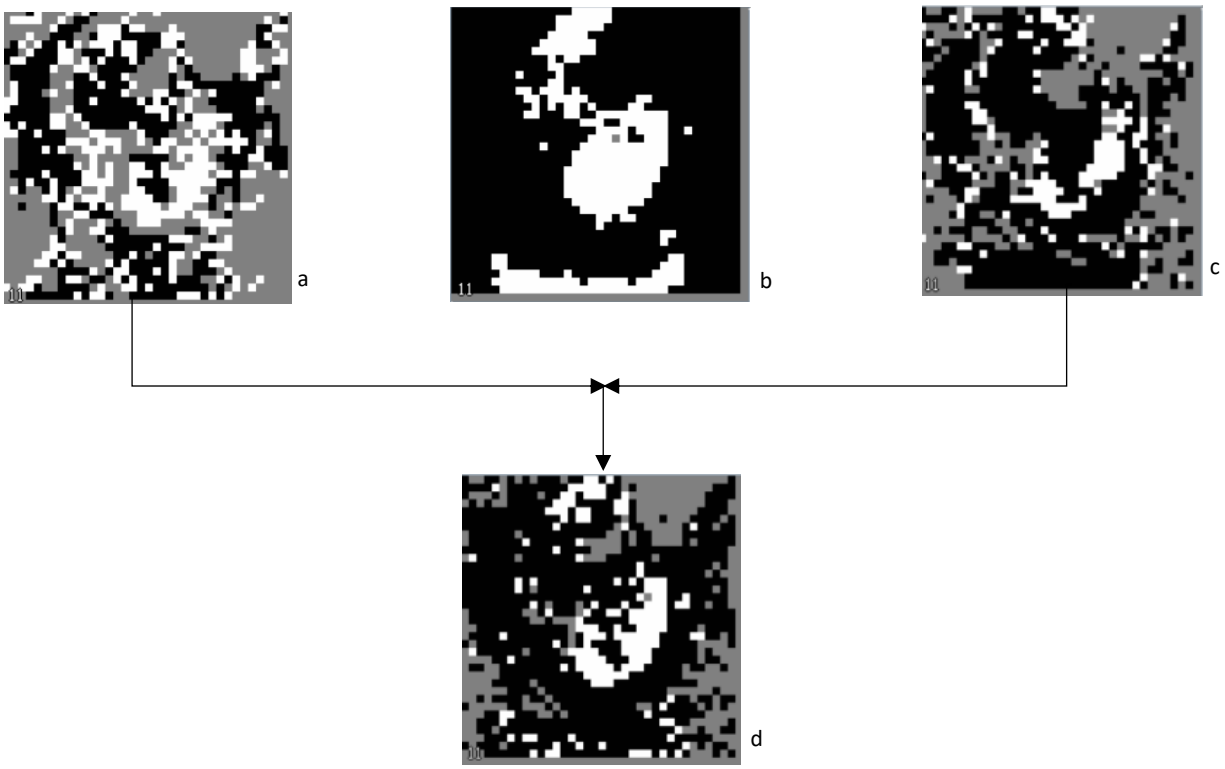


Figure 13: Application of the majority voting on the mask of the T2w sequence (a), DWI B1000 sequence (b) and ADC sequence (c) thus to obtain the final segmentation mask (d)

The networks implemented are all equal, and the structure (fig.14) consists of two subsequent Convolutional layers followed by Batch-Normalization layer, another Convolutional layer followed by the output layer. The first three Convolutional layers use 3x3 kernels and ReLU activation functions, while the output layer uses the *Softmax* function. The optimizer is Adam, with learning rate 0.001, loss function *Categorical Cross-entropy*. All the models are implemented on Python 3.7.0 with Keras (Theano backend).

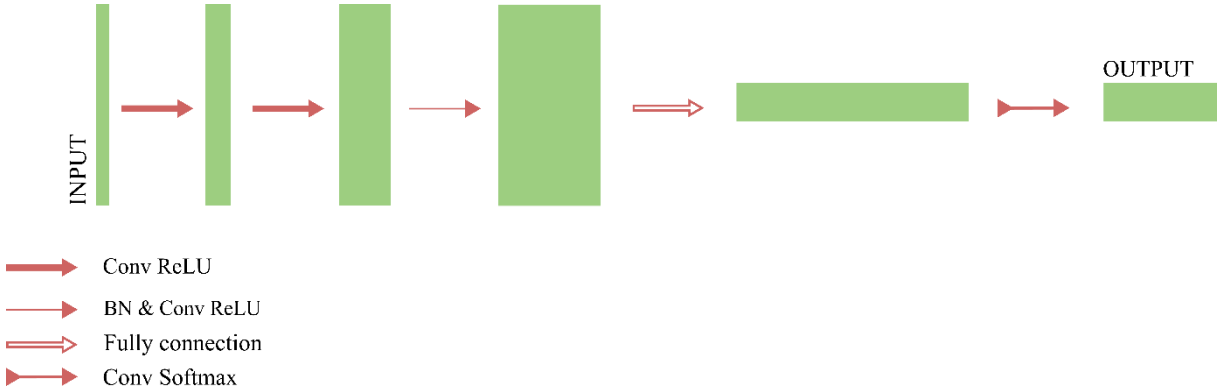


Figure 14: Structure of the CNN

All the networks have been implemented on Python 3.7.0 with Keras (Theano backend).

Post-processing

The post-processing phase aims to reduce the false positive elements. Firstly, a mask is created considering only the ROIs tumor classified (class 2). Then, three different hypothesis are applied on the predicted mask slices:

- The tumoral object must have an area higher than 100 pixels and lower than half-area of the cropped image;
- The tumoral object must not belonging to the area next to the border of the image;
- The tumoral object must be connected on at least three slices.

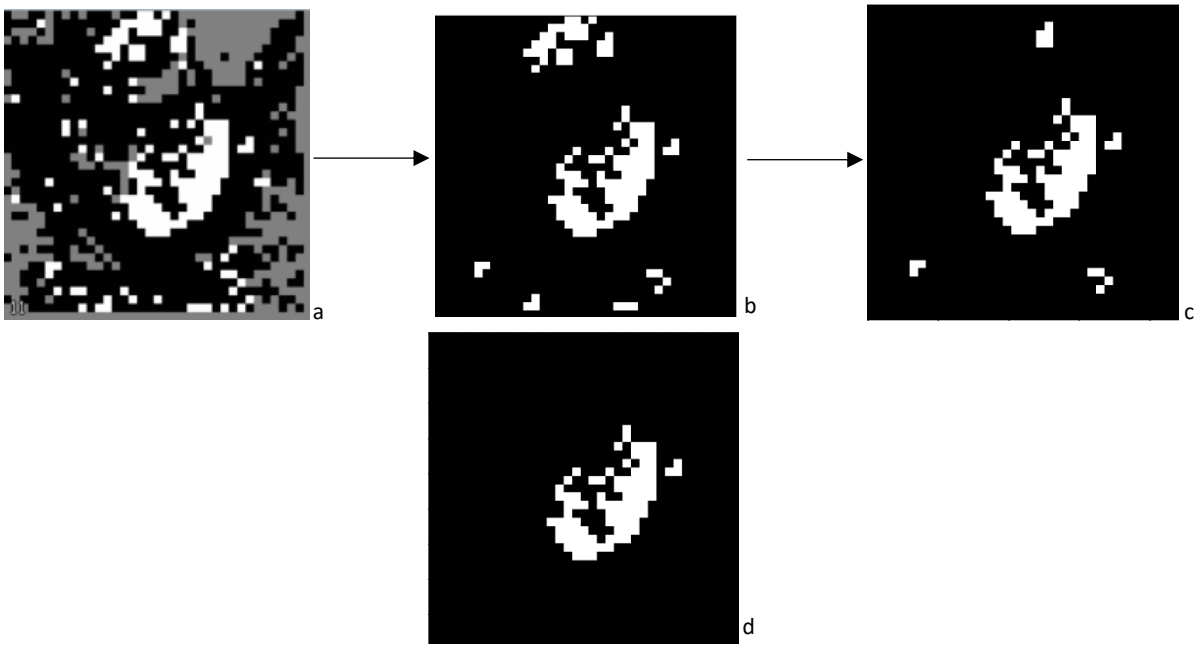


Figure 15: Post-processing phases. From the (a) mask obtained by the system a binary mask is obtained (b), then the areas which are lower than 100 pixels and which are next to the edge are removed (c). among the remaining object only the one which is connected on at least three slices creates the final mask (d)

Results

To check the performances of all the networks, the considered parameters are: Dice coefficient, Precision and Recall. Their formulas are shown in the eq.08 (14).

Equation 8: Dice Coefficient, Precision and Recall formuals

$$Dice\ coefficient = \frac{2\,TP}{2TP+FP+FN} \qquad Precision = \frac{TP}{TP+TP} \qquad Recall = \frac{TP}{TP+FN}$$

The first CNN system consists of three CNNs which classify 3x3 ROIs from the T2w, DWI B1000 and ADC sequences. The optimizers used for each CNN is the Adam, with learning rate of 0.001 and the loss function is the *categorical cross-entropy*. All the three CNNs have been trained with 150 epochs.

In the following table (tab.1) there are the values related to the already specified parameters, evaluated considering all the mask and the mask containing only the tumoral connected object.

Table 1: Results of the performances CNNs 3x3 ROIs classifier system in terms of Dice Coefficient, Precision and Recall considering all the mask and only the tumoral object identified by the system

CNN 3X3	ALL			
	Train		Test	
	Mean ± std	Median | 25th | 75th	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.60 ± 0.20	0.65 | 0.50 | 0.73	0.53 ± 0.19	0.54 | 0.39 | 0.69
PRECISION	0.63 ± 0.24	0.64 | 0.50 | 0.81	0.53 ± 0.24	0.57 | 0.32 | 0.71
RECALL	0.65 ± 0.22	0.70 | 0.51 | 0.84	0.65 ± 0.21	0.68 | 0.49 | 0.82
	ONLY TUMOR			
	Train		Test	
	Mean ± std	Median | 25th | 75th	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.61 ± 0.19	0.66 | 0.53 | 0.73	0.54 ± 0.18	0.54 | 0.40 | 0.69
PRECISION	0.65 ± 0.22	0.65 | 0.54 | 0.81	0.54 ± 0.24	0.57 | 0.33 | 0.72
RECALL	0.65 ± 0.22	0.70 | 0.51 | 0.84	0.65 ± 0.21	0.68 | 0.49 | 0.82

Analysing the obtained values, it is possible to notice that the Dice Coefficient is around 0.61, the Precision around 0.65 and the Recall around 0.65, which are not high enough to be satisfying. Despite of this it is possible to notice that differences between the values obtained considering all the mask and only the tumoral object are reasonably low, which means that thanks to the post-processing phase almost all the *False Positive* objects are removed.

It is possible to analyse also the system performances by each patient considering the graphs below (fig.16, fig.17 and fig.18).

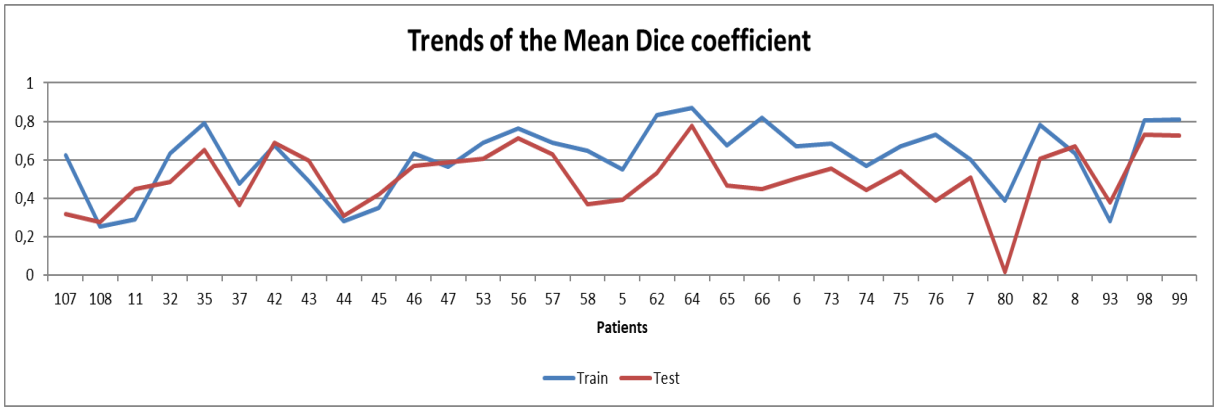


Figure 16: Mean Dice Coefficient's trend of the CNNs 3x3 ROIs classifier system, considering all the mask and only the tumoral object identified by the mask

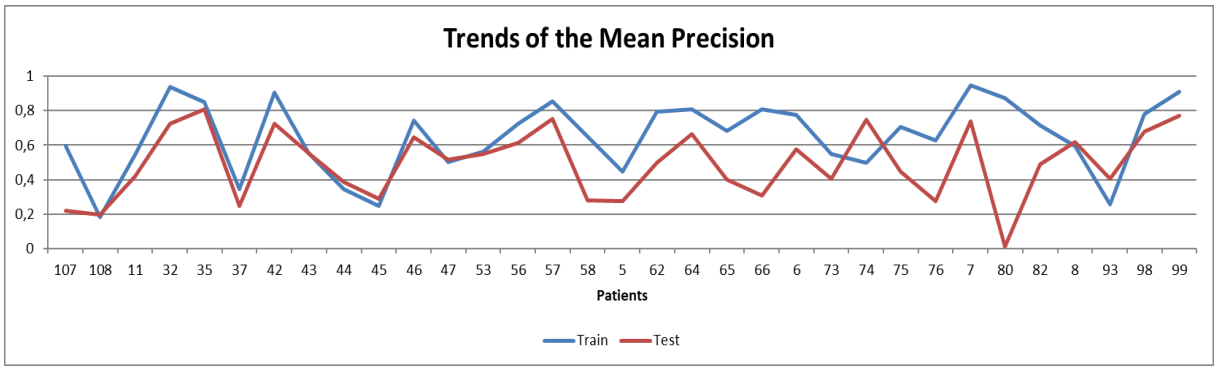


Figure 18: Mean Precision's trend of the CNNs 3x3 ROIs classifier system, considering all the mask and only the tumoral object identified by the mask

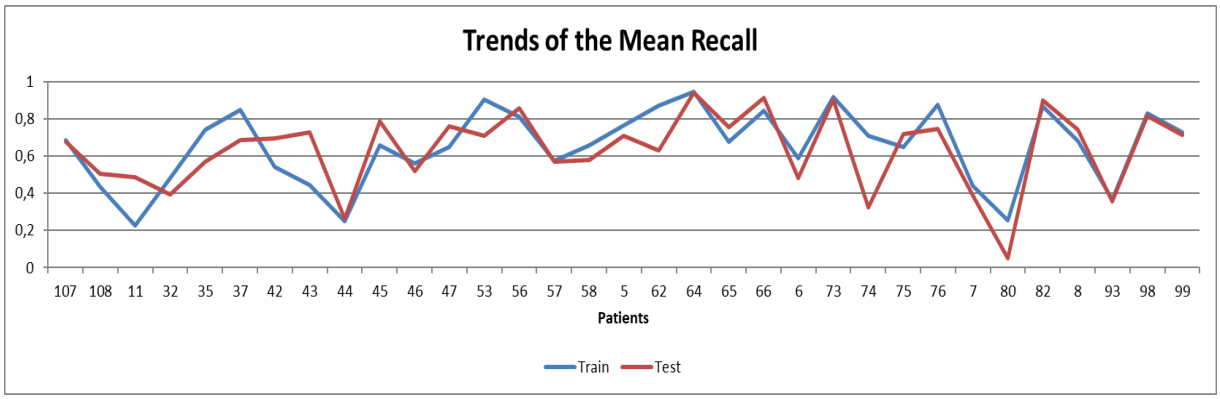


Figure 17: Mean Recall's trend of the CNNs 3x3 ROIs classifier system, considering all the mask and only the tumoral object identified by the mask

Analysing the graphs, it is possible to notice that the trends of the Recall (considering all the mask and only the tumoral object) are the same, which means that the network is able to identify correctly the tumoral area. The trends of the Dice Coefficient and the Precision are slightly different for patients 5 and 11. This means that the effect of the *False Positive* elements is relevant, despite the post-processing.

Another analysis has been made on each obtained mask related to the three CNNs. It is important to notice that the results related to the CNNs which classify the ROIs of the DWI B1000 and ADC sequences are obtained using the same dataset used for the training. Below the table (tab.2, tab.3, tab.4) with the results of the CNNs which classify the 3x3 ROIs of the T2w, DWI B1000 and ADC sequences.

Table 2: Results of the performances of the CNN 3x3 ROIs classifier related to the T2w sequence in terms of Dice Coefficient, Precision and Recall

CNN 3X3 T2W	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.42 ± 0.23	0.41 | 0.21 | 0.62
PRECISION	0.31 ± 0.20	0.30 | 0.13 | 0.49
RECALL	0.86 ± 0.21	0.95 | 0.83 | 0.99

Table 3: Results of the performances of the CNN 3x3 ROIs classifier related to the DWI B1000 sequence in terms of Dice Coefficient, Precision and Recall

CNN 3X3 B1000	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.64 ± 0.23	0.76 | 0.46 | 0.82
PRECISION	0.61 ± 0.27	0.70 | 0.40 | 0.81
RECALL	0.87 ± 0.20	0.96 | 0.87 | 0.99

Table 4: Results of the performances of the CNN 3x3 ROIs classifier related to the ADC sequence in terms of Dice Coefficient, Precision and Recall

CNN 3X3 ADC	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.35 ± 0.20	0.36 | 0.16 | 0.53
PRECISION	0.30 ± 0.21	0.28 | 0.11 | 0.50
RECALL	0.57 ± 0.21	0.57 | 0.45 | 0.73

It is possible to notice that the CNN which classify the DWI B1000 sequences has high value of Recall, which means that correctly segment the tumoral area in those images, while the performances of the networks related to the T2w and ADC sequences are very poor.

Below there are some graphs (fig. 19, fig. 20 and fig. 21) which shows how the CNN related to the DWI B1000 sequence behaves with the different patients, in comparison with the CNN related to the ADC sequence and T2w sequences' behaviours.

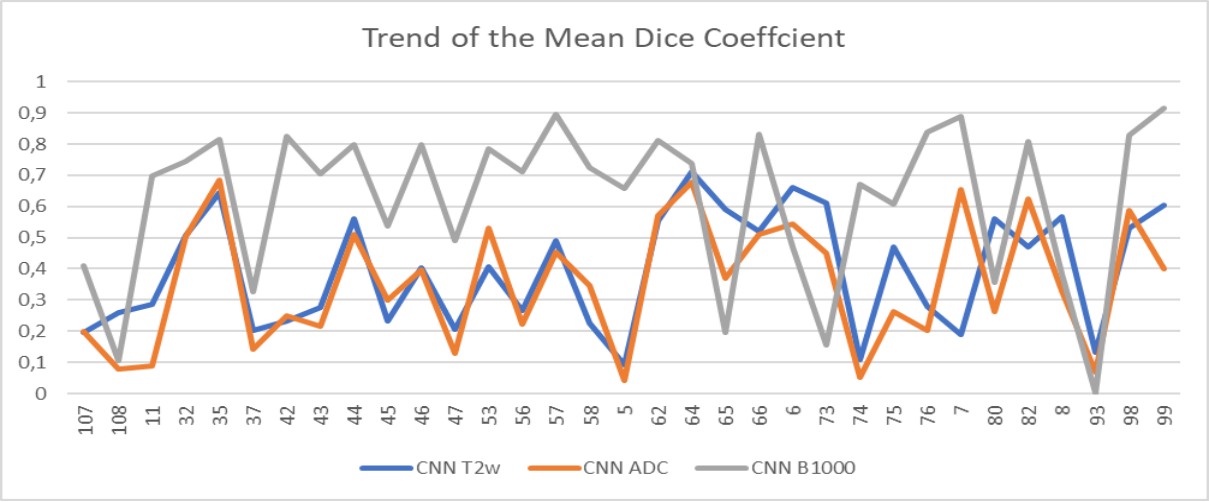


Figure 21: Comparison between the Mean Dice Coefficient of the CNN related to the T2w sequence, the CNN related to the DWI B1000 sequence and CNN related to the ADC sequence

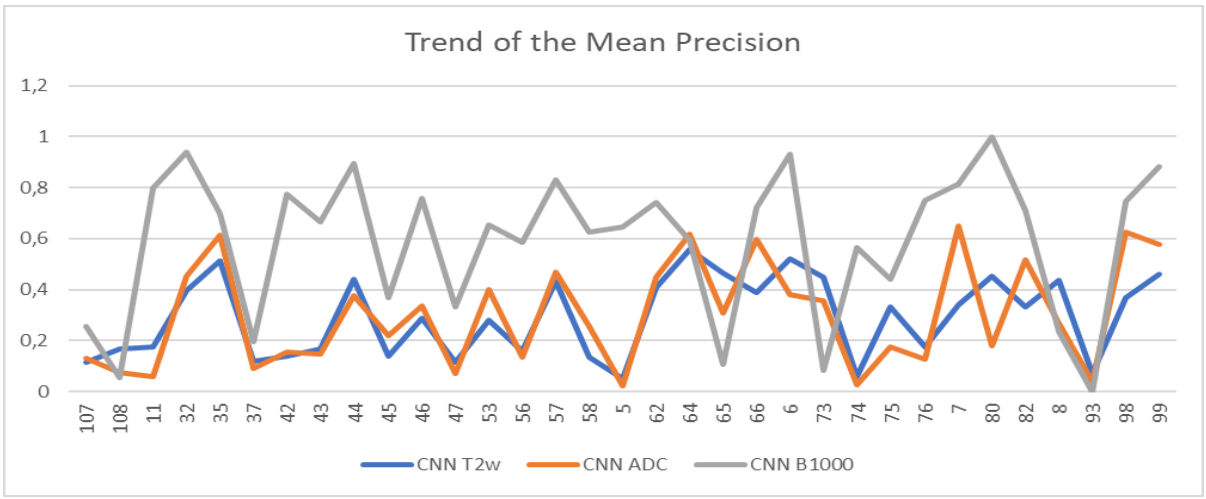


Figure 20: Comparison between the Mean Precision of the CNN related to the T2w sequence, the CNN related to the DWI B1000 sequence and CNN related to the ADC sequence

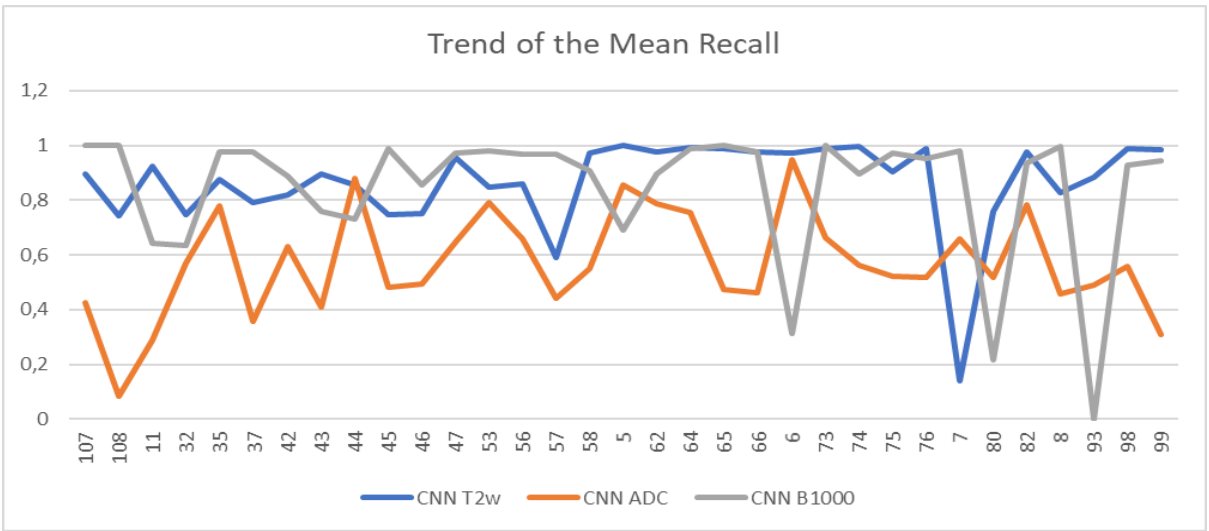


Figure 19: Comparison between the Mean Recall of the CNN related to the T2w sequence, the CNN related to the DWI B1000 sequence and CNN related to the ADC sequence

It is possible to notice that the CNN related to the DWI B1000 has higher performances, but in some patients its accuracy is very poor: for the patient 93 the value of Dice Coefficient, Precision and Recall is 0. Overall, the performances of the CNNs related to the DWI B1000 sequence has good performances, in comparison with the other CNNs. A possible way to improve the accuracy of the system is to make the prediction of the net related to the DWI more influent in the majority voting algorithm.

Here some example of well segmented slices and badly segmented (fig.22, fig.23).

Well segmented patient:

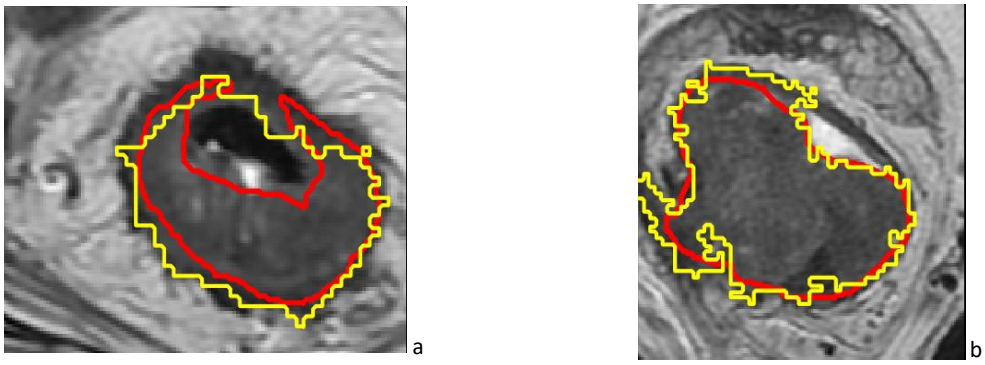


Figure 22: Paient 62 (a) and Patient 64 (b) - manual segmentation (red) and CNNs 3x3 ROIs segmentation (yellow)

Badly segmented patient:

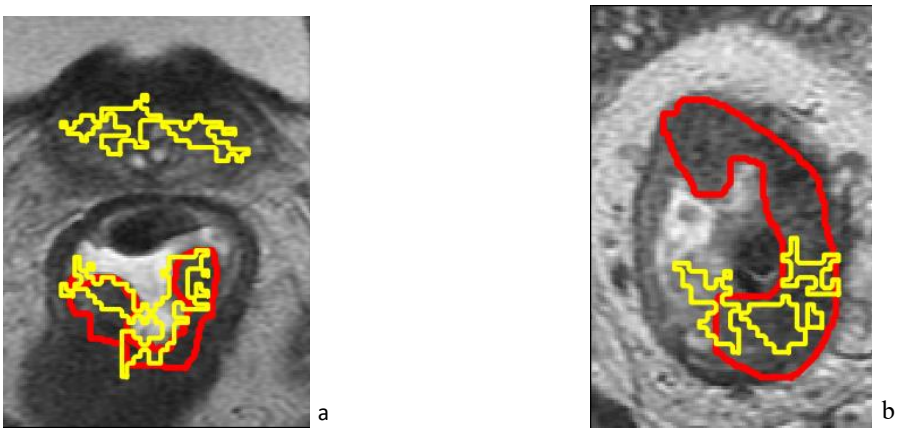


Figure 23: Patient 11 (a) and Patient 80 (b)- manual segmentation (red) and CNNs 3x3 ROIs segmentation (yellow)

The second system consists of the same CNNs, but they classify 6x6 ROIs. Also in this case, the optimizer used is Adam, with learning rate of 0.001. The loss function is the *categorical cross-entropy*. Moreover, the CNNs have been trained with 150 epochs.

Table 5: Results of the performances CNNs 6x6 ROIs classifier system in terms of Dice Coefficient, Precision and Recall considering all the mask and only the tumoral object identified by the system

CNN 6X6	ALL			
	Train		Test	
	Mean ± std	Median | 25th | 75th	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.58 ± 0.19	0.62 | 0.45 | 0.73	0.51 ± 0.19	0.54 | 0.38 | 0.66
PRECISION	0.60 ± 0.23	0.63 | 0.45 | 0.80	0.51 ± 0.24	0.54 | 0.29 | 0.70
RECALL	0.62 ± 0.20	0.64 | 0.52 | 0.77	0.62 ± 0.21	0.63 | 0.49 | 0.78
	ONLY TUMOR			
	Train		Test	
	Mean ± std	Median | 25th | 75th	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.58 ± 0.18	0.63 | 0.46 | 0.73	0.51 ± 0.19	0.54 | 0.39 | 0.66
PRECISION	0.62 ± 0.22	0.63 | 0.48 | 0.80	0.51 ± 0.24	0.55 | 0.30 | 0.70
RECALL	0.62 ± 0.20	0.64 | 0.52 | 0.77	0.62 ± 0.21	0.63 | 0.49 | 0.78

By analysing the parameters, it is possible to notice that the performance is decreased, due to the larger area of the considered ROI, which implies a lower resolution. Moreover, the values of the parameters evaluated considering only the tumoral object are very poor, which means that the net does not properly recognize the tumoral area. By the following graphs (fig.24, fig.25 and fig.26) it is possible to notice how the performances change among the different patients.

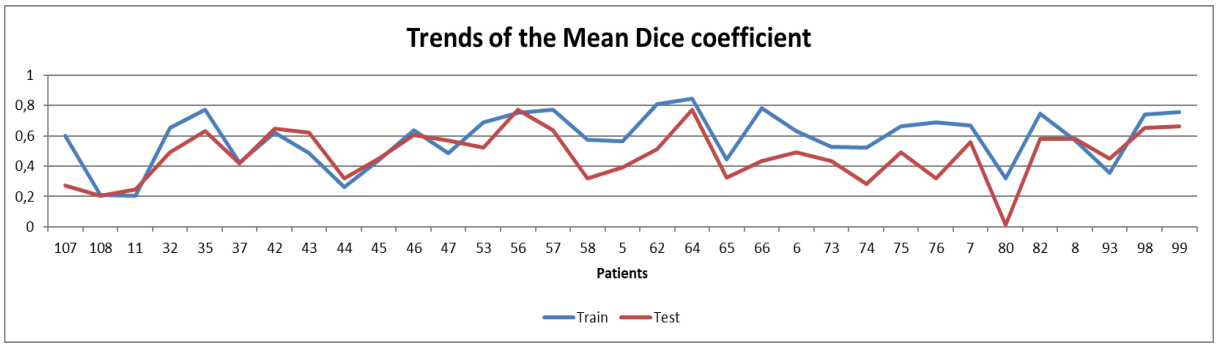


Figure 25: Mean Dice Coefficient's trend of the CNNs 6x6 ROIs classifier system, considering all the mask and only the tumoral object identified by the mask

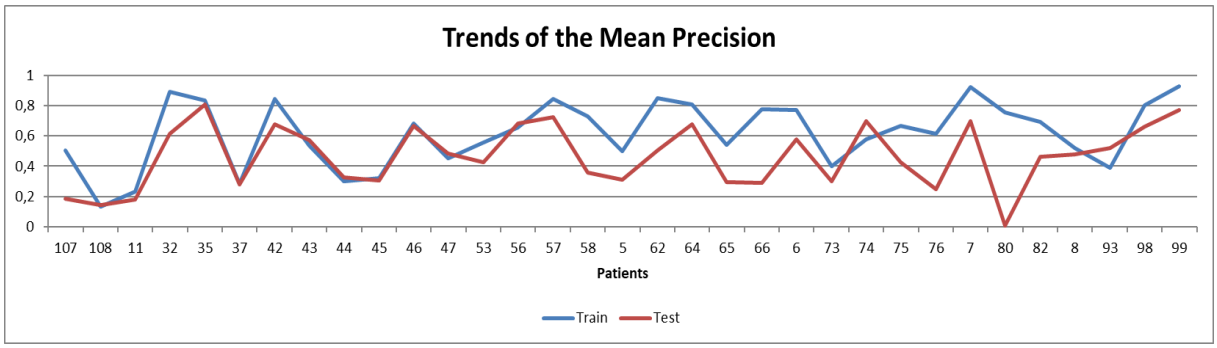


Figure 24: Mean Precision's trend of the CNNs 6x6 ROIs classifier system, considering all the mask and only the tumoral object identified by the mask

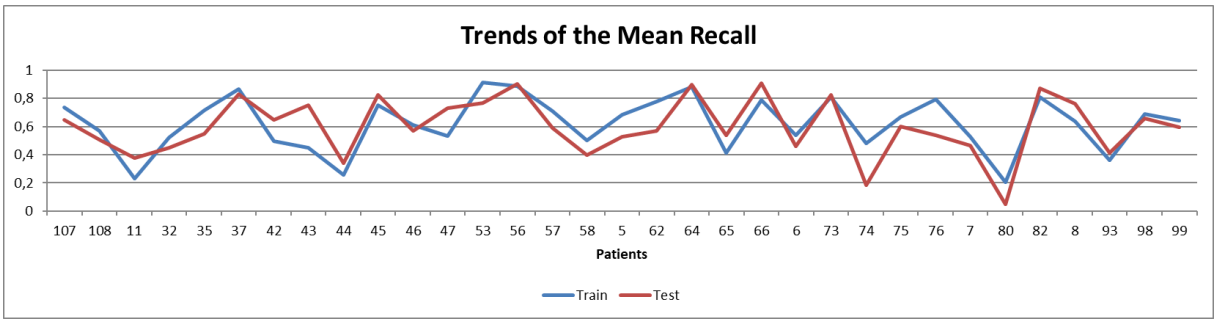


Figure 26: Mean Recall's trend of the CNNs 6x6 ROIs classifier system, considering all the mask and only the tumoral object identified by the mask

By analysing he trends it is possible to notice that the differences among the parameters evaluated considering all the mask and only the tumoral object are more relevant than in the previous case. This means that despite the post-processing the False Positive elements affect the performances of the system.

Another analysis has been made on each obtained masks related to the three CNNs. It is important to notice that the results related to the CNNs which classify the ROIs of the DWI B1000 and ADC sequences are obtained using the same dataset used for the training. Below the table (tab.6, tab.7, tab.8) with the results of the CNNs which classify the 3x3 ROIs of the T2w, DWI B1000 and ADC sequences.

Table 6: Results of the performances of the CNN 6x6 ROIs classifier related to the T2w sequence in terms of Dice Coefficient, Precision and Recall

CNN 6X6 T2W	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.32 ± 0.19	0.30 | 0.15 | 0.46
PRECISION	0.21 ± 0.15	0.18 | 0.09 | 0.33
RECALL	0.91 ± 0.17	0.98 | 0.90 | 1.00

Table 7: Results of the performances of the CNN 6x6 ROIs classifier related to the DWI B1000 sequence in terms of Dice Coefficient, Precision and Recall

CNN 6X6 B1000	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.62 ± 0.24	0.71 | 0.44 | 0.81
PRECISION	0.62 ± 0.27	0.71 | 0.40 | 0.82
RECALL	0.80 ± 0.22	0.89 | 0.75 | 0.94

Table 8: Results of the performances of the CNN 6x6 ROIs classifier related to the ADC sequence in terms of Dice Coefficient, Precision and Recall

CNN 6X6 ADC	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.43 ± 0.20	0.47 | 0.27 | 0.58
PRECISION	0.41 ± 0.24	0.40 | 0.21 | 0.63
RECALL	0.58 ± 0.21	0.59 | 0.44 | 0.74

It is possible to notice that also in this case the CNNs which classify the DWI B1000 sequences has high value of Recall, which is lower than the previous case, but with higher value of Dice Coefficient and Precision. Again, the CNNs related to the T2w and ADC sequences have poor performances, which are slightly higher than in the CNN 3x3 system.

Below there are some graphs which shows how the CNN related to the T2w sequence, to the DWI B1000 sequence behave with the different patients, in comparison with the CNN related to the ADC sequence's behaviour.

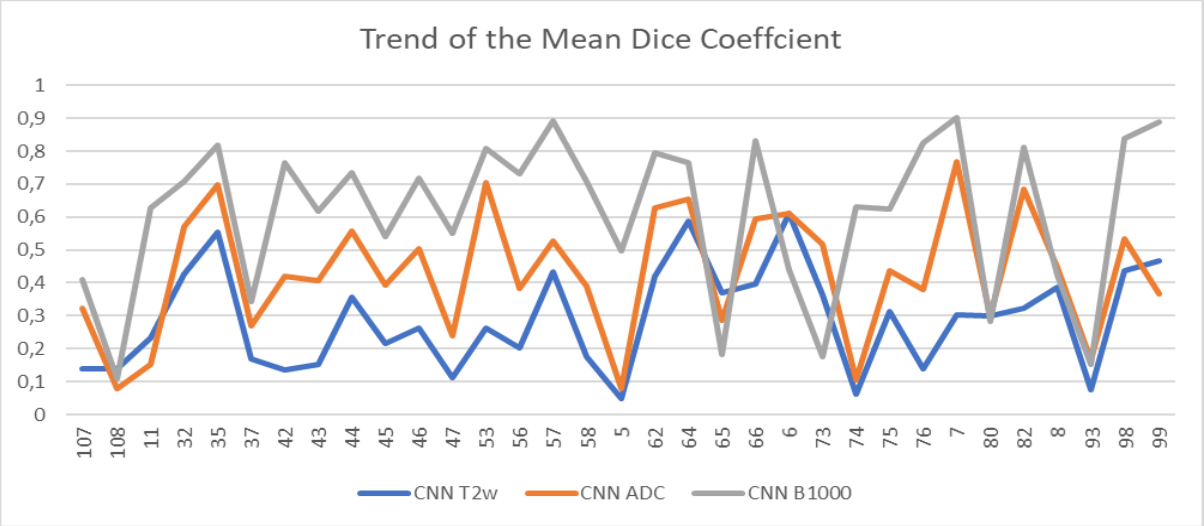


Figure 27: Comparison between the Mean Dice Coefficient of the CNN related the DWI B1000 sequence and CNN related to the ADC sequence

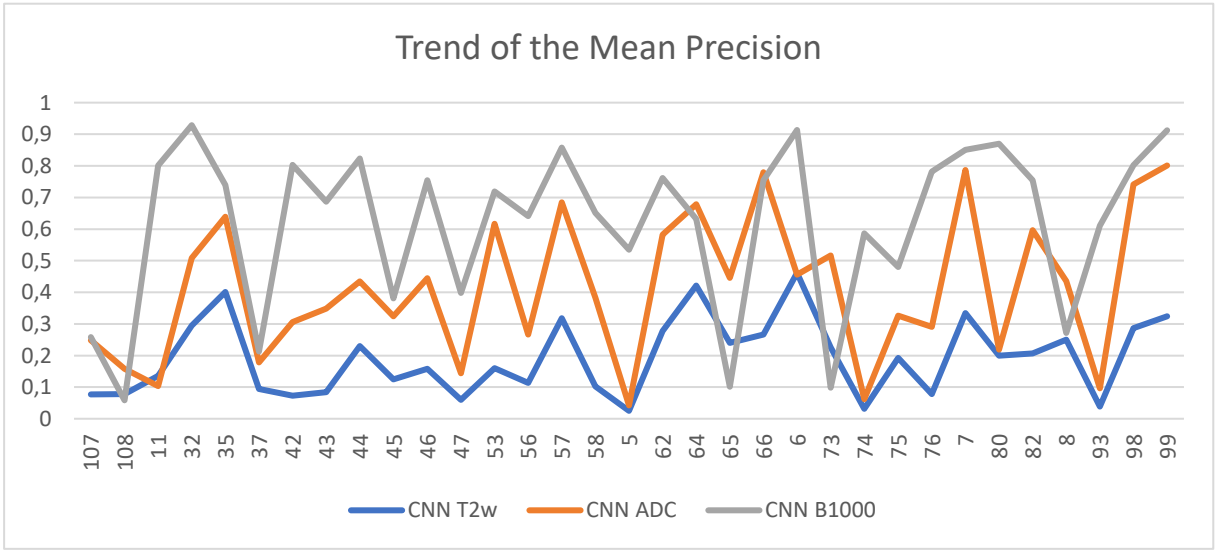


Figure 28: Comparison between the Mean Precision of the CNN related the DWI B1000 sequence and CNN related to the ADC sequence

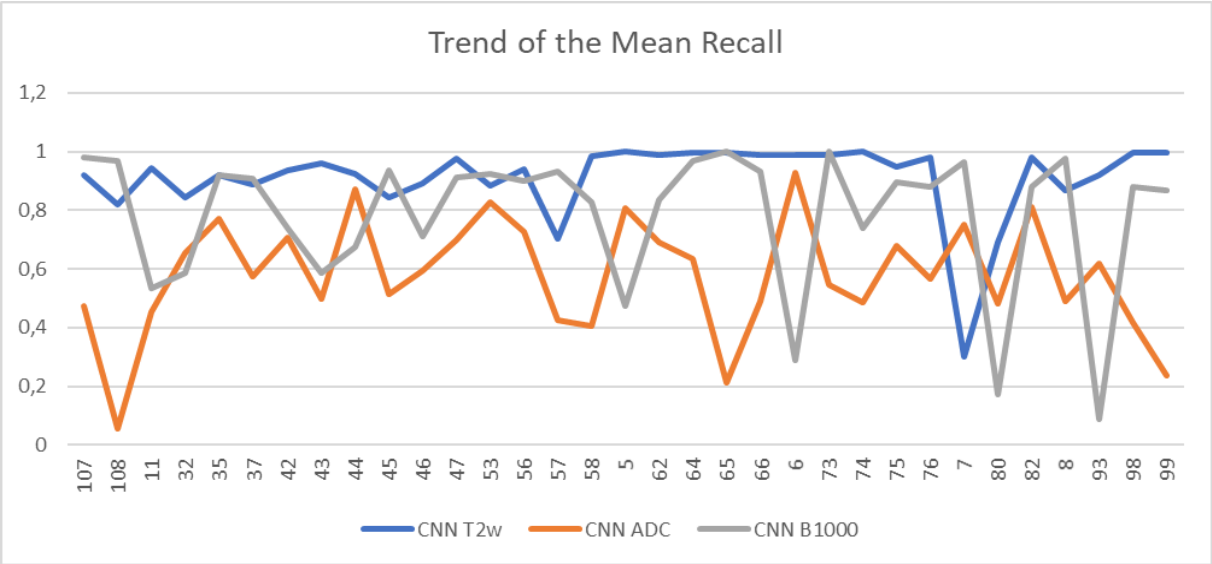


Figure 29: Comparison between the Mean Recall of the CNN related to the T2w sequence, the CNN related to the DWI B1000 sequence and CNN related to the ADC sequence

Here some example of well segmented slices and badly segmented (fig.30, fig.31).

Well segmented patient:

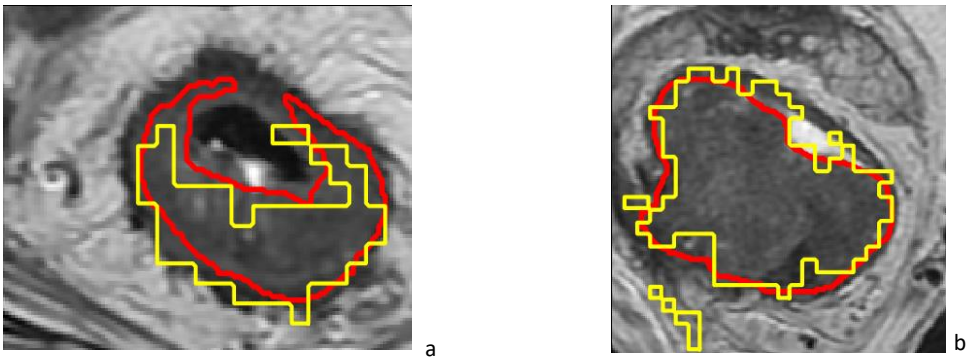
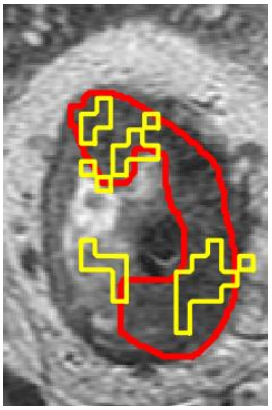


Figure 30: Patient 62 (a) and Patient 64 (b) - manual segmentation (red) and CNNs 6x6 ROIs segmentation (yellow)

Badly segmented patient:

a



b

Figure 31: Patient 11 (a) and Patient 80 (b)- manual segmentation (red) and CNNs 3x3 ROIs segmentation (yellow)

The last system implemented consists of three CNNs which classify 9x9 ROIs of the three sequences T2w, DWI B1000 and ADC. The other parameters of the net are the same as in the previous systems.

Table 9: Results of the performances CNNs 6x6 ROIs classifier system in terms of Dice Coefficient, Precision and Recall considering all the mask and only the tumoral object identified by the system

	ALL			
CNN 9X9	Train		Test	
	Mean ± std	Median | 25th | 75th	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.50 ± 0.20	0.55 | 0.38 | 0.66	0.43 ± 0.19	0.45 | 0.29 | 0.57
PRECISION	0.60 ± 0.27	0.65 | 0.38 | 0.83	0.51 ± 0.27	0.53 | 0.27 | 0.74
RECALL	0.50 ± 0.20	0.51 | 0.42 | 0.63	0.49 ± 0.21	0.49 | 0.34 | 0.63
	ONLY TUMOR			
	Train		Test	
	Mean ± std	Median | 25th | 75th	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.52 ± 0.20	0.56 | 0.39 | 0.66	0.45 ± 0.18	0.46 | 0.31 | 0.58
PRECISION	0.62 ± 0.25	0.66 | 0.42 | 0.83	0.51 ± 0.26	0.53 | 0.28 | 0.73
RECALL	0.51 ± 0.20	0.51 | 0.45 | 0.64	0.50 ± 0.21	0.50 | 0.36 | 0.64

As expected, the performances of the last system are poorer that the others, since the area of the considered ROIs is much bigger. Indeed, the Dice Coefficient is around 0.50, the Precision around 0.60, and the Recall around 0.50, values that confirms that, considering the parameters used for the implementation of the CNNs, larger the ROIs poorer the performances.

It is possible to notice the trend of the paraments considering all the mask and only the tumoral object thank to the following graphs (fig.32, fig.33 and fig.34).

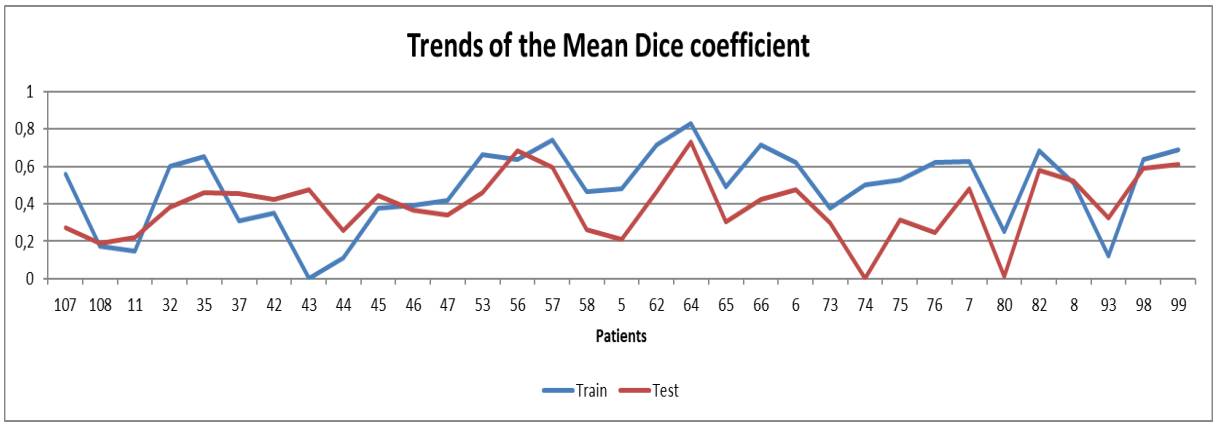


Figure 34: Mean Dice Coefficient's trend of the CNNs 9X9 ROIs classifier system, considering all the mask and only the tumoral object identified by the mask

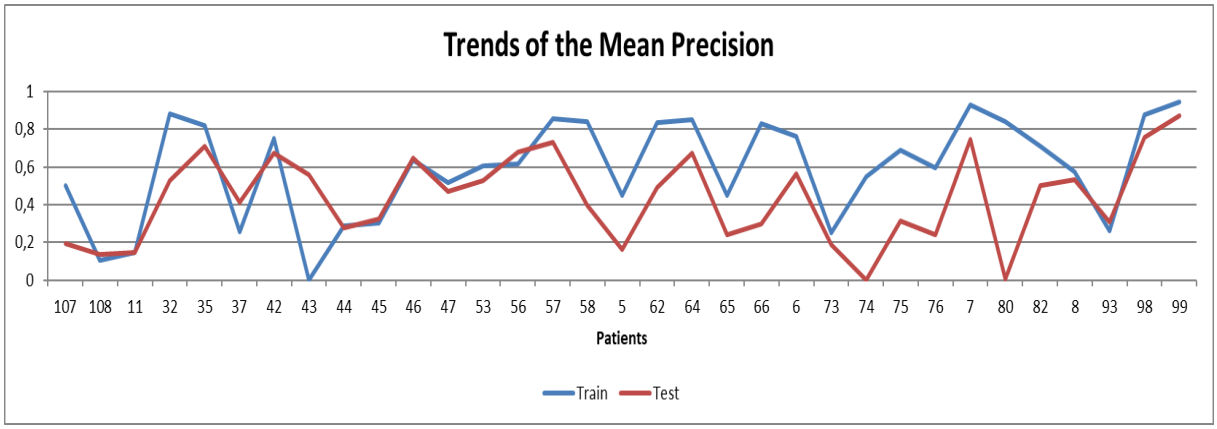


Figure 33: Mean Precision's trend of the CNNs 9X9 ROIs classifier system, considering all the mask and only the tumoral object identified by the mask

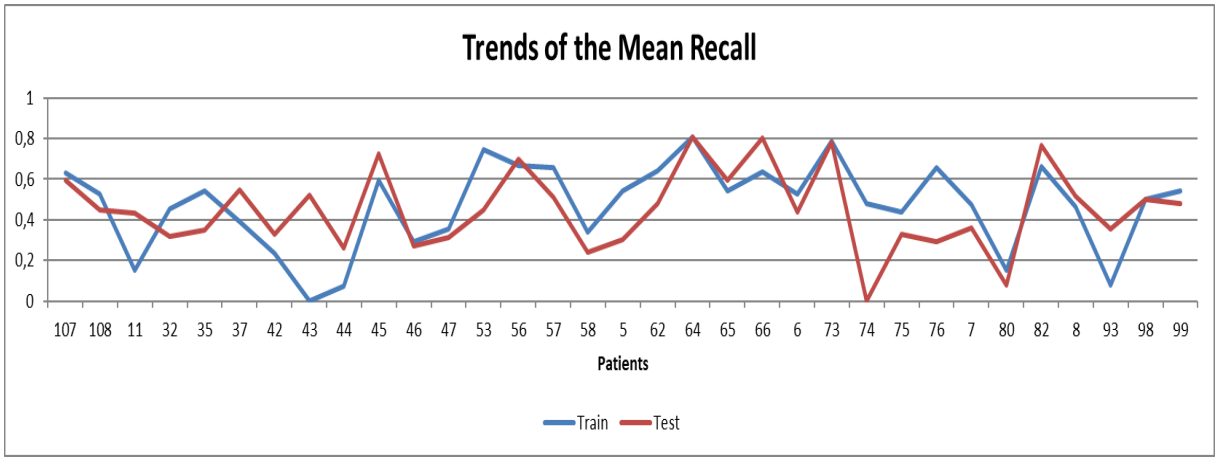


Figure 32: Mean Recall's trend of the CNNs 9X9 ROIs classifier system, considering all the mask and only the tumoral object identified by the mask

As expected the trends are considerably different in different patients, such as 5, 32, 42. This means that the net wrongly classifies several objects in the image, thus having a large number of *False Positive* elements. Moreover, there are several patients on which the performances of the system are very poor (11, 43, 44, 74, 80, 107, 108).

Another analysis has been made on each obtained mask related to the three CNNs. It is important to notice that the results related to the CNNs which classify the ROIs of the DWI B1000 and ADC sequences are obtained using the same dataset used for the training. Below the table (tab.10) with the results of the CNNs which classify the 3x3 ROIs of the T2w, DWI B1000 and ADC sequences.

Table 10: Results of the performances of the CNN 9X9 ROIs classifier related to the T2w sequence in terms of Dice Coefficient, Precision and Recall

CNN 9X9 T2W	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.30 ± 0.19	0.27 | 0.15 | 0.45
PRECISION	0.20 ± 0.15	0.16 | 0.08 | 0.31
RECALL	0.92 ± 0.15	0.99 | 0.92 | 1.00

Table 11: Results of the performances of the CNN 9X9 ROIs classifier related to the DWI B1000 sequence in terms of Dice Coefficient, Precision and Recall

CNN 9X9 B1000	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.51 ± 0.24	0.54 | 0.30 | 0.70
PRECISION	0.69 ± 0.30	0.82 | 0.47 | 0.94
RECALL	0.53 ± 0.25	0.54 | 0.38 | 0.72

Table 12: Results of the performances of the CNN 9X9 ROIs classifier related to the ADC sequence in terms of Dice Coefficient, Precision and Recall

CNN 9X9 ADC	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.32 ± 0.19	0.31 | 0.15 | 0.50
PRECISION	0.29 ± 0.21	0.26 | 0.13 | 0.48
RECALL	0.46 ± 0.21	0.45 | 0.33 | 0.59

It is possible to notice that the CNNs which classify the DWI B1000 sequences has high value of Recall, which means that correctly segment the tumoral area in those images, while the performances of the network related to the ADC sequence are very poor.

Below there are some graphs which shows how the CNN related to the DWI B1000 sequence behaves with the different patients, in comparison with the CNN related to the ADC sequence's behaviour.

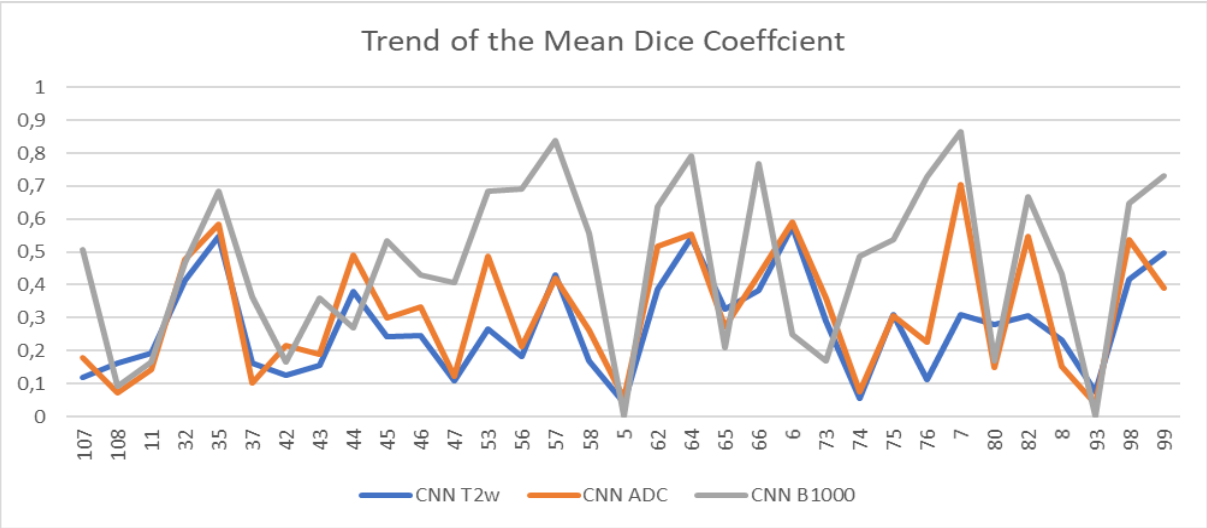


Figure 35: Comparison between the Mean Dice Coefficient of the CNN related to the T2w sequence, the CNN related to the DWI B1000 sequence and CNN related to the ADC sequence

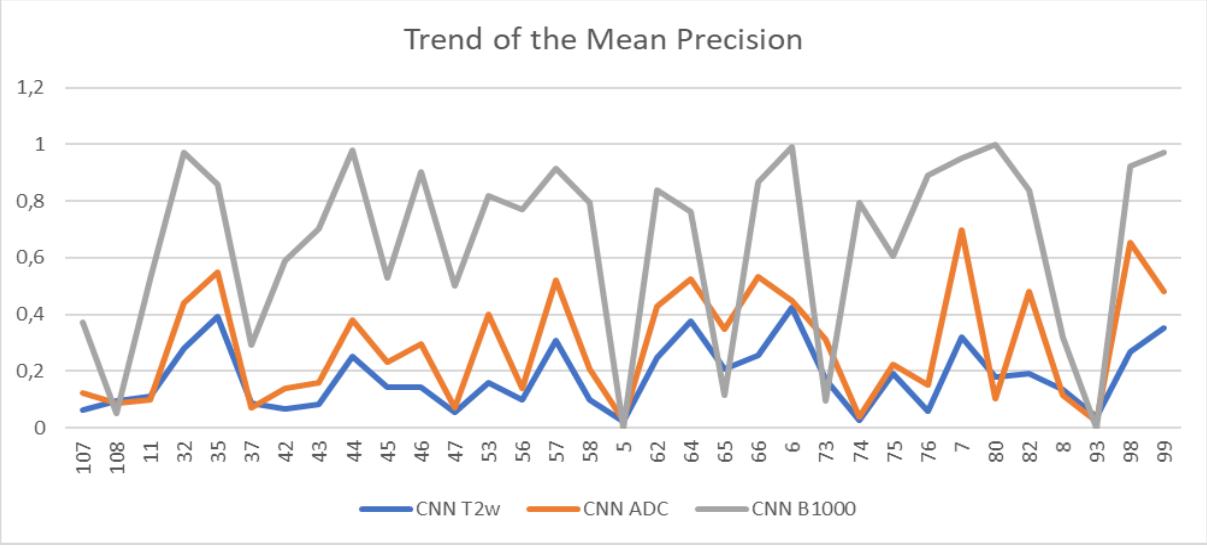


Figure 36: Comparison between the Mean Precision of the CNN related to the T2w sequence, the CNN related to the DWI B1000 sequence and CNN related to the ADC sequence

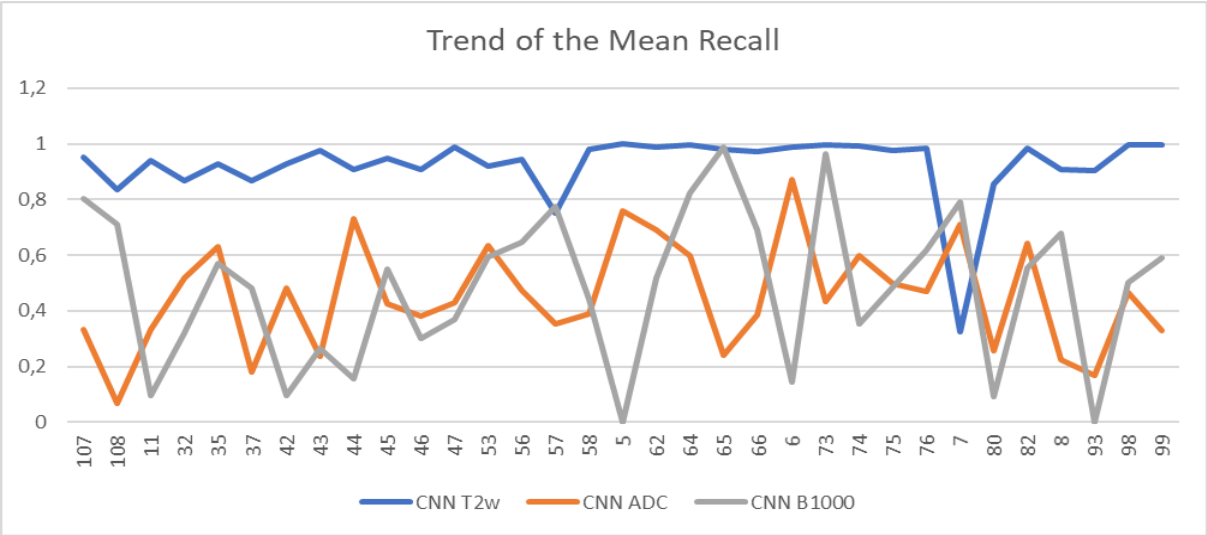


Figure 37: Comparison between the Mean Recall of the CNN related to the T"w sequence, the CNN related to the DWI B1000 sequence and CNN related to the ADC sequence

By the graphs it is possible to notice that in general the CNN which classify the DWI B1000 sequences has good performances, but there are several patients where the accuracy is very poor.

Here some example of well segmented slices and badly segmented (fig.38, fig39).

Well segmented patient:

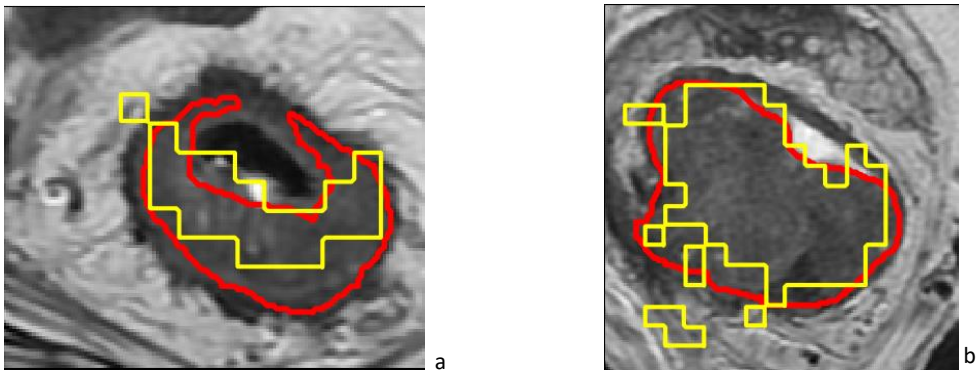


Figure 38: Patient 62 (a) and Patient 64 (b) - manual segmentation (red) and CNNs 9x9 ROIs segmentation (yellow)

Badly segmented patient:

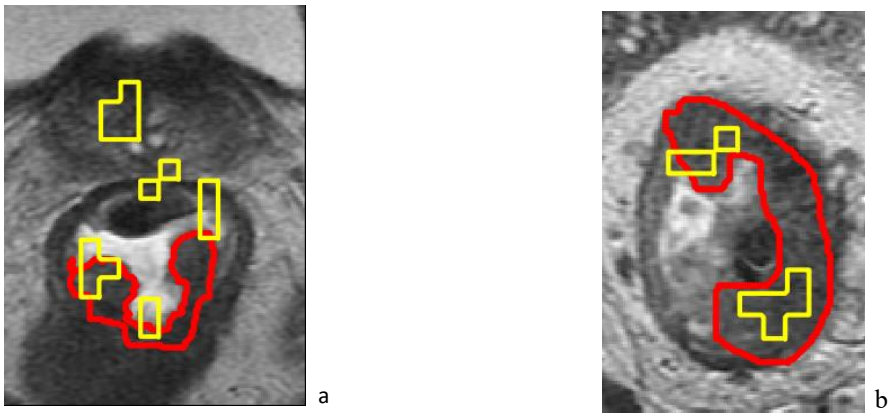


Figure 39: Patient 11 (a) and Patient 80 (b)- manual segmentation (red) and CNNs 9x9 ROIs segmentation (yellow)

Overall, the system with the best performance is the 3x3 ROIs classifier. In fact, it has Dice Coefficient 0.53±0.14, Precision 0.53±0.17 and Recall 0.64±0.19. The second system, the 6x6 ROIs classifier has Dice Coefficient 0.41±0.13, Precision 0.55±0.18 and Recall 0.44±0.15.

Validation

The purpose of the validation is to test the systems' ability to predict new data that are not used in the training set in estimating it. In this case the validation has been done using the Leave-one-out method. This method consists on partitioning a sample of data into complementary subsets, performing then the training on one subset, which become the new training set, and the other is used as the testing set. in this case the generated test set consists of the slices of the three sequences related to one patient, while the training set contains all the slices of the other patients. Since the last system has shown poor performances, the validation has been done considering only the first two implemented systems.

Thanks to the following graphs (fig.40, fig.41,fig.42) it is possible to notice how the performances of the net changes due to the lack of information of a specific patient. Also in this case the evaluated parameters are the Dice Coefficient, Precision and Recall, again considering all the mask and only the tumoral object.

- **SYSTEM OF CNNS. 3X3 ROIS CLASSIFIER**

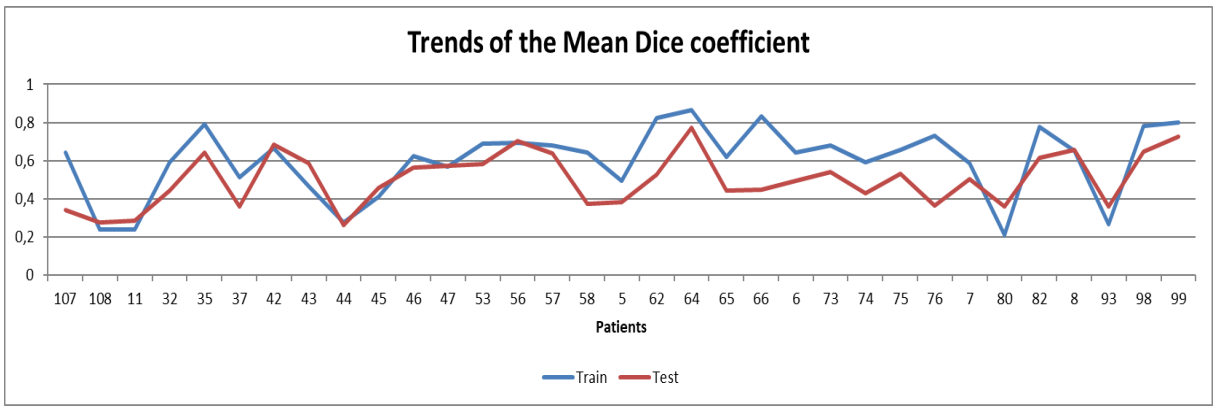


Figure 42: Mean Dice Coefficient 's trend of the CNNs 3X3 ROIs classifier system, considering all the mask and only the tumoral object identified by the mask

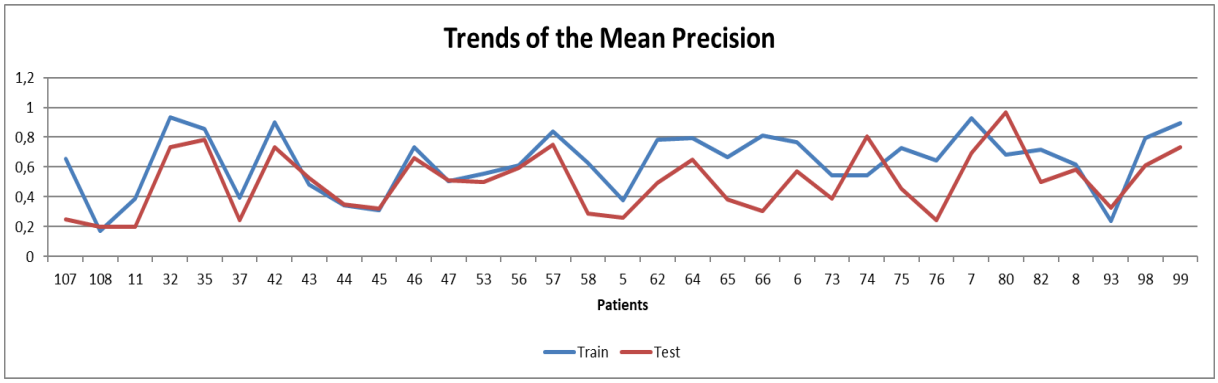


Figure 41: Mean Precision's trend of the CNNs 3x3 ROIs classifier system, considering all the mask and only the tumoral object identified by the mask

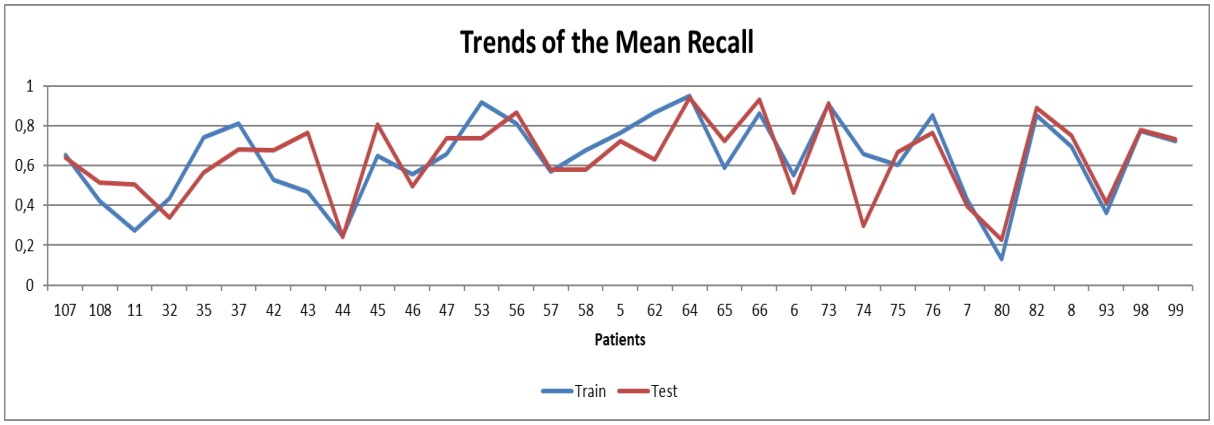


Figure 40: Mean Recall 's trend of the CNNs 3X3 ROIs classifier system, considering all the mask and only the tumoral object identified by the mask

By analysing the graphs, it is possible to notice that the performances are almost the same considering all the mask and only the object, except for the patients 11, 80, 74 which shows relevant differences, due to the *False Positive* elements.

- **SYSTEM OF CNNS 6X6 ROIS CLASSIFIER**

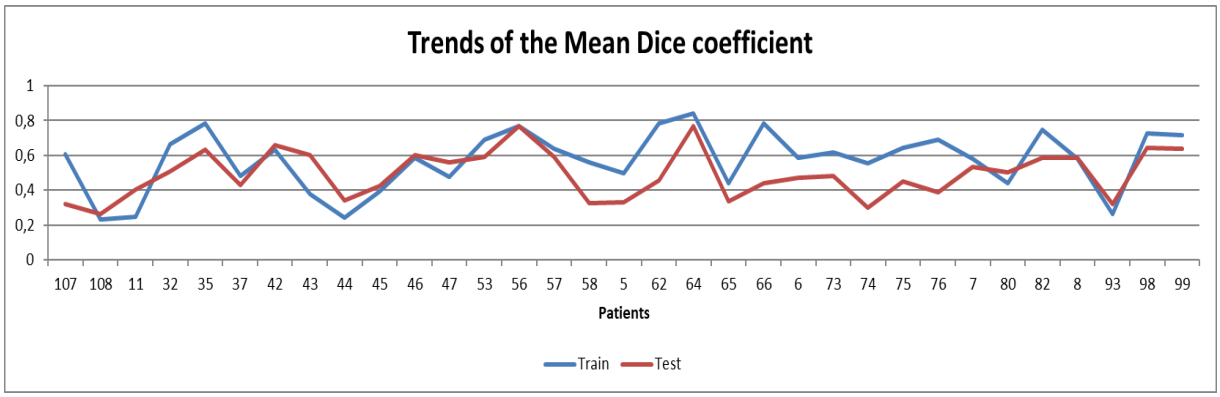


Figure 43: Mean Dice Coefficient 's trend of the CNNs 6x6 ROIs classifier system, considering all the mask and only the tumoral object identified by the mask

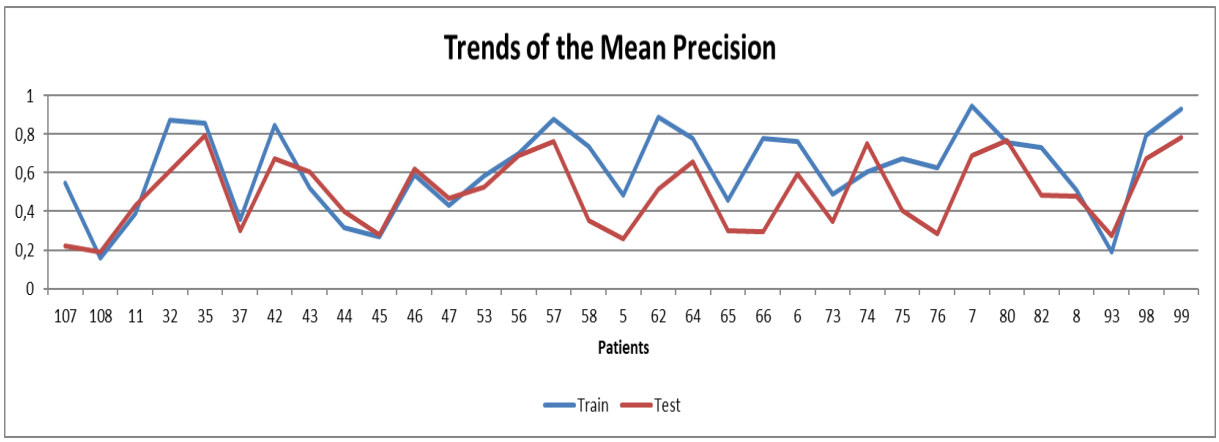


Figure 44: Mean Precision's trend of the CNNs 6x6 ROIs classifier system, considering all the mask and only the tumoral object identified by the mask

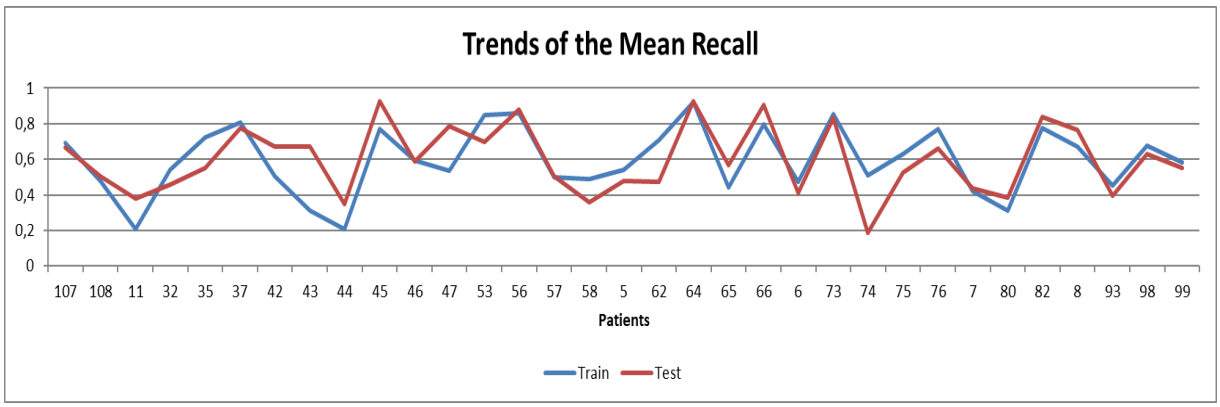


Figure 45: Mean Recall's trend of the CNNs 6x6 ROIs classifier system, considering all the mask and only the tumoral object identified by the mask

By analysing the graphs, it is possible to notice again that the trends are almost similar to each other, except for the patient 5, which shows relevant differences related to the Dice Coefficient ad the Precision. This means that the lack of information of this patient considerably affects the performance of the net, both for the CNNs 3x3 ROIs classifier and CNNs 6x6 ROIs classifier.

Results comparison

It has been made also a comparison between the results of this study with the ones that are in the literature (the comparison is among the average values):

Table 13: Comparison of the values of Dice Coefficient, Precision and Recall and other parameters between the CNN systems implemented and the literature.

	DICE COEFFICIENT	PRECISION	RECALL	SPECIFICITY	SENSITIVITY	HAMMOUNDE DISTANCE
IRVING ET AL. (4)	0.65 ±0.15	\	\	\	\	\
JLAN ET AL. (3)	0.84	0.83	0.97	0.88	0.27	8.2
TREBESCHI ET AL. (7)	0.69±0.01	\	\	\	\	\
HUANG ET AL. (8)	0.74±0.15	\	0.75±0.19	\	\	\
SOOMRO ET AL. (12)	0.94	\	\	\	\	\
HUANG ET AL. (10)	0.72±0.14	\	\	\	\	\
3X3 ROIS CLASSIFIER	0.61±0.19	0.65±0.22	0.65±0.22	\	\	\
6X6 ROIS CLASSIFIER	0.58±0.18	0.62±0.22	0.62±0.20	\	\	\
9X9 ROIS CLASSIFIER	0.50±0.20	0.60±0.27	0.50±0.20	\	\	\

It is possible to notice that the performances of the implemented systems in this study present lower performances than the ones in literature, but it is important to notice that these results are obtained considering the fact that some studies ((4), (3), (7), (8), (12) , (10)) use the U-Net algorithm, which classify each pixel of the image, not the ROI, and moreover the database in literature is much larger than the one used in this case.

U-Net

Material and Methods

Subject and Study Dataset

For this study 33 patients from the Candiolo Cancer Institute (IRCC Candiolo) with proven locally colorectal carcinoma were chosen. Among them there are 22 males and 11 females with adenocarminomas (28 cases) and mucinous carcinomas (5 cases). All patients have undergone multiparametric (mp) MRI, consisting of T2 weighted and diffusion weighted imaging (DWI), both axially angled. The diffusion sequence was performed using b-values B0 and B1000.

For each patient is an initial mask was created using a k-mean algorithm, and three slices of the tumoral volume were then manually adjusted by a radiologist. The manually segmented slices were used as ground truth. The final dataset consisted of 99 slices from the T2w sequence and corresponding segmentation masks, used for the creation of the training set and part of the test set. In fact, the test set consists of all the slices of the patients, including the not manually modified ones.

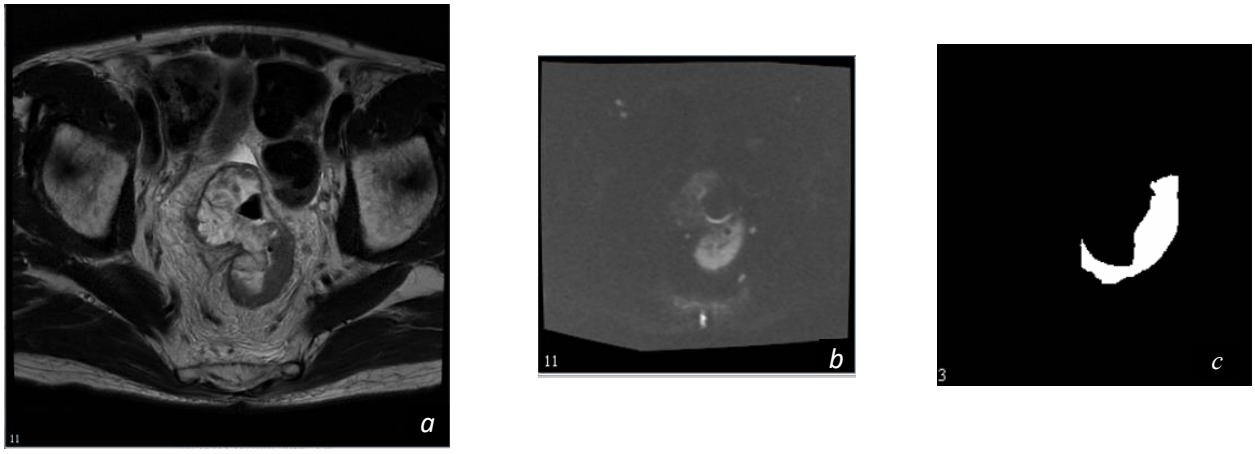


Figure 46: Some examples of T2w (a), DWI B1000 (b) and manual mask used (c)

Pre-processing

The pre-processing consists of the cropping phase, very similar to the previous phase explained. In fact, the cropping phase aims to identify the region where is the tumor. For doing so, a Fuzzy-c-means clustering is applied on each DWI B1000 slice. For each slice the algorithm creates four clusters, obtaining four different centroids. A first mask related to the tumor is created considering only the clusters having the value of the centroid between

the 50-percentile and 85-percentile, to minimize the artefacts. Then all the identified objects during the previous steps which are close to the edge of the image, and not centered are removed, since the colorectal tumor is known to be in the colorectal area. From the obtained mask, which contains the tumor, the region of interest is evaluated and applied on the T2w sequence. As in figure 47.

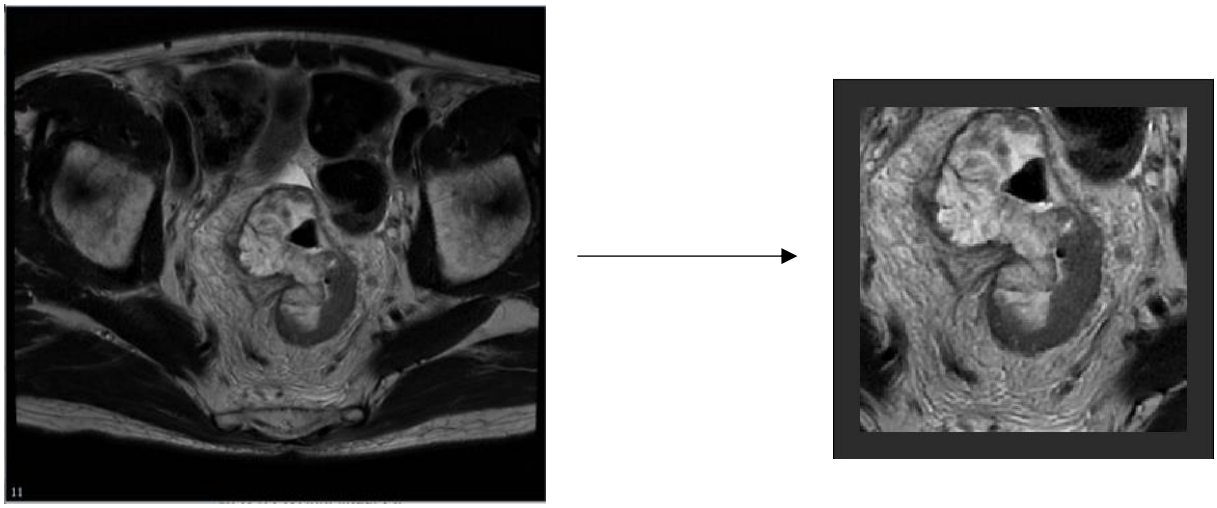


Figure 47: Example of T2w cropped

The cropped sequence is then subject to the standardization process [eq.05], thus to have the intensities of the images between 0 and 1.

Equation 9: Standardization formula

$$img_{\text{standardized}} = \frac{img - mean}{standard\ deviation}$$

After the standardization process, the cropped sequence in centered in the 256x256 matrix, which will be the input of the U-Net network. The final dataset consists of 99 standardized T2w 256x256images. The training set consists of all the dataset, while the testing set of the dataset and the slices of the T2w sequence which are not proven by the radiologist. The main issue of the training set for this network is the fact that is unbalanced, since the number of pixels belonging to the background is considerably higher than the number of pixels belonging to the tumoral area.

Method

Basic concepts of U-Net

The U-Net network belongs to the Fully Convolutional Network (FCN) class (15) (15). The main difference between the FCNs and CNNs is the absence of the Fully Connected layer, which is replace with the Up-Sampling layer and Deconvolutional layer. Thus, instead of obtaining a probability score to each class to classify the whole image, FCNs crate a score map for each class which has the same sizes of the input image, classifying each pixel.

The U-Net structure is divided in two parts, the **contracting path** and the **expansive path** (16).

The contracting path follows the typical architecture of a Convolutional neural network, with several Convolutional blocks applied on different levels of resolution, which decreases step by step. Each Convolution block consists of two subsequent Convolutional layers followed by a Max-Pooling layer. Due to this
architecture, the contracting path aims to extract the features related to the identification of the object of interest.
The expansive path has a symmetrical architecture of the contracting path. The Up-Sampling layer increases the resolution of the feature maps step by step. In this path, the results of the Up-Sampling layers and of the corresponding Convolution blocks of the first path are concatenated, and subject to new Convolution blocks, so-called Deconvolutional layers, in order to improve the predication accuracy . As a matter of fact, it is possible to extract the features regarding the positions on the image of the objects of interest. The last layer consists of a Convolution layer which creates the score map, from which the segmentation mask is extracted.

As for the previous neural network, there are several parameters which affects the performances of the network and its training phase. The optimizer, which evaluates the gradient error for the backpropagation process, plays an important role for the training phase. The most commonly used optimizers are *Adam* (*Adaptive moment estimation*) and *SGD* (S*tochastic Gradient Descent)* (17). The learning rate defines how the weights of the neurons are modified, affecting the computational time of the training phase. The loss function affects how the network learns. The most commonly used are the *binary* and *categorical cross-entropy* [eq 7]. Due to the unbalanced training dataset (in the mask the number of pixels belonging to the background is considerably higher than the number of the pixels belonging to the tumoral area, Huang et al. (8) have proposed to use the *Dice loss function*, to overcome this issue. The general formula which can be used, since it must be minimized [eq.10]:

Equazione 10: Dice Loss Function

$$Dice\ Loss\ Function = 1 - \frac{2 \cdot |A \cap B|}{|A| + |B|}$$

Where A is the ground truth (the manually segmented mask) and B is the precited segmentation mask by the network.

Moreover, the number of Convolution blocks in the contracting path, and consequently in the expansive path, affect the way the network learns. In fact, if the contracting path analyses too deeply the input image, the network may not be able to identify small objects, and the probability of extracting redundant features increase (15) (6) .

In this study the first U-Net architecture implemented consists of 5 Convolution blocks belonging to the contracting phase, where all the Convolutional layers use a 3x3 kernel with ReLU activation function, but the output which uses the 1x1 kernel with Sigmoid function. All the Max-Pooling and Up-Pooling layers uses 2x2 kernels. For the training of this network the loss function used is the *binary cross-entropy*, with *Adam* optimizer and lr = 0,0001 on 40 epochs. The database has been divided into training set (90%) and test set (10%). The training set has been used for training the net, while the test set for checking for model overfitting.

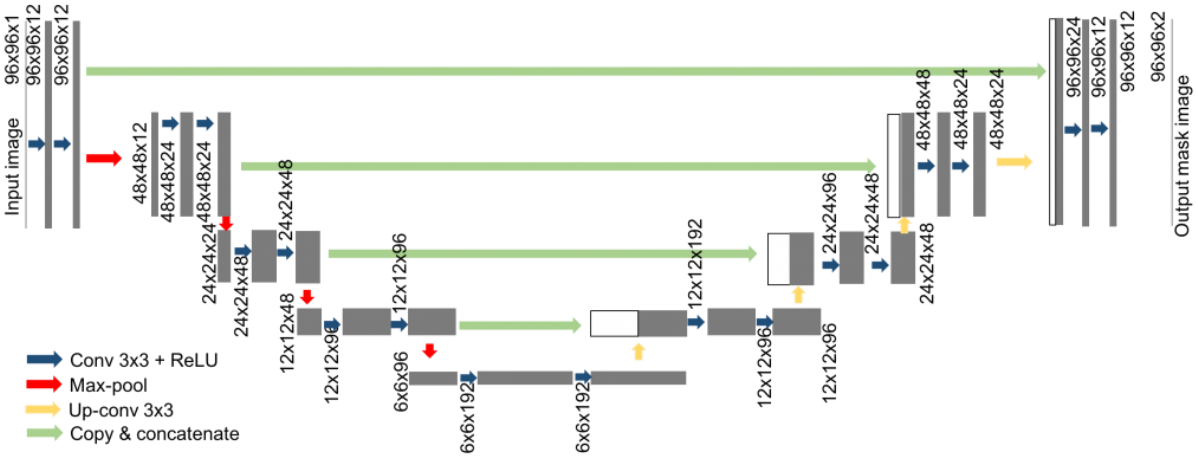


Figure 48: Architecture of the U-Net 1

The second U-Net implemented is very similar to the previous one, with the only difference regarding the number of Convolution blocks in the contracting path. In fact, for this net the number of Convolution block is 6, which means that the network ore deeply analyses the input image than the previous one.
The last implemented U-Net has the same structure as the previous net but has been trained on 60 epochs than 40.

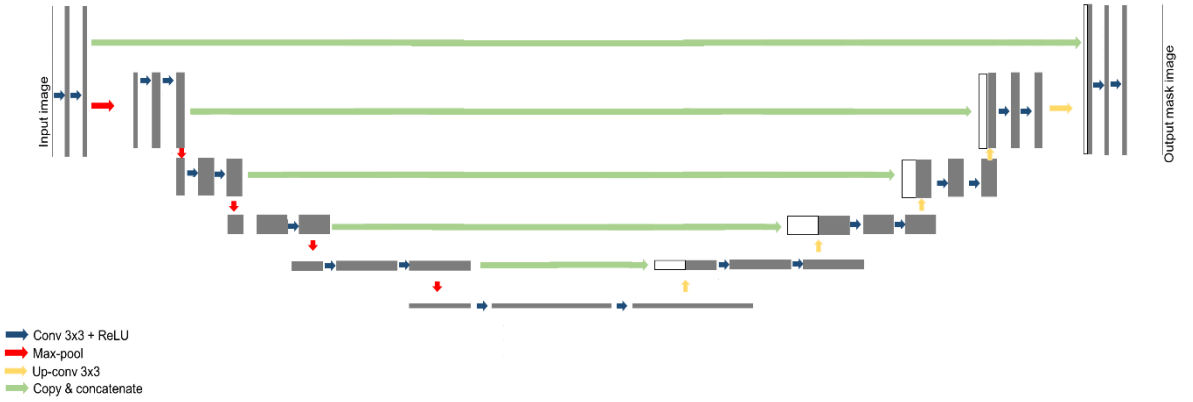


Figure 49: Architecture of the 2nd and 3rd U-Net

All the networks have been implemented on Python 3.7.0 with Keras (Theano backend).

Post-processing

The post-processing phase aims to reduce the false positive elements. Firstly, since the obtained mask from the network contains values which represent the probability of belonging to the tumoral class. Using the Ostu's thresholding (18) it is possible to extract

the binary mask. In order to reduce the False Positive elements, two different hypotheses are applied on the predicted mask slices:

- The tumoral object must have an area higher than 100 pixels and lower than half-area of the cropped image;
- The tumoral object must be connected on at least three slices.

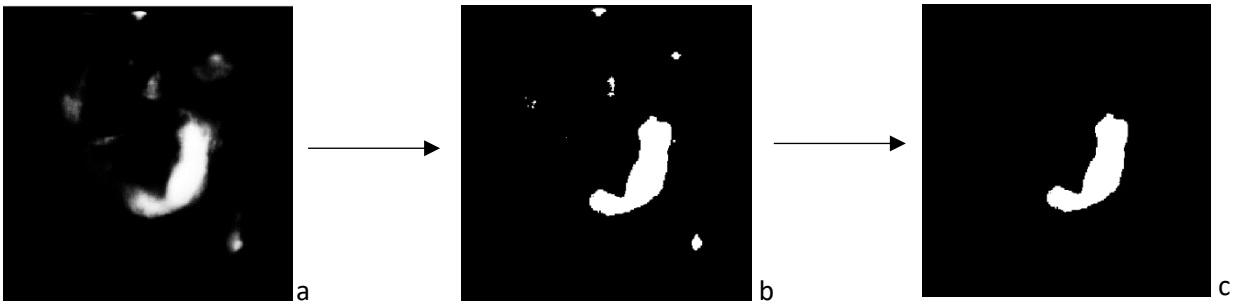


Figure 50: Post-processing steps: from the obtained mask (a) by the U-Net, a binary mask (b) is obtained using the Otsu thresholding. All the objects with area lower than 100 pixels and not connected on at least three slices are removed, thus to obtain the final mask (c)

Results

For the valuation of the performances of the networks several parameters have been evaluated: the dice coefficient, the precision and the recall, which equations are reported below [eq8]:

Equation 11: Dice Coefficient, Precision and Recall formuals

$$Dice\ coefficient = \frac{2\,TP}{2TP + FP + FN} \qquad Precision = \frac{TP}{TP + TP} \qquad Recall = \frac{TP}{TP + FN}$$

Where TP are the True Positive (all the pixel correctly classified as tumor), FP are the False Positive (all the pixels wrongly classified as tumor) and FN are the False Negative (all the pixels wrongly classified as background).
For each parameter the mean, standard deviation, median, 25-percentile and 75-percentile have been evaluated. Moreover, this analysis has been done firstly considering all the mask, thus all the false positive connected object wrongly identified as tumor from the net, and then considering only the connected object which correctly identifies the tumoral area. By doing so, it is possible to verify how much the net wrongly classifies.

The first implemented U-Net (U-Net 1) presents 4 descending levels, the loss function is the binary cross-entropy, the used optimizer is Adam, the learning rate is 0.0001, and it has been trained on 90% of the dataset with 40 epochs.
In the following tables (tab??, tab??) there are the values related to the Dice coefficient, Precision and Recall.

Table 14: Performances of the U-Net 1 in terms of DIce Coefficient, Precision and Recall, considering all the mask and only the tumoral object identified by the system

U-NET 1	ALL			
	Train		Test	
	Mean ± std	Median | 25th | 75th	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.70 ± 0.20	0.78 | 0.61 | 0.84	0.53 ± 0.21	0.55 | 0.35 | 0.71
PRECISION	0.72 ± 0.22	0.78 | 0.61 | 0.89	0.55 ± 0.27	0.56 | 0.34 | 0.79
RECALL	0.73 ± 0.22	0.82 | 0.68 | 0.88	0.65 ± 0.25	0.70 | 0.45 | 0.84
	ONLY TUMOR			
	Train		Test	
	Mean ± std	Median | 25th | 75th	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.71 ± 0.19	0.78 | 0.62 | 0.84	0.55 ± 0.21	0.56 | 0.42 | 0.71
PRECISION	0.75 ± 0.20	0.80 | 0.64 | 0.91	0.57 ± 0.26	0.59 | 0.37 | 0.80
RECALL	0.73 ± 0.22	0.82 | 0.68 | 0.88	0.65 ± 0.25	0.70 | 0.45 | 0.84

From the previous tables it is possible to notice that the parameters are higher than the ones related to the CNNs system. In fact, the Dice coefficient value is around 0.60, the Precision around 0.63, the Recall around 0.68.

To better analyse how the images affect the network's performance, the following graphs show the trend of the mean Dice Coefficients, mean Precisions and mean Recalls, all evaluated considering all the prediction mask and only the predicted tumoral object.

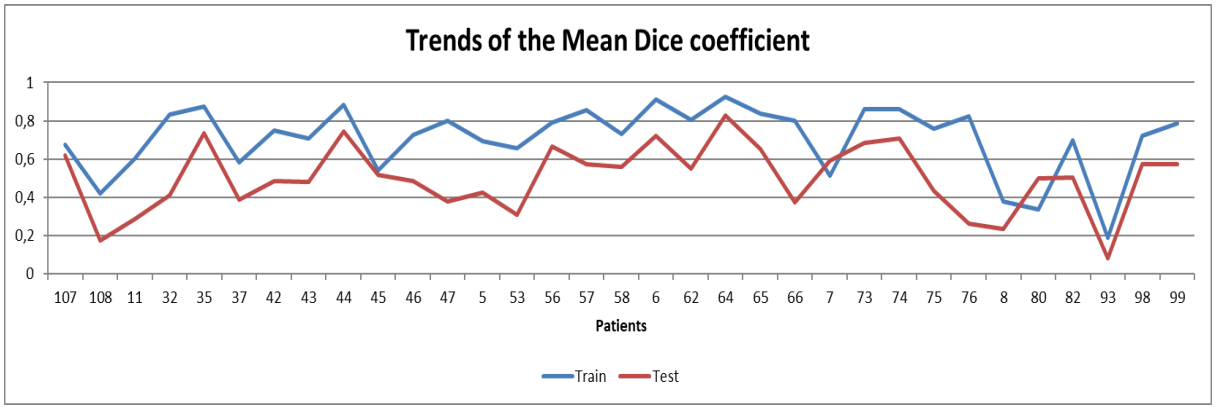


Figure 52: Mean Dice Coefficient's trend

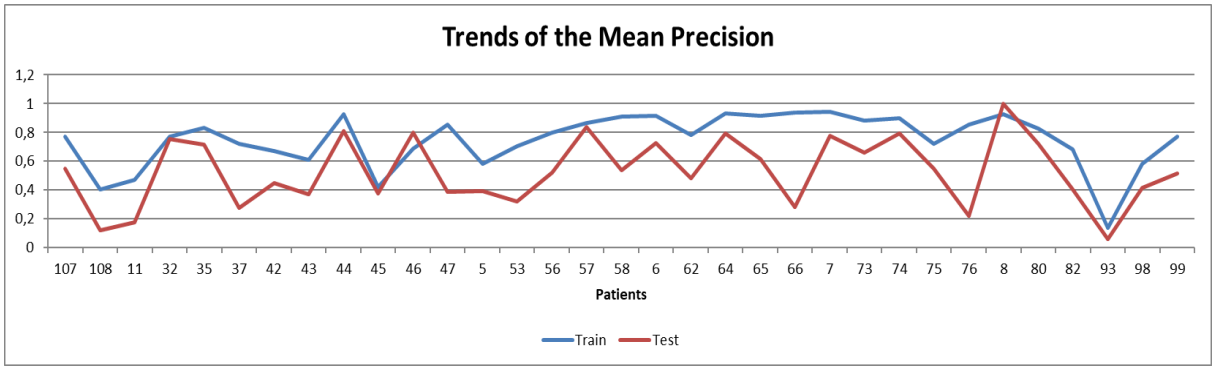


Figure 51: Mean Precision's trend

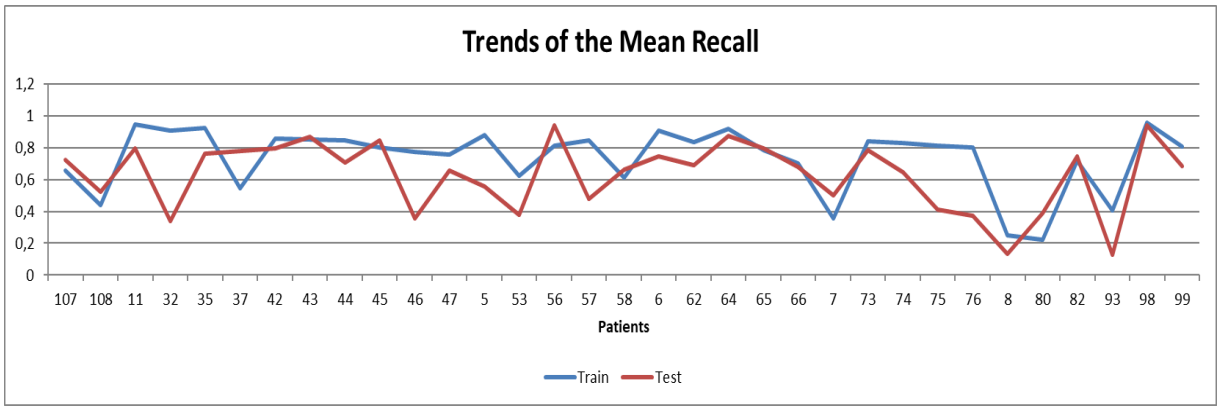


Figure 53: Mean Recall's trend

From the graphs it is possible to notice that trend of all the parameters are very similar considering both all the mask and only the tumoral object, except for the patients 5, 93, 107. In fact, the values are visibly different. Overall, it is possible to say that the net does not wrongly classify many connected object (low number of False Positive). It is important to notice that, even if the values of Dice Coefficient, Precision and Recall are higher than the one of the previous system, this net does not have satisfying performances.

Here some example of well segmented slices and badly segmented (fig.55, fig.56).

Well segmented patient:

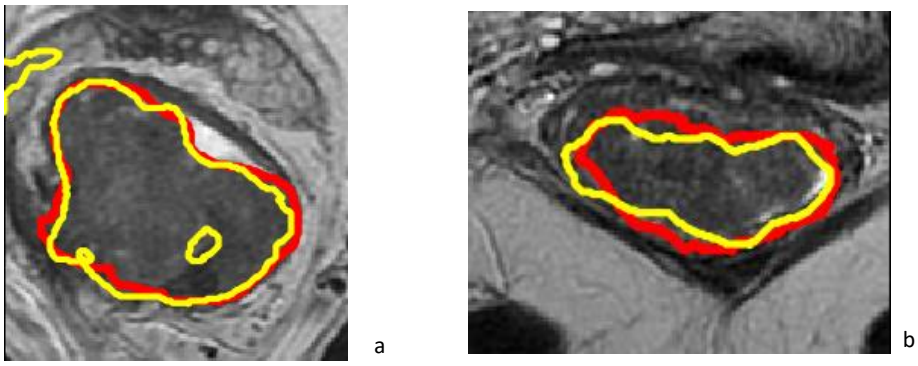


Figure 55: Patient 64 (a) and Patient 99 (b) - manual segmentation (red) and U-Net 1 segmentation (yellow)

Badly segmented patient:


a

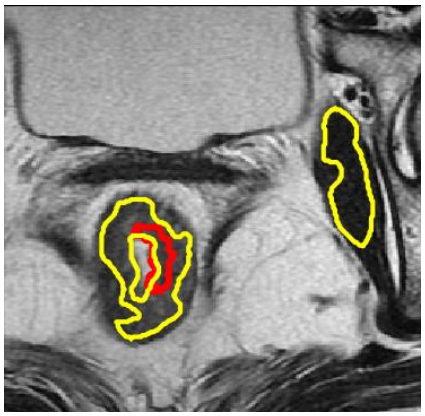

b

Figure 56: Patient 76 (a) and Patient 93 (b)- manual segmentation (red) and U-Net 1 segmentation (yellow)

In order to improve the accuracy, the number of descending levels has been increased. Indeed, the second implemented U-Net (U-Net 2) presents 5 descending levels, the binary cross-entropy as loss function, the Adam optimizer, the learning rate 0.0001, and again it has been trained for 40 epochs on the 90% of the dataset.

In the following tables (tab.15) show the values of Dice Coefficient, Precision and Recall considering all the mask, and only the predicted tumoral object.

Table 15:Performances of the U-Net 2 in terms of Dice Coefficient, Precision and Recall, considering all the mask and only the tumoral object identified by the system

U-NET 2	ALL			
	Train		Test	
	Mean ± std	Median | 25th | 75th	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.73 ± 0.20	0.81 | 0.66 | 0.87	0.54 ± 0.21	0.56 | 0.38 | 0.72
PRECISION	0.75 ± 0.18	0.80 | 0.68 | 0.88	0.57 ± 0.27	0.57 | 0.34 | 0.82
RECALL	0.78 ± 0.23	0.87 | 0.75 | 0.93	0.66 ± 0.26	0.72 | 0.48 | 0.89
	ONLY TUMOR			
	Train		Test	
	Mean ± std	Median | 25th | 75th	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.73 ± 0.20	0.81 | 0.67 | 0.87	0.56 ± 0.21	0.58 | 0.40 | 0.74
PRECISION	0.75 ± 0.18	0.80 | 0.68 | 0.88	0.59 ± 0.26	0.60 | 0.39 | 0.80
RECALL	0.78 ± 0.23	0.87 | 0.75 | 0.93	0.67 ± 0.26	0.73 | 0.49 | 0.90

By analysing the obtained values, it is possible to notice that by adding one descending layer the performance is improved in terms of correctly identified tumoral area. Indeed, the mean, median, 25-percentile and 75-percetile of all the parameters related are slightly increased, especially related to the Recall, but the standard deviation. For this system the Dice coefficient value is around 0.59, the Precision around 0.65, the Recall around 0.70.

It is possible to observe the trend of the mean of the Dice Coefficient, the Precision and the Recall.

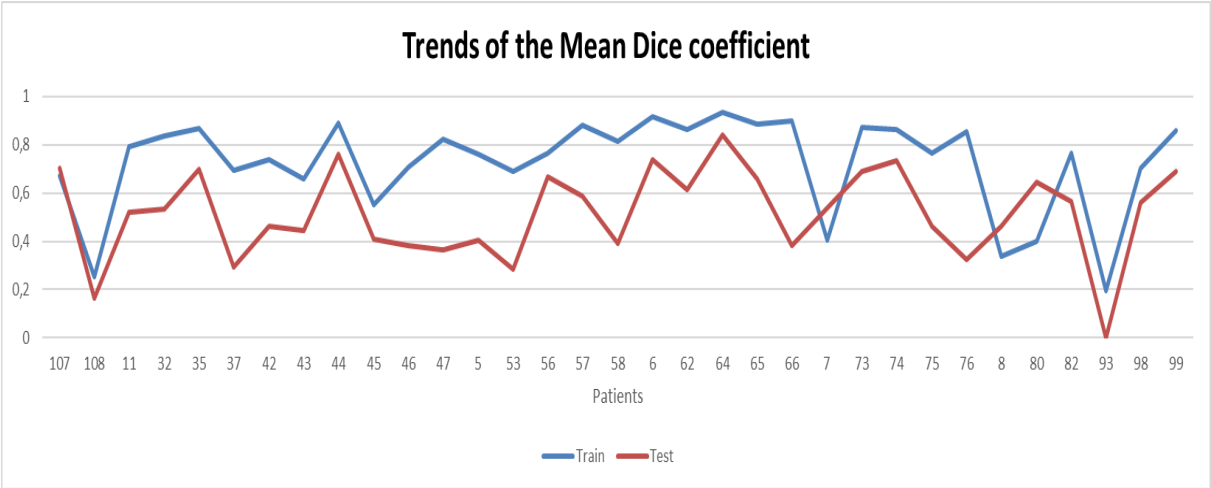


Figure 58: Mean Dice Coefficient's trend

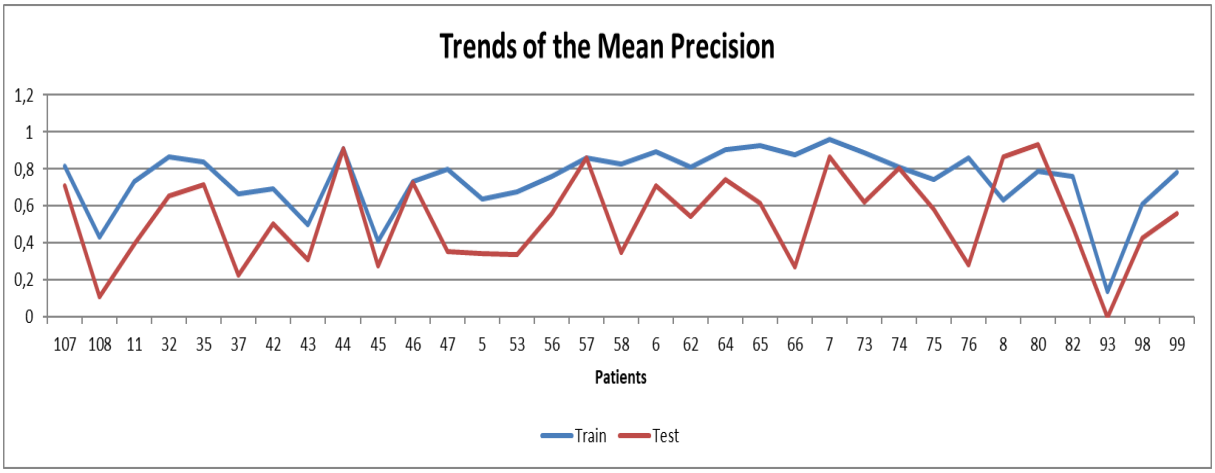


Figure 57: Mean Precision's trend

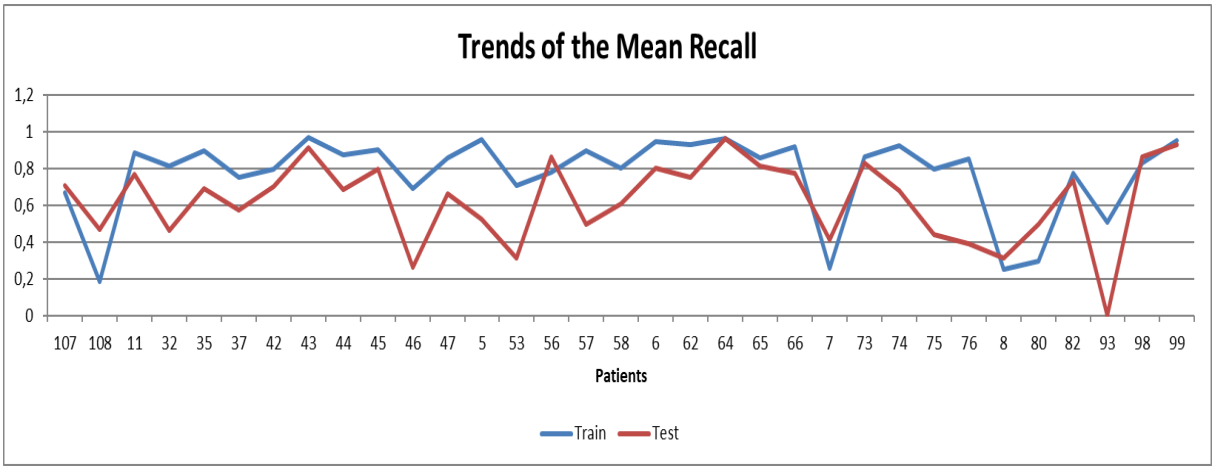


Figure 59: Mean Recall's trend

From the graphs it is possible to notice that trends of all the parameters are again similar considering both all the mask and only the tumoral object, but for some patients there are

visible differences. Overall, it is possible to say that the net still does not wrongly classify many connected objects among all the dataset, but in some cases the effects of the *False Positive* objects are relevant. It is important to notice that, even if the values of Dice Coefficient, Precision and Recall are higher than before, this net does not have satisfying performances.

Here some example of well segmented slices and badly segmented (fig.59, fig.60).

Well segmented patient:

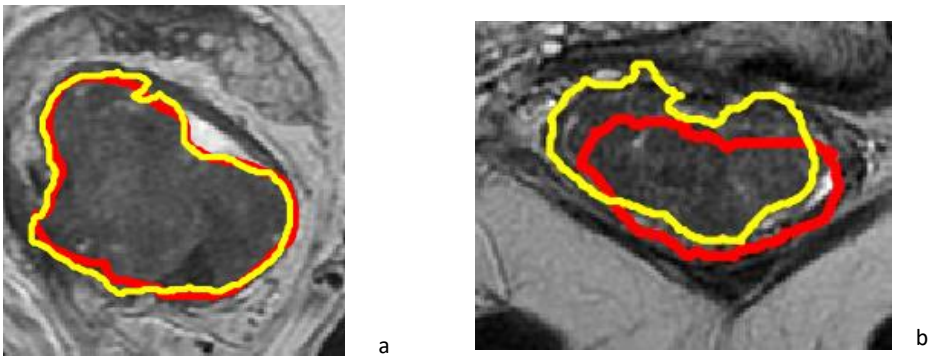


Figure 60: Patient 64 (a) and Patient 99 (b) - manual segmentation (red) and U-Net 2 segmentation (yellow)

Badly segmented patient:

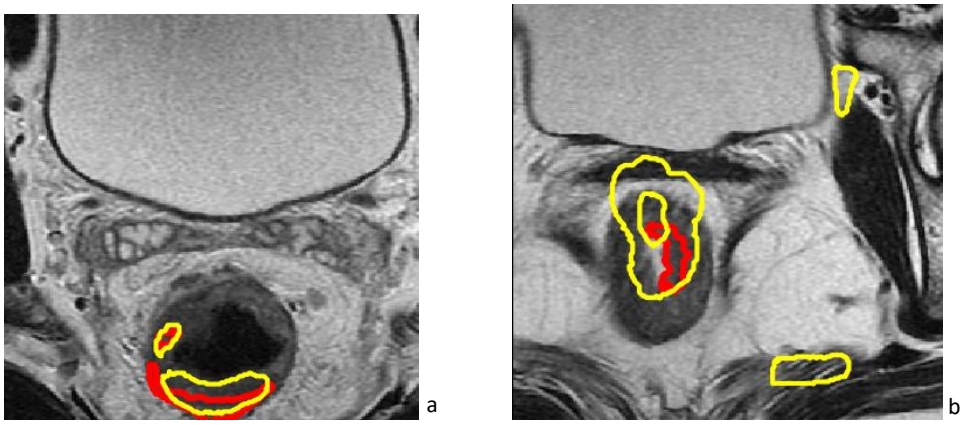


Figure 62: Patient 76 (a) and Patient 93 (b)- manual segmentation (red) and U-Net 2 segmentation (yellow)

Since the performances show an increment thanks to the additional layer, the third implemented network (U-Net 3) presents the same architecture of the previous network, but it has been trained on 60 epochs. In the following tables (tab.16) there are the values related to the already specified parameters.

Table 16:Performances of the U-Net 3 in terms of Dice Coefficient, Precision and Recall, considering all the mask and only the tumoral object identified by the system

U-NET 3	ALL			
	Train		Test	
	Mean ± std	Median | 25th | 75th	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.89 ± 0.15	0.92 | 0.88 | 0.95	0.58 ± 0.21	0.61 | 0.43 | 0.75
PRECISION	0.90 ± 0.16	0.95 | 0.91 | 0.97	0.61 ± 0.26	0.65 | 0.39 | 0.83
RECALL	0.88 ± 0.13	0.92 | 0.87 | 0.95	0.64 ± 0.24	0.69 | 0.48 | 0.85
	ONLY TUMOR			
	Train		Test	
	Mean ± std	Median | 25th | 75th	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.89 ± 0.14	0.92 | 0.88 | 0.95	0.59 ± 0.20	0.64 | 0.45 | 0.76
PRECISION	0.91 ± 0.15	0.95 | 0.91 | 0.97	0.63 ± 0.26	0.68 | 0.45 | 0.85
RECALL	0.88 ± 0.13	0.92 | 0.87 | 0.95	0.65 ± 0.24	0.69 | 0.49 | 0.85

Thanks to the increased number of epochs, all the parameters are significantly increased. In fact, the median Dice coefficient is around 0.70, the median Precision around 0.5 and the median Recall 0.76. These results are reasonably satisfying, since the network is trained only on the T2w sequences, and the parameters adopted are the ones already implemented in Keras.

It is possible to observe also the trend of the mean of the dice coefficient, the precision and the recall (fig.63, fig.64 an fig.65).

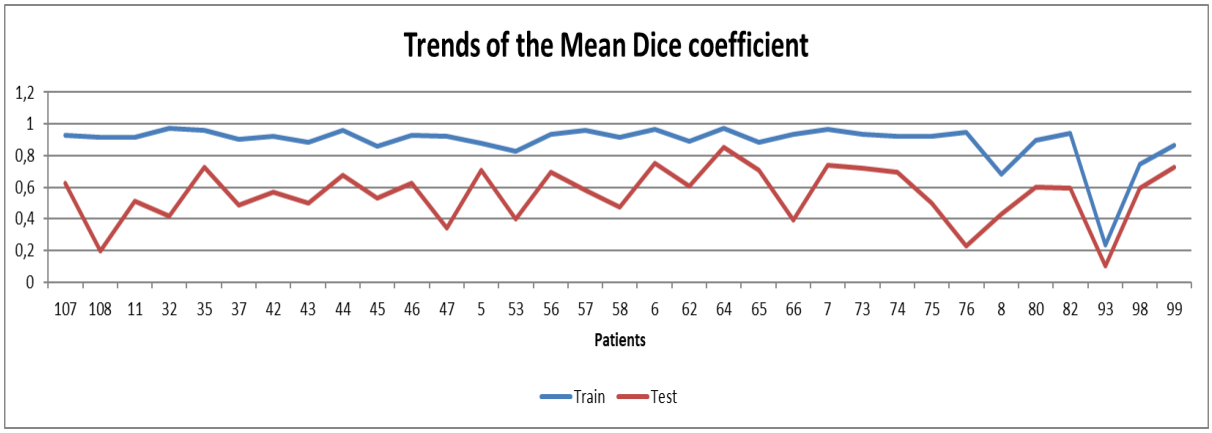


Figure 63: Mean Dice Coefficient's trend

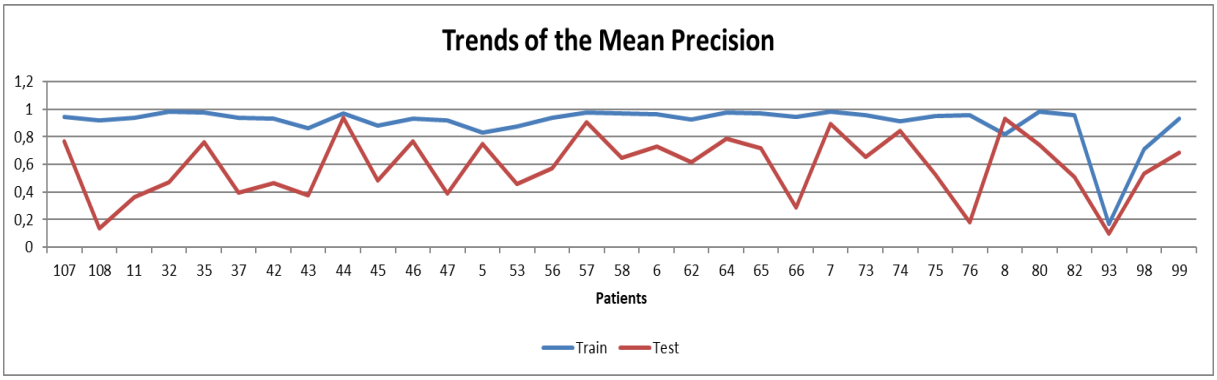


Figure 64: Mean Precision's trend

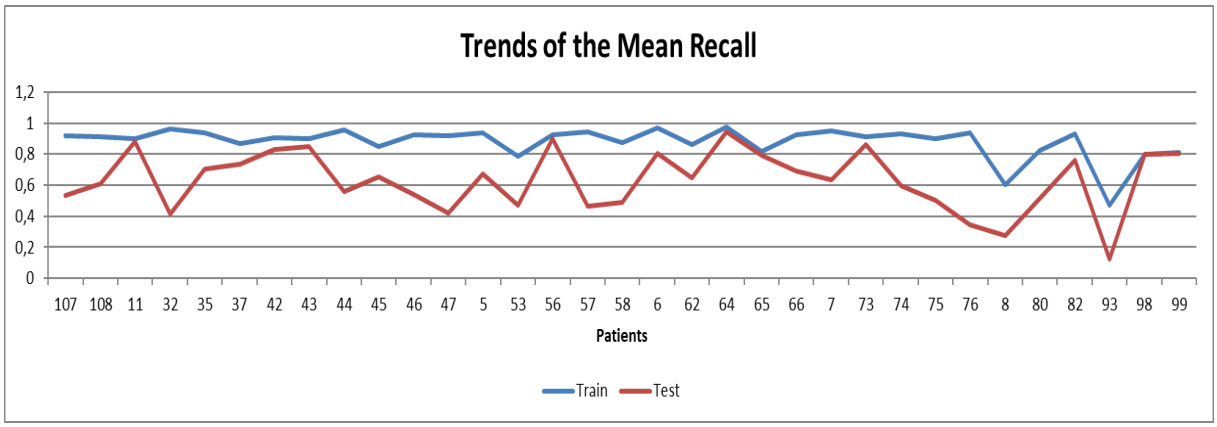


Figure 65: Mean Recall's trend

By analysing the trends it is possible to notice that the trends are almost the same considering all the mask and only the tumoral object, but for few patients, such as 32, 8, and 93. This means that the net has a little number of False Positive elements.

Here some example of well segmented slices and badly segmented (fig.64, fig.65).

Well segmented patient:

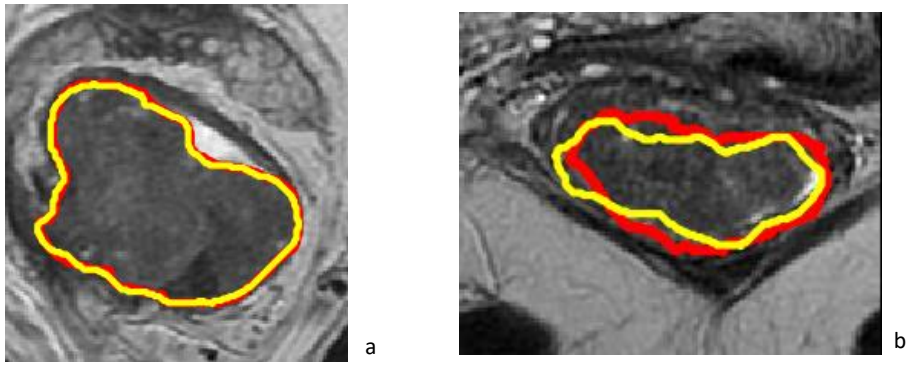


Figure 66: Patient 64 (a) and Patient 99 (b) - manual segmentation (red) and U-Net 3 segmentation (yellow)

Badly segmented patient:

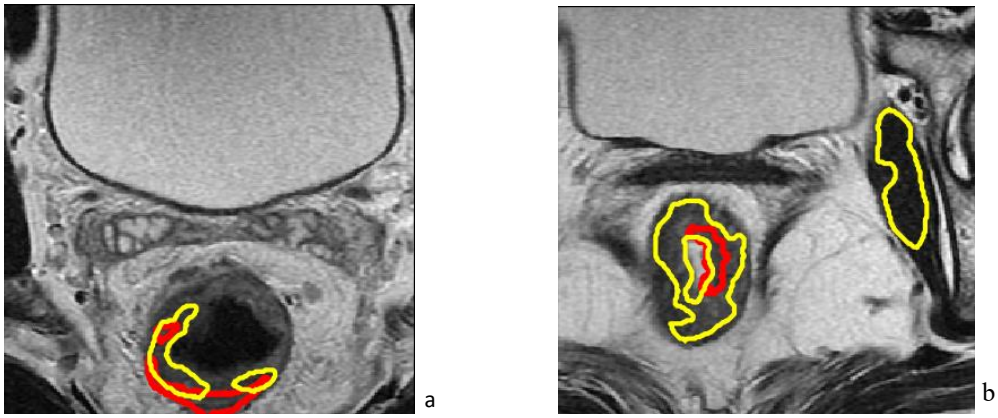


Figure 69:Patient 76 (a) and Patient 93 (b)- manual segmentation (red) and U-Net 3segmentation (yellow)

In conclusion it is possible to say that the last U-Net has the best performances (Dice Coefficient=0.67±0.14; Precision=0.71±0.14; Recall=0.75±0.14).

An additional neural network has been implemented. It shares the structure and the parameters of the *U-Net 3*, since it shows the best performances. The main difference is that the input images are characterized with three layers, one for the T2w, one for the DWI B1000 and the last one for the ADC. This system has been implemented considering the segmentation process carried by the radiologists. Indeed, the T2w sequence gives morphological informations, while the DWI B1000 and ADC sequences give localization and pathological informations of the cancer. This consideration has been made also for the CNNs systems, but in this case there is just one network analysing all the input informations, instead of three nets for the specific kind of sequence. Considering the multi-layered input image, the network will be called *Multi-layer U-Net*.

The pre-processing consists of pasting the cropped images in 256x256 zero matrices and standardizing them. Here some examples of input image (fig.66)

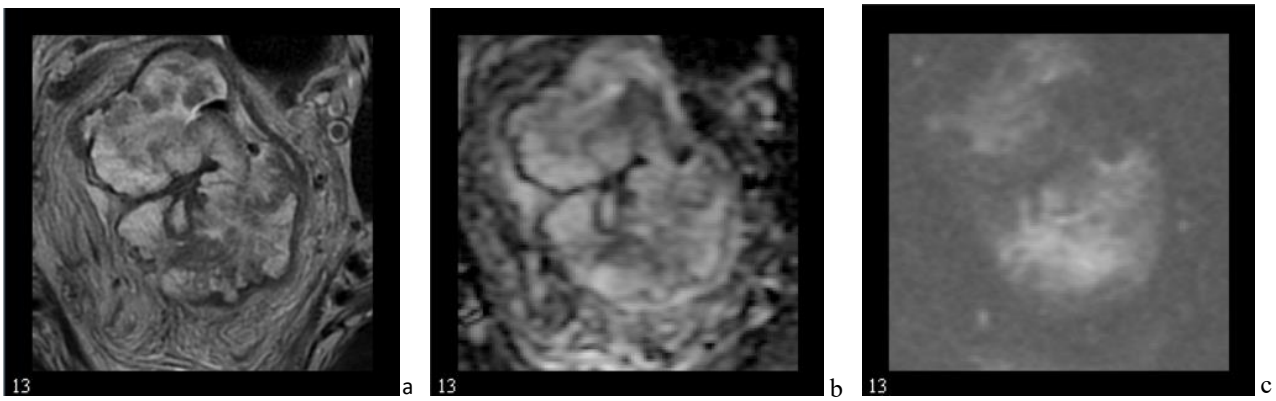


Figure 70: Input image's layers. T2w (a), DWI B1000 (b) and aDC (c) for the multi-layer U-Net

In the following tables (tab.17) there are the evaluated values related to the Dice coefficient, Precision and Recall considering all the mask and only the identified tumoral object.

Table 17: Performances of the last network implemented in terms of Dice Coefficient, Precision and Recall, considering all the mask and only the tumoral object identified by the system

	ALL			
MULTI-LAYERED U-NET	Train		Test	
	Mean ± std	Median | 25th | 75th	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.89 ± 0.14	0.93 | 0.88 | 0.96	0.61 ± 0.18	0.64 | 0.53 | 0.76
PRECISION	0.87 ± 0.15	0.92 | 0.85 | 0.95	0.57 ± 0.21	0.59 | 0.41 | 0.73
RECALL	0.91 ± 0.13	0.97 | 0.92 | 0.99	0.77 ± 0.21	0.83 | 0.67 | 0.93
	ONLY TUMOR			
	Train		Test	
	Mean ± std	Median | 25th | 75th	Mean ± std	Median | 25th | 75th
DICE COEFFICIENT	0.89 ± 0.14	0.93 | 0.88 | 0.96	0.62 ± 0.18	0.64 | 0.53 | 0.76
PRECISION	0.87 ± 0.15	0.92 | 0.85 | 0.95	0.57 ± 0.21	0.60 | 0.41 | 0.73
RECALL	0.91 ± 0.13	0.97 | 0.92 | 0.99	0.77 ± 0.21	0.83 | 0.68 | 0.93

From the previous tables it is possible to notice that the performances of the *Multi-layered U-Net* and the *U-Net 3* are very similar between each other. Despite the use of different sequences, the U-Net 3, which relies only on the T2w, shows slightly better performances.

Possible reasons behind this are:

- the not appropriate structure of the network;
- the number of training epochs, which probably should be higher than 60;
- the noisy DWI B1000 and ADC images which could add useless informations to the network.

To better analyse how the images affect the network's performance, the following graphs (fig.67, fig.68, fig.69) show the trend of the mean Dice Coefficients, mean Precisions and mean Recalls, all evaluated considering all the prediction mask and only the predicted tumoral object.

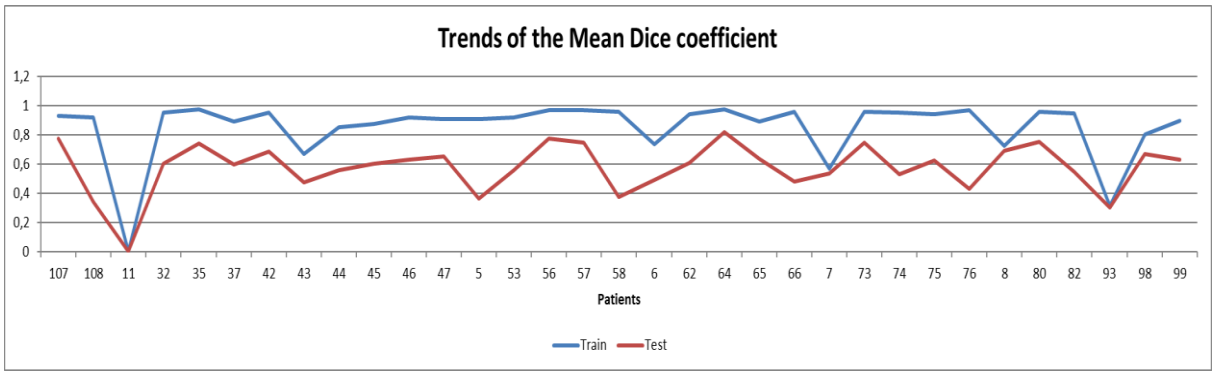


Figure 74: Mean Dice Coefficient's trend

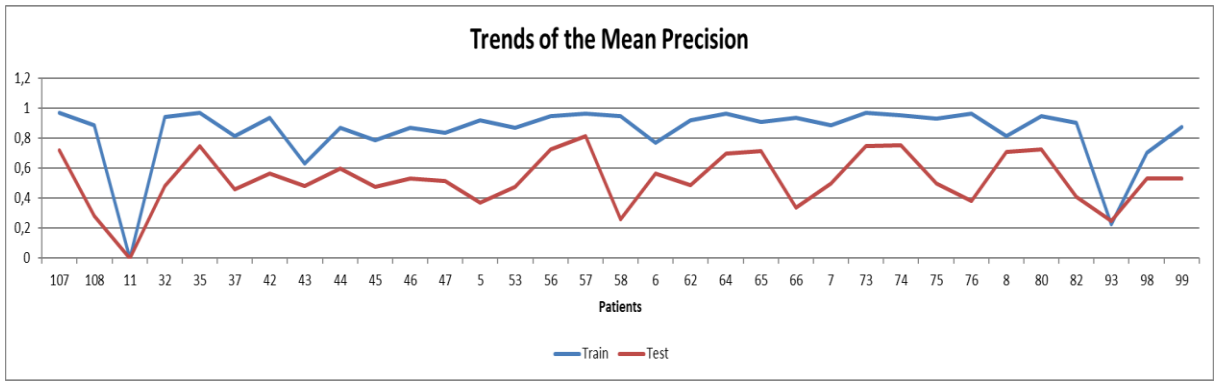


Figure 72: Mean Precision's trend

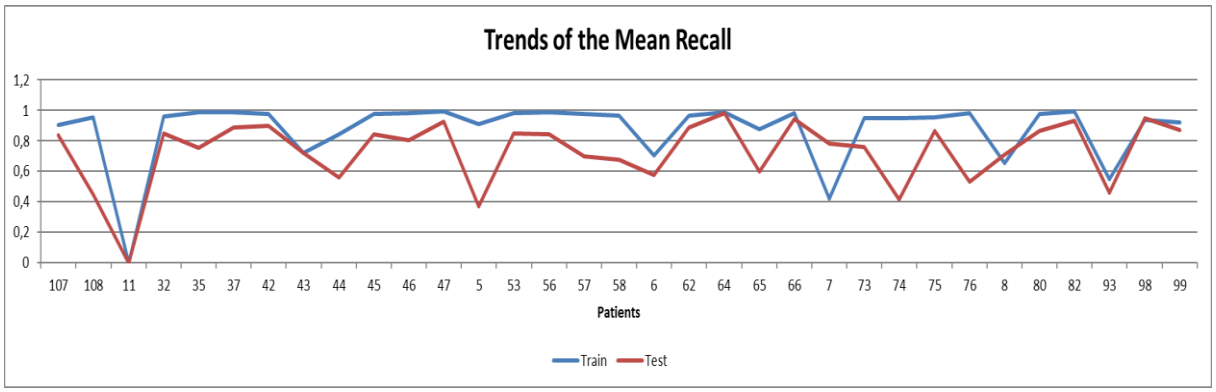


Figure 73: Mean Recall's trend

From the graphs it is possible to notice that trend of all the parameters are very similar considering both the train and test set, except for the patients 11, 5, 58. In fact, the values are visibly different. Overall, it is possible to say that the net is able to identify the tumoral area in almost all the patients, except for the patient 11. It is important to notice that, even if the values of Dice Coefficient, Precision and Recall are higher than the one of the previous system, this net has promising performances.

Here some example of well segmented slices and badly segmented (fig.70, fig.71).

Well segmented patient:

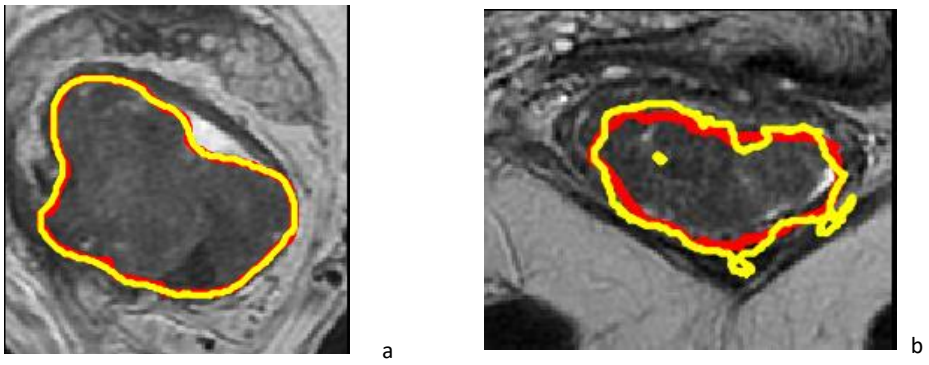


Figure 75: Patient 64 (a) and Patient 99 (b) - manual segmentation (red) and U-Net 1 segmentation (yellow)

Badly segmented patient:

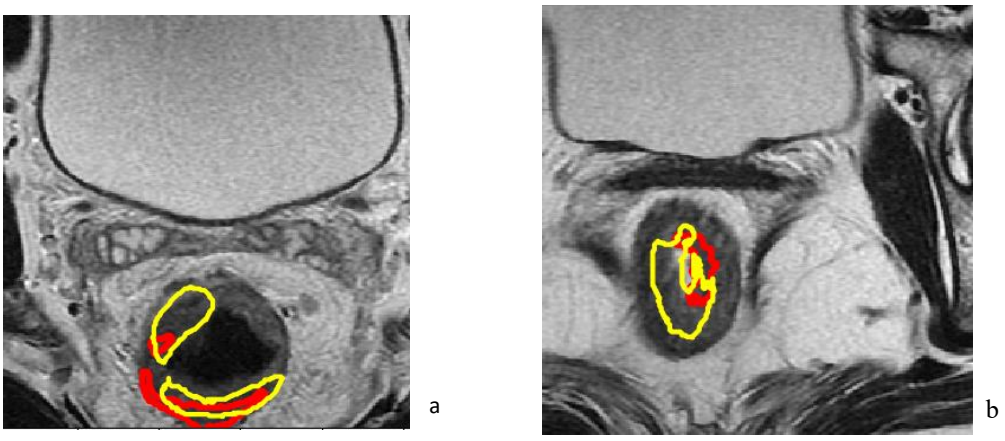


Figure 78: Patient 76 (a) and Patient 93 (b)- manual segmentation (red) and U-Net 1 segmentation (yellow)

Validation

The purpose of the validation is to test the systems' ability to predict new data that are not used in the training set in estimating it. In this case the validation has been done using the *K-fold cross validation*. This method consists on partitioning a sample of data into complementary subsets, performing then the training on one subset, which become the new training set, and the other is used as the testing set. In this case the generated test set consists of all the slices from the different sequences of three different patients, while the training set contains all the slices of the other patients. Since the first two systems have shown poor performances, the validation has been done considering only the last implemented U-Net (3rd U-Net).

Thanks to the following graphs (fig.72, fig.73, fig.74) it is possible to notice how the performances of the net changes due to the lack of information of a specific patient. Also in this case, the evaluated parameters are the Dice Coefficient, Precision and Recall, again considering all the mask and only the tumoral object.

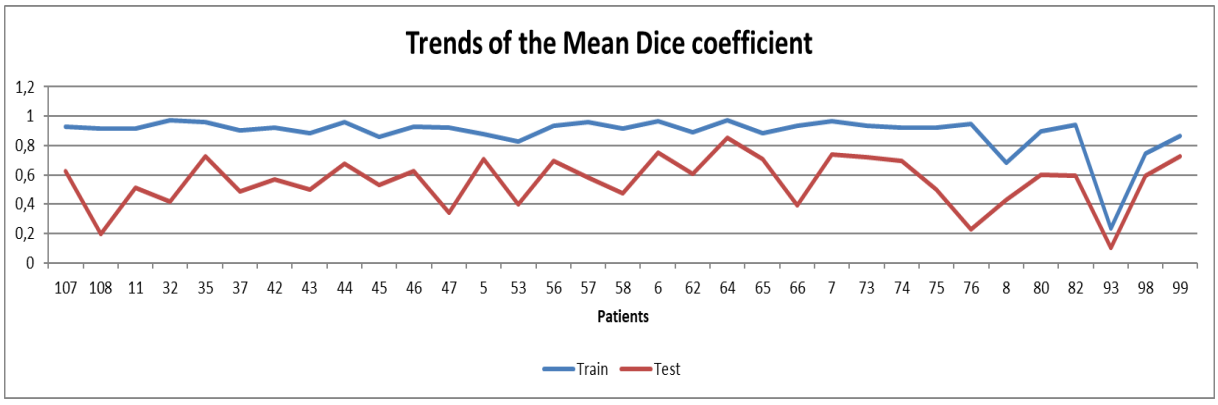


Figure 81: Mean Dice Coefficient's trend for the Validation of the U-Net 3

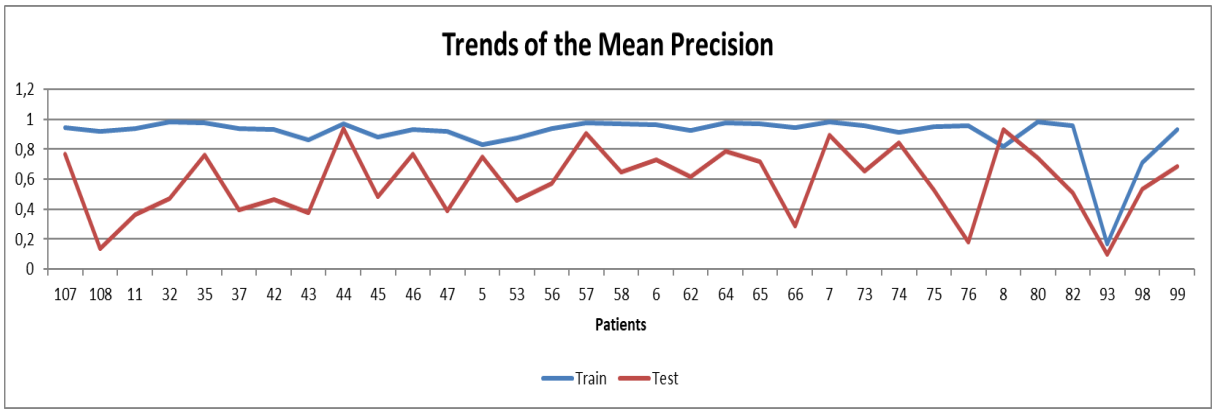


Figure 82 Mean Precision's trend for the Validation of the U-Net 3

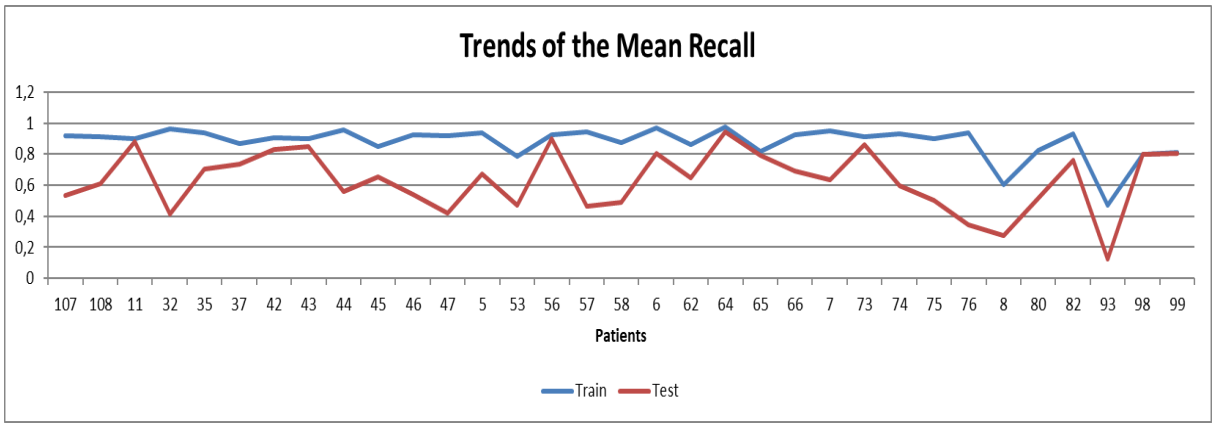


Figure 83: Mean Recall's trend for the Validation of the U-Net

By analysing the graphs, it is possible to notice that the performances are almost the same considering all the mask and only the object, except for the patient 93 which shows relevant differences, due to the False Positive elements.

By analysing the graphs, it is possible to notice again that the trends are almost similar to each other, except for the patient 76, which shows relevant differences related to the Dice Coefficient ad the Precision. This means that the lack of information of this patient considerably affects the performance of the net.

Results comparison

It has been made also a comparison between the results of this study with the ones that are in the literature (the comparison is among the average values):

Table 18: Comparison of the values of Dice Coefficient, Precision and Recall and other parameters between the three U-Nets implemented and the literature.

	DICE COEFFICIENT	PRECISION	RECALL	SPECIFICITY	SENSITIVITY	HAMMOUNDE DISTANCE
IRVING ET AL. (4)	0.65 ±0.15	\	\	\	\	\
JLAN ET AL. (3)	0.84	0.83	0.97	0.88	0.27	8.2
TREBESCHI ET AL. (7)	0.69±0.01	\	\	\	\	\
HUANG ET AL. (8)	0.74±0.15	\	0.75±0.19	\	\	\
SOOMRO ET AL. (12)	0.94	\	\	\	\	\
HUANG ET AL. (10)	0.72±0.14	\	\	\	\	\
1st U-Net	0.63±0.20	0.66±0.23	0.69±0.24	\	\	\
2nd U-Net	0.67±0.21	0.67±0.22	0.73±0.25	\	\	\
3rd U-Net	074 ± 0.17	0.77± 0.21	0.77± 0.17	\	\	\

It is possible to notice that the performances of the implemented systems in this study are reasonably similar to the ones in literature. In particular, the performance of the *U-Net 3* is higher than the one of the Irving et al. (4) study.

It is important to notice that in literature the dataset used is larger than the one used in the study. Moreover, most of the studies have used the whole tumoral volume, while in this study the provided dataset consists of several two-dimensional information.

Comparison between the Convolutional Neural Networks and the U-Net network

From this study it is possible to observe that among all the implemented networks the ones with best performances are the *CNN 3x3 system* and the *U-Net 3* netowork. Despite the higher performances of the latter than the first system, there are some issue that are common. Both of them rely on the pixels intensities which cause several false positive elements, despite the normalization/standardization applied during the pre-processing.

The following images shows the difference between the two systems segmentation.

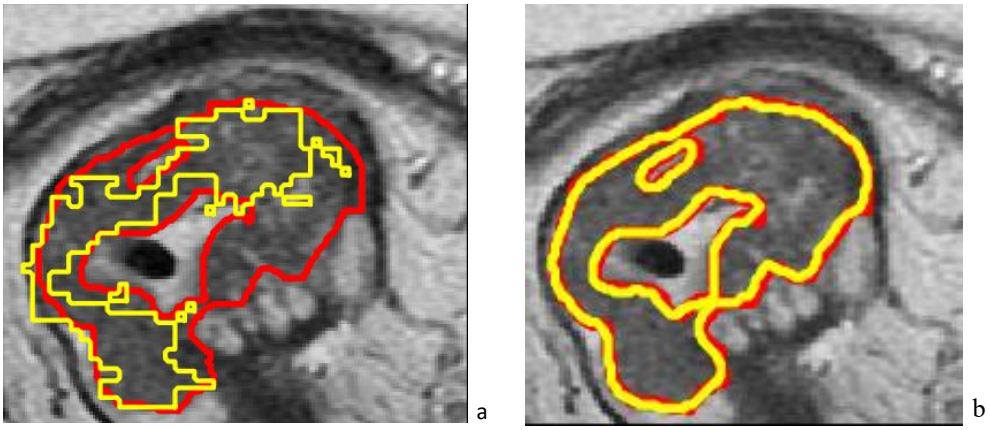


Figure 84: Patient 32 segmented by the CNN 3x3 system (a) and the U-Net 3(b). The red line is the manual segmentation, the yellow one the segmentation of the system

The first example shows a case with a tumor with a shape characterized with holes. It is possible to observe that both the network doesn't segment the holes as tumor, but the U-Net 3 is able to correctly segment all the area, while the CNN 3x3 a smaller area.

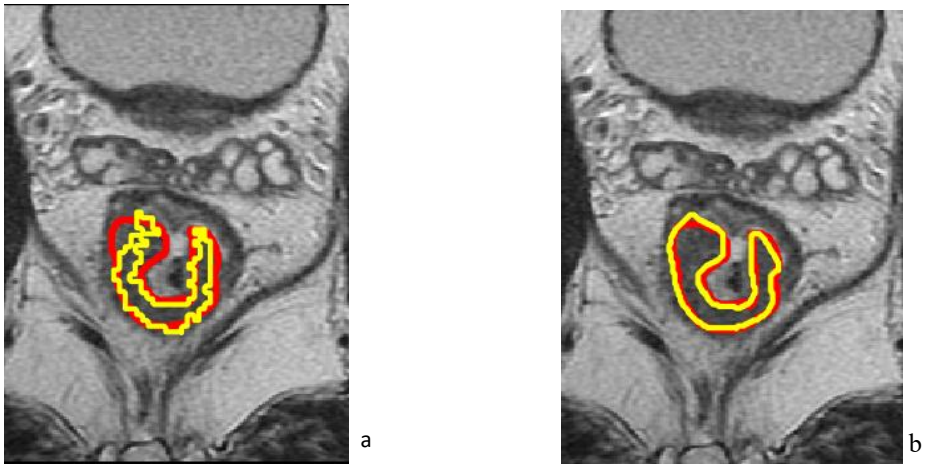
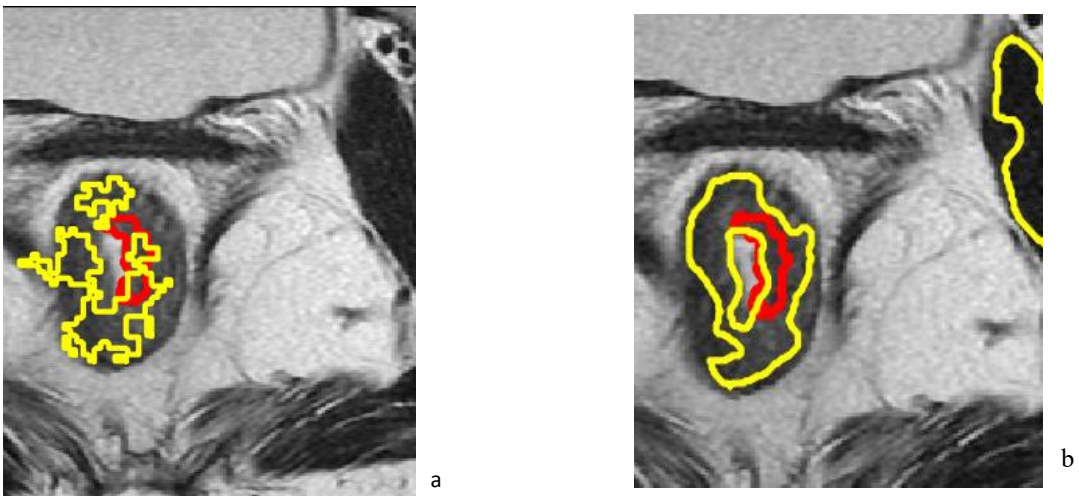


Figure 86: Patient 42 segmented by the CNN 3x3 system (a) and the U-Net 3(b). The red line is the manual segmentation, the yellow one the segmentation of the system

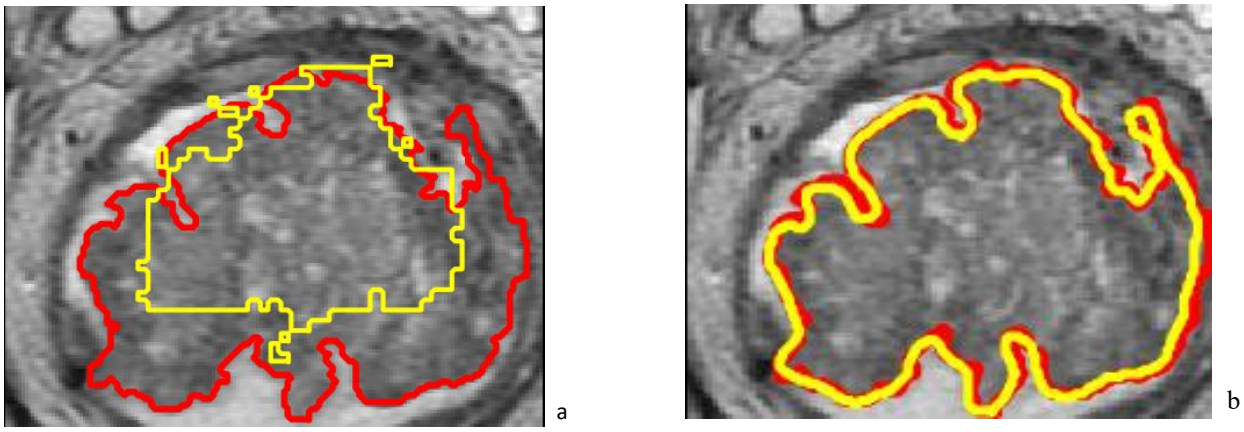
The second example shows another irregular shape of the tumor. Thanks to the homogeneity of the pixels both the networks correctly segment the area of the cancer. These examples shows that the networks rely mostly on the characteristics of the tumor, not its shape.



a b

Figure 87: Patient 93 segmented by the CNN 3x3 system (a) and the U-Net 3(b). The red line is the manual segmentation, the yellow one the segmentation of the system

The third example shows the case which is not correctly segmented by both the networks. The reason is that the area surrounding the tumoral tissue presents almost the same characteristics of the cancer, thus the wrong segmentation. The presence of biological components with very similar pixel intensity to the cancer affects the performances of the neural network, causing wrong segmentation (19).



a b

Figure 88: Patient 7 segmented by the CNN 3x3 system (a) and the U-Net 3(b). The red line is the manual segmentation, the yellow one the segmentation of the system

The fourth case shows the segmentation of the mucinous carcinoma. The CNN 3x3 system correctly system a smaller area included in the manual segmentation, while the U-Net 3 correctly segment all the tumoral tissue.

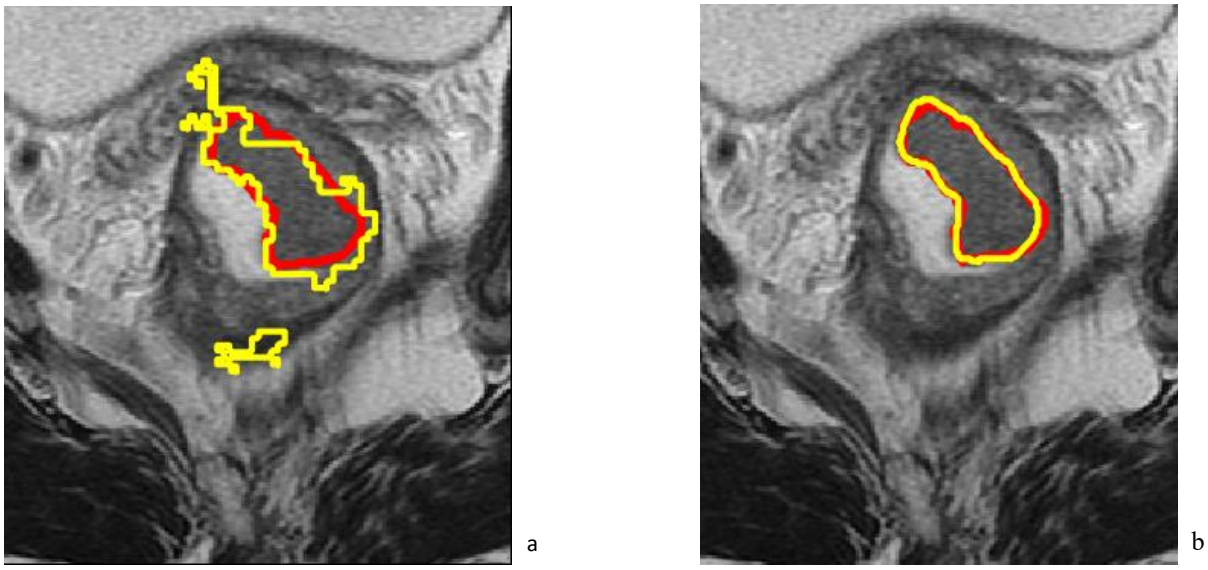


Figure 90: Patient 56 segmented by the CNN 3x3 system (a) and the U-Net 3(b). The red line is the manual segmentation, the yellow one the segmentation of the system

The last case shows a classic adenocarcinoma kind of colorectal cancer. Both the networks segment all the tumoral area. The CNN 3x3 shows different false positive areas. This is due to the fact that the characteristics of these areas are very similar to the one related to the malignant tissue.

Discussions

Nowadays the use of Deep Learning algorithms is increasing, and their applications in the medical field are promising, especially in order to provide a support to the radiologists. The main problems related to the Neural Networks are the extremely computational expensive training phase, the need of large dataset, which are very unbalanced between the positive and negative cases, and the fact that are still not clearly understood.

In this study different models have been implemented and compared, thus to first better understand how each parameter of the network affects its training and its performance, and to identify which architecture better work with the provided dataset, which includes T2w, DWI and ADC sequences.

Thanks to the evaluation of the Dice Coefficient, Precision and Recall, it is possible to confirm that among all the implemented networks, the one with the highest performances is the *U-Net 3*. Its architecture consists of 5 descending levels and has been trained for 60 epochs. The net is able to recognize properly the tumoral tissue, both adenocarcinomas and mucinous carcinomas cases, with a lower number of false positive elements than the CNN 3x3 system, which shows a high number of mis-classified pixels. This is because the U-Net structure allows to classify each pixel, instead of ROIs, and because it is able to collect features related to the position and the characteristics of the object of interest. The main issue regarding the network is related to the fact that in the T2w sequence there are several biological components, such as prostate, anus, muscles, and in general area surrounding the cancer which present very similar pixel intensity to the one belonging to the tumoral tissue. Thus, it is possible in several cases to obtain false positive elements. Overall, the obtained results are promising, considering the low number of training element.

A possible way to improve the implemented network is trying to create a structure being able to mimic the radiologist's tumor detection process, using all the sequences provided during an MRI exam. It could be done by modifying the number and the characteristics of the convolutional layers, changing the functions related to the backpropagation process during the training. Moreover, a larger dataset could improve the performances, thus including in the training phase the informations related to the whole volume of the cancer.

Ringraziamenti

A conclusione di questo lavoro di tesi, è doveroso porre i miei più sentiti ringraziamenti alle persone che ho conosciuto e che mi hanno accompagnato in questo periodo importante della mia vita. È difficile ricordare in poche righe tutte le persone che in modi diversi hanno contribuito a rendere meraviglioso questo momento.

Un ringraziamento sentito va alla prof.ssa Gabriella Balestra, per avermi dato l'opportunità di cimentarmi con strumenti sempre più innovativi. La mia stima per lei è dovuta, oltre che alla sua profonda esperienza e conoscenza, alla sua umanità con la quale ha saputo incoraggiarmi durante i momenti di difficoltà.
Un ringraziamento particolare va alla mia correlatrice Valentina Giannini, al suo entusiasmo, ai suoi consigli, e alla sua fiducia. Le sono profondamente grata per la stima, l'amicizia e per avermi dato l'opportunità di conoscere il suo gruppo di lavoro, incrementando così il mio entusiasmo per la ricerca.

Non possono certo mancare tutti coloro che mi hanno vista crescere in questo periodo. Primi tra tutti mio fratello Nikola, mamma e tata, che oltre a darmi supporto materiale, mi hanno sempre incoraggiato nei vari momenti di sconforto e mi hanno accompagnato nei momenti di soddisfazione. Un infinito grazie per i vostri consigli, le vostre critiche, e per il vostro amore.

Ringrazio mia zia Jovanka per le sue parole dolci, e ringrazio Davide per essermi sempre vicino, e ringrazio Andrea Maria Vittoria per il suo affetto travolgente.
Ringrazio i miei amici storici per il loro tempo e le loro parole di supporto, tutti coloro con cui ho condiviso oltre gli studi momenti indimenticabili, e coloro con i quali condivido la passione del teatro. Non vi cito uno ad uno, siete tantissimi, ma sappiate che siete tutti qui. Ringrazio tutti voi per avermi fatto capire che potevo farcela, e di aver raggiunto questo traguardo.

Hvala djede!

Jovana

References

1. Siegel RL, Miller KD and Jemal A: Cancer statistics, 2019. CA Cancer J Clin 69: 7–34, 2019.
2. Hu Z, Tang J, Wang Z, Zhang K, Zhang L and Sun Q: Deep learning for image-based cancer detection and diagnosis − A survey. Pattern Recognit 83: 134–149, 2018.
3. Jian J, Xiong F, Xia W, *et al.*: Fully convolutional networks (FCNs)-based segmentation method for colorectal tumors on T2-weighted magnetic resonance images. Australas Phys Eng Sci Med 41: 393–401, 2018.
4. Irving B, Cifor A, Papiez BW, Franklin J, Anderson EM, Brady SM and Schnabel JA: Automated colorectal tumour segmentation in DCE-MRI using supervoxel neighbourhood contrast characteristics. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 8673 LNCS: 609–616, 2014.
5. Krizhevsky A, Sutskever I and Hinton GE: ImageNet Classification with Deep Convolutional Neural Networks. ImageNet Classif with Deep Convolutional Neural Networks: 1097–1105, 2012.
6. Yasaka K, Akai H, Kunimatsu A, Kiryu S and Abe O: Deep learning with convolutional neural network in radiology. Jpn J Radiol 36: 257–272, 2018.
7. Aerts HJWL, Lahaye MJ, Parmar C, *et al.*: Deep Learning for Fully-Automated Localization and Segmentation of Rectal Cancer on Multiparametric MR. Sci Rep 7: 1–9, 2017.
8. Huang Y-J, Dou Q, Wang Z-X, *et al.*: 3D RoI-aware U-Net for Accurate and Efficient Colorectal Tumor Segmentation., 2018.
9. Soomro MH, De Cola G, Conforto S, *et al.*: Automatic segmentation of colorectal cancer in 3D MRI by combining deep learning and 3D level-set algorithm-a preliminary study. Middle East Conf Biomed Eng MECBME 2018–March: 198–203, 2018.
10. Huang YJ, Dou Q, Wang ZX, *et al.*: HL-FCN: Hybrid loss guided FCN for colorectal cancer segmentation. Proc - Int Symp Biomed Imaging 2018–April: 195–198, 2018.
11. Birlik B, Obuz F, Elibol FD, *et al.*: Diffusion-weighted MRI and MR- volumetry - in the evaluation of tumor response after preoperative chemoradiotherapy in patients with locally advanced rectal cancer. 33: 201–212, 2015.
12. Le Bihan D, Poupon C, Amadon A and Lethimonnier F: Artifacts and pitfalls in diffusion MRI. J Magn Reson Imaging 24: 478–488, 2006.
13. Yamashita R, Nishio M, Do RKG and Togashi K: Convolutional neural networks: an overview and application in radiology. Insights Imaging 9: 611–629, 2018.
14. Fraggetta F, Garozzo S, Zannoni GF, Pantanowitz L and Rossi ED: Routine Digital Pathology Workflow: The Catania Experience Filippo. J Pathol Inform 8: 129–132, 2017.
15. Harouni A, Karargyris A, Negahdar M, Beymer D and Syeda-Mahmood T: Universal multi-modal deep network for classification and segmentation of medical images. Proc - Int Symp Biomed Imaging 2018–April: 872–876, 2018.

16. Ronneberger O, Fischer P and Brox T: U-net: Convolutional networks for biomedical image segmentation. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 9351: 234–241, 2015.

17. Milletari F, Navab N and Ahmadi SA: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. Proc - 2016 4th Int Conf 3D Vision, 3DV 2016: 565–571, 2016.

18. Zhang M, Li X, Xu M and Li Q: Image Segmentation and Classification for Sickle Cell Disease using Deformable U-Net., 2017.

19. Joshi N, Bond S and Brady M: The segmentation of colorectal MRI images. Med Image Anal 14: 494–509, 2010.