POLITECNICO DI TORINO

Master's Degree in Biomedical Engineering

Politecnico di Torino

Master's Degree Thesis

Longitudinal Radio-Anatomical Modeling and Discovery of Prognostic Factors via Artificial Intelligence: an Ablation Study

Supervisors

Prof. Filippo MOLINARI

Dr. Stefano TREBESCHI

MSc. Laura ESTACIO CERQUIN

Candidate

Valerio PUGLIESE

July 2023

Summary

Response evaluation is a crucial aspect in the field of oncology as it allows clinicians to assess the effectiveness of anti-cancer treatments, make adjustments to management plans, and determine the overall prognosis of patients. The widely adopted quantitative tool for this purpose is the Response Evaluation Criteria In Solid Tumors (RECIST), which classifies therapy response based on one-dimensional diameter measurements of target lesions, categorizing them as partial response, stable disease, or progressive disease. However, RECIST has certain limitations, including inter- and intra- observer variability, as well as reliance on one-dimensional measurements only. These limitations can impact the accuracy of assessments and subsequently affect patient prognoses. Therefore, there is a need for a new method to overcome these drawbacks. Inspired by the classic radiological reporting approach that identifies all changes throughout the entire body, we can formulate the problem as an image-to-image registration task using neural networks. In this framework, anatomical changes between follow-up scans of the same patient are represented as deformation fields, and these deformations are utilized to predict survival, assuming that they hold valuable prognostic information. While this has been proposed in a few pilot studies, yielding significant results, it remains unclear whether the network's ability to model deformation fields is directly correlated with its ability to predict survival. This thesis aims to address this question through an ablation study, wherein different components of the network architecture are removed or modified to introduce variations in registration quality and examine their impact on survival prediction. The study design includes four experiments, plus an additional one, each analyzing different combinations of network components. These include variations in network size, expressed as features number, inclusion of skip layers, realism of reconstruction implemented via Generative Adversarial Networks (GANs), representation via Vision Transformers, and influence of embedding vectors via latent-space similarity. Survival prediction of the resulting models has been applied to an internal dataset consisting of thoraco-abdominal CT scans from patient who underwent immunotherapy between 01/01/2013 and 31/12/2018 at The Netherlands Cancer Institute-Antoni van Leeuwenhoek Hospital (NKI-AVL; Amsterdam, The Netherlands).

Table of Contents

List of Tables

List of Figures

Acronyms

1D 1-Dimensional

3D 3-Dimensional

AI Artificial Intelligence

ANN Artificial Neural Network

BN Batch Normalization

CNN Convolutional Neural Network

CHF Cumulative Hazard Function

CT Computed Tomography

DL Deep Learning

DNA DeoxyriboNucleic Acid

DSC Dice Similarity Coefficient

ECIS European Cancer Information System

FBP Filtered Back-Projection

GAN Generative Adversarial Network

GN Group Normalization

GPU Graphics Processing Unit

HU Hounsfield Unit

IR Image Registration

ML Machine Learning

NSCLC Non-Small-Cell Lung Cancer

NSD Normalized Surface Distance

PAM Prognostic AI Monitoring

RECIST Response Evaluation Criteria in Solid Tumors

ReLU Rectified Linear Unit

RSF Random Survival Forest

SSIM Structural Similarity Index Measure

TCIA The Cancer Imaging Archive

TV Total Variation

WL Window Level

WW Window Width

Chapter 1

General clinical background

1.1 Cancer characteristics

A tumor, or neoplasm, is an abnormal mass of cells capable of growing and dividing uncontrollably despite restriction of space, nutrients shared by other cells, or signals sent from the body to stop reproduction [1]. Tumors can be categorized as benign or malignant tumors according to their ability to migrate to distal sites, or creating metastases. Benign tumors tend to grow slowly and do not generally invade the neighbouring tissues remaining in their primary location [2]. Often, they have clear distinct boundaries. Consequently to their characteristics, they are not particularly harmful to the body, and surgery is commonly used for their removal [3].

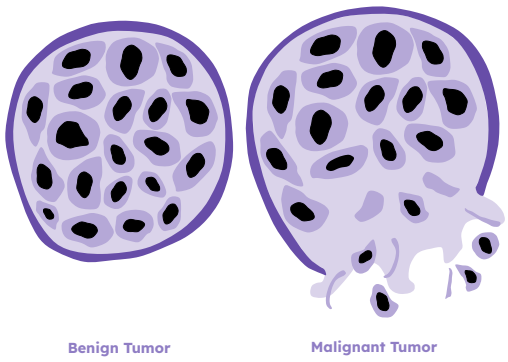


Figure 1.1: Illustration of benign and malignant tumors

On the other hand, a malignant tumor, also called cancer, is able to evade its site of origin, invading adjacent tissues, and migrate to distal sides via the bloodstream or the lymphatic system. Often, malignant tumors grow much faster. This spread

is called *metastatic seeding*, and it can occur anywhere in the body but it is most commonly found in the liver, lungs, brain, and bones [3].

The appearance of cancers can be traced to genetic mutations in proliferation-controlling genes, and the presence of environmental factors favourable to their growth, known as the six hallmarks of cancer [4]. Eleven years after the first publication, in 2011, Hanahan and Weinberg updated it adding two emerging hallmarks, and two enabling characteristics [5]. As a result, it is expected that these can grow in the future to include more hallmarks, as our understanding of cancer biology deepens [6]. The principal hallmarks indicated in [5] can be summarized as:

- *Genome instability* describing the increased tendency of the DNA genes to mutate;
- *Neo-angiogenesis* describing the ability of the tumor to generate a new vascular network for blood supply;
- *Activating invasion and metastasis* describing mechanism by which tumor cells expand into nearby environments;
- *Tumor-promoting inflammation* describing the mechanism by which tumor benefits from inflammatory actions, leading anti-cancer cells to secrete pro-survival, pro-migration, and anti-detection factors;
- *Enabling replicative immortality* describing the ability of the tumor to replicate unlimitedly;
- *Avoiding immune destruction* describing the mechanism by which tumors hijack immune system detection and destruction actions, as the immune checkpoint control;
- *Evading growth suppressors* describing the ability of the tumor to resist inhibitory signals that might stop their growth;
- *Sustaining proliferative signaling* describing ability of the tumor to start the proliferative cascade even without a growth-factor signal from neighboring cells;
- *Deregulating cellular energetics* describing ability of the tumor to reprogram cellular metabolism;
- *Resisting cell death* describing the ability of the tumor to evade apoptosis by altering the proper signaling.

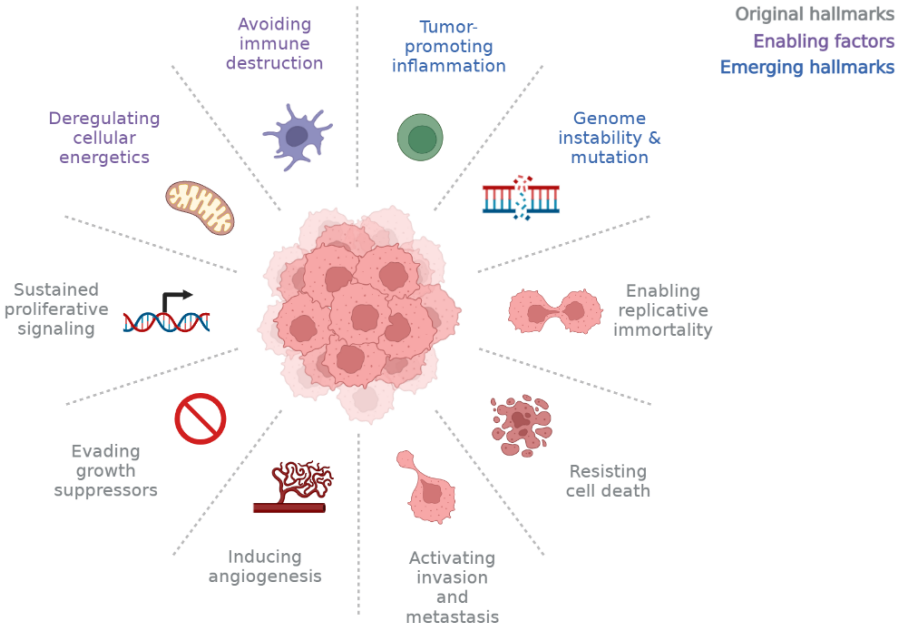


Figure 1.2: Illustration of cancer hallmarks. Image created with Biorender.com and adapted from [5]

1.2 Cancer epidemiology

Cancer remains among the leading causes of death worldwide and it is an important barrier to increasing life expectancy [6]. The estimated number of new cancer cases in 2020 in the European Union (+EFTA) was 2.76 million, with an estimated 1.29 million mortality. Moreover, according to the European Cancer Information System (ECIS) the incidence is expected to increase by 12% and 21% in 2030 and 2040 respectively [7].

Among all types of cancer, lung cancer is recognized as the fourth most frequently occurring cancer in European countries after breast, colon-rectum and prostate cancer. In 2020, 326 thousand of all newly diagnosed cancer were lung cancer, and its incidence is projected to increase by 23% in 2040. More importantly, it accounts approximately for the 23% and the 15% of all cancer mortality respectively in male and female patients; it represents the most lethal type of cancer disease, with an expected increase of almost 27% of mortality rate during the next 20 years [7]. According to [7], the incidence of kidney, liver, pancreas and stomach cancer is between 3.5 and 5 times less that the already mentioned lung cancer, and their expected rates in 2040 will increase by 20 to 30%.

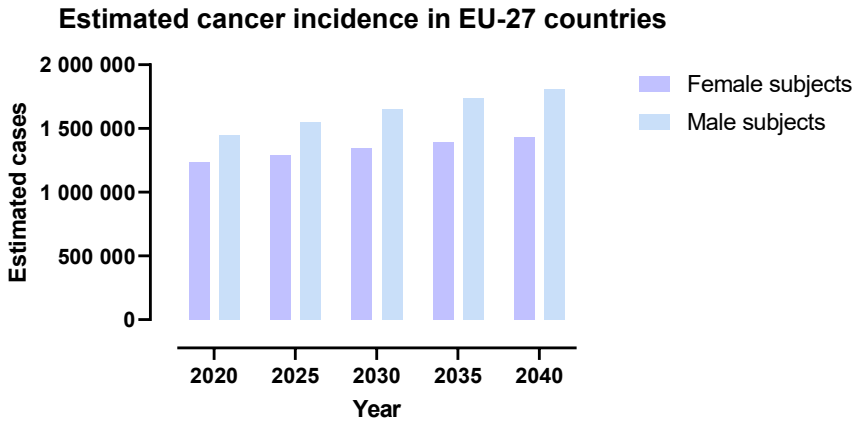


Figure 1.3: Overview of the expected incidence of cancer by year in EU-27 countries. Image adapted from [7]

1.3 Cancer treatment methods

Cancer treatment is the process of using surgery, radiotherapy, medical drugs, or other novel treatment options to cure, shrink, or stop the progression of cancer [6]. Prescribing the appropriate treatment option by the treating physician depends on several factors, such as the type and stage of the cancer, its location, its size and extent, and the presence of any comorbidity, as well as the wish of the patient [8]. Treatments can be broadly divided into two sub-categories: local and systemic therapies.

1.3.1 Local therapy

The most widespread techniques are surgery and radiotherapy. Surgery entails the removal of the tumor, as well as, when feasible, a part of the surrounding tissue, called *margin*, which should be clear of cancer cells to reduce the chances of recurrence [9].
Radiation therapy, or radiotherapy, on the other hand, leverages the interaction between high-energy photons or particles (as protons, neutrons or electrons) and the target. Using either photons or charged particles, the purpose is to damage the DeoxyriboNucleic Acid (DNA) of cancer cells leading to cellular death, i.e apoptosis [10].

1.3.2 Systemic therapy

Contrary to local therapy, a systemic treatment does not focus on a single specific target area but aims to eliminate cancerous cells affecting the whole body, therefore trying to counteract any active process of seeding of the tumor, as well as distal metastases in the same way as the primary tumor [11]. There are various types of systemic treatment available.
Chemotherapy is the most common treatment, based on the administration of drugs that can interfere with the mechanisms of cell proliferation. The majority of anti-neoplastic drugs used in this kind of treatment act specifically in processes such as DNA synthesis, or block the synthesis of DNA precursors or damage the integrity of DNA and prevent its transcription [12]. Chemotherapy acts on cells with a high rate of proliferation, a key feature of cancer cells. It however does not have the ability to distinguish between healthy and cancerous cells. As a result, healthy cells with high replication rates, such as hair bulbs, mucous membranes and bone marrow are also affected. The main side-effects of this therapy, in fact, are hair loss, anemia and digestive disorders such as nausea and vomiting [13].

Another branch of systemic treatments are the targeted therapies, which act selectively on molecular pathways of proliferation [14]. This is the chain of proteins in the cells that controls when and if the cells should replicate. Gene mutations cause change in the protein, which disrupts the signaling pathway, and pushes the cell to uncontrolled proliferation [15]. Targeted therapies are molecules developed to interrupt this pathway, stopping the cell from replicating [16]. In a similar fashion, other signaling pathways can be targeted other than the proliferation pathway, as the neo-angiogenesis signaling pathway [17].

Immunotherapy is a type of treatment that uses the patient's own immune system to fight the spread of cancer, preventing the tumor from inhibiting anti-cancer activity carried out from the immune system [18]. There are different types of immunotherapy mechanisms available [19]. For example, by increasing the ability of immune cells to present tumor antigens to the immune system [20]; blocking the pathway that tumor cells used to disable immune cells [20]; promoting the proliferation of certain immune cells, like T-helper and B cells [21]; or increase the cytotoxic activity of effector cells such as natural killer cells [20]. Several types of immunotherapy are already used in clinical practice for different cancers as melanoma, Non-Small-Cell Lung Cancer (NSCLC) and bladder cancer, commonly treated with antibodies for inhibitory immune checkpoints CTLA-4 and PD-1 [22].

1.4 Treatment monitoring

After the diagnosis, and the treatment start, the patient is monitored, in a process commonly referred as *follow-up*: it aims to observe and evaluate structural, functional, physiological and biochemical changes in the disease over time. In this way it is possible to understand whether the current treatment is giving the expected results, and consequently to take the appropriate countermeasures promptly [23]. Medical imaging plays an integral part during monitoring, diagnosis, and treatment planning [24]. The increasing relevance of medical imaging is due to the possibility of providing a more comprehensive view of the body, and therefore of the entire tumor burden, in a minimally-invasive fashion.
There are different imaging modalities, which differ mainly for the physical principles underlying the creation of the image, and for the quantification of anatomical structures or physio-functional maps [25]. A description of Computed Tomography technology is given in the following subsection, being the type of oncological image used in the proposed work.

1.4.1 Computed Tomography

Computed Tomography (CT) is a X-ray-based imaging technique which uses a motorized X-ray source that rotates around the circular opening of a donut-shaped structure called a gantry [25].

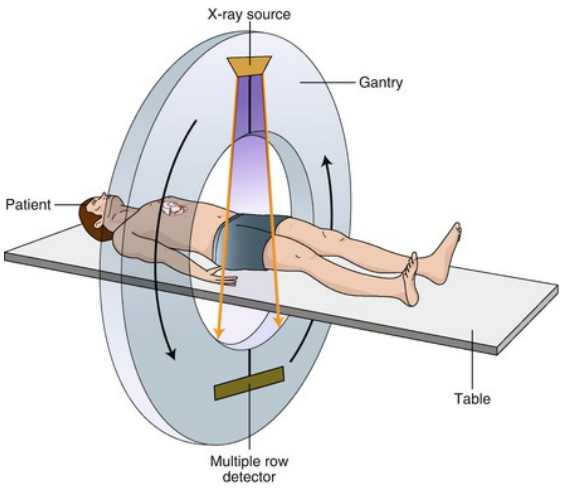


Figure 1.4: Example of a CT scanner. Image retrieved from [26]

Image Acquisition

During image acquisition, the patient lies on the bed, and slowly moves through the gantry. In the gantry the emitter (X-ray tube) revolves around the patient, emitting a collimated beam of rays, that passes through the patient's body and is collected by digital detectors. To date, most CT scanners have an array of detectors that can cover the entire ring, so as to always ensure the presence of sensors contralateral to the X-ray tube [27]. The movement of the emitter is continuous and is ensured by the *slip ring* technology: the current and voltage supply is given by the circular track on which the tube rotates, whereas the contact friction and the electromagnetic disturbances are minimized. Once hit by the incident rays, the sensors produce a signal proportional to the photons' intensity [25]. The denser the material in the middle, the less photons make it through to the detector. In particular, the attenuation follows Lambert-Beer law which relates the variation in the number of X-photons, after hitting a material, with a specific linear attenuation coefficient:

$$N = N_0 \cdot e^{-\mu x} \tag{1.1}$$

where N represents the number of photons emitted by the source, N_0 is the number of photons after passing through the material, x is the width of the material and, finally, μ represent the attenuation coefficient.
In Equation 1.1 μ is assumed constant along the line that connects source and detector, but in reality it is variable and the exponent of Lambert-Beer law is equal to the following integral:

$$\int_{scan-line} \mu(x, y) \cdot ds \tag{1.2}$$

Therefore, the final image will be dependent on the attenuation coefficient, or radio-density, of the points belonging to the anatomical areas scanned.
As can be seen from the above equation, with a single scan it is not possible to trace the radio-densities of the individual pixels; for this reason angular sampling is used, by radiating the region of interest from different angles.
Next, Filtered Back-Propagation (FBP) algorithms are used to assess the μ function, and therefore to know the attenuation coefficient of individual points of interest. These algorithms rely on the calculation of the Radon transform (which is equal to Equation 1.2), on its inversion and on the application of convolutional filters to minimize noise in the final reconstruction [28].

As mentioned above, the patient-support moves through the gantry to acquire the volume of interest, so this process is not limited to the reconstruction of a single slice but extends to a 3-Dimensional (3D) space. Typically 30 to 40 slices of the patient are acquired during an exam, of thickness of 0.5 mm. Obviously, for really big volumes (as in the Full-body CT scan) the thickness of the single slice can be

higher or it's possible to increase the gap between adjacent slices, also called *pitch factor*; these strategies aim to reduce the radiation dose to which the patient is subjected.
Finally, the radio-densities are normalized and expressed in Hounsfield Units (HU):

$$\mu(HU) = 1000 \cdot \frac{\mu - \mu_{H_20}}{\mu_{H_20}} \tag{1.3}$$

where μ_{H_20} represents the linear attenuation coefficient of water, and μ the one of a generic biological tissue. This scale is adimensional, starts from the negative value of -1000, representative of the air, and it is unbounded at the top.

Image interpretation & read-out

Once the CT scan is acquired, the radiologist extracts semantic information from it, e.g. whether there is a tumor, where it is located, how big it is, what are the risk for the patient if the tumor were to grow, etc. This is done qualitatively, by scrolling through the scan and assessing the structures seen. In this case, the radiologist will have to adjust the visualization settings, as the windowing: it is an essential operation to highlight particular structures and eliminate the not relevant biological tissues.

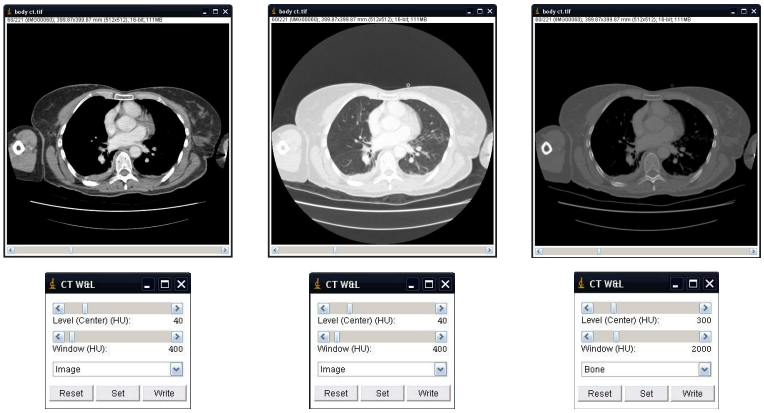


Figure 1.5: Application of different windows to a chest CT. Image retrieved from [29]

Generally the screen visualization is limited by the available grey-levels (typically 256, if 8 bits are assigned to each pixel). Windowing, on the other hand, helps the read-out by selecting the range of interest in HU and redistributing the grey tones only in the chosen window. The key parameters for the visualization are the Window Level (WL) and the Window Width (WW): the first is the mean value of the range, and the second is the number of HU contained in the same range.

1.4.2 Clinical decision making

As mentioned in the beginning of this section, medical imaging plays a key role in the evaluation of anti-cancer therapies and, therefore, in clinical decision making. In fact, the response evaluation criteria, used by radiologists in the follow-up phase to quantify the response to treatment, rely on repeated acquisitions over time, such as those of CT or MRI. In particular, the most used and recognized criteria are the Response Evaluation Criteria in Solid Tumors (RECIST), developed in 2000 [30], and the following updates and variants published over time, such as RECIST 1.1 and iRECIST (specifically adapted for immunotherapy).

Response Evaluation Criteria in Solid Tumors

RECIST uses different image acquisitions overtime and 1-Dimensional (1D) measurements to estimate the total tumor burden and to categorize the response to treatment.
As first step, the operator has to distinguish between measurable and non-measurable lesions. Measurable lesions are lesions with a maximal diameter of at least 10 mm and should allow reproducible and repeated measurements, or lymph nodes if their maximum short axis diameter exceeds 15 mm. Non-measurable lesions are defined as the smallest ones or as lymph nodes with a short axis diameter between 10 and 15 mm [31]. Then, they have to distinguish between target and non-target lesions on the axial plane. Target lesions should be chosen based on their size, representing all affected organs, and should also be suitable for consistent and repeatable measurements [32]. A maximum of 5 lesions in total, and of maximum 2 per organ can be selected as target. Then, the Sum of the longest Diameter (SoD) for the target lesions is calculated and reported. SoD functions in this case as estimate of the total tumor burden.
The measurement is repeated for the *baseline* scan, which should be performed as close as possible to the treatment start and not more than 4 weeks before, and for each *follow-up* scan, usually acquired 4 weeks after the previous one [32]. To assess the response to treatment the SoD over time is observed:

- if it is increased by at least 20%, or new lesions are detected, it is progressive disease (PD);
- if it is decreased by at least 30% and no new lesions are detected, it is partial or complete response to treatment (PR);
- in neither, the disease is stable (SD).

Figure 1.6 shows a possible application of RECIST criteria: the areas in light-blue represent the target lesions while those in violet the non-target ones. Some lesions

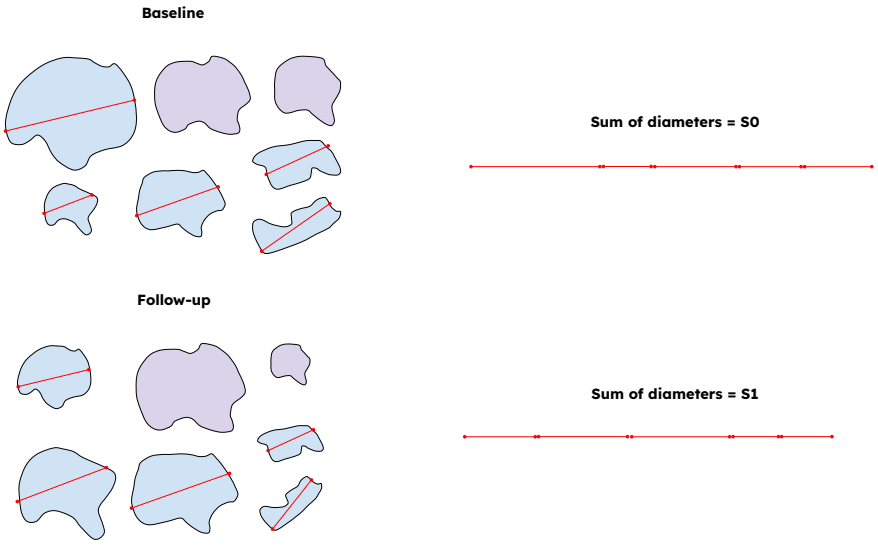


Figure 1.6: Application of RECIST criteria

increased their size between baseline and follow-up, while for others the maximum diameter decreased. In this example there are no new lesions and the sum of the diameters ($S1$) is smaller than the reference ($S0$), so it can be either a stable or in-response disease.

Limitations

Since target lesions are selected and uni-dimensional longest diameters are measured manually, discrepancies within multiple readings or between different individuals can cause inconsistency in response categorisation [33]. In fact, finding the maximum dimension of a tumor can be difficult, especially in irregularly shaped lesions, resulting in different measurements of the same lesion; different operators may end up choosing different target lesions, which might respond different to treatment, resulting in different outcome classes - the same patient classified as PD and PR depending on the operator who is doing the readout [33].
In addition, RECIST criteria are intrinsically constrained by relying on uni-dimensional measures relating exclusively to the lesions selected. In this way, many (potentially prognostic) imaging features are ignored. For example, in the case of a disease evolving in multiple distal sites, tumors with different locations may be characterized by a different microenvironment. In turn, different microenvironments may affect differently the response to treatment. RECIST chooses the target lesions regardless of their location, and by studying only the tumor diameter cannot take into account this additional information. On the contrary, a

method that does not choose a priori the lesions to be monitored, but quantifies all morphological changes over time, could also take into account factors external to the tumor, as angiogenesis or lymphocytic infiltrations [34].
Finally, selecting and measuring all the target lesions result in a time consuming task being completely manual.

Chapter 2

General technical background

2.1 Artificial Intelligence

Marco Somalvico, one of the pioneers of Artificial Intelligence in Italy, defines it as: "the discipline belonging to computer science that studies the theoretical foundations, the methodologies and techniques that allow the design of hardware and software systems capable of providing the computer with performance that, to a common viewer, would seem to be of exclusive pertinence to human intelligence" [35]. By this definition, the first attempts tried to emulate the processes of human reasoning for solving certain tasks by the computer. The attempt to schematize human decision-making in a series, albeit complex, of logical conditions and mathematical operations proved to be unsuccessful especially in complex and open contexts. Therefore, this discipline begins to impose itself and develop when the underlying paradigm changes: the goal is not to decipher the human mental process and make it available to the machine, but that the machine itself can develop its own decision-making process.
This concept is the basis of Machine Learning (ML), which was born using a large number of examples (consisting of both input data and expected responses) to generate its own rules of learning, then applicable to new data to produce new responses [36]. Going down through the hierarchy showed in Figure 2.1 we have Deep Learning (DL), a subfield of ML, based on Artificial Neural Networks.

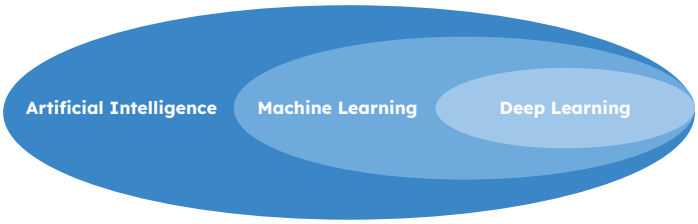


Figure 2.1: Relation between Artificial Intelligence, Machine Learning and Deep Learning

2.2 Machine Learning

Machine Learning is a sub-field of AI and comprises all the techniques, based on statistical learning, which aims to learn a function/task from input data, without being programmed to do so [37]. Although learning is done automatically, solely based on the patterns that the data contains, human intervention plays a crucial role. In fact, in traditional ML methods the input to the model is structured data: the developer has to "clean up" the data and provide only the information deemed necessary for the task.

Once the data is structured, the algorithm learns to build a function that minimises the error (loss function) between output and expected results. This particular type of learning, where the outcome variable is known, is called supervised. If the outcome variable is not known, the model learns a representation of the data instead, in a procedure called unsupervised learning [38].
The most common architectures are the Artificial Neural Network (ANN), decision tree, genetic algorithm, the Bayesian Network and the support-vector machine.

2.3 Deep Learning

Deep Learning has emerged as a prominent field within Machine Learning, Artificial Intelligent, data science, and analytics, due to its remarkable ability to learn from available data and extract valuable insights [39].
The first major innovation underlying DL is the automation of manual operations: while input data in ML is carefully studied and structured before being provided to the model, DL algorithms receive raw data (such as images or text) and independently recognize the features and patterns useful to the task. In addition, DL algorithms outperform traditional ML on large volumes of data [39].

Similar to other Machine Learning models, Deep Learning networks are based on ANNs: they are a family of computational models that mimic the behavior of biological neurons during the learning phase. A biological neuron can communicate with the others of the network through its terminations, called *dentrities* and *axon*: the first are tiny fibers where the electrical signals come from the surrounding space, the second is that part of the neuron that carries the output signal away. Between dendrites and axon, the *nucleus* receives the current, elaborates the output and allows the propagation of the signal (i.e of the information) outwards.

The fundamental artificial neuron used in the ANNs, also called *perceptron*, recreates this structure. The electrical signal is here replaced by the input value, which is connected with the perceptron through a weight, which measures its significance. Once all inputs are been multiplied by their weights, a weighted sum is operated to obtain an activation-value; according to its activation function and to the activation-value, the perceptron can either output a signal or stay silent.

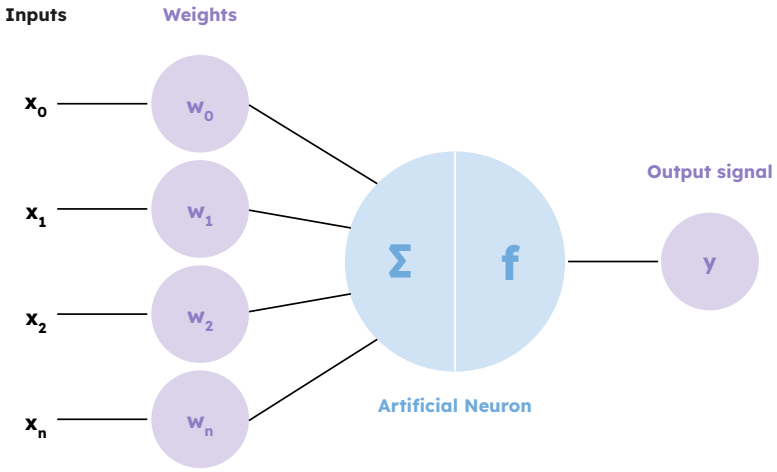


Figure 2.2: Schematic overview of a single artificial neuron

Since the inputs and the activation function are determined before the training phase, the only parameters that can actively change to obtain a change in the output signal are the weights. Hence, neural networks are adaptive systems we modify their structure during training to minimize the error between output and desired value. A single neuron is not suitable to solve complex tasks, so the common ANNs are composed of multiple neurons, divided into several layers.
Each neuron in the *input layer* receives input data from outside and sends the

output to each neuron in the next layer. Then there are the *hidden layers*, the number of which varies from case to case. Finally, the *output layer* contains a number of neurons dependent on the specific problem being addressed.
For example, in a classification problem with two classes, a single output neuron is enough: if the input belongs to the first class (codified with the value 0), the output value will be 0, otherwise it will be 1.

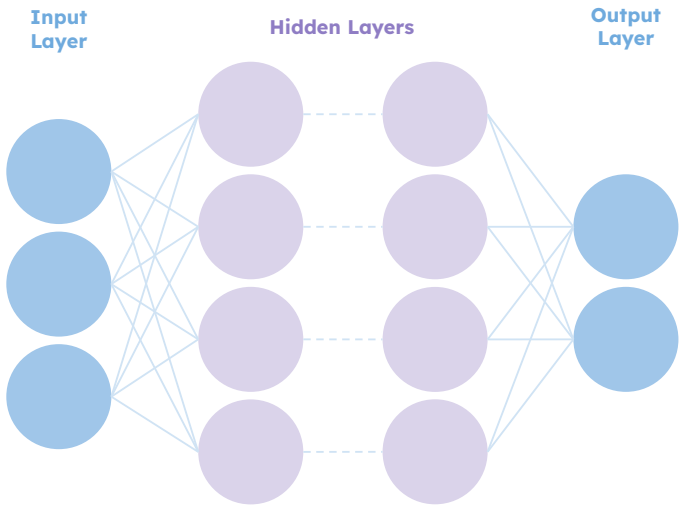


Figure 2.3: Schematic overview of a generic Artificial Neural Network

Another difference between the networks of ML and DL lies in the number of hidden layers: the greater this number, the greater the depth of the model, the greater the describable patterns present in the data. Hence the name Deep Learning.

2.4 Convolutional Neural Networks

Convolutional Neural Network (CNN) is another class of Deep Learning model, which is designed to deal with input on a regular grid, such as an image, where different objects are present, and processed focusing on each one of them individually, a concept inspired by animal visual cortex [40]. A digital image is nothing more than a matrix of pixels that indicate the color intensity of the various points of the image. The value of each pixel, therefore, represents an input data.
With this type of data, the use of neural networks described above (i.e ANN) is sub-optimal. In fact, having to associate for each pixel a neuron in the input

layer, it would end up in a excessively high number of nodes. As a result, the number of connecting weights between the input layer and the hidden layer would also increase dramatically. Furthermore, considering each pixel as a separate and independent value leads to the loss of the spatial hierarchies of features contained in the data. It is precisely the relationship between adjacent points that allows a correct analysis, interpretation and extraction of patterns in an image. These limitations are overcome by CNNs thanks to their characteristic convolutional layers: they use the linear operation of convolution to extract relevant features in the image.

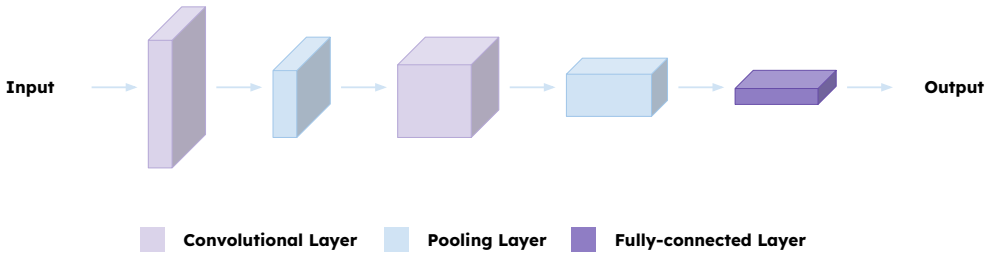


Figure 2.4: General architecture of a Convolutional Neural Network

2.4.1 Layers

Similar to ANNs, CNNs are also divided into input, hidden and output layers. Here, the hidden layer is replaced by a block generally composed of 4 levels: a convolution layer, an activation layer, a pooling layer and a normalization [41]. Commonly, in classification tasks, the network ends with a fully connected layer, which receives the features extracted in the previous steps and maps them to final outputs [40].

Convolutional layer

The convolutional layer is the core element of CNNs and aims to extract relevant patterns in the image. To this end, a small array of numbers, called a kernel, usually small in spatial dimensionality, spreads along the entirety of the depth of the input [41]. In particular, the input is element-wise multiplied with the kernel and then summed, and its result is put at the corresponding position in the output tensor.

This process is visualized in Figure 2.5 for a better understanding: in the figure it is shown the application of the kernel only on the first upper-left set of pixels, but the procedure is repeated until every pixel is considered.

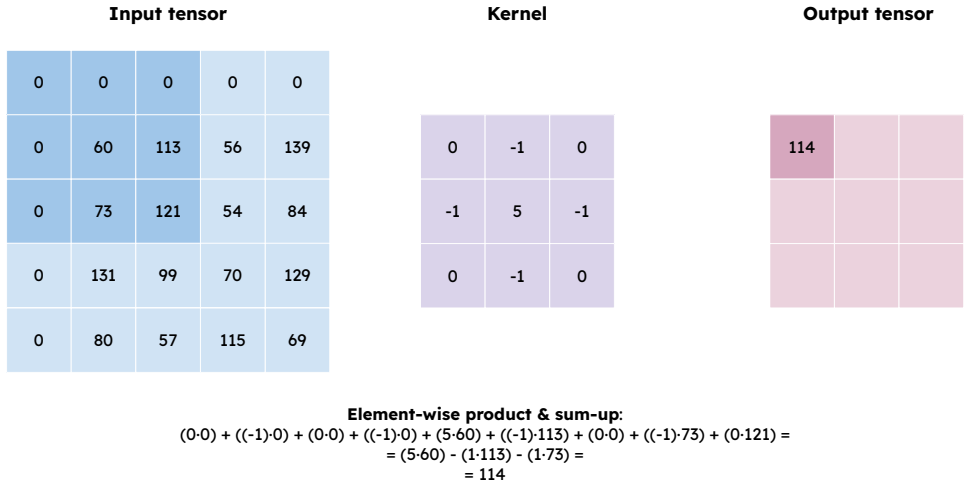


Figure 2.5: Visualization of the convolution operation between the input and the kernel

The most important hyperparameters, which actually define the size of the output, are the kernel size, the stride, the number of kernels and the zero-padding. The first is typically $3x3$. The number of kernels determines the depth of the resulting feature maps. The stride is the distance between two successive kernel positions; its most common value is 1, even if it can be larger to achieve downsampling of the feature maps. Finally, the zero-padding is a technique that addresses the loss of information at the edges of the matrix occurring when a kernel is applied: rows and columns of zeros are added on each side of the input tensor, so as to fit the center of the kernel on the outermost element and keep the same in-plane dimension [40].

Activation Layer

The activation layer typically follows the convolutional layer and it is essential to introduce a non-linear component in the model and, therefore, to aid the network to solve complex tasks. The most common activation function is Rectified Linear Unit (ReLU): it sets the negative values to 0 and leaves the positive values unchanged following the Equations 2.1 and 2.2.

$$ReLU(x) = max(0, x) \tag{2.1}$$

$$\frac{d}{dx}(x) = \{1 \quad if \quad x > 0, \quad 0 \quad otherwise\} \tag{2.2}$$

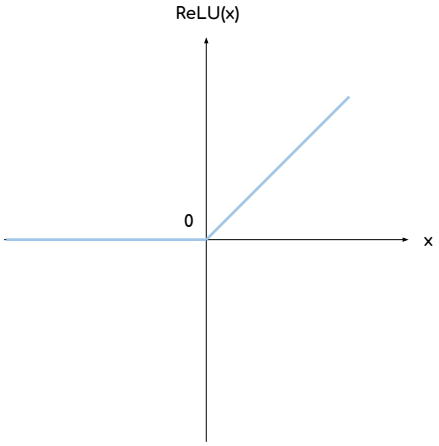


Figure 2.6: Visualization of the ReLU function

Pooling Layer

The pooling layer provides a downsampling operation which reduces the in-plane dimensionality of the feature maps, unchanging their depth. Its goal is to cut spatial information to force it to be semantic, by decreasing the number of subsequent learnable parameters and by keeping only the essential features extracted. It uses a filter that spreads along the input matrix, but does not contain any learnable parameters. The result of the process depends on the filter size, the stride and the padding.

The dimensionality reduction is commonly done either through the max-pooling filter or the global-average-pooling filter. A max-pooling layer, with a typical filter size of $2x2$ and a stride of 2, divides the input in patches and outputs the maximum value in each patch [40]. A global-average-pooling layer operates an extreme reduction, producing a $1x1$ array just taking the average of all the elements in each feature map.

Fully connected Layer

The fully connected layer, also called a dense layer, is typically used at the end of CNNs to map the output of previous steps to the final outputs of the model. Before applying it, the last feature maps of convolutional or pooling layers have to

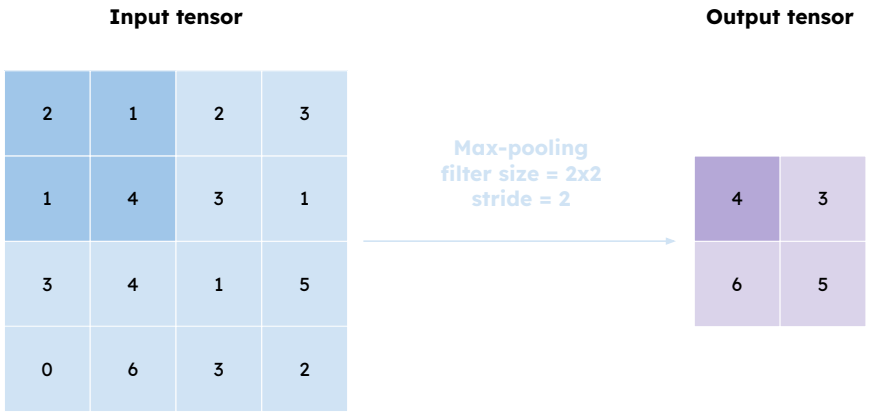


Figure 2.7: Visualization of max-pooling layer

be flattened. It comprises a series of layers of artificial neurons, where each unit is connected to all the neurons in the previous and next layer. So given all the connections, this layer is computationally very expensive [42]. One of its most used versions is the Multi-Layer Perceptron.

Normalization Layer

Normalization is a processing technique used to standardize data, in order to make the network unbiased to outliers, to speed up the learning [43] and to help the algorithm convergence.
The most widespread method is the Batch Normalization (BN), which normalizes features by mean and variance of a batch, and performs better with large batch sizes. On the other hand, Group Normalization (GN) works within groups of channels. GN's computation is independent of batch sizes, and its accuracy is stable in a wide range of batch sizes [44]. In Figure 2.8 N is the batch axis, C is the channel axis and H, W are the height and width of the tensor.

2.4.2 Training

The training of a CNN, or in general of a Neural Network, is an optimization-problem based process: given a set of inputs and the desired outputs, the goal of the model is to minimize the difference between the real and the predicted outcome.

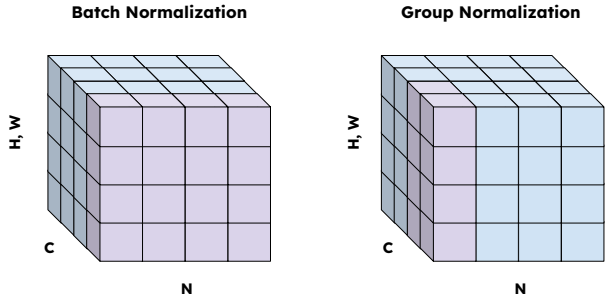


Figure 2.8: Difference between Batch and Group Normalization. Image adapted from [44]

To measure this difference a loss function is used, and its choice is strictly dependent on the specific task. For example, cross-entropy loss is suggested for multi-class classification problems, while mean squared error loss is the gold-standard for regression tasks [40]. As mentioned in 2.3, training is an iterative process where the weights are actively adjusted, at each iteration, to reach a minimum of the loss function. It is commonly characterized by two main steps: feed-forwarding and back-propagation.
Feed-forwarding is based on the calculation of neurons activation with the current value of weights and biases, and it proceeds from the input layer to the end of the model, outputting the predicted outcome.
Back-propagation proceeds from the output backwards, and minimises the loss of the prediction following the gradient (i.e derivative) of the loss function with reference to the weights of the network. This algorithm is known as gradient descent.

2.4.3 Applications in medical imaging

Recently, CNNs are being widely used by the medical imaging research community because of their outstanding performance in medical image analysis, and the advent of Graphics Processing Unit (GPU) [45].
The use of automatic systems in clinical practice has several advantages, such as the consistency of the result in the presence of the same input data, independence from human factors such as fatigue, and the possibility of being trained with a wide amount of images. For these reasons, their application can benefit the different stages of oncology practice, i.e diagnosis, follow-up and prognosis. To date there are artificial intelligence systems that, despite their limitations, have performance

comparable to that of a specialist [46].
The commonly addressed tasks by CNN models in medical imaging can roughly be divided into accomplishing the four main tasks – classification, detection, segmentation and registration [45]. We are going to address image registration in detail, for the purpose of this thesis.

2.5 Image Registration

Various clinical applications involving disease diagnosis and monitoring, image-guided treatment delivery, and post-operative assessment, utilise Image Registration (IR) [47]. It is the process of identifying a spatial transformation that maps two (or more) images to a common co-ordinate frame or, in other words, a voxel-wise correspondence. Considering two input images, I_1 and I_2, the goal of image registration is to find a displacement field f_{12} such that:

$$I_1(x) \approx I_2(x + f_{12}(x)) \tag{2.3}$$

The field f_{12} defines a function where each voxel in I_1 is in I_2. I_1 and I_2 are respectively called *fixed* and *moving* image, suggesting that the latter is transformed during the process to minimize the differences with the former.
Defining $warp(I_2, f)(x) = I_2(x + f(x))$ as the moving image warped according to f, the goal can be rephrased as finding f maximizing the similarity between I_1 and $warp(I_2, f)$ [48].

Over the last decades many different techniques have been proposed, as B-splines and radial basis functions [49], commonly described by a set of parameters and iteratively updating it following a transformation quality metric. These methods, also called *traditional*, are computationally heavy and take a long time to produce the result. Deep Learning models can also be used for registration, and they can be broadly divided into supervised and unsupervised networks. The first category requires ground-truth fields, and their quality and availability directly affect the result, being dense and ambiguous quantities that are almost impossible to be labeled manually [48]. In general, target values are obtained by either estimating them using traditional registration methods or using simulated images with known ground-truth fields. As such, supervised methods are hardly applicable.
On the other hand, unsupervised methods overcome the limitation of obtaining plausible ground-truth transformations [47] and are trained to maximize a similarity metric between the input images. These metrics can also be accompanied by regularization terms, so that the obtained deformation field meets certain criteria, such as invertibility or smoothness.

Chapter 3

Introduction

3.1 Problem Statement

To address the limitations of the current response evaluation criteria (i.e RECIST), highlighted in Section 1.4.2, novel methods are currently under research. While there have been attempt to automatise RECIST via computer algorithms, either by computer assisted measurements [50] or by tumor burden Artificial Intelligence (AI) - based segmentation [51], these would always fall short of the standard radiologist's work up.
In the pursue to imitate classical radiological reporting, the concept of Prognostic AI Monitoring (PAM) has been studied by Trebeschi, Loohuis, van der Loo et al. in [34], [51], [37], [52]. It is an AI tool that aims to estimate treatment response overtime and to predict survival, using longitudinal CT-imaging of patients. The modelling of changes that occur during follow-up is performed via unsupervised image registration. The output of this operation is a vector representation of morphological changes over time, which is then used for prognostication through a classifier, assuming that these changes hold an important prognostic value, often ignored by the methods currently in use in clinical practice [34].

The already mentioned pilot studies yielded significant results, but it is yet not clear whether the ability of the network to model the deformation field is directly proportional to the ability of the network to predict survival. This thesis aims to answer this question via an ablation study, where different components of the network architecture are removed or replaced, to monitor their impact on survival prediction. Figure 3.1 shows the basic components of PAM model, i.e the image registration (affine and elastic networks) and the prognostication modules.Figure 3.1 shows the basic components of PAM model, i.e the image registration (affine and elastic networks) and the prognostication modules.

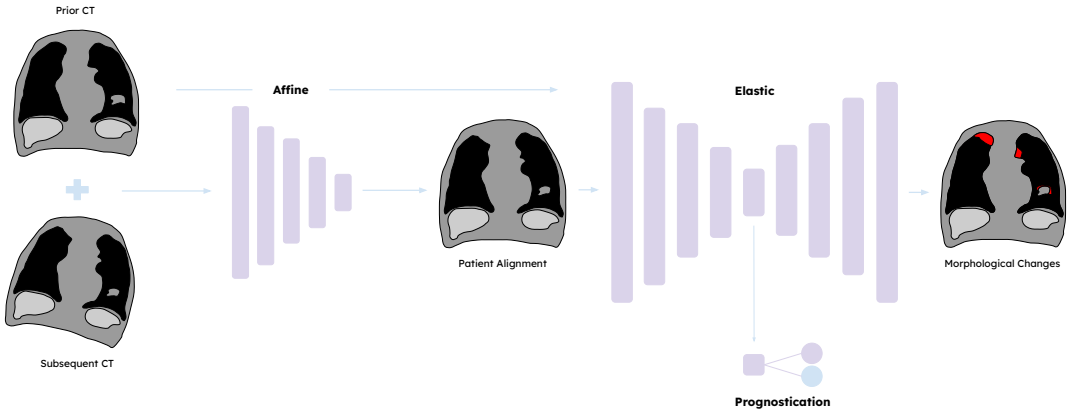


Figure 3.1: Representation of Prognostic AI-monitoring framework

3.2 Research question

The objective of this thesis is to investigate the relationship between the ability of the network to model the deformation field and its ability to predict survival, or, in other words, the impact that the vector representation of treatment response has to the survival prediction of oncological patients.

To do that, this research is designed as an ablation study, to investigate the contribution of the single components to the overall system. The experimental design involves the definition of different registration models: they share the basic architecture but differ from each other in the use of specific strategies, such as attention mechanism (via Vision Transformer), and adversarial loss (via Generative Adversarial Networks or GANs).
In particular, four experiments are executed, and each is performed via three sub-experiments, that test different size (number of filters) and capacity (skip layers) of the elastic network. Details of all the experiments are listed in Section 3.3, while size and capacity parameters of individual sub-experiments are shown in Table 3.1. As visible from the table, the elastic network of the first two sub-experiments follows a U-Net shape [53], implementing skip connections, whereas in the last one it takes up the structure of an autoencoder.
By performing several sub-experiments and by implementing different strategies, it is possible to highlight any registration performance trend, quantified via quality metrics such as Dice Similarity Coefficient, Normalized Surface Distance or Structural Similarity Index. Then, prediction and analysis of survival are performed by

Elastic network version	Set of filters	Skip layers
Big	[16, 32, 64, 128, 256]	yes
Small	[4, 8, 16, 32, 64]	yes
Big-no-skip	[16, 32, 64, 128, 256]	no

Table 3.1: Elastic network parameters in different sub-experiments

extracting the trained features of the different models and by leveraging predictive and associative methods such as Random Survival Forest or Cox Time-Varying Regression model. Finally, by evaluating prognostication quality via C-index metric, it is possible to highlight any predictive performance trend and to link registration and prognostication abilities for the different experiments.

3.3 Thesis structure

This thesis resumes the work accomplished at the Radiology department of Netherlands Cancer Institute (Amsterdam, NL), between October 2022 and June 2023, under the supervision of Stefano Trebeschi and Laura Estacio Cerquin, within the department of Radiology, chaired by Prof. Regina Beets-Tan. The work is structured as follows:

- Chapter 1, *General clinical background*, introduces the reader to the clinical context in which the proposed work fits. In particular, it focuses on the anatomy and epidemiology of tumors, on the most widespread oncological therapies and on treatment monitoring.

- Chapter 2, *General technical background*, introduces the preliminary technical concepts necessary for a complete understanding of the following experiments, focusing on Deep Learning and Image Registration.

- Chapter 3, *Introduction*, introduces the Prognostic AI Monitoring project, which serves as the basis for all the following experiments.

- Chapter 4, *Varying Model Size and Capacity*, explains in detail the registration and prognostication modules contained in the baseline model, and reports the results of Experiment 1.

- Chapter 5, *Implementing the Adversarial Learning via Generative Adversarial Network*, explains in detail the integration of Adversarial Learning in the model, and reports the results of Experiment 2.

- Chapter 6, *Implementing the self-Attention Mechanism via Vision-Transformer*, explains in detail the integration of Vision-Transformer in the model, and reports the results of Experiment 3 and Experiment 4.
- Chapter 7, *Additional Experiment: Enforcing Similarity in the latent-space*, explains in detail the modifications applied to the baseline model to enforce the latent-space similarity, and reports the results of Experiment 5.
- Chapter 8, *Linking registration to prognostic performance*, resumes the achieved results, outlines the relationship between image registration and survival prediction tasks, and highlights the limitation of this work and potential further improvements.

Chapter 4

Varying Model Size and Capacity

4.1 Introduction

This chapter contains the description of the *Experiment 1*, where the response-to-treatment modeling is performed using the so-called baseline model, inspired by the one in [34].
Since the impact of image registration quality on survival prediction performance has not yet been established, the experiment studies this correlation by defining three different models. The size of the network, here parameterised by the number of features, and its capacity, here defined as the presence of skip-layers in the U-Net architecture (elastic sub-network, Figure 3.1), are the aspects under investigation. We expect that higher capacity and size of the network will lead to higher registration quality, which will lead to higher ability of modeling deformations, which will lead to higher survival prediction performance.
More specifically, the lower the number of features, the lower the number of abstractions the network is able to extract from data; consequently, by decreasing them, a greater difficulty in modeling the deformation field is expected. Similarly, worse registration performance is expected if skip-connections are not included in the architecture: their implementation help to recover the spatial information lost during down-sampling and to stabilize training.
Once obtained the registration trend, we want to verify whether this is also reflected in the predictive ability of the model.

The first section of the chapter is dedicated to the description of the dataset used for the two different tasks; the second includes the explanation of the sub-networks and, finally, the last one shows and comments on the results of the

registration and survival prediction tasks. These results will serve as a reference for subsequent experiments, which are described in detail in Chapters 5 and 6.

4.2 Materials

4.2.1 Image Registration dataset

The registration network is trained with a large dataset from The Cancer Imaging Archive (TCIA). All available datasets that could contain CT scans were extracted from the TCIA, resulting in a set with tens of thousands of scans. To clean it up and maintain only qualitative scans, including the desired anatomical regions, a filtering was applied. First only the scans with more than 50 slices and with a difference between first and second axial coordinates in the range [0.1, 5.0] mm were selected. After that, only scans that actually contained entirely thorax and abdomen were retained. To automatically extract all slices between the lower-neck and the lowest part of the pelvis, the method of Zhang et al. was used [54]. Finally, the volumes were resampled to $2x2x2$ mm voxels, cropped to a final dimension of $192x192x300$, and clipped between -120 (fat) and 300 (cancellous bone) HU to help reduce computational memory [52].

4.2.2 Prognostication dataset

All patients that started immunotherapy between 01/01/2013 and 31/12/2018 at The Netherlands Cancer Institute-Antoni van Leeuwenhoek Hospital (NKI-AVL; Amsterdam, The Netherlands) were included. Immunotherapy was defined as any treatment including anti-PD1, PDL1, or CTLA4. The dataset was first filtered, to retain only the scans with more than 50 slices and with a slice thickness below 1 mm; then it was processed following the same protocol used for TCIA dataset.
For each patient included in the study, scans were paired (obtaining prior and subsequent scans) if acquired over a period of time between 30 and 120 days.

4.3 Methods

4.3.1 Image Registration module

The PAM model uses image registration to assess all morphological changes that occur between two follow-up images of the same patient during treatment, as shown in Figure 3.1. Specifically, the images are CT volumes, therefore featured by the height and width of each slice, and the number of slices. The registration module consists of two sub-networks, i.e the *affine* and the *elastic* network.
The first is the affine, whose goal is to roughly align the patient between the two

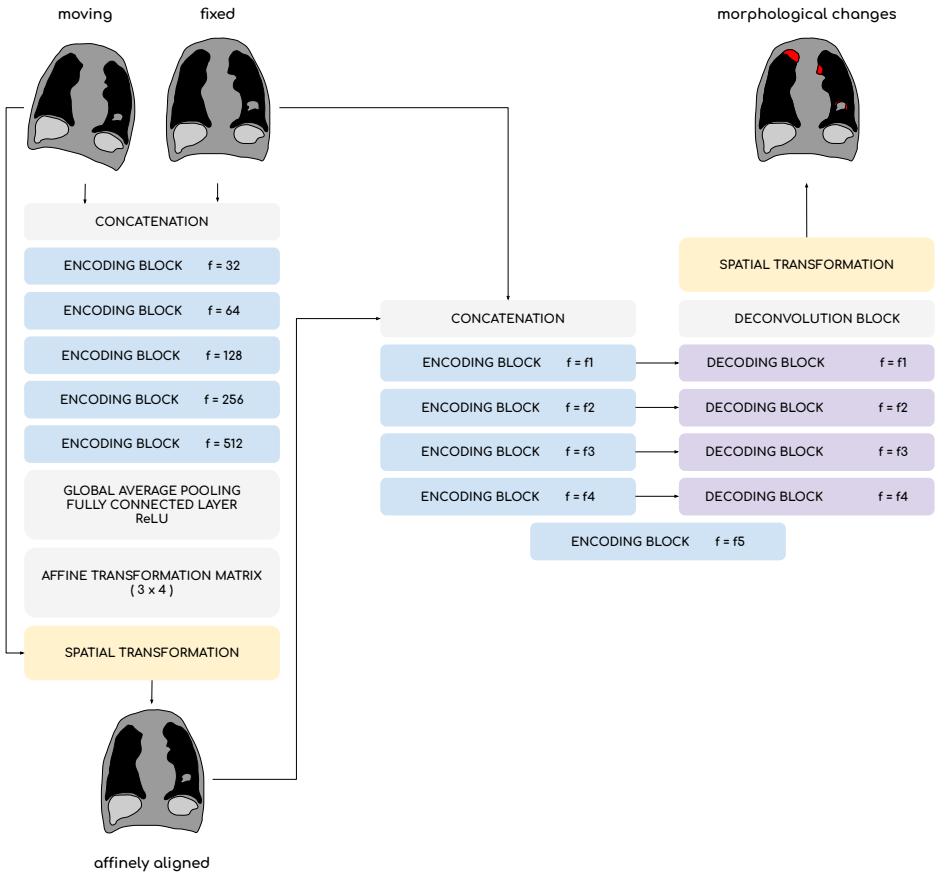


Figure 4.1: Architecture for the affine (on the left) and the elastic (on the right) networks, adapted from [34]

scans, and correct for different positions they might have assumed during acquisition [52], making use of translation, rotation, shearing, scaling, and reflection.
It includes six convolutional blocks followed by a fully connected layer, which regresses the 12 parameters of the affine transform between fixed and moving image [34]. These parameters represent a 3×3 transform matrix A and a 3-dimensional displacement vector b. Then, the affine transform is applied to the moving image via a spatial transformation layer (implementation taken from [55]), obtaining the affinely warped image.

The elastic sub-network takes in input the fixed image and the output of the previous network, and aims to identify morphological changes during the course of the treatment (i.e., longitudinal tracking) [34]. It follows a U-Net architecture, which can be divided into two parts: the encoding, or contracting, path and the

decoding, or expanding, path. Skip-connections are added to concatenate the feature maps resulting from the encoder to the corresponding decoder layer [56]. There are respectively five down- and four up-sampling layer, and a single deconvolution block. Each encoding block consists of a convolutional layer, a GN layer and a ReLU activation function. In the decoding path, each block consists of concatenation, deconvolution, GN and ReLU function. Finally, the single deconvolution block will output the dense flow field, a volume feature map with three channels (x, y, z displacements) of the same size as the input [48].

Both affine and elastic sub-networks are trained together in an unsupervised manner, trying to minimize the dissimilarity between the moving image warped by the spatial transformer and the fixed image. To quantify it, the Pearson correlation coefficient, or simply called correlation coefficient, is used: it is based on the covariance between the image volumes V_1 and V_2, which is defined as follows.

$$Cov[V_1, V_2] = \frac{1}{|\Omega|} \sum_{x \in \Omega} V_1(x) \cdot V_2(x) - \frac{1}{|\Omega|^2} \sum_{x \in \Omega} V_1(x) \cdot \sum_{y \in \Omega} V_2(y) \tag{4.1}$$

In Equation 4.1 Ω represents the grid on which V_1 and V_2 are defined. The correlation coefficient is defined as:

$$CC[V_1, V_2] = \frac{Cov[V_1, V_2]}{\sqrt{Cov[V_1, V_1] \cdot Cov[V_2, V_2]}} \tag{4.2}$$

It can assume all the values in the range $[-1, 1]$, and if it is equal to one of the ends it means that the two image volumes are linear functions of each other [48]. The correlation coefficient loss used in the training is defined as:

$$L_{CC}(V_1, V_2) = 1 - CC[V_1, V_2] \tag{4.3}$$

To enforce the estimated deformation fields to be spatially smooth, the Total Variation (TV) loss has been employed as a penalty by penalizing large differences, as in [57] and [48]. It is defined as:

$$L_{TV} = \frac{1}{3|\Omega|} \sum_{x} \sum_{i=1}^{3} (f(x + e_i) - f(x))^2, \tag{4.4}$$

where $e_{1,2,3}$ represent the natural basis of $\mathbf{R}^3$.
So, finally, the final loss function for the whole registration module is:

$$L = L_{CC}^{affine} + \alpha \cdot L_{TV}^{affine} + L_{CC}^{elastic} + \beta \cdot L_{TV}^{elastic} \tag{4.5}$$

In Equation 4.5 α and β are the weights for the regularization penalties, and they are set to $\frac{1}{100}$ and $\frac{1}{10}$ respectively. A higher value has been assigned to β to contain too large elastic deformations, which can result in a final image not anatomically plausible.

Evaluation metric

To assess registration quality, Dice Similarity Coefficient (DSC), Normalized Surface Distance (NSD) and Structural Similarity Index Measure (SSIM) were used.
The internal NKI dataset was segmented by using TotalSegmentator tool [58]: it is an algorithm based on nnUNet [59] which outputs the segmentations of a wide range of anatomical structures, providing also a statistics file containing the volume (in mm^3) and the mean intensity of each class. Once obtained the segmentations, only the biggest (i.e liver) and the smallest (i.e 11^{th} left rib) ones were studied, as we assume on average intermediate performance for all intermediate volume labels. For each scan pair, the oldest scan has been used as fixed image and the most recent in time as moving, and registration was performed. The resulting deformation field was then applied to the segmentation mask of the moving image to obtain the warped segmentation. Finally, DSC and NSD were calculated between fixed and warped segmentation masks of liver and the 11^{th} left rib.
DSC is a overlap-based metric that quantifies the similarity between two volumetric samples as defined below:

$$DSC = \frac{2 \cdot |X \cap Y|}{|X| + |Y|}, \tag{4.6}$$

where $|X|$ and $|Y|$ represent the number of elements in each sample.
NSD, on the other hand, does not focus on the entire volume but it is a boundary-distance-based method that quantifies which fraction of a segmentation boundary is correctly predicted. It requires the definition of a threshold, which was set to 3 mm. A surface element is considered correctly predicted if the closest distance to the ground-truth surface is smaller than or equal to the threshold.
This metric fits our application, since it focuses on the localisation of the overall anatomical structure (i.e segmentation) and on the alignment of fixed and warped surfaces [58]. The choice to use both DSC and NDS wants to give a complete view of the registration performance, keeping in mind the inherent limitations of the individual metrics. Volumetric overlap-based metrics, such as DSC, are highly sensitive to the object size [60], since a False Negative (FN) voxel penalizes more the final score for smaller volumes. Contrarily, distance-based metrics, such as NSD, are highly sensitive to object shape [61] and to its surface-to-volume ratio. In fact, the likelihood that the boundaries of a thin and elongated volume (such as the rib) are contained within the set threshold is higher than for a object with a lower surface-to-volume ratio (such as the liver).
Since the limitations of the two metrics are somewhat complementary, considering the size and structure of the two anatomical structures considered, it was decided to use them together. A not-segmentation-based metric, i.e SSIM, was implemented to assess the similarity between entire volumes, independently from specific anatomical structures. It performs a localized comparison rather than a global similarity,

making it suitable for image registration, where the alignment needs to be accurate at a fine-grained level. Moreover, it is a perceptually meaningful metric, taking into account features as luminance, contrast and structural information that underpin human perception.

4.3.2 Prognostication module

The trained registration network is used to extract the quantitative vectorial features from the scan pair, describing the deformation field, and serving as input for prognostication. In particular, the features of the deepest layer of the elastic sub-network are extracted: according to the size of the network set in the sub-experiments, they can include either 256 or 64 feature maps. Global average pooling is applied in order to transform the features from a tensor shape to a vector. The resulting vector represents the input of the core of the prognostication module, i.e the Random Survival Forest (RSF): it is a survival regression model that operates by splitting the dataset into different groups with the same mortality hazard, and additionally it is able to deal with censored data [62].
As its precursor (i.e Random Forest) it consists in several decision trees, which are trained on different parts of the training set, to increase the generalisation ability. The RSF outputs a score representing the risk that the failure event occurring: the final score is dependent on the output of each and every decision tree.

In statistics, survival analysis represents the collection of methods that aim to estimate (and predict) the amount of time until one event occurs. Therefore, it is necessary to define an event, when that event has taken place, in days, and the observation-time. For the purposes of the thesis, the event is defined as the death of the patient, the observation-time is equal to 1 year after the date when the last CT was acquired, and the survival-time is defined as the days until the event occurs [63]. When a patient does not experience the event during the observation interval, it is called *right-censored*.

The RSF model, as mentioned in Section 4.2, was trained with an internal NKI dataset. For each scan pair in the training set, in addition to the features extracted from the registration task, the RSF takes in input also the time-interval between prior and subsequent scan (in days) and the time-interval between prior scan and the start of treatment date (in days).

To better understand the risk score predicted, let's consider a single tree. According to the input x_i, the sample ends up in one of the leaves of the tree, called h. Given h containing N items, each item is characterized by a event-time T and a event-status δ ($= 1$ if event occurred, else $= 0$). So, there exist N event-times

possible in the h leaf. For a given event-time T, it is possible to define two parameters: (I) d_T is the number of deaths occurred until that time, (II) Y_T is the number of patients at risk of death until that time [64]. The Cumulative Hazard Function (CHF) can be calculated for h:

$$CHF_h = \hat{H}_h(t) = \sum_{T<t} \frac{d_T}{Y_T} \tag{4.7}$$

The above quantity is also known as the Nelson-Aalen estimator; it can be formulated also as the time-integral of the hazard function $h(t)$, which describes how the probability of the event to occur changes over time.
Since all the items in the same leaf share the same risk, the prediction of the decision tree for the i^{th} sample is equal to the CHF calculated for its leaf:

$$H(t|x_i) = \hat{H}_h(t) \tag{4.8}$$

The overall RSF prediction is defined as the mean over all single decision tree predictions, and it is called Ensemble CHF:

$$CHF_{ensemble} = H_e(t|x_i) = \frac{1}{n_{trees}} \cdot \sum_{j=1}^{n_{trees}} H_j(t|x_i) \tag{4.9}$$

The actual RSF output represents the expected number of deaths if all cases in the dataset were similar to i^{th} sample:

$$r(x) = \sum_{j=1}^{J} H_e(t_j|x_i), \tag{4.10}$$

where $(t_1, ..., t_j, ..., t_J)$ are the entire set of unique event-times for the learning data [65]. The output score is analyzed performing predictive and associative evaluations.

Predictive analysis

The predictive analysis is based on the Harrell's concordance index (C-index). The C-index is an evaluation metric that quantifies the discrimination power of the model: in particular, it outputs the probability that, in a random pair of samples, the sample that experiences first the event had a worst predicted risk score [62]. It can range between 0 and 1, and a value equal to 0.5 represents a random prediction. C-index has a similar interpretation to Area under the ROC Curve (AUC), but it is able to deal with censored data. The prognostic value of the AI risk-score is also assessed via logrank-test, after splitting the scores in two groups according to the median, to evaluate whether the differences between the two classes are statistically significant or not. For the most performative model the Kaplan-Meier curves were also used, by maintaining the same splitting and obtaining the *high risk* and *low risk* curves.

Associative analysis

A Cox time-varying regression analysis is performed to evaluate the relationship between AI risk-score and the patient survival, correcting for cofactors in the data. These cofactors are informative of pathology and therapy, and include: cancer-type, the intake of opiods, corticosteorids or immunosuppressants, brain radiotherapy, bone radiotherapy, or other types of radiotherapy. As for the Kaplan-Meier curves, the associative analysis is performed only for the model showing the higher C-index.

4.4 Results

4.4.1 Study Cohort

The preprocessing applied to TCIA dataset, used to train the image registration module, resulted in a final dataset of 2185 CT volumes, later divided into a training set (80% = 1747) and a test set (20% = 438).

	patients	scan pairs	age	M/F ratio	% mortality	survival months SoT	survival months 2_{nd} scan
Total	861	3036	58.8	1.07	46.8	25.1	15.4
subset							
Melanoma	477	1601	57.1	1.48	41.1	28.6	17.4
Lung cancers	119	473	59.3	0.56	52.6	25.6	15.3
Breast cancers	73	304	57.1	0.00	68.7	16.9	10.3
Kidney cancers	81	239	62.4	2.92	36.0	23.7	13.7
non-Melanoma	40	131	64.8	1.43	38.2	38.1	26.0
Training set	431	1463	59.9	1.27	48.0	25.1	15.2
Test set	430	1573	57.7	0.91	45.6	25.1	15.5

Table 4.1: Patient characteristics for the training set and the independent test set

For the prognostication dataset, a cohort of 861 patients was collected: it contained 5044 CT volumes and 3068 pairs. Scans were paired if acquired over a period of time between 30 and 120 days. Since multiple acquisitions of the same patient in the same date were performed, for example by varying the CT acquisition parameters, we define unique pairs (= 3036) as the number of scan pairs taking into account a single acquisition per date. The cohort was later divided into a training set and a independent test set based on the patient identifier. Patients with even ID numbers were assigned to the training set, patients with odd ID numbers to the test set. Table 4.1 shows patient average characteristics in both training and test set and by splitting the cohort by cancer type. Only the most recurrent five cancer types are reported: melanoma (C 4.3), lung (C 3.4), breast (C 5.0), kidney (C 6.4) and

skin non-melanoma (C 4.4). It is important to notice that these are all the tumor types that have been reported for these patients, but are not necessarily the tumor types they were receiving treatment for. The column *M/F ratio* refers to the ratio between male and female subjects. The two columns related to survival contain the survival months after the start of treatment (SoT) and after the subsequent scan date (2_{nd} scan). The entire process that led to the final dataset is shown in the consort diagram of Figure 4.2.

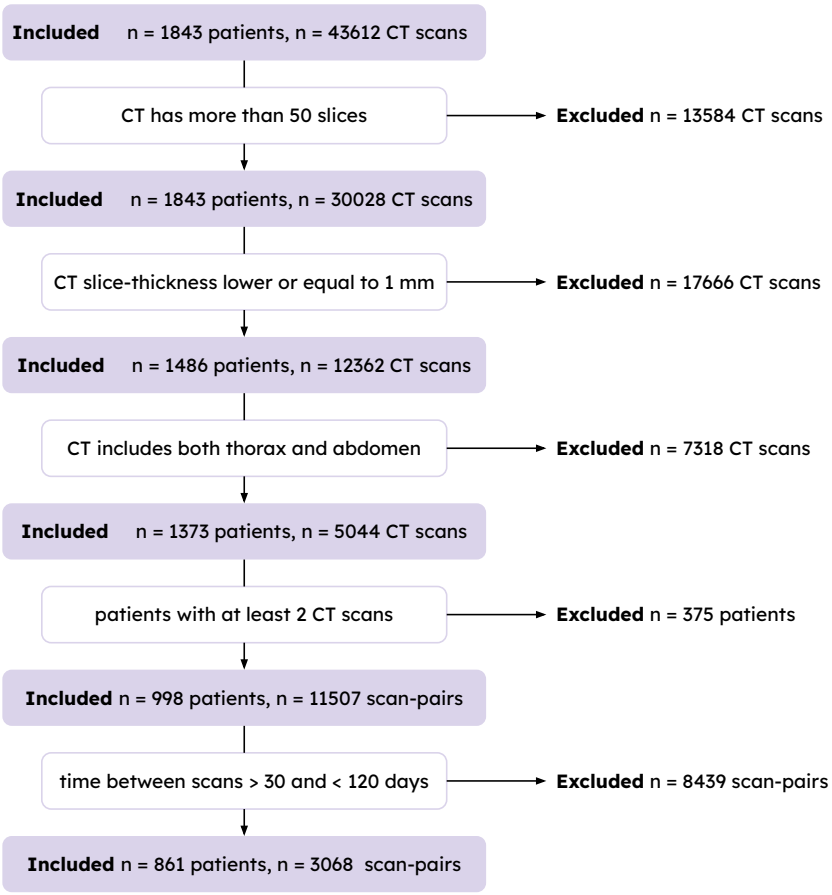


Figure 4.2: Consort diagram of patient, scan and scan-pair selection

4.4.2 Image Registration results

The registration accuracy was assessed by performing the registration of the NKI dataset scan-pairs and by calculating SSIM, and DSC and NSD between the fixed and warped segmentation masks for the biggest (i.e liver) and the smallest (i.e

11^{th} left rib) volumes. The metrics values averaged along the dataset are shown in Table 4.2. The results of the three different models, which differ from each other by size and capacity of the elastic sub-network, show the same trend: all network configurations share higher DSC values for bigger structures and higher NSD values for smaller structures. The highest performance in terms of DSC_{liver}, NSD_{liver} and NSD_{rib} are reached by the *Big-no-skip* model.

	DSC_{liver}	DSC_{rib}	NSD_{liver}	NSD_{rib}	$SSIM$
Big	$89.9 \pm 5.9\%$	$66.6 \pm 21.7\%$	$72.7 \pm 13.9\%$	$81.1 \pm 24.5\%$	$83.9 \pm 4.4\%$
Small	$88.6 \pm 5.7\%$	$52.6 \pm 23.6\%$	$69.4 \pm 13.4\%$	$73.4 \pm 27.2\%$	$75.3 \pm 5.7\%$
Big-no-skip	$92.5 \pm 3.6\%$	$60.8 \pm 20.0\%$	$83.4 \pm 11.9\%$	$87.2 \pm 24.6\%$	$73.8 \pm 6.1\%$

Table 4.2: DSC, NSD and SSIM mean and standard deviation values for the three sub-experiments of *Experiment 1*

Contrarily, the highest DSC_{rib} is related to the *Big* model. A lower DSC_{rib} and an higher NSD_{rib} can be explained by taking into account the size of the anatomical structure. Since the total segmentation covers a limited volume, the misclassification of a few voxels can significantly bring down the final score (= 60.8%), although the remarkable alignment between fixed and registered surfaces (= 87.2%).
Since the overall performance cannot be explained exhaustively by DSC and NSD values, a qualitative evaluation of the registration of the three models was also made, supported by SSIM values.
In Figure 4.3 an example of qualitative behavior for a random TCIA-dataset scan pair is given. For this example, the DSC_{liver} has been calculated between fixed and warped masks: the trend is the same as in Table 4.2, so it is a good example of the average differences between the sub-experiments. Fixed and moving scans don't belong to the same patient, which is why they look very different from each other. The choice to show such an example, far from the actual purpose of the registration module (receiving in input scans of the same patient to model the differences), was made to emphasize more easily the advantages and limitations of the three models.
At first glance, the *Big-no-skip* network creates less detailed images than the other two models, with more nuanced contours. Despite the lack of detail, it is able to reconstruct medium-large anatomical structures' shapes effectively. We assume that the good modeling ability is due to the size of the model, sufficiently high to allow the decoding of the main shapes present in the image. At the same time, the absence of skip-connections probably does not allow the model to retain all the information necessary for a meticulous reconstruction, leading to a lower SSIM (= 73.8 %)
Contrarily, the *Big* network reconstructs more faithfully the structures present in the fixed image, managing to maintain a high quality overall. In fact, it shows

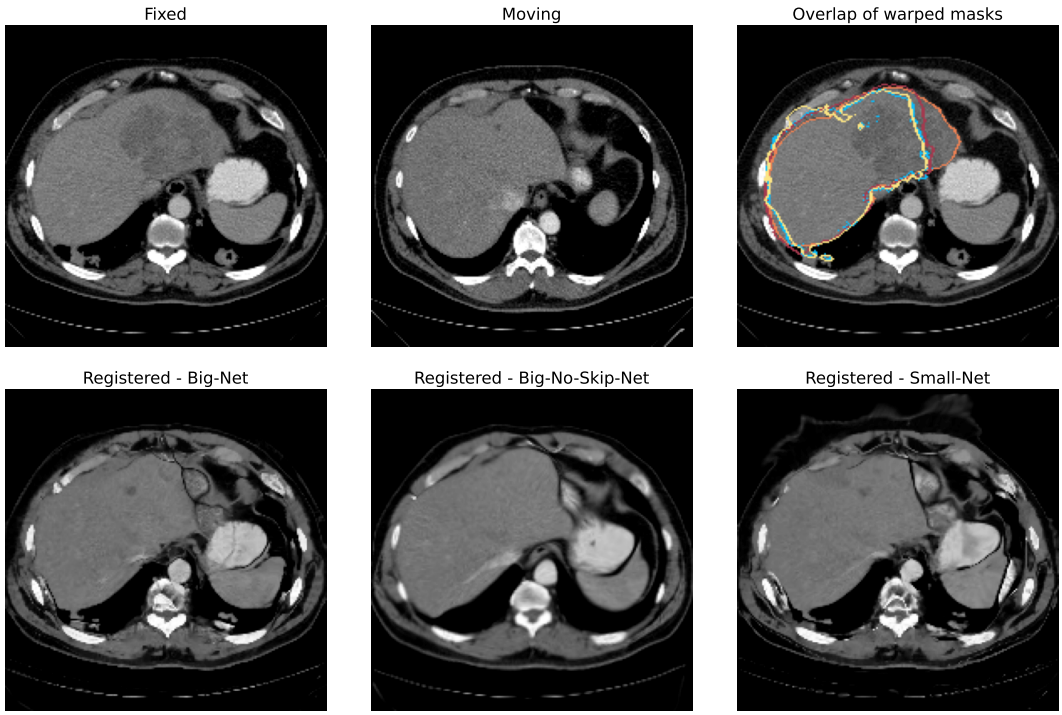
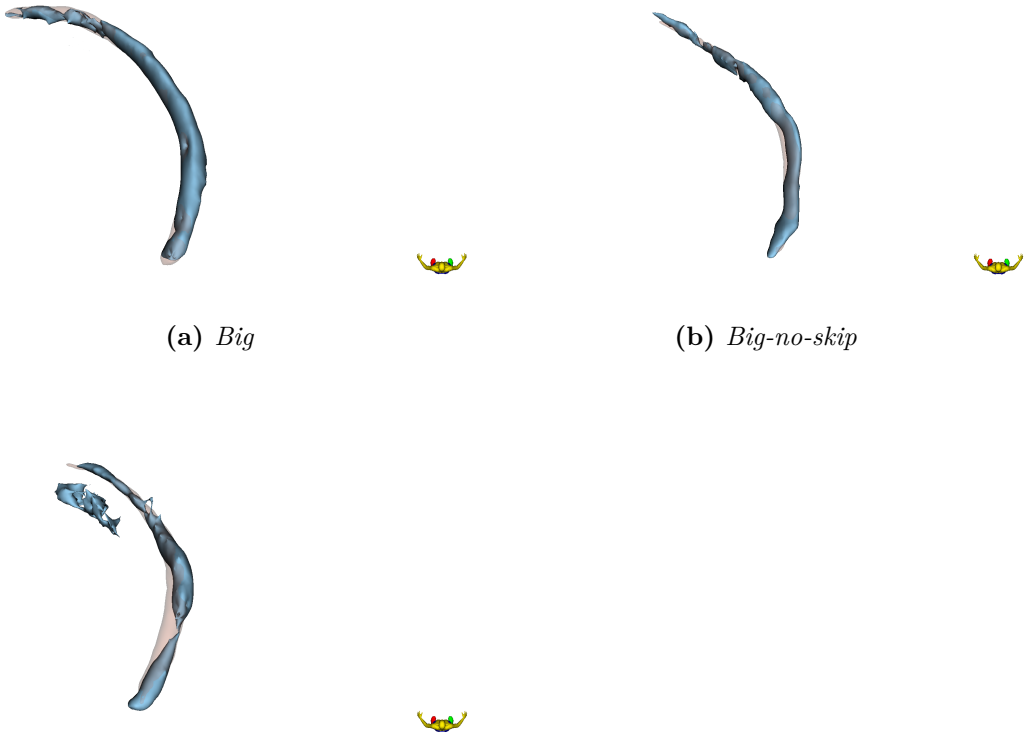


Figure 4.3: Qualitative comparison of registration performance for the three sub-experiments of *Experiment 1*. The figure shows an example of fixed - moving scan pair, the registered scan outputted by each model, and the contours of the liver segmentation masks registered by each model. The orange line is the ground-truth, blue refers to *Big*, amaranth to *Big-no-skip* and yellow to *Small*.

the highest SSIM value throughout the experiment. The evaluation of the entire registered volume allows some considerations that, analyzing only the liver mask, would not be possible. In fact, the contour of the segmentation mask outputted by the *Big* model (blue line) is very similar to that of the *Big-no-skip* (amaranth line), and both do not include the upper right portion of the liver. Nevertheless, the *Big-no-skip* does not reconstruct that area at all, while the *Big* manages to model it (roughly) but does not recognize it as belonging to the liver. Finally, the *Small* network manages to preserve the anatomical likelihood, thanks to the implementation of skip-connections, but fails to model some deformations, probably due to the lower number of filters. Despite the slightly nuanced edges, it overall manages to align the segmentation of the organ (yellow line).

For completeness, after having shown in Figure 4.3 the liver masks produced by the three registration models, in Figure 4.4 a comparative example between

the 11^{th} left rib masks is shown. For a better visualization of the anatomical structure (very small compared to the total scanning volume) we chose a volumetric rendering, instead of a single axial slice. For each model in the figure a fixed-moving pair has been selected from the NKI dataset having a DSC_{rib} value equal to the average along the entire test dataset. This way it is possible to visualize the average behavior of the three networks. Following the trend of DSC and NSD shown in Table 4.2, *Big* and *Big-no-skip* models have similar performance for the rib, managing to reconstruct the volume, while the *Small* one underperforms with a greater number of false positive and false negative voxels.



(a) *Big* **(b)** *Big-no-skip*

(c) *Small*

Figure 4.4: Rendering of 11^{th} left rib volumetric segmentation masks for the three sub-experiments of *Experiment 1*. In each subplot the blue volume represents the warped mask, which is superimposed to the light orange fixed mask. The symbol on the bottom right of the figures depicts the point of view in space: the subject is analyzed on the transverse plane, from top to bottom. The caption of each subplot refers to the network used for the registration. Image created with 3D-Slicer.

4.4.3 Prognostication results

A RSF was trained to predict survival from the imaging features extracted from the registration module, which in turn inputs two scans of the same patient acquired in different time-points, called *prior* and *subsequent* scans. The regression model outputs a mortality hazard related to the 1-year survival from the date of subsequent scan. To assess the prediction quality the C-index was used and the dataset was filtered and processed. Only scan-pairs where prior scan was performed before SoT and the event was experienced at least after 2 months after SoT were retained. Then, imaging features of multiple pairs acquired on the same date have been averaged in order to obtain only unique pairs. As a result, the survival prediction was conducted by analyzing the first two scans acquired during the monitoring stage, i.e baseline (BL) and first follow-up (FU1). The final test set counted 138 scan pairs. This choice was made by assuming that the time interval right after the start of immunotherapy is the clinically most relevant and in which the treatment effects are most evident. Results of the survival AI-score, in terms of C-index and statistical significance assessed via log-rank test, are shown in Table 4.3. Confidence intervals were estimated via bootstrapping using repeated sampling with replacement (100 times).

	C-index	*p-value*
Big	0.65 (0.57 - 0.73)	3.4e−2
Small	0.69 (0.63 - 0.76)	3.2e−3
Big-no-skip	0.63 (0.54 - 0.70)	1.6e−2

Table 4.3: C-index, with relative confidence intervals, and p-values for the three sub-experiments of *Experiment 1*

Table 4.3 shows an inverted trend compared to that shown in Table 4.2 for DSC and NSD metrics. In fact, the higher the overlap between fixed and warped segmentation masks of liver and rib, the lower the survival prediction accuracy. However, if SSIM is considered to assess registration quality, no clear link is shown between registration performance and C-index values. All models are statistically significant ($p < 0.05$), emphasizing the prognostic value of the AI-score in correctly dividing patients into two groups with different risk.

4.5 Discussion

In this chapter, the baseline architecture was utilized to model the radio-anatomical changes occurring during the follow-up of oncological patients and to predict their 1-year survival. To introduce variations in image registration quality and evaluate the

corresponding impact on prediction performance, an ablation study was conducted, manipulating the size and capacity of the network. The behavior of the three defined models partially aligned with expectations. The *Big* model, equipped with highest number of computational units and connections (i.e, convolutional filter and skip layers), demonstrated superior registration accuracy based on SSIM evaluation. This finding emphasized the advantages of deeper models in extracting abstractions from data, resulting in more accurate and detailed warped images. Additionally, the *Big-no-skip* model exhibited the lowest SSIM value, indicating the beneficial role of the information exchange between the encoding and decoding paths within the elastic network for the task. However, the DSC and NSD results did not entirely correspond with the SSIM trend. The model without skip-connections displayed higher overlap of anatomical structures between fixed and warped images. The *Small* model yielded lower performance compared to the *Big* model. The DSC and NSD results suggested that skip-connections were necessary to obtain qualitatively and detailed reconstructed images, but they could be omitted when the primary objective was correct structure alignment. Nonetheless, since the *Small* network exhibited lower values in DSC, NSD, and SSIM evaluations, it can be concluded that reducing the number of convolutional filters resulted in coarser registration.
Survival prediction was conducted by extracting imaging features from the bottleneck of the elastic network and utilizing them to train a RSF regression model, which estimated the 1-year survival risk factor. Testing the predictor solely on the BL-FU1 scan-pairs (138 pairs) from the internal NKI dataset demonstrated the prognostic value of the proposed AI framework. All models achieved C-index values between 0.63 and 0.69. In the Introduction section, the hypothesis was posited that more accurate registration and greater overlap between anatomical structures of fixed and warped images (representing prior and subsequent scans) would enhance the robustness of the prognostication task. This assumption was rooted in the goal of recreating a quantitative alternative to radiological reporting, wherein all deformations are identified and accounted for during the assessment of treatment response. Accordingly, if all deformations were accurately modeled through precise registration, it became feasible to evaluate the prognostic significance of these changes, thereby improving survival prediction performance. However, the results did not support our initial hypothesis and instead revealed an unexpected trend. The *Small* network, which was described by a 64-features vector (instead of 256 as for the other models) exhibited the highest performance, followed by the *Big* and the *Big-no-skip* models, contradicting the trends observed in the DSC and NSD metrics. No apparent correlation was found between SSIM and C-index values.

Chapter 5

Implementing the Adversarial Learning via Generative Adversarial Network

5.1 Introduction

This chapter contains the description of the *Experiment 2*, where the baseline model is modified by the addition of a discriminator, in order to leverage the adversarial learning. This mechanism, which is explained in detail in Section 5.2, is the basis of Generative Adversarial Networks (GANs) and promotes the obtaining of realistic images.
In *Experiment 1* the training of the registration module was carried on employing a penalty (i.e Total Variation loss) on large unrealistic deformations. Despite its usefulness, it is possible that its implementation prevents the network to model large, potentially clinically significant deformations [52]. As a result, this constraint may limit the field of effectiveness of the registration, and consequently limit the predictive ability of the model. Because of that, we expect that by lowering the contribution of the smoothness penalty, and replacing it by implementing the adversarial loss, it is possible to maximize the capacity of the network to model morphological changes while assuring realistic deformations, therefore increasing its performance in the prediction of survival.
As well as the first experiment, this one is also designed as an ablative study, performing three sub-experiments by modifying size and capacity of the elastic

network. The objective of *Experiment 2* is therefore to highlight a possible registration trend, to verify that this is reflected in the prediction, and to confirm the effectiveness of the implementation of the adversarial loss by comparing the results of the models with and without discriminator.

5.2 Technical background

GANs, formulated by Goodfellow et al. in [66], belong to the family of generative models and are composed of two neural networks that continuously try to beat each other in a so called minimax game.
The two components of the model are known as the *generator* and the *discriminator*. The first, starting from random noise, has the task of generating synthetic data that can be as similar as possible to the real data present in the training set to fool the discriminator [67]. At the same time, its opponent has to distinguish the real data from the ones generated by the generator.
During training, the generator is constantly trying to outsmart the discriminator by synthesizing better and better fakes, while the discriminator is working to correctly classify real and fake data. The equilibrium of this game is achieved when the generator is able to generate "perfect" samples, bringing the discriminator to a success rate of 50% [68].

More formally, let x be data representing an image, and $D(x)$ be the discriminator network which outputs the probability that x came from training data rather than the generator. By defining z as the latent space vector from which the generator creates fake samples, $G(z)$ represents the generator function which maps z into data-space [66]. Hence, D is trained to maximize the probability it classifies correctly ($logD(x)$); G is trained to minimize $log(1 - D(G(z)))$. The overall loss function is defined as follows:

$$\min_G \max_D (D, G) = E_x[logD(x)] + E_z[log(1 - D(G(z)))] \tag{5.1}$$

In Equation 5.1 E_x and E_z denote the mean likelihood over all original data and synthetic data respectively.

5.2.1 GAN in Image Registration

GANs are also a common component of Deep Learning Image Registration approaches [47]. These models, widely used in the medical domain as tools for data augmentation and segmentation [47], are effective in improving the overall image alignment, as showed by Fu et al. in [69].
The generator is a registration network which predicts the deformation field, and

consequently the warped moving image. The discriminator receives in input the fixed and the warped moving image, and judges whether images are well aligned and feeds misalignment information to the generator during training [70].
Hence, the addition of a discriminator offers a learnable mechanism to evaluate the similarity between two images, and simplifies the task of choosing a suitable similarity metric [47].

5.3 Methods

The datasets used in this experiment are the same of those of *Experiment 1*, described in Section 4.2.

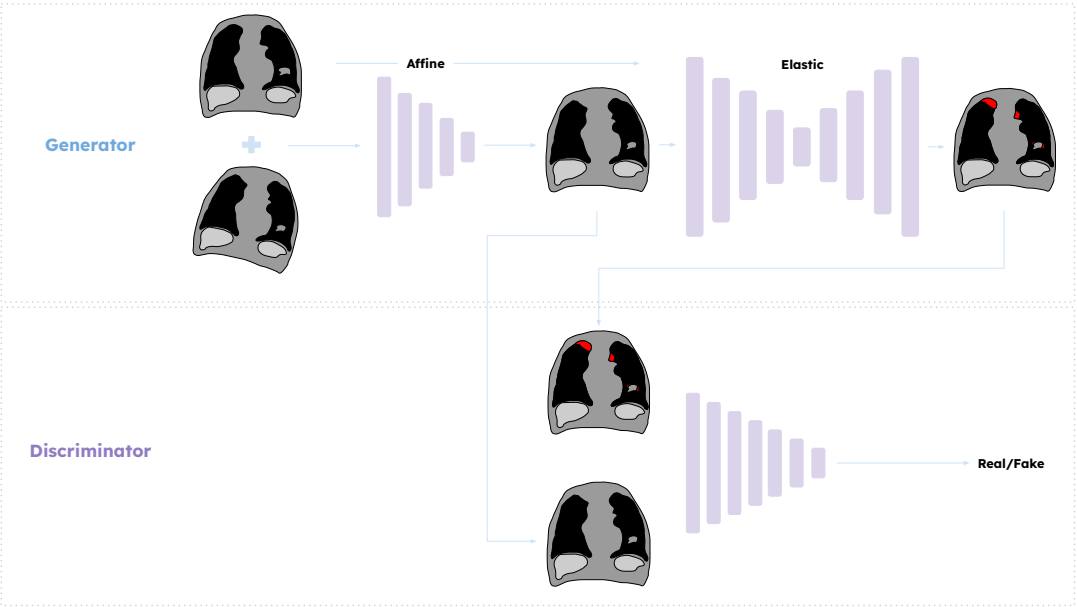


Figure 5.1: Representation of Adversarial PAM

To effectively implement the adversarial learning to the model, the affine and elastic networks are ideally grouped in a generator, and the discriminator is added. It is made up of seven convolutional blocks, followed by an adaptive average pooling layer, a linear layer and a final sigmoid activation function. Every convolutional block applies a convolution, a GN and a ReLu activation function to the incoming feature map. Block by block the feature maps double their depth, from a value of 8 up to 512. The discriminator needs two inputs, a real and a generated image. For the sake of brevity we will use the term *real* to indicate the affinely aligned image, and the term *fake* to indicate the warped moving image.

The affinely aligned image has been used as a reference for the discriminator, instead of the fixed one, because it is assumed that any unrealistic deformations to be corrected come mainly from the elastic network [52], and in order to avoid trivial solutions. The affine network eventually produces some artifacts at the edges of the image (zero-padding to maintain the same shape after the transformation), that are in turn passed on to the final warped image. Choosing the fixed image as a reference, exempt from these artifacts, would have led the discriminator to rely exclusively on this distinction instead of focusing on the real morphological details. The resulting network is called Adversarial PAM.
Consequently, the loss that drives the training of the generator now includes an extra component, i.e adversarial loss.

$$L_{generator} = L_{CC}^{affine} + \alpha \cdot L_{TV}^{affine} + L_{CC}^{elastic} + \beta \cdot L_{TV}^{elastic} + \gamma \cdot MSE_{fake}^{1} \tag{5.2}$$

The first 4 terms are the same of those in Equation 4.5, and the last one measures the ability of the generator to fool the discriminator. The subscript denotes what the prediction should be compared to, and the superscript represents a real image (1) or a fake image (0).
In other words, the last term quantifies the mean-squared error between a real image and the output of the discriminator when it has in input a fake image.
γ is set to $\frac{1}{10}$, and β is ten times higher than in Experiment 1, so equal to $\frac{1}{100}$. The contribution of the regularization term has been lowered for the elastic network (β) because of the presence of the adversarial learning term, which has the task of penalizing unrealistic deformations. The loss function of the discriminator is:

$$L_{discriminator} = \frac{1}{2} \cdot (BCE_{real}^{1} + BCE_{fake}^{0}) \tag{5.3}$$

It quantifies discriminator's ability to classify real from generated samples. BCE represents the binary cross-entropy.

5.4 Results

The data filtering and preprocessing applied are the same of those of *Experiment 1*, so the resulting cohorts used for image registration (TCIA dataset) and prognostication (NKI dataset) are the ones already described in Section 4.4.1 and in Table 4.1.

5.4.1 Image Registration results

The registration accuracy was assessed by performing the registration of the NKI dataset scan-pairs and by calculating SSIM, and DSC and NSD between the fixed

and warped segmentation masks for the biggest (i.e liver) and the smallest (i.e 11^{th} left rib) volumes. The metrics values averaged along the dataset are shown in Table 5.1. The results of the three different models, which differ from each other

	DSC_{liver}	DSC_{rib}	NSD_{liver}	NSD_{rib}	$SSIM$
Big	$84.5 \pm 7.3\%$	$50.0 \pm 26.2\%$	$59.5 \pm 12.8\%$	$64.0 \pm 30.2\%$	$82.9 \pm 3.7\%$
Small	$84.8 \pm 6.9\%$	$29.6 \pm 23.8\%$	$60.1 \pm 12.8\%$	$51.6 \pm 32.3\%$	$74.0 \pm 5.5\%$
Big-no-skip	$90.0 \pm 4.2\%$	$43.7 \pm 24.4\%$	$72.5 \pm 12.9\%$	$73.1 \pm 32.6\%$	$68.0 \pm 6.6\%$

Table 5.1: DSC, NSD and SSIM mean and standard deviation values for the three sub-experiments of *Experiment 2*

by size and capacity of the elastic sub-network contained in the generator, show the same trend found in *Experiment 1*. In fact, the most performative model in terms of DSC and NSD is the *Big-no-skip*, followed by the *Big* and the *Small*. In addition, the *Big* performs the most accurate registration overall, showing highest SSIM, followed by *Small* and *Big-no-skip* models.
From the comparison between the *Experiment 1* and *Experiment 2*, the metrics related to the liver differ just by a few percentage points. This reveals that the ability to register large volumes is maintained even by implementing the adversarial learning. In contrast, the three models of *Experiment 2* show values of DSC and NSD less than about 15% for the 11^{th} left rib. It is assumed that the reduction recorded for smaller structures is attributable to how the adversarial learning has been implemented. As visible in the generator's loss function formula (Equation 5.2), the last term includes the MSE between a real image and the discriminator's output when it receives a fake image input. The MSE is undoubtedly an intuitive metric, but it focuses on the overall features of the samples, not local. Therefore, we hypothesize that, since the majority of the image is occupied by larger structures, such a loss focuses on increasing the similarities of such structures, paying less attention to the smaller details. It is possible that, by implementing loss functions that aim at obtaining a more visually qualitative image, such as the Fréchet Inception Distance (FID), this limitation can be minimized. In addition, the decision to use the affinely warped image as the reference image for the discriminator, to prevent it from focusing on potential artifacts introduced by the affine transformation, may have influenced the performance. While this approach helps mitigate the discriminator's sensitivity to artifacts, it introduces an inherent error, since the affinely warped image will always contains some degree of registration error and artificiality (not present in the real fixed image). As in the previous chapter, for a complete understanding of the registration performance of the three models, a qualitative evaluation of the task is carried out. The same randomly extracted scan pair from TCIA dataset is shown in Figure 5.2. For this example, the DSC_{liver}

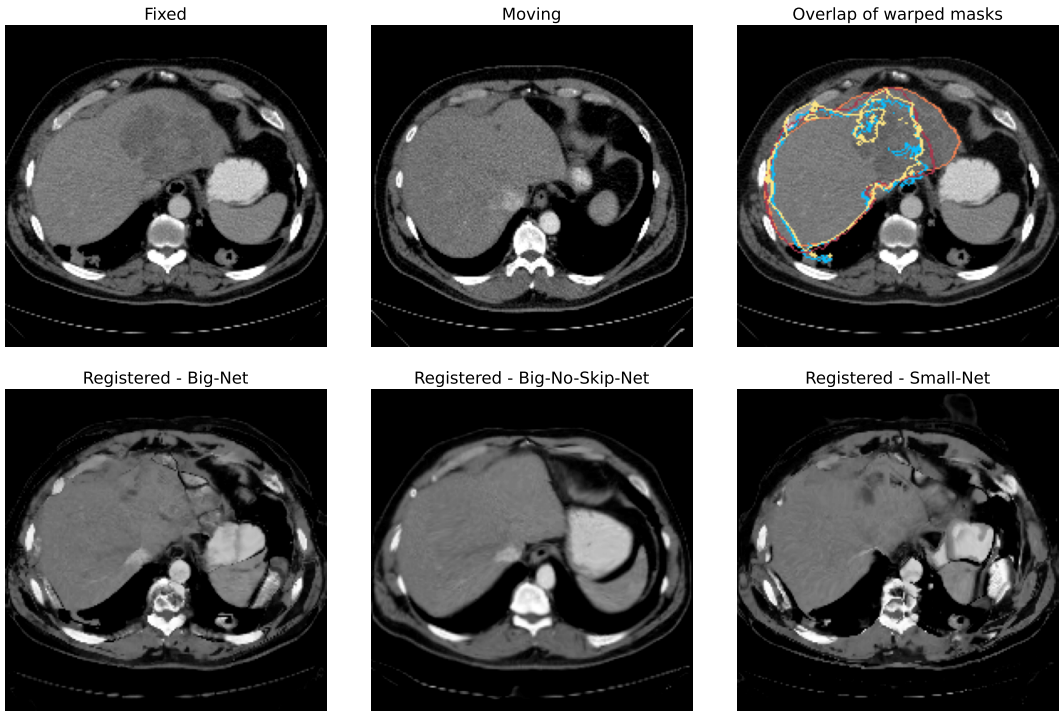
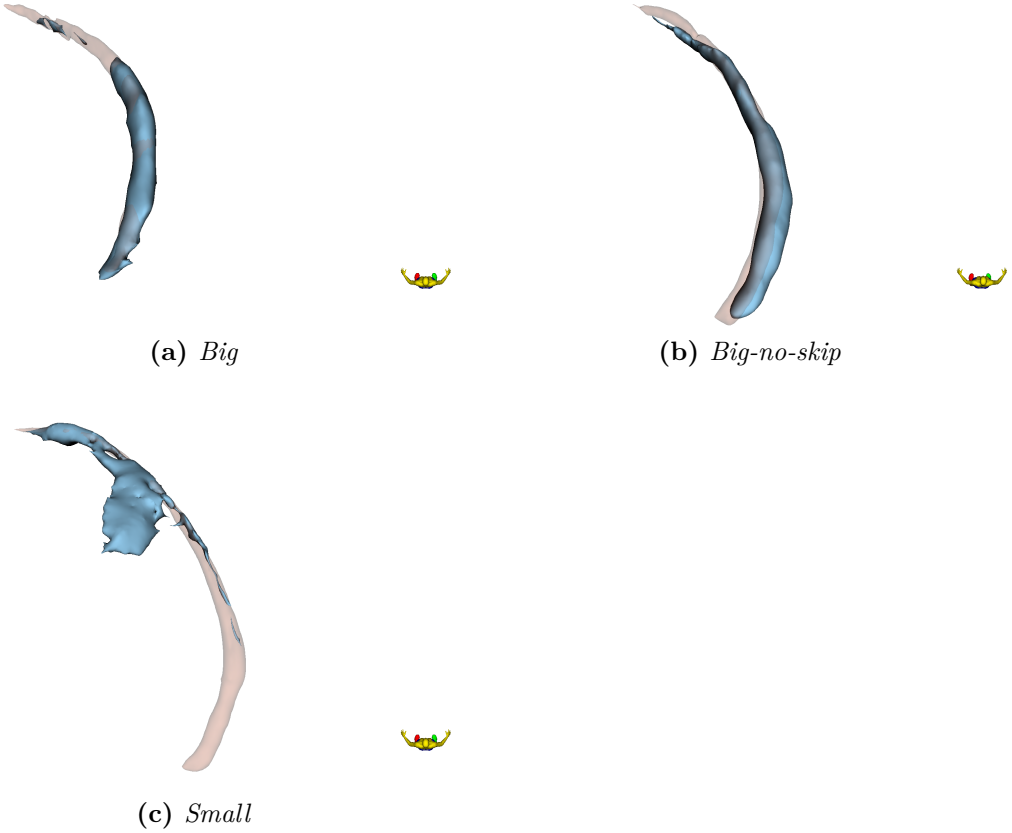


Figure 5.2: Qualitative comparison of registration performance for the three sub-experiments of *Experiment 2*. The figure shows an example of fixed - moving scan pair, the registered scan outputted by each model, and the contours of the liver segmentation masks registered by each model. The orange line is the ground-truth, blue refers to *Big*, amaranth to *Big-no-skip* and yellow to *Small*.

has been calculated between fixed and warped masks: the trend is similar to the one shown in Table 5.1, so it is a good example of the average differences between the sub-experiments.
Following the behavior shown in the absence of adversarial learning, the *Big-no-skip* model is visually less detailed, focusing on registering the main shapes present in the scan. For this reason, it is able to identify correctly the largest structures and to produce low noise segmentation masks. The *Big* model, on the other hand, thanks to the skip connections, is able to recreate the anatomical detail and to generate a more realistic overall image. However, the search for such detail leads the network to fragment the structures too much, as visible from the too noisy segmentation mask of the liver (in blue). The performance of the *Small* model does not differ too much from that of the *Big* one, but the registered image appears unrealistic because of many contours not precise enough. A common feature of the three Adversarial PAM models, deducible from the qualitative analysis, is the

limited possibility of deformation in respect to *Experiment 1*, especially for smaller structures such as the stomach (adjacent to the liver) and the spleen (adjacent to the stomach). The addition of the loss term probably contained too many variations in the deformation field to preserve the overall realism.

A visual evaluation between the 11^{th} left rib warped masks outputted by the models, shown in Figure 5.3, concludes the qualitative assessment of registration.



(a) *Big* **(b)** *Big-no-skip*

(c) *Small*

Figure 5.3: Rendering of 11^{th} left rib volumetric segmentation masks for the three sub-experiments of *Experiment 2*. In each subplot the blue volume represents the warped mask, which is superimposed to the light orange fixed mask. The symbol on the bottom right of the figures depicts the point of view in space: the subject is analyzed on the transverse plane, from top to bottom. The caption of each subplot refers to the network used for the registration. Image created with 3D-Slicer.

For a better visualization of the anatomical structure (very small compared to the total scanning volume) we chose a volumetric rendering, instead of a single axial slice. For each model in the figure a fixed-moving pair has been selected from the NKI dataset having a DSC_{rib} value equal to the average along the entire test dataset. This way it is possible to visualize the average behavior of the three networks. The performance worsening for smaller structures is clearly visible in the volumetric mask created by *Small* model, where only a third of the entire rib is considered and the overall shape is not maintained. The *Big* one is able to preserve the bone structure, but with lower performance in respect to *Experiment 1*. Finally, as reflected by DSC and NSD values, the *Big-no-skip* has the ability to effectively register the rib within a confidence interval of 3 mm.

5.4.2 Prognostication results

A RSF was trained to predict survival from the imaging features extracted from the registration module, which in turn inputs *prior* and *subsequent* scans of the same patient. The regression model outputs a mortality hazard related to the 1-years survival from the date of subsequent scan. To assess the prediction quality, the C-index was used and statistical significance was measured via log-rank test. As in the previous experiment, the performance were not assessed taking into account all the NKI independent test set, but only the baseline (BL) and first follow-up (FU1) unique scans. Unique pairs were obtained by averaging imaging features of all the multiple pairs acquired on the same date. Results of the survival AI-score, in terms of C-index and statistical significance, are shown in Table 5.2. Confidence intervals were estimated via bootstrapping using repeated sampling with replacement (100 times).

	C-index	*p-value*
Big	0.56 (0.46 - 0.63)	5.8e−1
Small	0.62 (0.55 - 0.69)	2.6e−3
Big-no-skip	0.61 (0.53 - 0.68)	1.9e−1

Table 5.2: C-index, with relative confidence intervals, and p-values for the three sub-experiments of *Experiment 2*

Table 5.2 shows a clear drop in predictive performance compared to the baseline. *Experiment 1* models showed C-index above 0.63 and up to 0.69, while the imaging features extracted from the adversarial architecture lead to values that range from 0.56 to 0.62. Moreover, only the RSF trained with the *Small* model features is statistically significant ($p < 0.05$). The trend found in the previous chapter is no longer respected here, as there is an inversely proportional relationship between registration quality and predictive accuracy. However, as visible from the

comparison between the Table 5.2 and 5.1, the model that overall shows the worst DSC and NSD values is the one better predicting survival.

5.5 Discussion

In this chapter, the integration of adversarial learning into the framework was accomplished via Generative Adversarial Networks, by introducing a discriminator into the image registration module previously presented in Chapter 4.
The decision to explore this architecture aimed to provide the model with a self-learnable mechanism for evaluating the similarity between two scans and penalizing unrealistic deformations [47]. This approach aimed to limit the use of strict predetermined penalties that could hinder the model's ability to capture large yet clinically significant changes. Consequently, the weight of the smoothness penalty used in the baseline model was reduced to produce more realistic and anatomically plausible warped images.
Similar to the previous chapter, three sub-experiments were conducted, wherein the size and capacity of the elastic network within the generator were systematically ablated. The behavior of the three models exhibited partial alignment with expectations. Qualitative examination revealed that the integration of the adversarial loss effectively increased realism and improved anatomical details in terms of contrast, brightness, and other perceptual features. However, all sub-experiments demonstrated limited accuracy in registering smaller structures, such as the stomach and spleen. Notably, DSC and NSD related to the rib exhibited a significant decline. As explained in Section 5.4, it could be attributable to how the adversarial learning has been implemented, i.e MSE between real and fake images. Despite the intuitiveness of the metric, MSE does not focus on local and smaller features. It is possible that, by implementing loss functions that aim at obtaining a more visually qualitative image, such as the Fréchet Inception Distance (FID), also fine-grained details could be accurately modeled.
In addition, the decision to use the affinely warped image as the reference image for the discriminator, to prevent it from focusing on potential artifacts introduced by the affine transformation, may have influenced the performance. While this approach helps mitigate the discriminator's sensitivity to registration artifacts, it introduces an inherent error, since the affinely warped image will always contains some degree of registration error and artificiality (not present in the real fixed image). Further investigation could address this problem by applying to the fixed image the same affine transformation encoded by the network and by selecting the resulting image as reference. It would be artificially modified to have the same artifacts that any image outputted from the affine network would have, so it could be possible to both prevent the model to focus on these artifacts and to drive the

adversarial learning to a realistic sample.
The results indicated that while the implementation of adversarial learning led to enhanced realism for larger regions, the modeling of deformations in smaller anatomical areas remained challenging. Consequently, the calculated metrics demonstrated a slight decrease in SSIM, and in DSC and NSD values for the liver. Comparing the outcomes to *Experiment 1*, the general image registration trend persisted even with the addition of adversarial learning: models equipped with skip layers achieved higher SSIM values, whereas the *Big-no-skip* model performed better in aligning larger regions, as reflected by higher DSC and NSD values.
Survival prediction was conducted by extracting imaging features from the bottleneck of the elastic network and utilizing them to train a RSF regression model, which estimated the 1-year survival risk factor. The predictor was tested only on the BL-FU1 scan-pairs (138 pairs) from the internal NKI dataset. The models achieved C-index values of 0.56 (*Big*), 0.61 (*Big-no-skip*) and 0.62 (*Small*); only the latter showed statistical significance. No correlation between image registration performance and C-index values was found. From the comparison with the baseline model, it can be concluded that the implementation of adverse learning has led to imaging features with lower prognostic value, and that led to distinguish the risk class of patients with less accuracy.

Chapter 6

Implementing the self-Attention Mechanism via Vision Transformer

6.1 Introduction

This chapter contains the description of *Experiment 3* and *Experiment 4*, where the registration module includes a Vision Transformer in the latent space of the elastic sub-network, in order to leverage the attention mechanism, which is explained in detail in Section 6.2. The baseline model is a CNN-based network, and consequently suffers of the inherent limitation of convolution operation, i.e the local receptive field. Despite its ability to actually perform the registration task, it is possible that the size of the local receptive field limits the performance of the model to establish the correspondence between the same anatomical structures of two images, especially when the same anatomical structure is distant [71].
Because of that, we assume that the integration of a attention-based model, which enables to model long-range spatial relations in data, can be beneficial to obtain more accurate deformation fields. In both *Experiment 3* and *Experiment 4* this mechanism is implemented, and they differ from each other in the use of adversarial learning. Similarly to the experiments presented in the previous chapters, they are also designed as an ablative study, performing three sub-experiments by modifying size and capacity of the elastic network.

The purpose of *Experiment 3* is to assess how the ability to consider features from spatially distant regions is reflected in registration performance, and whether this improvement actually leads to a more accurate survival prediction. Finally,

Experiment 4 is proposed to study the behavior of the architecture including CNNs, attention and adversarial mechanisms. For the already mentioned reasons, it is expected that the individual changes made to the network will improve the modelling ability of the deformation field. As a result, it is safe to assume that even their combination will benefit the ultimate goal.

6.2 Technical background

Transformers are self-attention-based architectures and in the recent years have become the models of choice in the field of Natural Language Processing (NLP), in particular in machine translation tasks [72].
These networks, proposed by Vaswani et al. in [73], mimic cognitive attention, whose purpose is to focus on the important parts of an input and to disregard information that is not relevant [74]. It is basically performed via dot product between different input subgroups (i.e, tokens), so specifically weighting the significance of each part of the input data: this is the key-operation of the self-attention mechanism.

Firstly input data are tokenized, so divided into multiple inputs $[x_1, x_2...x_i...x_n]$. Each input is multiplied with three matrices, W_Q, W_K and W_V, to derive three new vectors, called *queries*, *keys* and *values*, which will be used to calculate the attention weights. These matrices represent the controllable parameters during the training step. The attention function can be described as an operation that aims to weigh the importance of a token within the general set, and works sequentially on each token. Considering you want to determine the weight of the i-th input: the i-th query is multiplied by the j-th key vectors, the result w_{ij} is passed through a softmax, and then it is used in a weighted sum with the j-th values. In practice this operation is applied to all the queries simultaneously, packed together into the Q matrix. The keys and values are also packed together into matrices K and V. The mechanism can be summarized as:

$$Attention(Q, K, V) = softmax(\frac{Q \cdot K^T}{\sqrt{d_k}}) \cdot V \tag{6.1}$$

In Equation 6.1 d_k represents the dimensionality of queries and keys, and the scaling factor $\sqrt{d_k}$ is added to handle the cases of large d_k, which can lead to large magnitudes of dot products, pushing the softmax function into regions of extremely small gradients [73].
A variant of the above mentioned operation is the Multi-Head Self-Attention: the self-attention process is performed simultaneously on different *heads* (groups of tokens), and finally the results of each head are concatenated and projected to the initial dimension. The use of multiple heads allows the attention layer to attend to informations from different representation subspaces [73] and makes the process

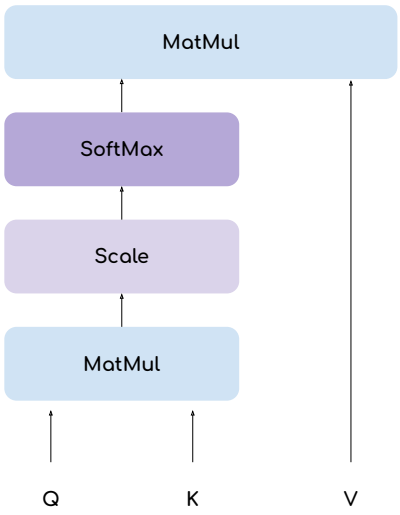


Figure 6.1: Illustration of attention mechanism. Image adapted from [73]

parallelizable, so faster.

Transformers are self-attention-based networks, whose blocks, repeated N times, essentially include multi-head self-attention operations, normalization layers, and feed-forward networks, as MultiLayer Perceptrons. These architectures can comprehend, as in CNNs, an encoder and a decoder part.

6.2.1 Transformers in Computer Vision

Inspired by the huge success of Transformers in NLP, in the recent years there has been an increasing interest in developing self-attention-based architectures in Computer Vision [75], as the so called *Vision Transformer*, published by [72]. As mentioned above, Transformers receive 1D sequence of token embeddings as input data. Vision Transformers, to meet this requirement, have to manipulate the input 2D images. The image are first divided in patches, and then each patch is flattened through a linear projection to a latent D-dimensional space. In order to retain positional information contained in the initial image, position embeddings are added to the patch embeddings [72]. Despite the preliminary data processing, Transformer architecture can be applied in the same way both in Computer Vision and in NLP. These models have proved their usefulness in different fields, such as that of Image Registration [75] [71] [76], showing better performances when integrating self-attention and CNN architectures: by implementing the self-attention mechanism it is possible to better model long-range spatial relations [72].

6.3 Methods

The datasets used in these experiments are the same of those of *Experiment 1*, described in Section 4.2. In this section we will discuss the methods of *Experiments 3* and *4*: as mentioned in Section 6.1 they differ from each other in the use of adversarial learning. In particular, in *Experiment 3* just a Vision Transformer is added in respect to the baseline model presented in Chapter 4, in *Experiment 4* both self-attention and a discriminator are used.

The choice of integrating a self-attention-based network with the baseline model creating a hybrid model, instead of replacing the entire convolutional structure with self-attention layers, is merely practical. In fact, the application of a naive Transformer to full-resolution volumetric images increase significantly memory and computational complexity [71]. So, it is added at the bottleneck of the elastic network, receiving in input the high-level features extracted from the deepest encoding layer. These feature maps have dimensions $fx12x12x18$: f is the number of features (dependent on the depth of the model of the sub-experiment performed), $12x12x18$ are the height, width and length of a single map as result of previous convolutions and max-pooling operations.

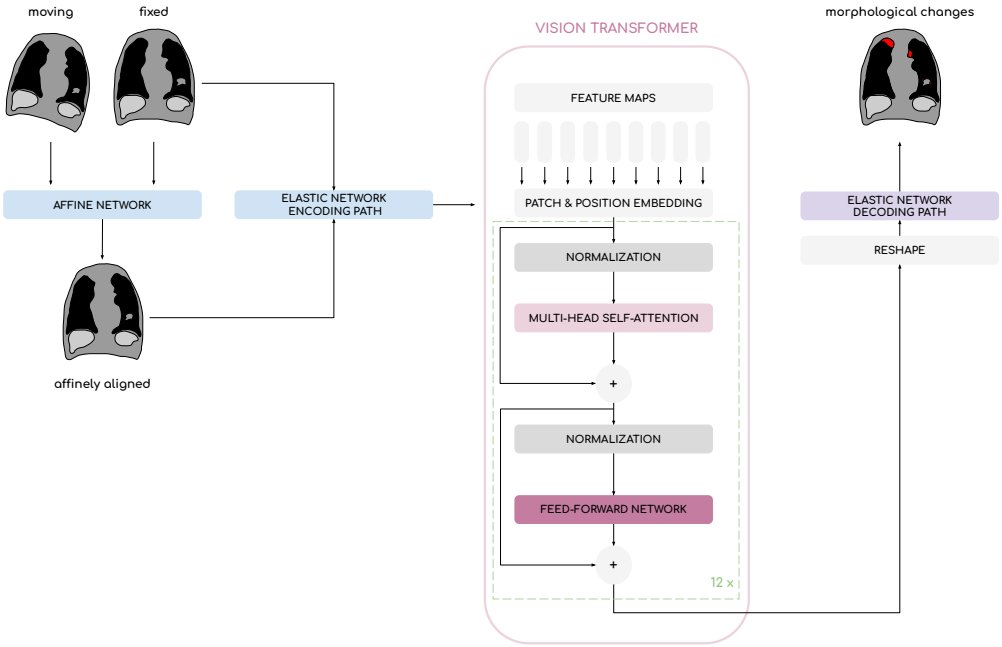


Figure 6.2: Representation of ViT-PAM

As visible in Figure 6.2, this block starts with patch and position embeddings.

Each feature map is split into $N = 12$ cubic patches of dimension $P^3 = 6x6x6$; then the patches are linearly projected into a D-dimensional space. D, also called embedding size, is dependent on the sub-experiment performed: for *Big* and *Big-no-skip* networks D was set to 4096, for the *Small* to 1024.
The choice of these values respects the proportionality between the depths of the models used in the ablative study: as well as the deeper features extracted in the big model are 4 times greater than those in the small one, here the embedding size chosen for the big model is 4 times greater than the one chosen for the small one.
After the linear projection, realized via a convolutional layer, position embeddings are added to retain positional information [75]. The resulting embeddings pass through 12 Transformer encoder blocks: each block is composed of two normalization layers, a Multi-Head Self-Attention layer, a Multi-Layer Perceptron and residual connections. The number of heads is set to 8.
At the end, the output of the Vision Transformer has dimensions NxD, so a reshape is performed to obtain the same shape of the input feature maps to start the decoding and enable the skip-connections, if applied. The reshape is realized via deconvolution, with a scale factor of (2,2,3). The resulting network is called ViT-PAM.

In *Experiment 3* this model is used for registration, and the loss function used during its training is the one showed in Equation 4.5. Otherwise, in *Experiment 4* the adversarial learning is used; so, as in *Experiment 2*, a discriminator is added and the combination of affine and elastic (here containing the ViT) is seen as a generator. The architecture of the discriminator and the loss function used are the same ones detailed in Chapter 5.

6.4 Results

The data filtering and preprocessing applied are the same of those of *Experiment 1*, so the resulting cohorts used for image registration (TCIA dataset) and prognostication (NKI dataset) are the ones already described in Section 4.4.1 and in Table 4.1.

6.4.1 Image Registration results - Experiment 3

The registration accuracy was assessed by performing the registration of the NKI dataset scan-pairs and by calculating SSIM, and DSC and NSD between the fixed and warped segmentation masks for the biggest and the smallest volumes. The metrics values averaged along the dataset are shown in Table 6.1.
The results of the three different models, which share the implementation of a Vision Transformer and differ from each other by size and capacity of the elastic sub-network, show a different trend than the architectures described in the previous

experiments. In both baseline and Adversarial PAM the most performative sub-experiment, in terms of DSC and NSD, was the one without the skip-connections. Here the *Big-no-skip* model suffers a noticeable drop in performance, not being able to effectively register smaller and/or larger anatomical structures. For the remaining models the metrics scores are comparable with those obtained in *Experiment 1*, with a slightly improvement in terms of SSIM. With the ViT-PAM architecture

	DSC_{liver}	DSC_{rib}	NSD_{liver}	NSD_{rib}	$SSIM$
Big	$89.6 \pm 6.1\%$	$65.3 \pm 23.8\%$	$72.5 \pm 14.1\%$	$78.2 \pm 27.2\%$	$84.4 \pm 4.1\%$
Small	$88.9 \pm 5.9\%$	$47.2 \pm 26.1\%$	$69.5 \pm 14.1\%$	$69.5 \pm 30.3\%$	$76.7 \pm 5.4\%$
Big-no-skip	$74.2 \pm 6.0\%$	$13.2 \pm 8.8\%$	$27.6 \pm 6.7\%$	$44.0 \pm 16.5\%$	$59.0 \pm 6.8\%$

Table 6.1: DSC, NSD and SSIM mean and standard deviation values for the three sub-experiments of *Experiment 3*

in mind, the trend reversal for the *Big-no-skip* model is not surprising. At the end of the encoding path of the elastic network, the transformer splits the input feature-maps into patches, uses them to weigh the correlations between different points of the image, and then outputs a NxD tensor. To start the decoding path with a tensor with the same dimension of the last feature-maps of the encoding path, a manual reshape was applied. It's possible that all the latent-space manipulations have somehow lost semantic information to the model, to preserve the spatial-relational information discovered by convolutional and self-attention layers. As a result, those models with skip connections had the chance to retrieve lost information to reconstruct semantically plausible images, while the *Big-no-skip* did not.

As in the previous chapters, it is advisable to verify that qualitatively the registrations follow the overall trend of the metrics calculated for the liver and the rib. Hence, in Figure 6.3 the registration of two randomly extracted scans from TCIA dataset is shown. For this example, the DSC_{liver} has been calculated between fixed and warped masks: the trend is similar to the one shown in Table 6.1, so it is a good example of the average differences between the sub-experiments.
The performance drop for the *Big-no-skip* model is confirmed also visually: it cannot distinguish different structures and fails in the task. The other two networks output warped images with similar features and drawbacks of the baseline ones: they provide an higher anatomical detail, leading to more realistic images, which at the same time can limit the uniformity of organs. This behavior is visible in the upper right part of the liver for both of them, and in the spleen for the *Small* network.

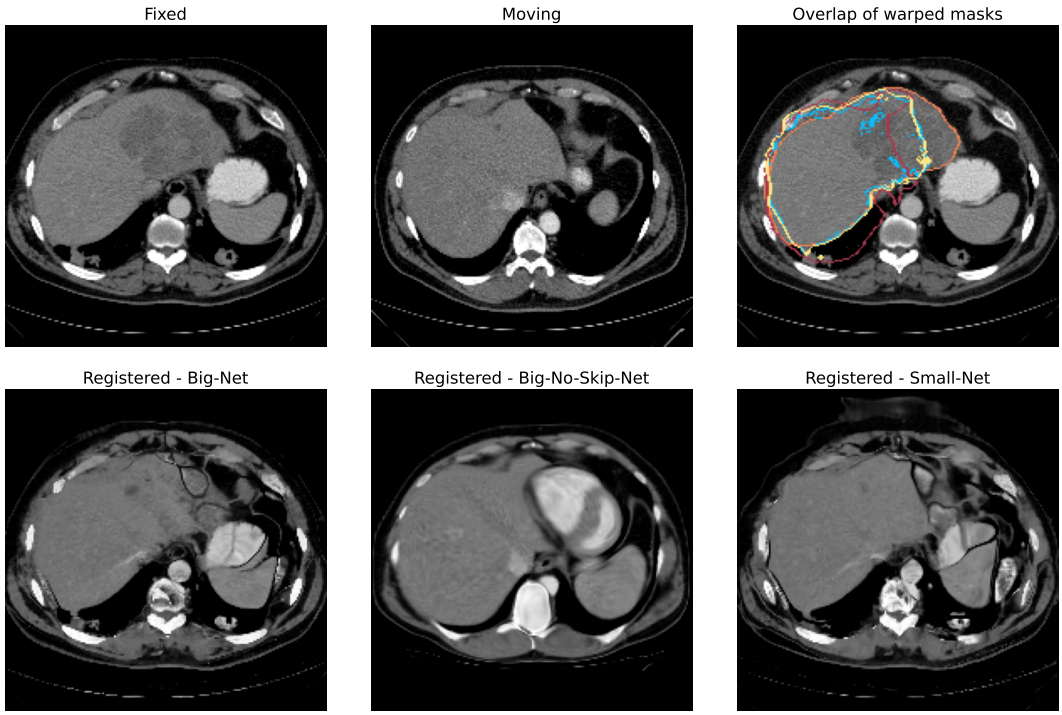
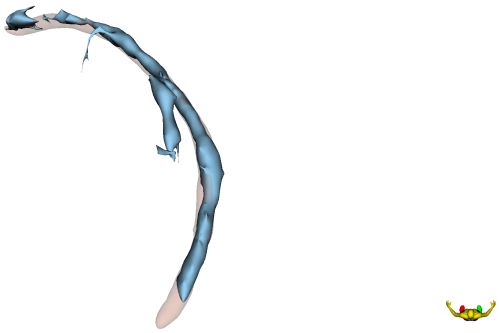


Figure 6.3: Qualitative comparison of registration performance for the three sub-experiments of *Experiment 3*. The figure shows an example of fixed - moving scan pair, the registered scan outputted by each model, and the contours of the liver segmentation masks registered by each model. The orange line is the ground-truth, blue refers to *Big*, amaranth to *Big-no-skip* and yellow to *Small*.

For completeness, in Figure 6.4 a qualitative comparative example between the rendering of 11^{th} left rib volumetric masks is shown. For each model in the figure a fixed-moving pair has been selected from the NKI dataset having a DSC_{rib} value equal to the average along the entire test dataset. This way it is possible to visualize the average behavior of the three networks.
Following the DSC and NSD metrics trends, the rib is better registered by the models implementing the skip connections: despite some false positive zones, where the predicted mask extends outside the ground-truth area, the warped segmentations are localized correctly. The *Big-no-skip* network shows a mean DSC_{rib} of 13.2% and a mean NSD_{rib} of 44.0%; the example displayed confirms its difficulty in registering the small structures, despite the appreciable similarity of shape and the spatial proximity of the two masks.

(a) *Big* **(b)** *Big-no-skip*



(c) *Small*

Figure 6.4: Rendering of 11^{th} left rib volumetric segmentation masks for the three sub-experiments of *Experiment 3*. In each subplot the blue volume represents the warped mask, which is superimposed to the light orange fixed mask. The symbol on the bottom right of the figures depicts the point of view in space: the subject is analyzed on the transverse plane, from top to bottom. The caption of each subplot refers to the network used for the registration. Image created with 3D-Slicer.

6.4.2 Prognostication results - Experiment 3

A RSF was trained to predict survival from the imaging features extracted from the registration module, which in turn inputs *prior* and *subsequent* scans of the same patient. The regression model outputs a mortality hazard related to the 1-years survival from the date of subsequent scan. To assess the prediction quality, the C-index was used and statistical significance was measured via log-rank test. As in the previous experiments, the performance were not assessed taking into account all the NKI independent test set, but only the baseline (BL) and first

follow-up (FU1) unique scans. Unique pairs were obtained by averaging imaging features of all the multiple pairs acquired on the same date. Results of the survival AI-score, in terms of C-index and statistical significance, are shown in Table 6.2. Confidence intervals were estimated via bootstrapping using repeated sampling with replacement (100 times).

	C-index	*p-value*
Big	0.57 (0.48 - 0.65)	6.2e−1
Small	0.56 (0.56 - 0.60)	4.2e−1
Big-no-skip	0.61 (0.50 - 0.67)	2.0e−1

Table 6.2: C-index, with relative confidence intervals, and p-values for the three sub-experiments of *Experiment 3*

Table 6.2 shows, as in the previous chapter, a clear drop in performance compared to the baseline. The imaging features extracted after the reshape of the ViT output, implemented in the bottleneck of the network, lead to C-index values between 0.56 and 0.61. All the models did not showed statistical significance ($p > 0.05$). There are no observable trends linking the registration and prediction tasks. For example, despite the different registration performance between the *Big* and *Small* models, highlighted by all the metrics used, their extracted features lead to very similar predictive accuracy. The model that does not implement skip connections, which fails in properly registering the scans, is the one providing the best prognostication.

6.4.3 Image Registration results - Experiment 4

The image registration quality was assessed by following the same protocol of the previous experiments. The metrics values averaged along the dataset are shown in Table 6.3.

	DSC_{liver}	DSC_{rib}	NSD_{liver}	NSD_{rib}	$SSIM$
Big	$86.0 \pm 6.8\%$	$50.8 \pm 23.7\%$	$62.9 \pm 13.7\%$	$63.8 \pm 26.6\%$	$81.3 \pm 3.9\%$
Small	$86.1 \pm 6.6\%$	$41.2 \pm 23.9\%$	$61.6 \pm 13.3\%$	$62.0 \pm 28.1\%$	$73.8 \pm 5.7\%$
Big-no-skip	$76.5 \pm 3.5\%$	$10.1 \pm 13.8\%$	$35.6 \pm 9.9\%$	$32.7 \pm 27.5\%$	$58.5 \pm 6.9\%$

Table 6.3: DSC, NSD and SSIM mean and standard deviation values for the three sub-experiments of *Experiment 4*

Table 6.3 shows that *Experiment 3* and *Experiment 4* share the same registration trend: the models implementing the skip-connections behave similarly in terms of DSC and NSD, while the *Big-no-skip* follows the same worsening of the ViT-PAM. The last column of the table, referring to SSIM values, shows the same trend of the

previous experiments: an higher ability of the network to model deformations is reached if skip connections are used. In Figure 6.5 the registration of two randomly extracted scans from TCIA dataset is shown, to perform a qualitative assessment. For this example, the DSC_{liver} has been calculated between fixed and warped

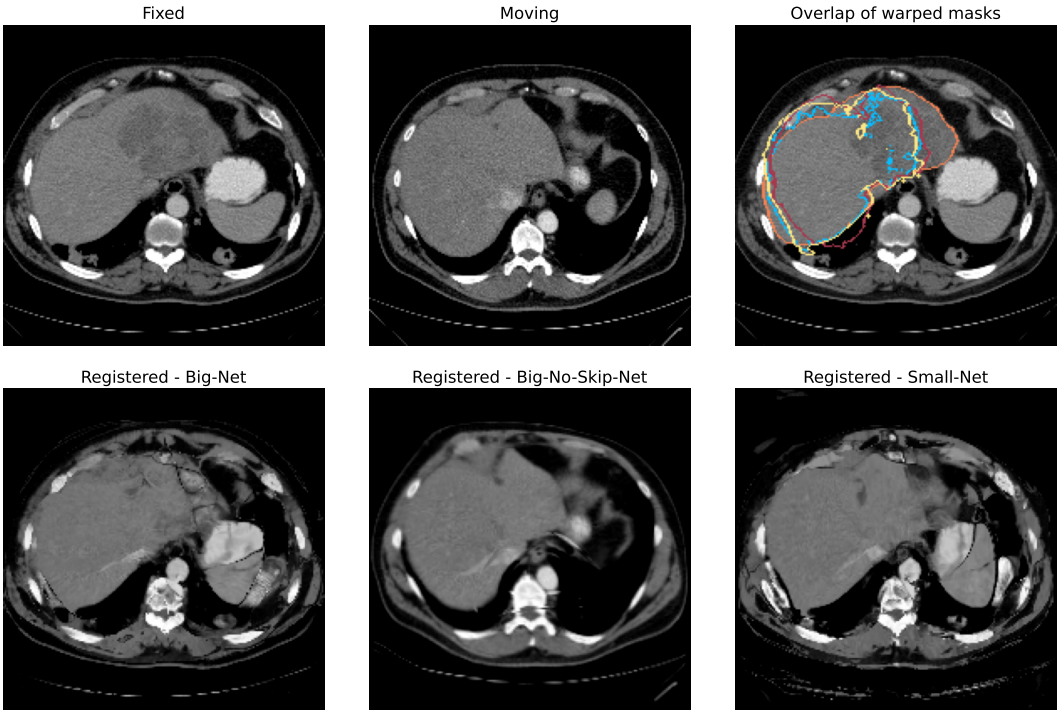
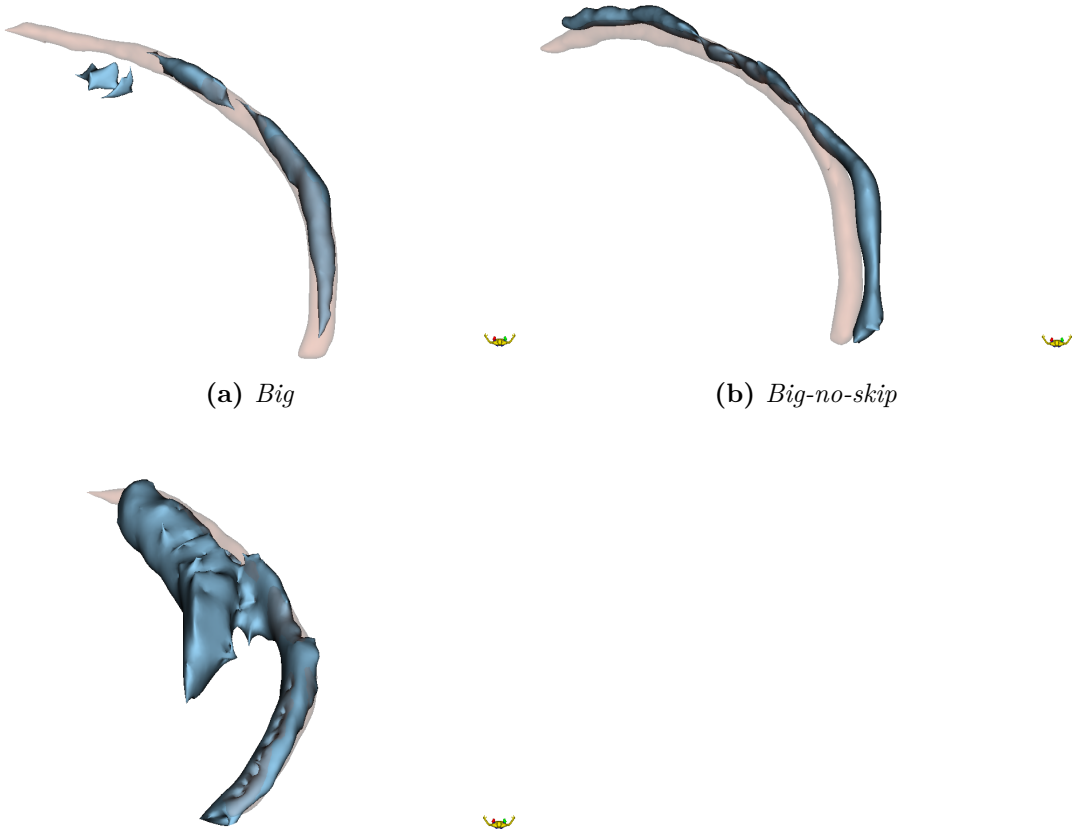


Figure 6.5: Qualitative comparison of registration performance for the three sub-experiments of *Experiment 4*. The figure shows an example of fixed - moving scan pair, the registered scan outputted by each model, and the contours of the liver segmentation masks registered by each model. The orange line is the ground-truth, blue refers to *Big*, amaranth to *Big-no-skip* and yellow to *Small*.

masks: the trend is similar to the one shown in Table 6.3, so it is a good example of the average differences between the sub-experiments. The performance drop for the *Big-no-skip* model is reflected also in the visual example, highlighting the difficulty of the network to reconstruct faithful and detailed images without skip layers, if self-attention mechanism is implemented. *Big* and *Small* models, as in *Experiment 3*, behave similarly, sharing an accurate registration of medium-large structures. However, despite the bigger network achieves higher realism in terms of structure integrity and detail, they both suffer from the same limitations described in *Experiment 2*: the integration of a discriminator in the model causes less plausible deformations for smaller structures, such as the stomach and the spleen.

For completeness, in Figure 6.6 a qualitative comparative example between the rendering of 11^{th} left rib volumetric masks is shown. For each model in the figure a fixed-moving pair has been selected from the NKI dataset having a DSC_{rib} value equal to the average along the entire test dataset. This way it is possible to visualize the average behavior of the three networks.



(a) *Big*

(b) *Big-no-skip*

(c) *Small*

Figure 6.6: Rendering of 11^{th} left rib volumetric segmentation masks for the three sub-experiments of *Experiment 4*. In each subplot the blue volume represents the warped mask, which is superimposed to the light orange fixed mask. The symbol on the bottom right of the figures depicts the point of view in space: the subject is analyzed on the transverse plane, from top to bottom. The caption of each subplot refers to the network used for the registration. Image created with 3D-Slicer.

The most performative version of the experiment is the *Big* one, with a mean DSC_{rib} of 50.8% and a mean NSD_{rib} of 63.8%, but despite its acceptable results it is unable to reliably reconstruct the bone. The qualitative examination shows the

limitations of the remaining two models: the DSC and NSD values of the *Small* network are obtained over-segmenting the rib with a high number of false-positive voxels, while the rib warped by *Big-no-skip* is anatomically plausible but not correctly aligned.

6.4.4 Prognostication results - Experiment 4

The survival prediction quality was assessed by following the same protocol of the previous experiments and by using the same RSF structure. Results of the survival AI-score, in terms of C-index and statistical significance, are shown in Table 6.2. Confidence intervals were estimated via bootstrapping using repeated sampling with replacement (100 times).

	C-index	*p-value*
Big	0.57 (0.48 - 0.63)	3.7e−1
Small	0.52 (0.43 - 0.59)	9.8e−1
Big-no-skip	0.66 (0.57 - 0.73)	3.0e−4

Table 6.4: C-index, with relative confidence intervals, and p-values for the three sub-experiments of *Experiment 4*

Table 6.4 shows similar results compared to the ones obtained by integrating just the ViT in the model. In fact, also here the *Big* and *Small* predictive performance are below 0.57, and the *Big-no-skip* model is the most performative one. Despite this, the addition of adversarial mechanism has led to appreciable changes in the prognostication. First, the difference in registration quality, in terms of SSIM, between the models implementing skip connections is also reflected in C-index values. The *Small* model, in fact, produces an almost random prediction. Moreover, although the *Big-no-skip* fails in the registration task, its features lead to an acceptable C-index value (0.66) and to statistical significance ($p < 0.05$).

6.5 Discussion

In this chapter, the integration of self-attention mechanism into the image registration network was achieved by incorporating a Vision Transformer in the bottleneck of the elastic network. This architectural choice aimed to address the inherent limitations of convolution-based networks, which are primarily designed to capture local features due to the size of the local receptive field. In contrast, self-attention mechanisms, as demonstrated by Vision Transformers [75] [71] [76], excel at capturing long-range spatial relationships [72]. Therefore, leveraging this mechanism within the convolutional elastic network was deemed advantageous for achieving

more accurate image registration by enabling the capture of correspondences between distant points in the images.
The chapter described two experiments: *Experiment 3* and *Experiment 4*, which differed in the presence of adversarial learning. It was hypothesized that both the self-attention mechanism and adversarial learning could individually enhance registration accuracy, and thus their combined integration (*Experiment 4*) was expected to further benefit the overall goal. Different models with varying size and capacity were defined for these experiments, as in the previous chapters. The results partially confirmed the assumptions made, shedding light on the complexity of the task at hand.
Firstly, both quantitative and qualitative results clearly demonstrated the failure of models lacking skip connections in achieving accurate registration. The *Big-no-skip* models exhibited a significant decline in metrics such as DSC, NSD, and SSIM, rendering them the least performant models. Conversely, the ViT proved beneficial for the task when skip connections were implemented, allowing the hybrid model to fully leverage the encoding capabilities of transformers without sacrificing semantic details. This improvement was evident in slightly higher SSIM values for both the *Big* and *Small* networks compared to the baseline. However, the integration of the adversarial loss, as observed in *Experiment 2*, did not significantly enhance the model's ability to obtain more realistic and accurate deformation fields. While visual examination of *Experiment 4* revealed more detailed images, suggesting an improvement in realism, it also highlighted the challenge of correctly registering small and localized areas. Consequently, the DSC and NSD values associated with the rib region were approximately 10% lower, and the SSIM values showed a slight decrease as well.
Survival prediction was conducted by extracting imaging features from the bottleneck of the elastic network, after the reshape of the ViT output embedding vector, and utilizing them to train a RSF regression model, which estimated the 1-year survival risk factor. The predictor was tested only on the BL-FU1 scan-pairs (138 pairs) from the internal NKI dataset. The results from the two experiments exhibited similar trends in terms of C-index values. In both *Experiment 3* and *Experiment 4*, the most performant model was the *Big-no-skip* architecture, achieving C-index values of 0.61 and 0.66, respectively. On the other hand, the *Big* and *Small* models demonstrated lower performance with C-index values below 0.60, which did not reach statistical significance. Despite the apparent trend that the model performing worst in the registration task was the most accurate in prediction, no correlation was found between the performance of the models in image registration and prognostication tasks.

Chapter 7

Additional Experiment: Enforcing Similarity in the latent-space

7.1 Introduction

This chapter contains the description of *Experiment 5*, where a new version of the baseline model is proposed, following the observations made in the previous experiments. In *Experiment 1* the registration network is trained mainly trying to minimize the dissimilarity between the moving image warped by the spatial transformer and the fixed image, along with a penalty term that enforces the estimated deformation field to be spatially smooth.
Despite the use of the overall loss function (Equation 4.5) has proven its effectiveness through satisfactory registration results, it is possible that driving the training by focusing only on warped images' features could limit the learning power of the network. Consequently, we assume that by adding a loss term related to the similarity between the encoded high-level features of reference (fixed) and target (registered) images, the registration network could reach a deeper learning and perform better. As in the previous chapters, also this additional experiment is designed as an ablative study, introducing variation in the registration quality and trying to link the registration performance to the ability of the model to predict survival. Unlike previous experiments, the three sub-experiments are defined by varying size of both affine and elastic sub-modules, instead of modifying size and capacity of just the elastic one. The objective of *Experiment 5* is therefore to confirm whether the addition of a loss term that enforces the similarity of latent-spaces leads to more accurate and faithful registrations.

7.2 Methods

The datasets used in this experiment are the same of those of *Experiment 1*, described in Section 4.2.

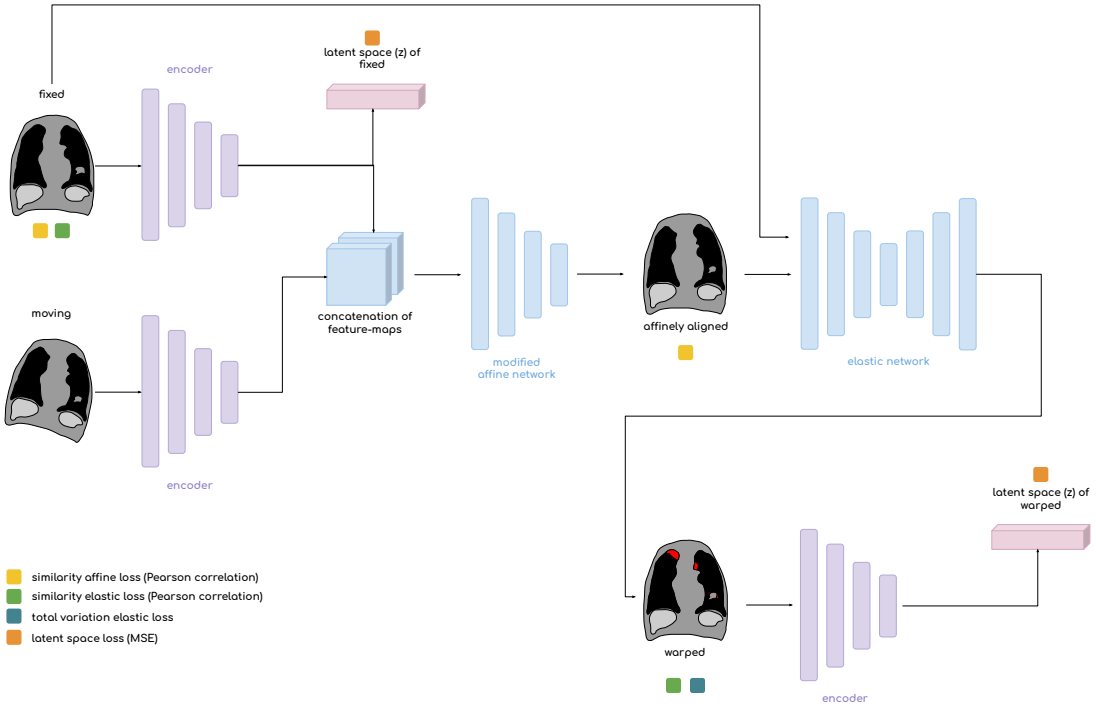


Figure 7.1: Representation of Split-Encoders PAM

In order to actively force the network to minimize the distance between the embedding spaces of fixed and warped image, the baseline architecture has been modified, especially in the block related to the affine transformation. In the model shown in Figure 4.1, the affine network was simply an encoder that received in input the concatenation of the images to register. In the proposed new version, called *Split-Encoders*, that initial block is replaced by two separate encoders, which encode separately the features of the images, and a final encoder deputy to regresses the 12 parameters of the rigid transform. Each of the two split encoders is made up of five convolutional blocks, followed by a fully connected layer, and outputs both a multi-dimensional feature map and a latent-space vector. As visible from Figure 7.1, the feature maps are concatenated together and input the affine network. The elastic network structure was not modified. Compared to the baseline, the new architecture also presents an encoder used to extract the latent-space of the final

warped image. Figure 7.1 shows also all the loss functions used in the training; the final loss function used to train the latent-space was the MSE.
The dimensions of the various blocks that make up the registration network depend on the specific sub-experiment performed. Table 7.1 shows an overview of the differences between the network parameters throughout the sub-experiments.

	Set of filters Split-Encoders	**Latent-space size**	**Set of filters Elastic network**
Big	[16, 32, 64, 128, 256]	512	[16, 32, 64, 128, 256]
Small	[16, 32, 64, 128, 256]	512	[4, 8, 16, 32, 64]
Smaller	[4, 8, 16, 32, 64]	128	[4, 8, 16, 32, 64]

Table 7.1: Split-Encoders and Elastic network parameters for the three sub-experiments of *Experiment 5*

As visible from Table 7.1, unlike the previous experiments, the ablation here does not focus exclusively on the elastic network size and capacity, but the aspects under investigation are the size of the embedding space and the number of elastic network's convolutional filters. As in the previous chapters, the *Big* version of the model is characterized by a higher number of convolutional filters. The *Small* one involves reducing the size of the elastic network while keeping the encoders' dimensionality unchanged. This sub-experiment serves two purposes. Firstly, it helps us examine how changes in the deformation representation size impact the quality of registration. Secondly, it allows us to measure the impact of the new architecture by comparing its results with those of the *Small* version of *Experiment 1*. On the other hand, the *Smaller* model involves further reducing the filters of the encoders used to obtain the latent-space of fixed, moving and warped images. This reduction results in a fourfold decrease in the length of the embedding vectors. We believe that comparing models with latent-spaces of different sizes is crucial as it helps us understand how deep the embeddings should be to capture the necessary informations for the task.
The influence of skip connections is not taken into account, as it was extensively investigated in the previous chapters, and as the core of the architecture does not rely on the elastic network.

7.3 Results

The data filtering and preprocessing applied are the same of those of *Experiment 1*, so the resulting cohorts used for image registration (TCIA dataset) and prognostication (NKI dataset) are the ones already described in Section 4.4.1 and in Table 4.1.

7.3.1 Image Registration results

The registration accuracy was assessed by performing the registration of the NKI dataset scan-pairs and by calculating SSIM, and DSC and NSD between the fixed and warped segmentation masks for the biggest and the smallest volumes. The metrics values averaged along the dataset are shown in Table 7.2.

	DSC_{liver}	DSC_{rib}	NSD_{liver}	NSD_{rib}	$SSIM$
Big	$92.0 \pm 4.0\%$	$75.3 \pm 13.1\%$	$78.5 \pm 11.5\%$	$90.4 \pm 13.8\%$	$85.8 \pm 3.7\%$
Small	$90.4 \pm 4.4\%$	$62.4 \pm 16.6\%$	$73.4 \pm 11.4\%$	$85.8 \pm 16.3\%$	$80.2 \pm 4.1\%$
Smaller	$89.4 \pm 4.8\%$	$49.7 \pm 23.1\%$	$69.4 \pm 12.4\%$	$72.9 \pm 24.6\%$	$81.2 \pm 4.3\%$

Table 7.2: DSC, NSD and SSIM mean and standard deviation values for the three sub-experiments of *Experiment 5*

According to Table 7.2, the most performative model is the *Big* one, which reached the highest values of all the metrics used. The accuracy in terms of DSC and NSD is directly proportional to the size of the overall network. In fact, the highest performances have been obtained with a greater number of convolutional filters, and these have suffered a decrease matching the downsizing of the network. This trend is also reflected in SSIM values, where a performance of 85.8% is reached for the *Big* model, and around 80% for less dense models.
The effectiveness of the new architecture and the additional latent-space-based loss function is easily seen from the comparison with the results obtained for the baseline model, showed in Table 4.2. In addition to an increase in performance in terms of SSIM and a more precise alignment of larger structures (such as the liver), there is a remarkable improvement in the reconstruction of smaller anatomical structures. In fact, for both *Big* and *Small* models, the results show DSC_{rib} and NSD_{rib} values increased by 10%.

As in previous chapters, for a complete understanding of the registration performance of the three models, a qualitative evaluation of the task is carried out. The same randomly extracted scan pair from TCIA dataset is shown in Figure 7.2. For this example, the DSC_{liver} has been calculated between fixed and warped masks: the trend is similar to the one shown in Table 7.2, so it is representative of the average behaviour of the three models. The SSIM results, which measure the registration quality of the entire volumes, are perfectly reflected in the visual example shown. All three sub-experiments are able to reconstruct the CT scan satisfactorily (with SSIM values above 80%), and the major differences between the models lie in the anatomical faithfulness of the smaller structures. In fact, although the liver is warped in a very similar way a priori by the size of the network, the stomach and the spleen appear more fragmented and less realistic when fewer

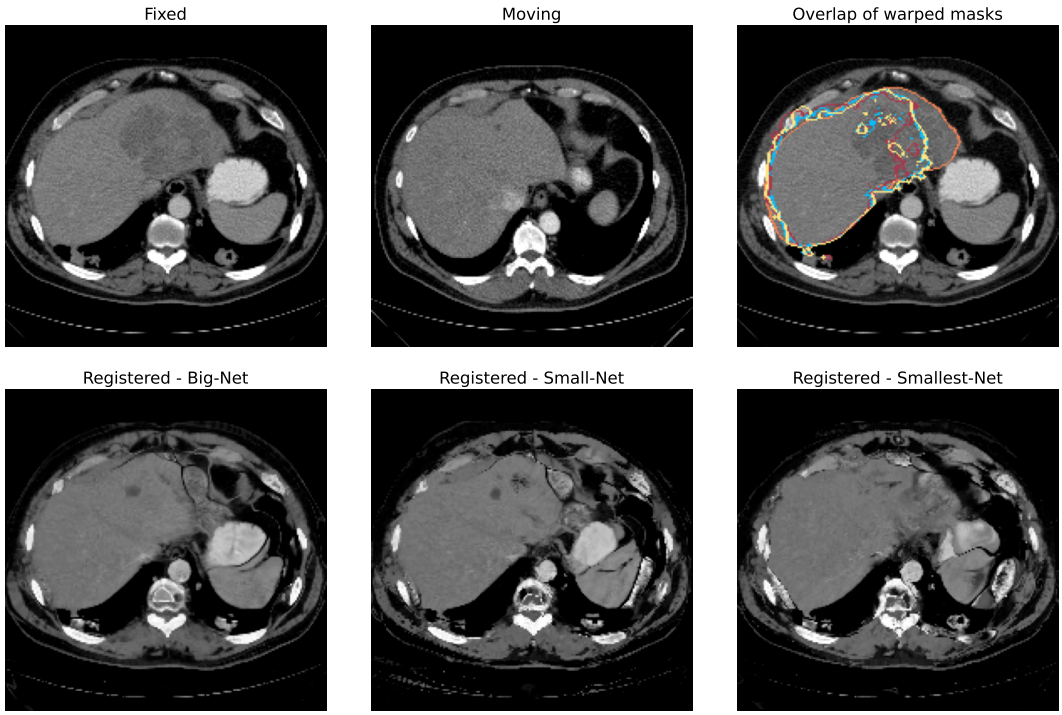
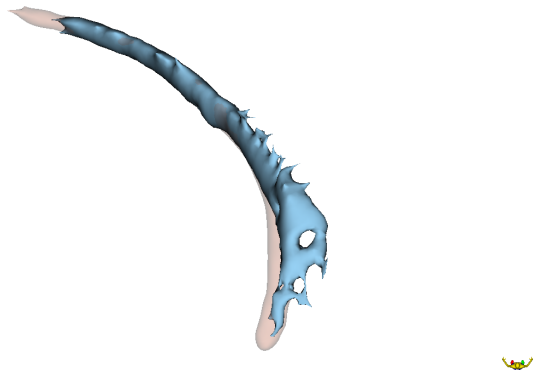


Figure 7.2: Qualitative comparison of registration performance for the three sub-experiments of *Experiment 5*. The figure shows an example of fixed - moving scan pair, the registered scan outputted by each model, and the contours of the liver segmentation masks registered by each model. The orange line is the ground-truth, blue refers to *Big*, amaranth to *Smaller* and yellow to *Small*.

convolutional filters are used for encoding features. However, Figure 7.2 confirms also qualitatively that the implementation of the new loss function is beneficial for the task and helps to overcome the performance obtained by the baseline.

For completeness, after having shown in Figure 7.2 the liver masks produced by the three registration models, in Figure 7.3 a comparative example between the 11^{th} left rib masks is shown. For each model in the figure a fixed-moving pair has been selected from the NKI dataset having a DSC_{rib} value equal to the average along the entire test dataset. This way it is possible to visualize the average behavior of the three networks. Following the trend of DSC and NSD shown in Table 7.2, *Big* model is able to realistically reconstruct the bone structure of the rib with high accuracy, while *Small* and *Smaller* underperform with a greater number of false negative voxels, even though the overall shape is maintained.

(a) *Big* **(b)** *Small*



(c) *Smallest*

Figure 7.3: Rendering of 11^{th} left rib volumetric segmentation masks for the three sub-experiments of *Experiment 5*. In each subplot the blue volume represents the warped mask, which is superimposed to the light orange fixed mask. The symbol on the bottom right of the figures depicts the point of view in space: the subject is analyzed on the transverse plane, from top to bottom. The caption of each subplot refers to the network used for the registration. Image created with 3D-Slicer.

7.3.2 Prognostication results

The survival prediction quality was assessed by following the same protocol of the previous experiments: a RSF regression model was used and the testing was performed solely on the BL-FU1 unique scan-pairs from the internal NKI dataset. Results of the survival AI-score, in terms of C-index and statistical significance, are shown in Table 7.3. Confidence intervals were estimated via bootstrapping using repeated sampling with replacement (100 times).

	C-index	*p-value*
Big	0.70 (0.63 - 0.78)	9.0e−4
Small	0.65 (0.56 - 0.75)	1.3e−2
Smaller	0.66 (0.57 - 0.77)	1.5e−1

Table 7.3: C-index, with relative confidence intervals, and p-values for the three sub-experiments of *Experiment 5*

Table 7.3 shows the effectiveness and usefulness of the new proposed architecture in the survival prediction task. In fact, all three experiments lead to a C-index value of more than 0.65, outperforming the baseline with the *Big* model and behaving similarly with the smaller ones. By splitting the test set into two groups with different risk levels (median of AI-scores), and using the log-rank test, it was possible to estimate that this division is statistically significant for *Big* and *Small* models ($p < 0.05$). The prognostication trend seems to follow that of registration quality, if SSIM is taken into account. Both SSIM and C-index values, in fact, reach a peak with the *Big* model, while have very similar performance for smaller ones. The large gap between *Small* and *Smaller* models in the registration of minor anatomical structures in terms of DSC and NSD (greater than 10%) has an influence on statistical significance but not on C-index values.

7.4 Discussion

In this chapter, an additional experiment called *Experiment 5* is introduced, aimed at improving the registration performance compared to the baseline model. Since none of the previous experiments achieved significantly higher registration results than the baseline, we decided to modify the underlying structure of the registration module to assess the impact of these changes on the performance. The baseline model primarily focused on minimizing dissimilarity between the moving and fixed images, without considering similarity in the latent-space features. We hypothesized that enforcing similarity in the high-level features of reference and target images could enhance learning for the task.
To achieve this, some adjustments were made to the architecture. The first block of the pipeline was replaced with two *Split Encoders*, each taking a single image (either fixed or moving) as input. These encoders produced an embedding vector and feature-maps after the encoding process. Additionally, a slightly modified version of the affine sub-network was used. The ablation in this experiment did not involve altering the capacity of the elastic network. Instead, three sub-experiments were conducted, varying size of both the embedding size and the deformation field representation.

Experiment 5 effectively surpassed the performance of the baseline model, indicating that incorporating a latent-space similarity loss, alongside an image similarity loss, contributed to a more robust registration process. Notably, the most significant improvement was observed in the registration accuracy of smaller regions, with DSC and NSD values of the rib increasing by 10% compared to the baseline for both the *Big* and *Small* models. Moreover, the DSC and NSD values for the liver and SSIM values also exhibited notable improvements, achieving the highest values in the entire study. A trend can be observed from the comparison of the three sub-experiments in this chapter: the performance in terms of DSC and NSD is directly proportional to the size of the network. In other words, a larger network capable of extracting more abstractions during feature encoding results in higher accuracy in structure alignment. Additionally, all three models achieved SSIM values above 80%.
Survival prediction was conducted by extracting imaging features from the bottleneck of the elastic network and utilizing them to train a RSF regression model, which estimated the 1-year survival risk factor. The predictor was tested only on the BL-FU1 scan-pairs (138 pairs) from the internal NKI dataset. The prognostication results aligned with the trend observed in the registration task, with larger models achieving higher prediction accuracy. The *Big* model attained a C-index of 0.70, followed by the *Small* and *Smaller* models with values of 0.65 and 0.66, respectively. All three versions demonstrated C-index values exceeding 0.65, surpassing the performance of previous experiments and emphasizing the prognostic value of the *Split Encoders* architecture.

Chapter 8

Linking registration to prognostic performance

Previous studies postulated the prognostic value of tracking all anatomical changes between follow-up scans of the same patient receiving anti-cancer treatment using image-to-image registration [34] [51] [52] [37]. In this thesis, we investigated the relationship between the ability of the network to model morphological changes through image-to-image registration in serial CT images of the same patient and the ability of the same registration features describing the deformation field to predict survival. To prove this hypothesis, we introduced variations in registration quality by means of ablation of the network architecture, and examine their potential correlation with the survival prediction accuracy. In particular, four experiments were performed, each implementing a different registration strategy, which included mechanisms of adversarial loss and self-attention mechanism, and each architecture was defined in three versions, varying the size and capacity of the elastic sub-network.

Image-to-image registration

The performance of image-to-image registration was assessed using Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD) between the fixed and warped segmentation masks of the liver and the 11^{th} left rib, and SSIM to evaluate the overall volumetric similarity. A qualitative assessment of registration quality was also conducted to evaluate the anatomical plausibility and realism of the registered images. As expected, the ablation of size and capacity of the registration module led to variations in the registration quality.
In *Experiment 1* the baseline model was employed, and it was observed that the *Big* network achieved the most accurate volumetric registration based on higher SSIM values. However, when it came to precise structure alignment, assessed by

DSC and NSD, the *Big-no-skip* model outperformed the others. This suggests that skip-connections may not be crucial for achieving precise alignment but are beneficial for overall volumetric registration accuracy and perceptive quality. *Experiment 2* introduced the adversarial learning, and its integration with the baseline model enhanced the anatomical realism as assessed through visual examinations, but slightly worsened the registration of smaller areas of the scan. *Experiment 3* and *Experiment 4* focused on the integration of a Vision Transformer in the bottleneck of the elastic sub-network, creating a hybrid model. Surprisingly, the *Big-no-skip* models in both experiments exhibited a drop in registration performance compared to the previous ones. This drop was evident in lower DSC, NSD and SSIM values, as well as in qualitative assessments. In contrast, the combination of self-attention and skip layers was beneficial for the task, since it allowed to fully exploit the encoding potentialities of Transformers without losing any semantic detail. It resulted in improved overall structural similarity, suggesting that self-attention mechanism can enhance the ability of the model to capture long-range dependencies and improve the alignment of structural features.

Since none of the described experiments reached significantly higher registration performance than the baseline, and following practical observation built during the development process, an additional experiment was executed, called *Experiment 5*. Unlike the previous experiments, which focused on variations in the elastic sub-network, *Experiment 5* sought to investigate the effects of the latent-space similarity loss on achieving accurate registration. To assess the influence of enforcing similarity in the latent-space of fixed and warped images, some adjustments were made to the registration architecture. The first block of the pipeline was replaced with two *Split Encoders*, each taking a single image (either fixed or moving) as input. These encoders produced an embedding vector and feature-maps after the encoding process. Additionally, a slightly modified version of the affine sub-network was used.
The ablation in this experiment did not involve altering the capacity of the elastic network. Instead, three sub-experiments were conducted, varying size of both the embedding size and the deformation field representation. *Experiment 5* effectively surpassed the performance of the baseline model, indicating that incorporating a latent-space similarity loss, alongside an image similarity loss, contributed to a more robust registration process. Notably, both large and small versions exhibited higher SSIM values, as well as improved DSC and NSD scores. A noticeable enhancement in the registration accuracy was showed for smaller structures, such as the rib.

By comparing all the experiments conducted, the image-to-image registration network introduced in the additional experiment, featuring the *Split Encoders* and the incorporation of latent-space similarity, overall outperformed the other

architectures. In particular, the most performant model was the *Big* model, which achieved a mean SSIM value of nearly 86% across the internal NKI dataset. Figure 8.1 provides a visual example, showcasing two scans of the same patient acquired at different time-points (fixed and moving), along with the affinely warped image and the final registered image generated by that model. The plot demonstrates the model's remarkable capability to perform accurate registration, faithfully recreating the shapes and details of large organs (such as the liver and stomach) as well as smaller organs (such as the spleen and colon) with high precision and realism.

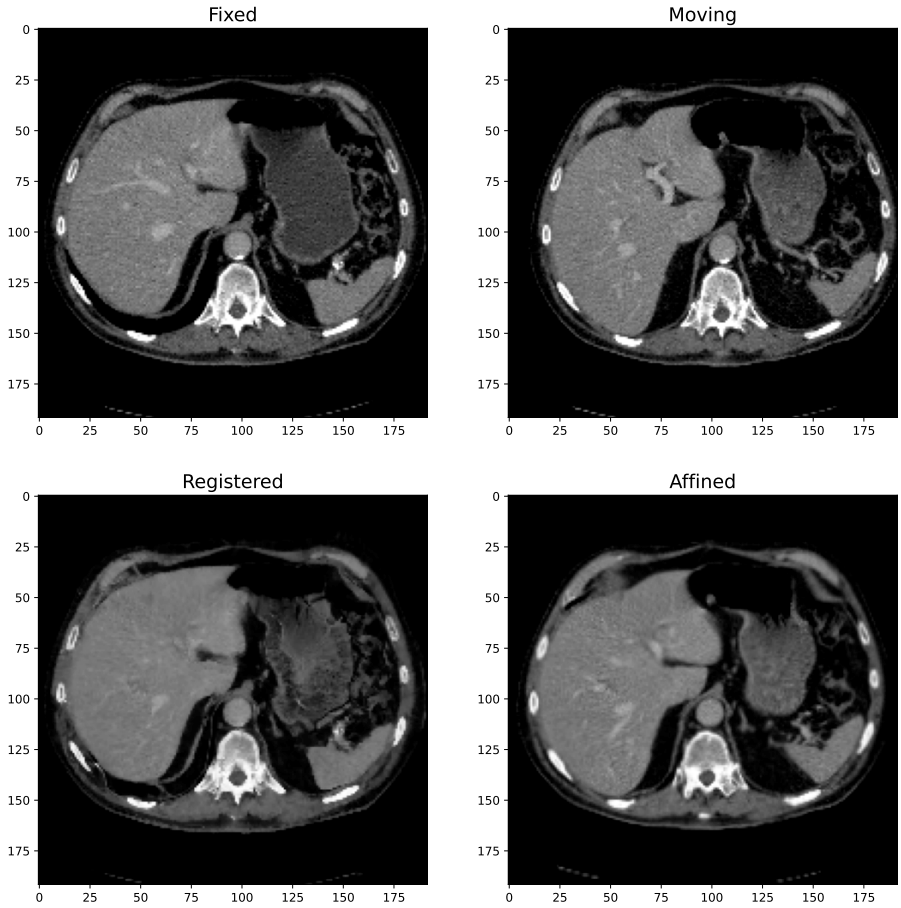


Figure 8.1: Visual example of registration performance for the *Big* model of *Experiment 5*. Fixed and moving images belong to the same patient

Survival prediction

The survival prediction task was carried out by training a RSF from the high-level imaging features extracted from the registration module, which in turn inputs two scans of the same patient acquired in different time-points, called *prior* and *subsequent* scans. The regression model predicted max 1-year survival from the subsequent scan date of the input prior-subsequent pair. To this end, an internal dataset embedded with follow-up images and survival times was used. For simplicity, and to obtain a clearer statistical signal, testing was performed on only BL-FU1 scans, which represent a critical time interval where the effects of immunotherapy are more pronounced and clinically relevant. To assess prognostication quality, the C-index was calculated and statistical significance was measured via log-rank test between the highest and lowest risk groups, defined by the median of the predictions.

The metric measured for the several experiments did not exhibit a consistent trend. The *Small* model demonstrated the highest predictive power in the first two experiments, achieving C-index values of 0.69 and 0.62, whereas the *Big-no-skip* models, when ViT was integrated, demonstrated superior accuracy in *Experiment 3* and *Experiment 4*, with C-index values of 0.61 and 0.66, respectively. *Experiment 5* sub-experiments led to highest predictive accuracy, reaching a C-index between 0.65 and 0.70. Comparing the influence of different architectures and mechanisms on the prognostication ability, both the baseline and *Split Encoders* structures showcased the highest performance, while the integration of the ViT resulted in performance deterioration.

In addition to the predictive analysis based on the C-index, the prognostic value of the most accurate model (*Big* model from *Experiment 5*, achieving a C-index of 0.70) was further investigated using Kaplan-Meier curves and Cox time-varying regression analysis. The Kaplan-Meier plot was generated by dividing the test set into two risk groups (high risk and low risk) based on the median AI-risk score generated by the RSF model. As shown in Figure 8.2, the plot indicates that the model effectively distinguishes patients at different risk levels based on imaging features.
The Cox time-varying regression model was employed to assess the relationship between the AI risk-score and patient survival likelihood, considering other relevant factors present in the data, such as pathology and therapy information. The results of the analysis are presented in Figure 8.3. The figure demonstrates that the AI risk-score exhibits a significant association with survival ($log(HR) > 0$), along with factors such as opioids-intake and cancer-type (breast cancer).
However, it is important to note that these findings do not necessarily imply that patients diagnosed with breast cancer and undergoing opioids-intake are the only

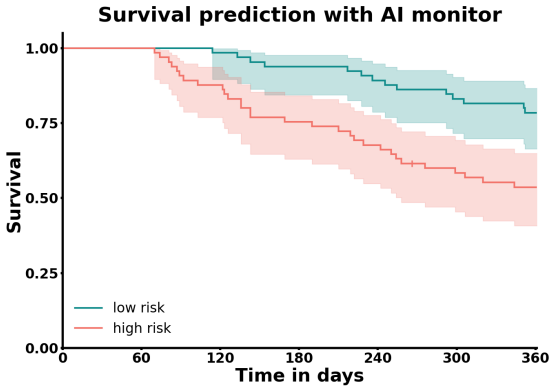


Figure 8.2: Kaplan-Meier curves for the three categories based on the risk score provided by the *Big* model of *Experiment 5*

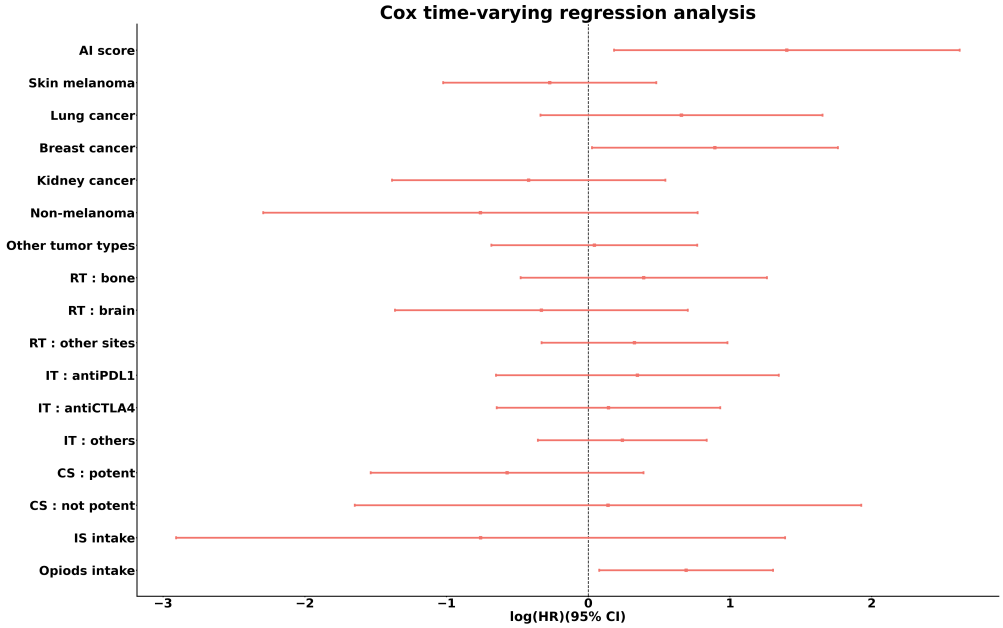


Figure 8.3: Cox time-varying regression analysis for the prognostic AI-risk score provided by the *Big* model of *Experiment 5*. Cofactors used in the analysis include: pathology description, radiotherapy (RT) site, immunotherapy (IT) medication, corticosteroid (CS) type, immunosuppressant (IS) and opiods intake

ones at a statistically higher risk of death. In reality, the cancer-types indicated in the analysis may not exclusively represent the tumors for which patients were receiving treatment; they could also include past pathologies that have already been treated. Therefore, since immunotherapy is generally not the primary treatment for breast cancer [77], it is likely that patients with breast cancer as a cofactor had a pre-existing oncological history and were simultaneously undergoing treatment for another pathology, which would increase their overall risk of death. Additionally, it should be noted that opioids are typically prescribed by physicians to patients with advanced stages of the disease to alleviate the pain associated with ongoing therapies and advanced cancer spread [78]. Given the complex and large dataset analyzed, their clinical correlations need to be scrutinised critically before drawing any conclusion.

Relationship between image registration and prognostication quality

To comprehensively evaluate the relationship between the network's capability to model radio-anatomical changes over time and its ability to predicting survival, a correlogram was utilized. It is showed in Figure 8.4. The correlogram analysis provided insights into the interdependencies of the different evaluations metrics used in the study, allowing for a rigorous examination of the inter- and intra- associations of the two tasks. All sub-plots, except those in the main diagonal of the matrix, depict a scatter plot and the resulting linear regression. The figures in the main diagonal show the distribution of the individual variables. Not surprisingly, the metrics employed to measure registration performance exhibit a notable linear correlation, indicating that a strong capability to accurately register individual structures (high DSC and NSD) corresponds to a high accuracy for the entire volume (high SSIM), and vice versa. Similarly, the metrics associated with survival prediction accuracy and robustness, such as C-index and p-value (p), demonstrate a similar trend. Hence, a higher C-index implies a greater likelihood for the model to be statistically significant ($p < 0.05$). In Figure 8.4, instead of p values, $-log(p)$ are reported.

Despite this, as visible from the two last rows of the correlation plot, registration and prognostication qualities are uncorrelated, having a similar C-index or p for different registration qualities. These findings suggest that the factors influencing successful image registration may not directly align with the factors associated with accurate survival prediction. In other words, our results suggest that an improvement in the tracking of anatomical changes between serial images of the same patient will not result in an improvement of the prognostic performance. Although this result may be influenced by certain choices made in the study, such as the use of DSC, NSD and SSIM as evaluation metrics for registration performance, and further investigations

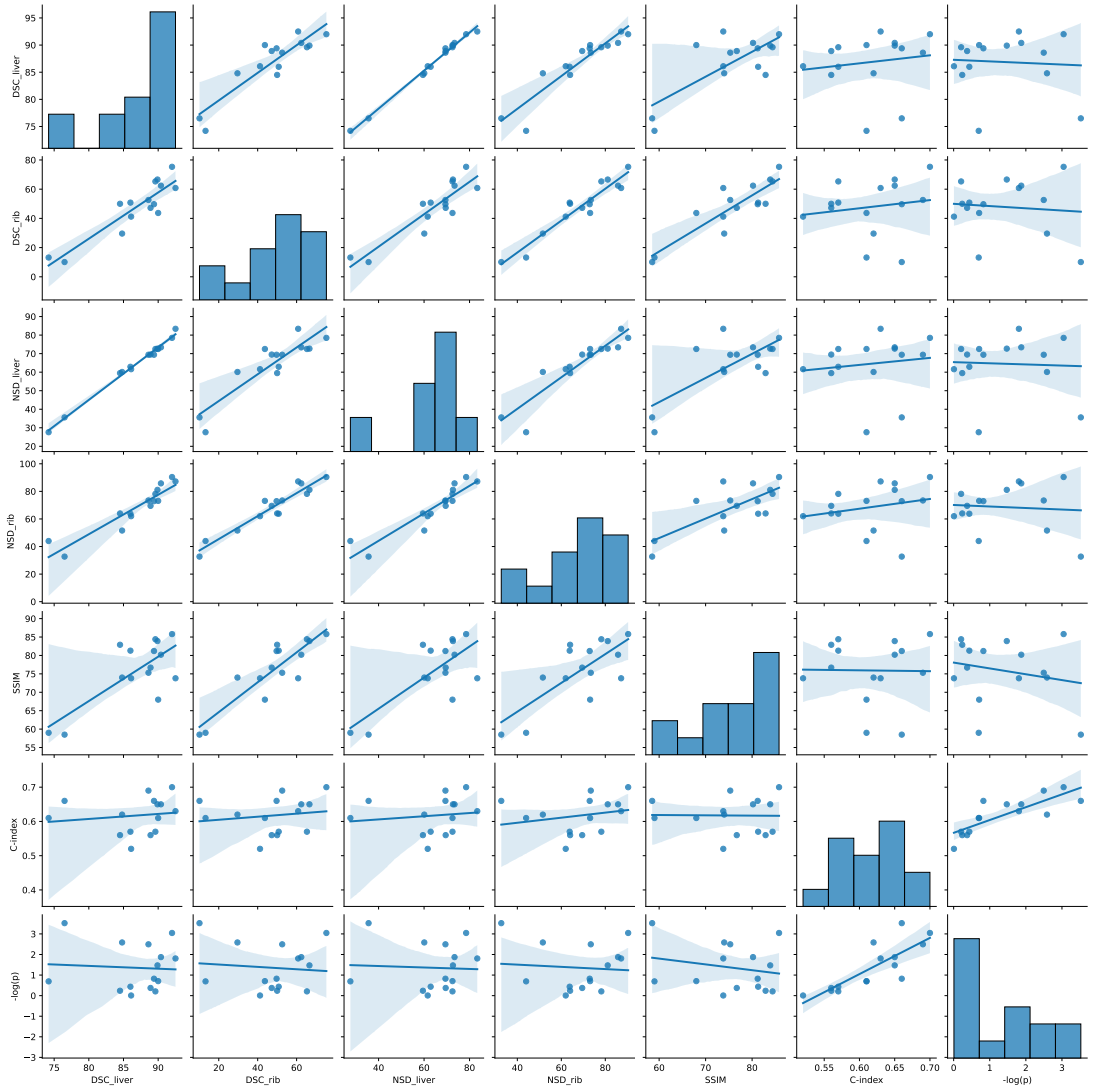


Figure 8.4: Correlation plot

could address it by using different parameters, it is reasonable considering the complexity and diversity of the two tasks. Even though prognostication involves the identification and the tracking of all the radio-anatomical changes that occur during the treatment, predicting patient survival entails analyzing a variety of clinical and biological factors that may extend beyond the morphological changes captured by image registration. Survival prediction often relies on a combination of clinical variables, genetic markers, treatment information, and other non-imaging data.

The proposed model used unsupervised image registration, meaning that no explicit information about the location and extent of tumors or other specific features is provided during the task. Using thoraco-abdominal CT scans, a multitude of changes can occur, but not all of these changes may have equal prognostic relevance. It is possible that the image registration module does not extract a sufficient number of prognostic features. The RSF classifier is responsible for distinguishing input features that are more or less relevant for survival. However, if very large volumes are registered compared to the size of the tumor lesion, or if all scan areas are given equal weight without considering the localization of the tumor, the vector input to the classifier may contain an excessive number of non-prognostic factors. As a result, the lack of linear link between registration performance and survival prediction accuracy could be attributed to the loss of potentially prognostic local features in the registration module, which are not considered during classification.

Future work

Despite the excellent results obtained in terms of registration ($SSIM = 85.8\%$, $DSC_{liver} = 92.0\%$, $NSD_{rib} = 90.4\%$) and the promising survival prediction results ($C - index = 0.70$), these findings suggest that the framework currently in use cannot lead to significant prognostication improvements. A direct link between image registration and prediction quality would have suggested to focus on a more optimal and accurate registration, as an improvement would have brought an improvement in the final task also. Since this linear correlation has not been demonstrated, it is difficult to predict how to improve the prognostication algorithm by continuing to work on the modelling of radio-anatomical changes via image-to-image registration. Therefore, to support treatment-response assessment in oncology more robustly, it may be necessary to find alternatives to the unsupervised image-to-image registration model presented in this study.

A potential improvement could involve providing the model with additional information in a semi-supervised manner. For example, incorporating information about the location of tumor burden or other clinically relevant features during the registration process could enhance the model's ability to capture the most important morphological changes related to prognosis.
Or, alternatively, future developments could focus on applying the pipeline used in this study to specific patches containing cancer, combining segmentation and registration operations together. Segmentation of the area of the scan containing the tumor mass, provided by an operator or automatically obtained from an AI model, would allow to choose with accuracy the patch to be registered. As said, it is possible that the registration of such extensive scans can lead to a deformation field

rich in non-prognostic changes. In contrast, identifying critical areas a priori may increase the likelihood of detecting deformations directly related to the patient's survival.
The segmentation operation could also be used independently of the registration operation, proposing a totally different approach than the one proposed. The lesions present in the scans could be segmented, also possibly in combination with masks related to specific target organs, and could be used to carry out volumetric monitoring over time. As a result, the measurement of segmented volumes would be used to predict prognosis and survival, extending RECIST criteria to the entire tumor burden. The use of segmentation would certainly allow a more intuitive and fast approach, as well as more interpretable and explainable.

Bibliography

[1] *Cancer*. URL: https://stanfordhealthcare.org/medical-conditions/cancer/cancer.html.

[2] AdmacOncology. *What are the differences between malignant and benign tumours?* Oct. 2021. URL: https://www.admaconcology.com/2021/06/10/malignant-vs-benign-tumor-know-the-differences/.

[3] Aisha Patel. «Benign vs malignant tumors». In: *JAMA oncology* 6.9 (2020), pp. 1488–1488.

[4] Douglas Hanahan and Robert A Weinberg. «The hallmarks of cancer». In: *cell* 100.1 (2000), pp. 57–70.

[5] Douglas Hanahan and Robert A Weinberg. «Hallmarks of cancer: the next generation». In: *cell* 144.5 (2011), pp. 646–674.

[6] Mehdi Astaraki. «Advanced machine learning methods for oncological image analysis». PhD thesis. KTH Royal Institute of Technology, 2022.

[7] *European Cancer Information System*. URL: https://ecis.jrc.ec.europa.eu/index.php.

[8] *Understanding your options and making treatment decisions*. URL: https://www.cancer.org/treatment/treatments-and-side-effects/planning-managing/making-decisions.html.

[9] Vincent T DeVita, Theodore S Lawrence, and Steven A Rosenberg. «DeVita, Hellman, and Rosenberg's cancer: principles & practice of oncology». In: (2015).

[10] *External beam radiation therapy for cancer*. URL: https://www.cancer.gov/about-cancer/treatment/types/radiation-therapy/external-beam.

[11] *Systemic therapy options for lung cancer*. Nov. 2020. URL: https://www.foxchase.org/clinical-care/conditions/lung-cancer/treatment-lung-cancer/systemic-therapy.

[12] Peter Nygren. «What is cancer chemotherapy?» In: *Acta Oncologica* 40.2-3 (2001), pp. 166–174.

[13] *Chemioterapia*. URL: https://www.airc.it/cancro/affronta-la-malattia/guida-alle-terapie/chemioterapia.

[14] Charles Sawyers. «Targeted cancer therapy». In: *Nature* 432.7015 (2004), pp. 294–297.

[15] *Targeted therapy for cancer*. URL: https://www.cancer.gov/about-cancer/treatment/types/targeted-therapies.

[16] Alexandre André da Costa, Dipanjan Chowdhury, Geoffrey I Shapiro, Alan D D'Andrea, and Panagiotis A Konstantinopoulos. «Targeting replication stress in cancer therapy». In: *Nature Reviews Drug Discovery* (2022), pp. 1–21.

[17] *Treatment for cancer: Cancer treatment options*. URL: https://www.cancer.org/treatment/treatments-and-side-effects/treatment-types.html.

[18] K Esfahani, L Roudaia, Net al Buhlaiga, SV Del Rincon, N Papneja, and WH Miller. «A review of cancer immunotherapy: from the past, to the present, to the future». In: *Current Oncology* 27.s2 (2020), pp. 87–97.

[19] *Immunotherapy*. URL: https://www.avl.nl/en/information-about-cancer/overview-of-all-treatment-options/immunotherapy/.

[20] Alex D Waldman, Jill M Fritz, and Michael J Lenardo. «A guide to cancer immunotherapy: from T cell basic science to clinical practice». In: *Nature Reviews Immunology* 20.11 (2020), pp. 651–668.

[21] Krupa Naran, Trishana Nundalall, Shivan Chetty, and Stefan Barth. «Principles of immunotherapy: implications for treatment strategies in cancer and infectious diseases». In: *Frontiers in microbiology* 9 (2018), p. 3158.

[22] Yiping Yang et al. «Cancer immunotherapy: harnessing the immune system to battle cancer». In: *The Journal of clinical investigation* 125.9 (2015), pp. 3335–3337.

[23] Malcolm Brigden and Michael McKenzie. «Treating cancer patients. Practical monitoring and management of therapy-related complications.» In: *Canadian Family Physician* 46.11 (2000), pp. 2258–2268.

[24] Henry N Wagner Jr and Peter S Conti. «Advances in medical imaging for cancer diagnosis and treatment». In: *Cancer* 67.S4 (1991), pp. 1121–1128.

[25] Raj Acharya, Richard Wasserman, Jeffrey Stevens, and Carlos Hinojosa. «Biomedical imaging modalities: a tutorial». In: *Computerized Medical Imaging and Graphics* 19.1 (1995), pp. 3–25.

[26] *Basic principles in Computed Tomography (CT)*. July 2016. URL: https://thoracickey.com/basic-principles-in-computed-tomography-ct/.

[27] Thomas Flohr. «CT systems». In: *Current Radiology Reports* 1.1 (2013), pp. 52–63.

[28] Marwa T Al Hussani and Mohammed H Ali Al Hayani. «The use of filtered back projection algorithm for reconstruction of tomographic image». In: *Al-Nahrain Journal for Engineering Sciences* 17.2 (2014), pp. 151–156.

[29] *CT Window Level*. URL: `https://imagej.nih.gov/ij/plugins/ct-window-level/index.html`.

[30] Patrick Therasse et al. «New guidelines to evaluate the response to treatment in solid tumors». In: *Journal of the National Cancer Institute* 92.3 (2000), pp. 205–216.

[31] *RECIST 1.1 - The Basics*. URL: `https://radiologyassistant.nl/more/recist-1-1/recist-1-1-1`.

[32] Elizabeth A Eisenhauer et al. «New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)». In: *European journal of cancer* 45.2 (2009), pp. 228–247.

[33] Soon Ho Yoon, Kyung Won Kim, Jin Mo Goo, Dong-Wan Kim, and Seokyung Hahn. «Observer variability in RECIST-based tumour burden measurements: a meta-analysis». In: *European journal of cancer* 53 (2016), pp. 5–15.

[34] Stefano Trebeschi et al. «Prognostic value of deep learning-mediated treatment monitoring in lung cancer patients receiving immunotherapy». In: *Frontiers in oncology* 11 (2021), p. 609054.

[35] Marco Somalvico et al. *Intelligenza artificiale*. Scienza & vita nuova, 1987.

[36] Francesca Pastore. «Sviluppo di un sistema ibrido basato su Machine Learning e Deep Learning per la classificazione di lesioni tumorali in immagini di mammografia sintetica = Development of a hybrid system based on Machine Learning and Deep Learning for breast lesions classification in synthetic mammography images». MA thesis. Politecnico di Torino, 2022.

[37] Iris van der Loo. «Prognostication from Longitudinal Multisequence Brain MRI using Artificial Intelligence». MA thesis. University of Twente, 2022.

[38] Pariwat Ongsulee. «Artificial intelligence, machine learning and deep learning». In: *2017 15th international conference on ICT and knowledge engineering (ICT&KE)*. IEEE. 2017, pp. 1–6.

[39] Iqbal H Sarker. «Deep Learning: a comprehensive overview on techniques, taxonomy, applications and research directions». In: *SN Computer Science* 2.6 (2021), pp. 1–20.

[40] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. «Convolutional neural networks: an overview and application in radiology». In: *Insights into imaging* 9.4 (2018), pp. 611–629.

[41] Keiron O'Shea and Ryan Nash. «An introduction to convolutional neural networks». In: *arXiv preprint arXiv:1511.08458* (2015).

[42] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. «Understanding of a convolutional neural network». In: *2017 international conference on engineering and technology (ICET)*. Ieee. 2017, pp. 1–6.

[43] Yann LeCun, Leon Bottou, Genevieve B Orr, Klaus-Robert Müller, et al. «Neural networks: Tricks of the trade». In: *Springer Lecture Notes in Computer Sciences* 1524.5-50 (1998), p. 6.

[44] Yuxin Wu and Kaiming He. «Group normalization». In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 3–19.

[45] Hang Yu, Laurence T Yang, Qingchen Zhang, David Armstrong, and M Jamal Deen. «Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives». In: *Neurocomputing* 444 (2021), pp. 92–110.

[46] Pimrada Potipimpanon, Natamon Charakorn, and Prakobkiat Hirunwiwatkul. «A comparison of artificial intelligence versus radiologists in the diagnosis of thyroid nodules using ultrasonography: A systematic review and meta-analysis». In: *European Archives of Oto-Rhino-Laryngology* 279.11 (2022), pp. 5363–5373.

[47] Xiang Chen, Andres Diaz-Pinto, Nishant Ravikumar, and Alejandro F Frangi. «Deep learning in medical image registration». In: *Progress in Biomedical Engineering* 3.1 (2021), p. 012003.

[48] Shengyu Zhao, Tingfung Lau, Ji Luo, I Eric, Chao Chang, and Yan Xu. «Unsupervised 3D end-to-end medical image registration with volume tweening network». In: *IEEE journal of biomedical and health informatics* 24.5 (2019), pp. 1394–1404.

[49] Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. «Deformable medical image registration: A survey». In: *IEEE transactions on medical imaging* 32.7 (2013), pp. 1153–1190.

[50] ABIM Foundation. *PET Scans After Cancer Treatment*. Dec. 2018. URL: `https://www.choosingwisely.org/patient-resources/pet-scans-after-cancer-treatment/`.

[51] Stefano Trebeschi et al. «Development of a prognostic AI-monitor for metastatic urothelial cancer patients receiving immunotherapy». In: *Frontiers in Oncology* (2021).

[52] Ingmar Paul Loohuis. «Exploring the Prognostic Value of Deep Learning Image-to-Image Registration for Immunotherapy Patient Monitoring». MA thesis. University of Twente, 2022.

[53] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. «U-net: Convolutional networks for biomedical image segmentation». In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241.

[54] Pengyue Zhang, Fusheng Wang, and Yefeng Zheng. «Self supervised deep representation learning for fine-grained body part recognition». In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE. 2017, pp. 578–582.

[55] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. «Voxelmorph: a learning framework for deformable medical image registration». In: *IEEE transactions on medical imaging* 38.8 (2019), pp. 1788–1800.

[56] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. «U-net: Convolutional networks for biomedical image segmentation». In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[57] Hongming Li and Yong Fan. «Non-rigid image registration using self-supervised fully convolutional networks without training data». In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. 2018, pp. 1075–1078.

[58] Jakob Wasserthal, Manfred Meyer, Hanns-Christian Breit, Joshy Cyriac, Shan Yang, and Martin Segeroth. «TotalSegmentator: robust segmentation of 104 anatomical structures in CT images». In: *arXiv preprint arXiv:2208.05868* (2022).

[59] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. «nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation». In: *Nature methods* 18.2 (2021), pp. 203–211.

[60] Annika Reinke et al. «Common limitations of image processing metrics: A picture story». In: *arXiv preprint arXiv:2104.05642* (2021).

[61] Silvia Seidlitz et al. «Robust deep learning-based semantic organ segmentation in hyperspectral images». In: *Medical Image Analysis* 80 (2022), p. 102488.

[62] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. «Random survival forests». In: (2008).

[63] *Understanding predictions in survival analysis*. URL: `https://scikit-survival.readthedocs.io/en/stable/user_guide/understanding_predictions.html`.

[64] Erica Tavazzi. «Random Survival Forests per la stratificazione del rischio in pazienti affetti da Sclerosi Laterale Amiotrofica». MA thesis. Universita' degli Studi di Padova, 2017.

[65] *Random survival forests*. URL: https://www.randomforestsrc.org/articles/survival.html.

[66] Goodfellow Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, and David Warde-Farley. «Generative adversarial nets." In Advances in neural information processing systems». In: (2014).

[67] Zhengwei Wang, Qi She, and Tomas E Ward. «Generative adversarial networks in computer vision: A survey and taxonomy». In: *ACM Computing Surveys (CSUR)* 54.2 (2021), pp. 1–38.

[68] *DCGAN*. URL: https://pytorch.org/tutorials/beginner/dcgan_faces_tutorial.html.

[69] Yabo Fu, Yang Lei, Tonghe Wang, Kristin Higgins, Jeffrey D Bradley, Walter J Curran, Tian Liu, and Xiaofeng Yang. «LungRegNet: An unsupervised deformable image registration method for 4D-CT lung». In: *Medical physics* 47.4 (2020), pp. 1763–1774.

[70] Jingfan Fan, Xiaohuan Cao, Zhong Xue, Pew-Thian Yap, and Dinggang Shen. «Adversarial similarity network for evaluating image alignment in deep learning based registration». In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 739–746.

[71] Mingrui Ma, Yuanbo Xu, Lei Song, and Guixia Liu. «Symmetric transformer-based network for unsupervised image registration». In: *Knowledge-Based Systems* 257 (2022), p. 109959.

[72] Alexey Dosovitskiy et al. «An image is worth 16x16 words: Transformers for image recognition at scale». In: *arXiv preprint arXiv:2010.11929* (2020).

[73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. «Attention is all you need». In: *Advances in neural information processing systems* 30 (2017).

[74] Mikaela Hardebro. «Transformer Based Object Detection and Semantic Segmentation for Autonomous Driving». MA thesis. Linköping University, 2022.

[75] Junyu Chen, Yufan He, Eric C Frey, Ye Li, and Yong Du. «Vit-v-net: Vision transformer for unsupervised volumetric medical image registration». In: *arXiv preprint arXiv:2104.06468* (2021).

[76] Yungeng Zhang, Yuru Pei, and Hongbin Zha. «Learning dual transformer network for diffeomorphic registration». In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*. Springer. 2021, pp. 129–138.

[77] *Treatment Options for Breast Cancer*. URL: `https://www.cancer.org/cancer/types/breast-cancer/treatment.html`.

[78] Judith A Paice et al. «Use of Opioids for Adults With Pain From Cancer or Cancer Treatment: ASCO Guideline». In: *Journal of Clinical Oncology* (2022).