Ingmar Paul Loohuis

Exploring the Prognostic Value of Deep Learning Image-to-Image Registration for Immunotherapy Patient Monitoring

Exploring the Prognostic Value of Deep Learning Image-to-Image Registration for Immunotherapy Patient Monitoring

Ingmar Paul Loohuis

A thesis presented for the degree of
Master of Science

UNIVERSITY OF TWENTE.

ANTONI VAN LEEUWENHOEK
NEDERLANDS KANKER INSTITUUT

University of Twente
The Netherlands
03-02-2022

Foreword

This thesis is the culmination of the work performed over the last year at the Department of Radiology at the Netherlands Cancer Institute. I am very grateful to have been hosted by Prof. Dr. Regina Beets-Tan at this department. This project gave me the knowledge and skills to apply AI to a diverse range of medical problems, for which I am grateful to all the people involved. I would like to thank Dr. Stefano Trebeschi and MSc. Teresa Tareco Bucho for their close supervision of the research and for continually challenging me. I would also like to express my gratitude to MD/MSc. Zuhir Bodalal for taking the time to help me with the clinical part of the research and arranging a host of clinical activities for me to participate in. I also want to thank Dr. Jelmer Wolterink for providing supervision and always providing me with interesting approaches. Also, a thank you to Prof. Dr. Christoph Brune for chairing the graduation committee, and to MSc. Bryan Wermelink for being the external member of the committee. And a very special thanks to MSc. Elyse Walter for being my mentor during the last two years, I've greatly appreciated all the talks we've had and the honest feedback you have always given. I hope you enjoy reading the thesis.

Contents

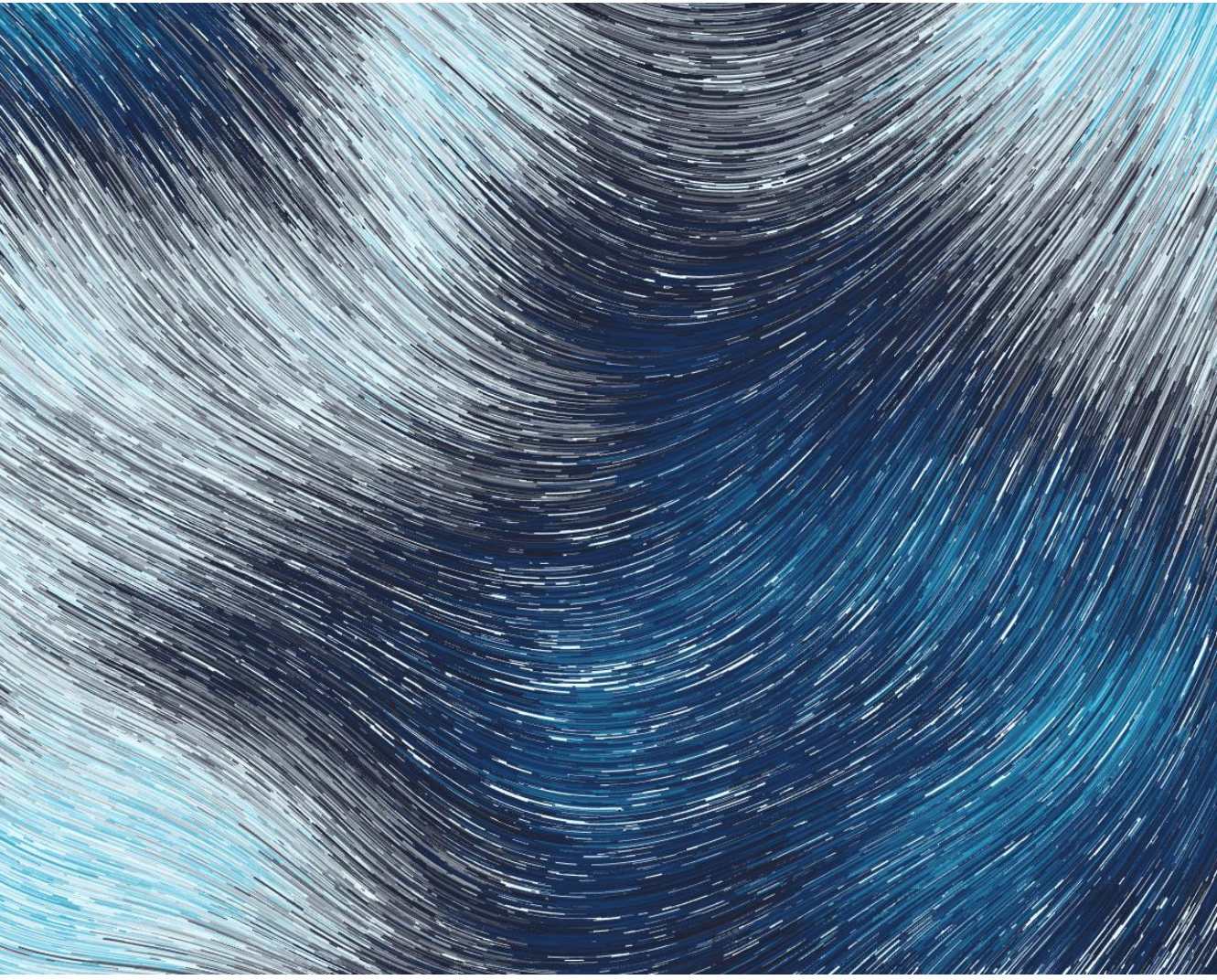
Clinical and Technical Introduction **1**

Chapter 1: Unsupervised Image Registration for the Quantification and Prognostication of Morphological Changes in Longitudinal CT-imaging **9**
1.1 Introduction 10
1.2 Methods 11
1.3 Results 15
1.4 Discussion 17
1.5 Conclusion 22

Chapter 2: Explainability of the Deformation Field via Disentangled Representation Learning **23**
2.1 Introduction 24
2.2 Methods 26
2.3 Results 28
2.4 Discussion 29
2.5 Conclusion 32

Chapter 3: Anatomical Fidelity and Realism in Morphological Changes via Adversarial Learning **33**
3.1 Introduction 34
3.2 Methods 36
3.3 Results 37
3.4 Discussion 38
3.5 Conclusion 41

General Discussion **43**



Introduction

Clinical Introduction

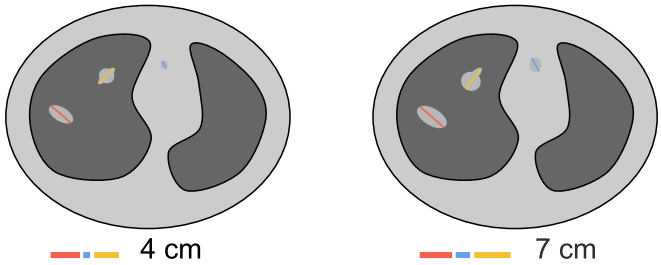


Figure 1: An example of a RECIST measurement for a patient at two different time-points. Visualized are two axial thorax CT slices with their respective RECIST measurements. The sum of the maximum diameters has increased by 75 percent, defining this as progressive disease. The diameters for a real RECIST measurement do not have to be located on the same slice.

Immunotherapy is a relatively new type of treatment for cancer compared to surgery, chemo- and radiotherapy[1]. The basic premise of immunotherapy is that the immune system can recognize and control tumor growth, and therefore it is possible to stimulate the immune system to fight the cancer. Several types of immunotherapy are already used in clinical care for different advanced-stage cancers such as melanoma, non-small cell lung cancer and bladder cancer. The most notable types of immunotherapy are antibodies for inhibitory immune checkpoints CTLA-4 and PD-1, which initiate an anti-tumor response [2]. As cancer care is highly personalized, patient monitoring is necessary for guiding further treatment.

Imaging is key for the monitoring of cancer patients receiving immunotherapy. Follow-up imaging, performed at a regular time interval, is the main method used for detecting disease progression, assessing the effect of treatment and related toxicity[3, 4]. Imaging modality is most often CT, but MRI is also used depending on the anatomical location of lesions. Assessment of imaging is mostly done qualitatively. Currently, the only standardized, quantitative method for image assessment is the Response Evaluation Criteria in Solid Tumors (RECIST). A slightly adapted version of RECIST was developed specifically for patients receiving immunotherapy, iRECIST[5, 6], namely accounting for phenomena like pseudo-progression, but not often used in clinical trials.

RECIST evaluates a maximum of five lesions quantitatively, using the change of the sum of maximum diameters between scans. Other lesions are assessed in a qualitative manner[5]. Treatment response is defined as a 30 percent decrease of the sum of diameters and disease progression is defined

as an increase of 20 percent. Any changes that cannot be characterized as either progression or response are defined as stable disease. Continuation of treatment, starting a different treatment, or stopping treatment altogether is partly based on this categorization[5]. A visualization of a RECIST measurement can be seen in Figure 1.

RECIST shows a high intra- and interobserver variability[7, 8]. A reason for the high variability is that different observers choose different tumors to measure, leaving much room for subjectivity. Criteria for the selection of target lesions are vague. Finding the maximum dimension of a tumor can be difficult, especially in irregularly shaped tumors. RECIST reduces the quantitative assessment to five unidimensional measurements, thereby not including information relevant for clinical decision making such as treatment toxicity, growth of high-risk lesions, and inflammation[9, 10]. As manual measurements are required in RECIST, it can be a time-consuming task. RECIST is therefore a suboptimal method for the quantitative assessment of treatment response or disease progression.

As important clinical decisions are being made using RECIST measurements, overcoming these limitations could lead to significant improvements for cancer patients. For example, earlier and more accurate identification of treatment response or disease progression would lead to prompt intervention, and higher chances of improvement of the conditions of the patient, while reducing effects caused by unnecessary treatments[11]. Waiting time and costs associated with radiological assessments can also be reduced by implementing a less time-consuming method. An automated quantitative method for image assessment would therefore have the potential to increase the quality of life for cancer patients in several ways.

To address the limitations inherent to RECIST, a new method should be:

- able to combine changes throughout the whole body quantitatively into a prognostic score;
- fast and fully-automatic, able to evaluate a scan in the order of seconds;
- explainable, in order to be deployed in clinical practice;
- able to handle heterogeneous data.

New methods are being developed that try to quantify morphological changes between scans to thereby tackle the limitations of RECIST. However, most of these methods are segmentation-based[11–14]. Tumors are segmented and the change of tumor volume is linked to clinical response.

Since ground-truth segmentations are very time-intensive to obtain, most segmentation-based methods are often only trained for one certain cancer type, making these methods not generalizable to other cancer types. Furthermore, using only the change of tumor volume disregards a lot of clinical information that is present outside of the tumor. Other methods use more information than just tumor volume, but they only include the local area around the tumor and therefore also need segmentations[13]. One of the methods used for response prediction uses FDG-PET/CT scans, where the PET scan provides physiological information[15]. As PET-scans are not taken routinely for treatment monitoring, a method is desired that only uses CT imaging.

A new method, the prognostic AI monitor (PAM), that can predict survival using longitudinal CT-imaging of patients receiving immunotherapy, has been proposed[16, 17]. Pilot studies showed that PAM could predict survival for patients with lung and bladder cancer receiving immunotherapy with an AUC of respectively 0.69 and 0.73. However, these pilot studies had several limitations including the usage of small datasets containing low-resolution CT-imaging, focussing on one cancer type only. The predictions of PAM lacked explainability, making clinical implementation difficult. In this thesis, these limitations of PAM will be addressed. Chapter 1 will present improvements to PAM making it capable of prognostication on a larger pancancer cohort of patients receiving immunotherapy. In Chapter 2, the problem of explainability is tackled by disentangling the features used for prognostication. Chapter 3 will improve the realism of the image registration by implementing adversarial loss for training.

Technical Introduction

Machine Learning

Machine learning is the process by which algorithms learn without explicit instructions by drawing inferences from patterns in data. Different statistical machine learning models exist such as decision trees, random forests, and neural networks. To explain the techniques used in this thesis, random forests and neural networks will be briefly introduced. To understand a random forest, an understanding of a decision tree is needed. Decision trees take an input that branches in several directions, based on a certain decision such as exceeding a threshold or by chance, where each branch can lead to one of the outputs. An ensemble of decision trees used on random subsets of the data can be used to increase the model's accuracy, a technique called random forest. In the last decade, neural networks have

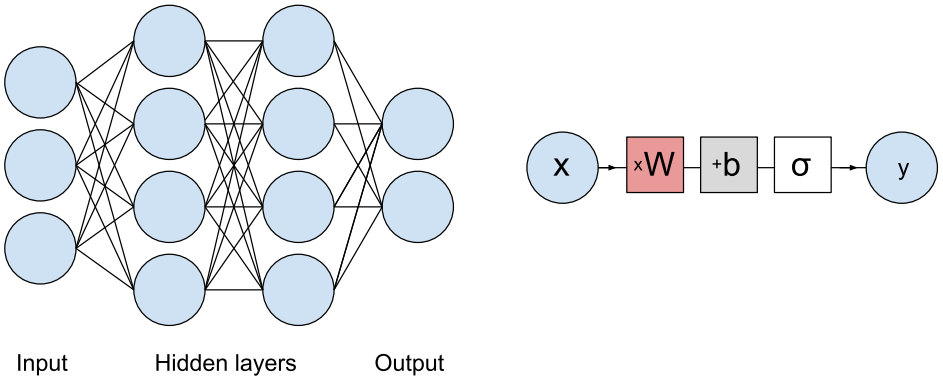


Figure 2: Left: Example of a neural network with the circles representing the nodes and the black lines the connections between the nodes. Right: Example of the mathematical operations that transform the input x to the output y. W indicates the weight, b the bias and σ a non-linearity.

become the staple for problems relating to unstructured data such as images and text. Neural networks take input values as separate nodes, such as words in a sentence or pixels in an image, and connect these to a subsequent layer of nodes. Connections represent a mathematical operation: $y = \sigma(Wx + b)$ where the weight W and bias b are trainable parameters. A node in a specific layer is connected to all nodes in the previous layer. The final layer of the neural network predicts the output. The neural network is optimized to minimize the difference between the predicted output and the desired output. The difference between predicted and the desired output is backpropagated through the neural network, changing the weights and biases. An example of a simple neural network is visualized in Figure 2. [18]

Due to improvements in computational power, neural networks have become larger and larger to capture more patterns present in the data. Training these larger models has been coined *deep learning*. Besides a specific model architecture, several other parts are needed for training a deep neural network: a loss function and an optimizer. The loss function maps the difference between the predicted and desired output to an optimizable metric. It is therefore a metric that indicates how well the model performs on the training data. Examples of loss functions are (binary) cross-entropy and mean square error. As the loss function indicates how well the model performs, an optimizer is needed to determine in what manner and when the weights of the model should be updated to optimize the value of the

loss function. Examples of optimizers are (stochastic) gradient descent and ADAM. [19]

In the field of computer vision, neural networks have also taken over more classical heuristic-based approaches for image classification, segmentation, and object detection. Especially a subtype of neural networks, convolutional neural networks (CNN), has made significant improvements to the state-of-the-art. CNNs use trainable convolutional kernels, which can be interpreted as small filters, to extract relevant features from images. Initial convolutional layers extract features such as edges and boundaries while subsequent convolutional layers extract features such as certain shapes. Different architectures of CNN exist, some of the most famous being VGG, Inception, and ResNet. [19]

A special kind of neural network used in computer vision is the autoencoder, which is composed of two parts. The first part, the encoder, encodes the input to a smaller-dimensional representation, a point in the latent space. The second part, the decoder, decodes a point in the latent space to a larger-dimensional output. This larger dimensional output can be a reconstruction of the input image. Autoencoders are therefore typically used for dimensionality reduction as the smaller dimensional latent space can be used as input features for different tasks, such as classification[19]. An architecture that looks similar to an autoencoder is the U-Net[20]. The U-net is also composed of an encoder and decoder, but aside from the latent space, there are other connections between the encoder and decoder. These skip layers transfer information, thereby reducing the bottleneck effect of the latent space, allowing for a higher resolution output. U-nets are typically used for image-to-image tasks, such as segmentation, denoising, and image registration.

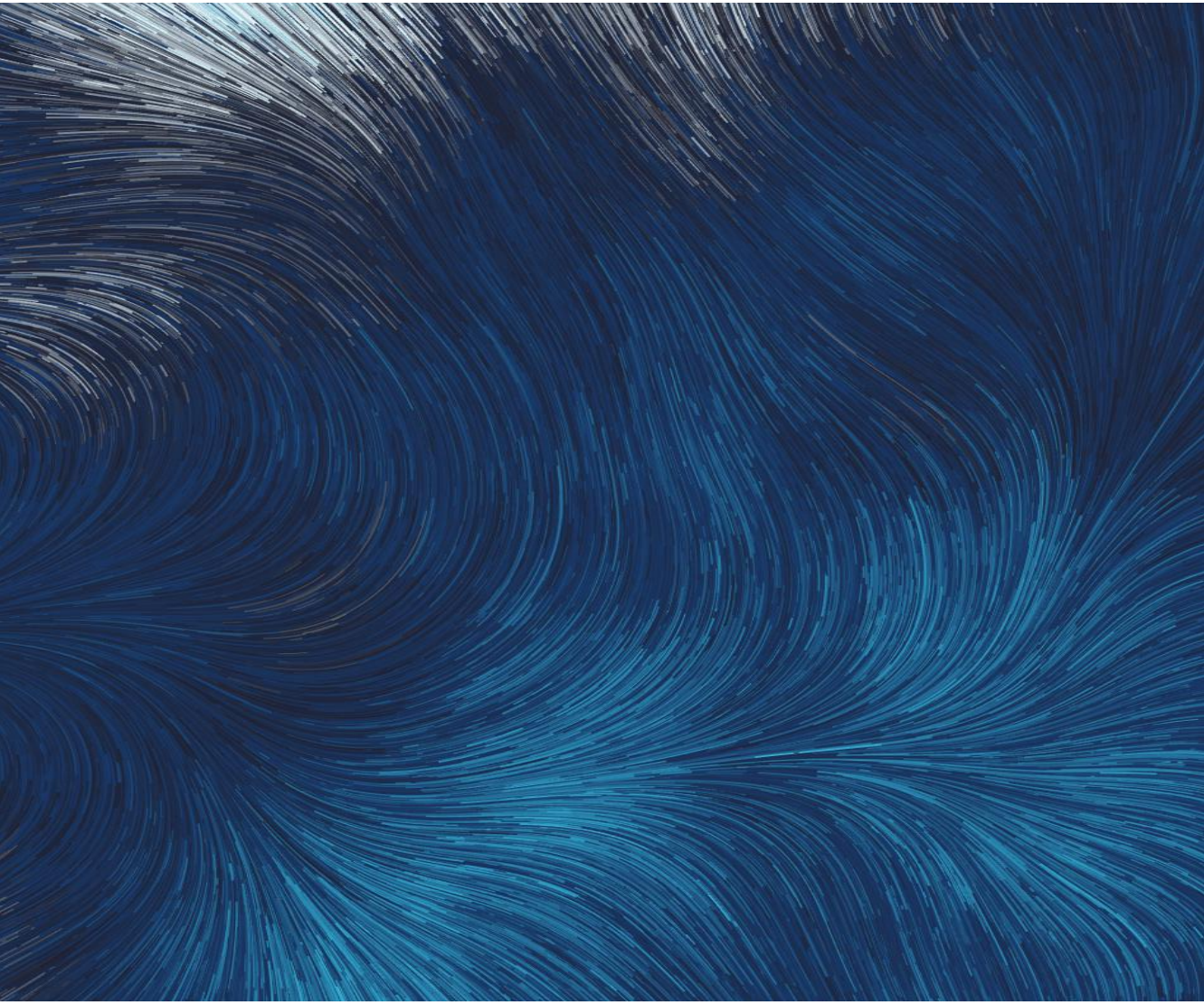
Image Registration

Image registration is the alignment of one image, the moving image, to the target image, the fixed image. It is often used in the medical imaging domain, especially for radiation treatment planning. Treatment plans are made on a pre-treatment scan, and image registration is used to translate those plans to scans obtained during treatment. Image registration is most often performed in two steps: first an affine transformation and then an elastic deformation. The affine transformation corresponds to a rough alignment of both images, making use of translation, rotation, shearing, scaling, and reflection. In the case of medical imaging, the images are a volume. The 3D affine transformation can be mathematically expressed as a 4x4 matrix, A. The affine transformation matrix can then be multiplied to a certain voxel's location $[x, y, z, 1]^T$, obtaining the location of the voxel

after the affine transformation.

Performing only a rigid affine transformation on medical images often yield suboptimal results, as physiological and pathological processes change the location, size, and shape of organs. To compensate for these changes, the moving image needs to be elastically deformed. Over the last decades, many different techniques have been proposed that model these elastic deformations. Examples of techniques used are B-splines and radial basis functions[21], with more recently neural networks obtaining state-of-the-art results[22]. Most of the techniques employing deep learning, model the elastic deformation using a dense displacement vector field, where for every voxel in the original volume a vector is given specifying the exact magnitude and direction of the deformation. [21, 22]

Deep learning can also be used for tasks such as image registration. A difference can be made between the supervised and unsupervised training of these models. [23] In the supervised setting, pairs of images are needed with known transformations between the scans. However, as these transformations are not known they need to be artificially generated, which could lead to the model only being able to capture artificial transformations. Another approach is to use a transformation made by an expert, however, these are time-consuming to create. Currently, unsupervised methods have pushed the state-of-the-art as large and real datasets can be used making the models able to capture clinically relevant deformations. The model is trained to maximize a similarity metric between both input images while regularization is used to constrain and guide the model. VoxelMorph is one such example of an algorithm that is tasked to perform image registration in an unsupervised manner. [24] The model calculates sequentially the affine transformation matrix and elastic deformation field between two input images. An example of the two sequential steps are partly visualized in Figure 3.



Chapter One

Unsupervised Image Registration for the Quantification and Prognostication of Morphological Changes in Longitudinal CT-Imaging

Abstract

CT imaging is performed for the monitoring of treatment response in cancer patients receiving immunotherapy. RECIST is currently used for prognostication but has several limitations. This study implements a novel deep learning approach on a large pancancer dataset that predicts survival by quantifying morphological deformations. The network was pretrained for image registration using thorax CT scans, and features were extracted from the latent space of the network. These features are then linked to 1-year survival using a classifier. n=1007 patients were included, resulting in n=5253 scan pairs. The classifier was able to predict 1-year survival with AUC = 0.61 ($p<0.0001$) for the test-set. Split by type of cancer, the highest AUC was achieved for patients with thoracic cancers AUC = 0.64 ($p<0.0001$). AUC varied during the treatment timeline for thoracic cancers, with an AUC of 0.78 for scans obtained between 17 and 43 weeks after the start of treatment. This study showed that features extracted from an image registration model can be linked to survival, resulting in a moderate prognostic performance.

1.1 Introduction

Longitudinal CT imaging is routinely performed for the monitoring of treatment response in cancer patients receiving immunotherapy. Quantitative assessment of these scans is commonly performed using RECIST, which calculates the change of the sum of diameters of a maximum of five lesions,

with the rest of the lesions assessed qualitatively[5, 25]. However, this method has high intra- and interobserver variability, is time-consuming, and ignores relevant clinical information such as treatment toxicity[7, 8].

An objective reader that can accurately assess longitudinal, whole-body patient imaging would overcome RECIST's limitations, resulting in more accurate clinical decision-making. Recently, two pilot studies have been performed based on artificial intelligence (AI) architectures that can be used for prognostication, called the prognostic AI monitor (PAM)[16, 17]. The architecture is visualized in Figure 3.

A first pilot study showed that PAM could predict 1-year survival from longitudinal chest CT images in patients with stage-IV non-small cell lung cancer, with an average AUC of 0.68[16]. A subsequent pilot study performed on abdominal CT-imaging from patients with stage-IV urothelial cancer obtained an AUC of 0.73[17]. However, these studies have several limitations: PAM was only trained on relatively small datasets containing one cancer type. To fully assess the prognostic performance of PAM, a pancancer cohort of patients is needed to capture the heterogeneity in imaging phenotype. Patient populations can then be identified where PAM is most effective. These populations can for example be based on cancer type, age, and treatment type. The pilot studies also used low-resolution CT scans. The low resolution of these input scans can cause clinical information to be lost, such as smaller lesions, which can negatively impact prognostic performance.

In this study, we will tackle these limitations by adapting PAM to use high-resolution CT scans as input and studying the prognostic performance of PAM on a wide variety of cancer types. A comparison will be performed between PAM and RECIST 1.1. To summarize, the unique contributions of this study are as follows:

- Implementing PAM on a larger pancancer dataset.
- Using higher resolution imaging as input.
- Providing a quantitative comparison between RECIST 1.1 and PAM.

1.2 Methods

Architecture

PAM is not directly trained to predict survival but pre-trained to perform image registration. Quantitative features are obtained from an embedding layer and subsequently linked to survival using a classifier. An overview of the method can be seen in Figure 3. More specifically, the first part of the

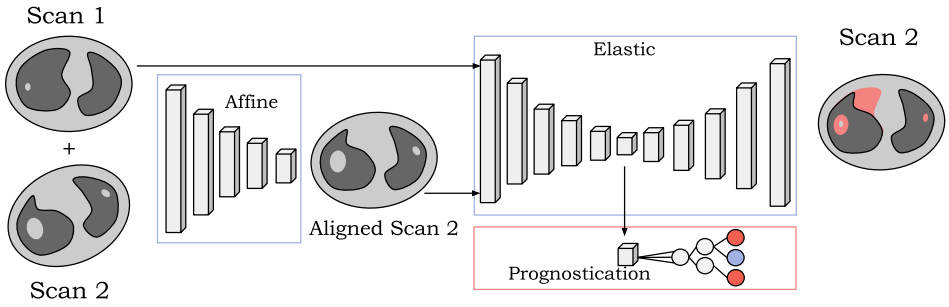


Figure 3: General overview of PAM, with the image registration (blue box) and prognostication (red box) modules, which are trained separately. The full model is comprised of an affine registration, an elastic registration, and the linkage of features to survival using a classifier.

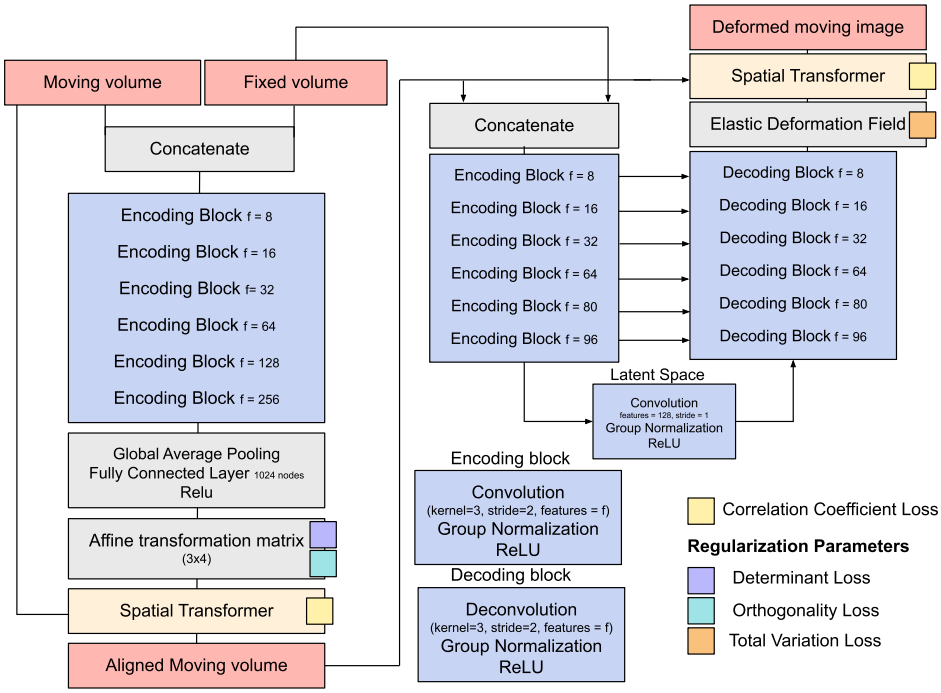


Figure 4: Model architecture used for unsupervised image registration adapted from [16].

network architecture consists of six convolutional blocks followed by a fully-connected layer, which calculated the affine transformation matrix between the moving image and the fixed image. The affine transformation matrix is

then applied to the moving image, resulting in a transformed image. The purpose of the affine transformation is to align the patient between the two scans, and correct for different positions they might have assumed during acquisition. The affine-transformed moving image is concatenated to the fixed image, serving as input to the second part of the network. The second part of the network generates a deformation field, which is applied to the aligned moving image which generates the deformed moving image. This allows the network to model anatomical changes between follow-ups. It has a U-Net architecture [20], with six down- and up-sampling layers, with skip layers in between the down- and up-sampling convolutional layers. The output of the network is an elastic deformation field. The affine and elastic parts of the network are trained together. The specifics of the architecture can be found in Figure 4.

Data for Image Registration

To train the unsupervised image registration model, thorax CT scans are needed. CT scans with a maximum axial resolution of 1.0 mm from The Cancer Imaging Archive (TCIA) uploaded before the 21st of April 2020 were obtained. The thorax, which included all slices between the lower-neck and the lowest parts of the diaphragm, was automatically extracted using the method of Zhang et al [26]. All scans were clipped between -120 (fat) and 300 (cancellous bone) Hounsfield Units (HU), to obtain only soft-tissue and help reduce computational memory. Scans were then normalized between 0 and 1. To avoid intensity stretching artifacts from the affine registration, each volume was padded with a 0-valued pixel, creating a 1-pixel edge at all borders of the volume.

Unsupervised Image Registration

Scans were randomly paired and training was performed using PAM with the ADAM optimizer[27] such that the cross-correlation loss between the fixed and affinely transformed moving image was minimized, as well as between the fixed image and elastically deformed moving image. Deformation penalties were implemented, on both the affine (orthogonality and determinant loss) and elastic transformation (total variation loss) that punishes large deformations as was done in[28]. The final loss function is:

$$L = L_{cc}^{affine} + 0.1(L_{det}^{affine} + L_{ortho}^{affine}) + L_{cc}^{elastic} + 0.1(L_{variation}^{elastic}) \quad (1)$$

To improve the training, it was performed using a curriculum method. The cross-correlation loss was computed on a smoothed version of the image,

gradually decreasing the amount of smoothing during training. Smoothing was implemented using average pooling, with the kernel k being the smoothing parameter that was decreased during training.

To assess registration quality, a simplified variant of the AQUIRC method was used [29]. Three scans are randomly selected from the dataset. PAM is used to calculate the transformation from scan 1 to scan 2: T_1^2. This is repeated to obtain the transformation T_2^3 and T_3^1. The three transformations are then applied to a randomly chosen point with location x in scan 1, resulting in $T_1^2(T_2^3(T_3^1(x)))$. The Euclidean distance can be calculated between the original point x and the transformed point. This is used as a measure for registration error. Registrations for circuits with large euclidean error distances were qualitatively inspected.

Data for Prognostication

All patients that started immunotherapy before January 1, 2019, at the Netherlands Cancer Institute were included. The study was reviewed and approved by the Institutional Review Board. Immunotherapy was defined as any treatment including anti-PD1, PDL1, or CTLA4. Cancer types were grouped based on the WHO classification, with only patients with the four most common groups of cancer being included. These were thoracic cancers (C3), skin cancers (C4), breast cancers (C5), and genito-urinary cancers (C6). Diagnostic thorax CT scans were collected and preprocessed using the same protocol as scans from the TCIA dataset. One baseline scan, performed before the start of treatment, and all follow-up scans were included. Scans of the same patient were paired if obtained within a time interval of 6 months, or less. Scans were equally split between training and independent test-set, on a patient basis in a stratified manner to ensure equal distribution of scans based on WHO classification, age, and overall survival.

Survival prediction

Features were extracted from the embedding layer of the network using scan pairs as input. A random forest classifier was trained to predict 1-year survival. Statistical analysis was performed using bootstrapping (n=1000 repeats), while results were corrected for multiple hypothesis testing using the false discovery rate method. Statistical significance, which was set at $p<0.05$, was assessed using the Mann-Whitney-U test. For the Kaplan-Meier curves, patients were split into three groups based on the predicted

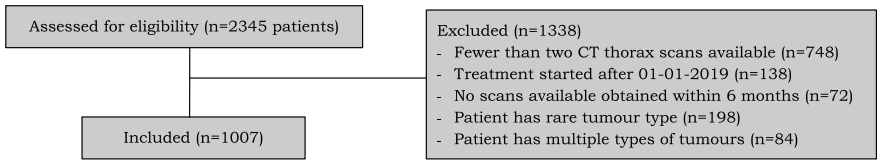


Figure 5: Diagram with patient exclusion criteria and number of patients included and excluded.

probability of survival, and statistical significance was tested using the log-rank test.

A radiologist qualitatively assessed a selection of scan pairs and gave a prediction for the survival of the patient based on these scan pairs. Predictions of the radiologist were compared to the predictions of PAM. A different radiologist performed a quantitative analysis using the RECIST 1.1 guidelines. Results for survival analysis were compared for the RECIST 1.1 analysis and PAM.

1.3 Results

A total of 10,294 scans from TCIA were included for unsupervised image registration, resulting in 5,462 scan pairs. To assess image registration performance, 100 image registration circuits were generated containing 3 scans per circuit. 100 points were randomly chosen in the first scan of the registration circuit, and these points were transformed through the whole registration circuit. The median error was 34 mm +/- 27 mm. Qualitative analysis of outliers showed that points with large deformations were most often located at the highest and lowest axial slices. This most often happened in scan pairs where the highest and lowest axial slices of each scan do not correspond to the exact same anatomical location. Other cases showed unrealistic deformations at areas where registration error was highest.

2345 NKI patients were screened for inclusion. 1007 patients were included, resulting in 5253 scan pairs. Figure 5 shows the number of patients excluded at each step. Table 1 shows patient characteristics in both the training and test set and split by cancer type. The scan pairs were fed through the registration network and features were extracted. These features were linked to survival using the random forest classifier. Prognostic results for the whole test-set showed AUC = 0.61 ($p<0.0001$). Split by type of cancer, results for thoracic cancers showed AUC = 0.64 ($p<0.0001$), skin cancers AUC = 0.55 ($p<0.01$), breast cancers AUC = 0.62 ($p<0.001$) and genito-urinary cancers AUC = 0.54 ($p=0.16$).

	patients	scan pairs	age	% mortality	months survival
Total	621	3362	59	32	12.2 (11.0)
Training set	503	2628	59	28	12.1 (10.9)
subset					
Thoracic cancers	187	1089	62	29	14.4 (12.5)
Skin cancers	209	1052	55	24	11.5 (10.3)
Breast cancers	45	244	57	40	11.6 (7.8)
Genito-urinary cancers	62	243	60	28	10.1 (6.7)
Test set	504	2625	60	28	11.2 (10.1)
subset					
Thoracic cancers	192	1082	63	34	11.0 (10.8)
Skin cancers	215	1001	55	19	12.7 (10.3)
Breast cancers	39	261	57	47	9.4 (7.7)
Genito-urinary cancers	58	281	64	25	10.6 (9.0)

Table 1: Patient characteristics for the training set and test set. Numbers in parentheses represent standard deviation.

To more accurately determine the prognostic performance, ROC-AUC was calculated by applying a sliding temporal window of 6 months during the treatment timeline. Results are shown in Figure 6 for the whole test-set and the two cancer types with most patients, as the others had too few patients to perform temporal analysis. Only for thoracic cancers, a trend was visible that AUC increased during treatment, with a maximum AUC of 0.78 that was achieved for the time interval of 17 to 43 weeks after the start of treatment.

Kaplan Meier curves were generated using the predictions of PAM, also shown in Figure 6. Differences between probability groups for the full test-set and the thoracic cancer cohort were all statistically significant ($p<0.001$). For the skin-cancer cohort, only the difference between the low- and high-probability was statistically significant ($p<0.05$).

Thirteen scan pairs of patients with thoracic cancers were assessed qualitatively by a radiologist. Qualitative assessment of the input scans showed that for cases where PAM correctly predicted the outcome, the radiologist predicted the same. For cases where the network did not correctly predict treatment outcome, the radiologist also in a few instances incorrectly predicted survival.

Baseline and the first scan during treatment were obtained for 304 patients. These scan pairs were assessed using RECIST 1.1 guidelines by a radiologist. Kaplan-Meier curves were generated where patients were split

according to the RECIST 1.1 classification: progressive disease, stable disease, or treatment response. 118 of these scan pairs were also available in the NKI dataset. The classifier of PAM was retrained to make sure all of the 118 patients were not seen during training of the random forest classifier. For the 118 patients, Kaplan-Meier curves were also generated for predictions using both the RECIST 1.1 guidelines and PAM. These curves are visualized in Figure 7.

1.4 Discussion

Results showed statistically significant prognostic performance on the independent test-set in both thoracic and skin cancers, but for genito-urothelial and breast cancers no statistically significant performance was shown. As only CT thorax imaging was used, it was expected that thoracic cancer would have highest prognostic performance as the tumors are located primarily in the thorax. This generally also holds for metastasis of skin cancers, which could explain the higher prognostic performance compared to abdominal cancers. Another explanation could be the lack of data for the genito-urothelial cancers, leading to worse prognostic performance.

PAM was pre-trained to perform image registration, so to assess image registration performance an adapted AQUIRC measure was used. A median registration error of more than 30 mm indicates that the registration could still be improved. However, the scans that were used to assess registration quality were from the TCIA dataset, which is very heterogeneous. Patients in this dataset differed in age, body size, and anatomy. Using these scans to assess registration performance overestimates registration error. As prognostication was done on scans from the same patient, the registration error would be expected to be lower. The voxels where registration error was highest, corresponded to voxels at the upper and lower edges of the volume. As the anatomical location of these edges can slightly differ, it is not strange that the registration error is higher at these locations. Due to time constraints, it was not possible to more accurately assess registration quality, by for example not taking into account registration accuracy at the edges of the volume.

Initially, the same network presented in the pilot studies was trained to perform image registration, however, the network did not converge to an adequate registration loss. It was hypothesized that due to the increase in image resolution, the network was too small to model all deformations. Therefore, to compensate for the increase in image resolution, the size of the network was increased compared to the pilot studies. This allowed the network to converge to a lower registration loss, which is hypothesized to have also increased the capability of the network to model morphological

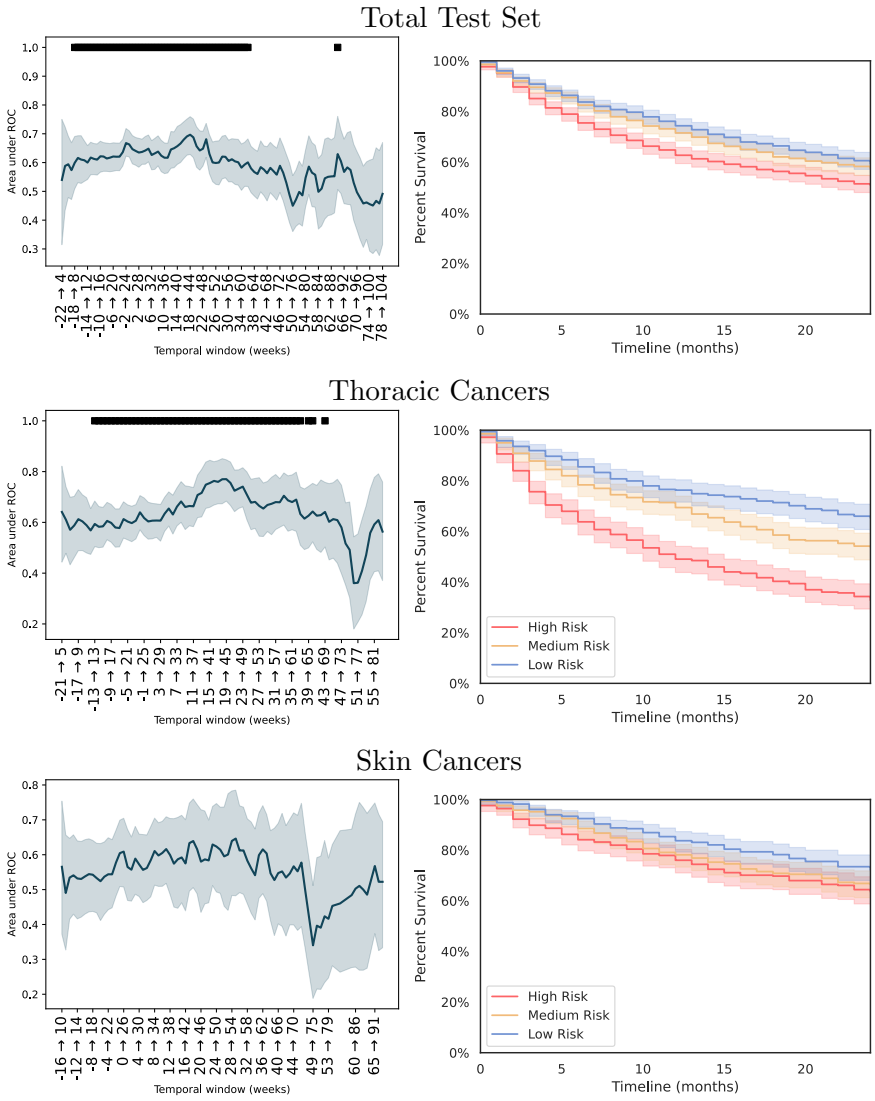


Figure 6: Left: ROC-AUC with a sliding temporal window. X-axis represent temporal window in which both CT scans were performed with respect to start of treatment. Black bar represent statistical significant results at that time-point. Right: Kaplan Meier curves, with the timeline representing months after second scan and the y-axis percent survivals. Patients were split into three groups based on the predictive probability of the patient not surviving: high, medium and low risk groups, splits were performed on the 33rd and 67th percentile. Lighter colors around lines indicate 95% confidence interval. Upper row represent results for the total test-set, the middle row for the thoracic cancer cohort and the bottom row for the skin cancer cohort.

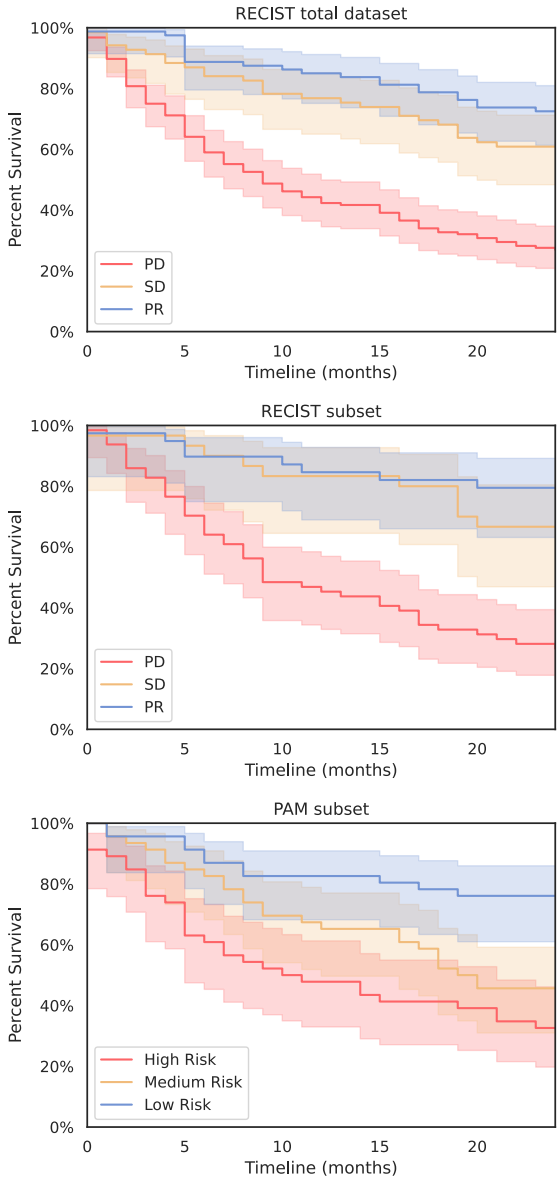


Figure 7: Upper row visualized the Kaplan-Meier for the RECIST 1.1 analysis on the total dataset of 304 patients. Middle row visualizes the results for the 118 patients that are also present in the NKI dataset. Lower row visualizes for this same subset the results obtained using PAM. For the RECIST evaluations patients were split into three groups: partial response, PR, stable disease, SD, and progressive disease, PD. For the results obtained with PAM patients were split into three groups based on the predictive probability of the patient not surviving: high, medium and low risk groups, splits were performed on the 33rd and 67th percentile.

changes. As the increase in image resolution also allowed for the inclusion of small lesions, there are extra morphological changes that can be modeled by the network. Both the increase of the network size and the inclusion of small lesions by the increased image resolution might have contributed to higher prognostic performance.

Our results show that prognostic performance for thoracic cancer increases further into treatment. This could be because the effects of immunotherapy are more visible further into treatment, making survival prediction easier as morphological changes are more present. This effect is less visible in the results from skin cancer patients and absent from the other cancer types. This trend can also be caused by a selection bias, as patients with a very poor short-term prognosis, and therefore large morphological changes between scans might choose to forego immunotherapy. The remaining patients, those that are included in this study, could therefore have fewer morphological changes between scans at the start of treatment.

The comparison between RECIST 1.1 and PAM shows that PAM has a similar prognostic performance as RECIST 1.1. Patients classified as progressive disease using RECIST 1.1 have a worse prognosis than the patients in the high-risk group using PAM. This indicates that RECIST 1.1 is currently still a better predictor for clinical outcomes. However, there are three points where PAM could easily be improved to obtain a better performance. Firstly, RECIST 1.1 evaluations were performed using thoracic and abdominal imaging, in contrast to PAM where only the thoracic imaging was used. Secondly, to better compare different architectures the classifier used to link features to survival was not optimized. Lastly, patients were split into risk groups using the 33rd and 67th percentile. This is a coarse method for classification, not taking into account the actual distribution of treatment outcomes. Tackling these three points might further improve prognostic performance.

The prognostic performance of PAM, even for thoracic cancers, is still moderate. A cause could be the lack of information about survival in the latent space features. This information is not guaranteed to be present in the features, as the latent space is optimized for image registration. The features are therefore not optimized for prognostic performance. To optimize the prognostic features, an approach has been implemented using the pre-trained network but swapping the elastic decoder to a neural network that predicts survival from the latent space. This network was then trained end-to-end for survival but did not produce any significant results. A better implementation would be to incorporate an extra regularization term that incorporates survival while still training the network to perform image registration. This method would maximize the chance that the features are

prognostically relevant. Due to time constraints, this was not implemented for this study.

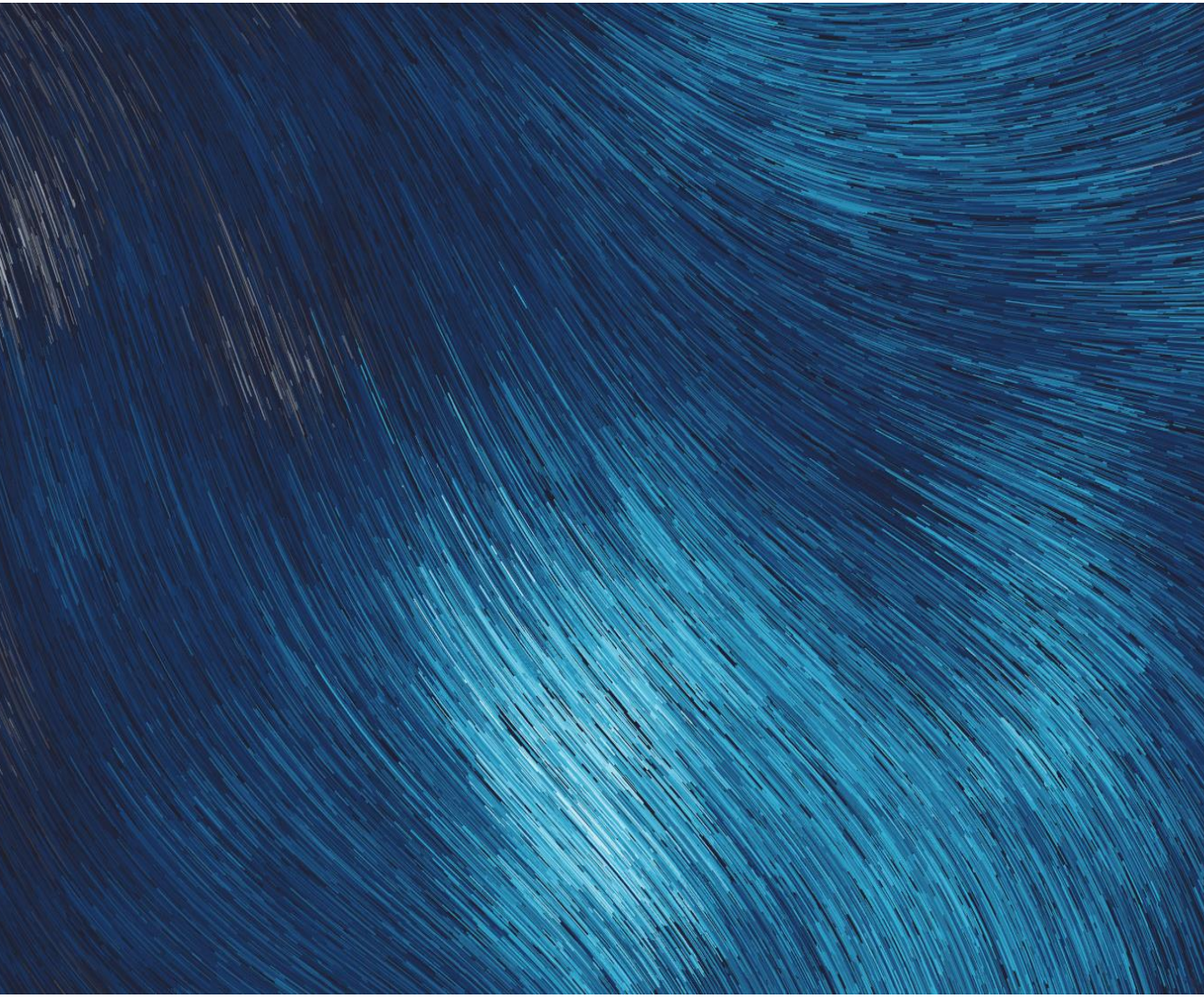
PAM was trained to predict survival; however, RECIST assesses response to treatment. With no ground truth available for the response assessment, it is not possible to train a network to predict treatment response. The current standard is RECIST, but using this as the ground-truth would just train the network to replicate RECIST, which as noted earlier is an inaccurate and subjective metric. Predicting one-year survival can be seen as a proxy for predicting treatment response, as a response to treatment is likely to correlate to longer survival. Survival however is a coarse metric to use, as there are a host of other factors that influence one-year survival, that might not all be present in imaging. Training a network to predict one-year survival is also difficult because of edge cases, changes in scans for a patient that survives for 11 months versus 13 months will not seem that different, this can cause training to be difficult. Future investigation should expand to more clinical and biological endpoints.

Overall prognostic performance of PAM on thoracic cancers shows low sensitivity but high specificity. The low sensitivity is caused by the presence of false negatives. A false negative corresponds to PAM predicting that the patient will survive for more than one year, but in reality, the patient did not. The low sensitivity could hint at the absence of prognostically relevant morphological changes in scans of patients that were classified as a false negative. The qualitative assessment showed that these false-negative scan pairs often did not exhibit any morphological changes between both scans. In contrast, scans pairs from patients where PAM correctly classified the patient as not surviving for more than one year exhibited large morphological changes such as increase of tumor size and increase of the amount of atelectasis.

The features that were linked to survival are, being generated by a neural network, not independent of each other. In other words, they are entangled. Entangled features cannot be interpreted in a similar way that other clinical parameters can be interpreted, which makes it difficult to assess why and to what extent these features are clinically relevant. Error analysis is therefore difficult to do, which limits the extent to which PAM can be deployed clinically as it is difficult to assess when and why PAM is failing. More interpretable features would overcome this problem and will allow for more accurately determining specific patients populations for which PAM is most suitable.

1.5 Conclusion

AI-analysis of longitudinal chest CT scans has prognostic value in patients receiving immunotherapy, especially for patients with thoracic cancer. The proposed method has a prognostic performance comparable to RECIST 1.1, the current gold standard for quantitative assessment of longitudinal imaging.



Chapter Two

Explainability of the Deformation Field via Disentangled Representation Learning

Abstract

Explainability is important for deploying machine learning into a clinical setting. Earlier research showed that a deep learning model pre-trained for image registration is capable of predicting survival for cancer patients receiving immunotherapy. However, it is not straightforward to understand what the features that are used to generate survival predictions represent. One of the reasons is that these features are not independent of each other, but rather "entangled". To enforce a disentangled feature representation, the Hessian Penalty (HP) was used during training. For comparison, two identical networks were trained, only one of which used the HP. The network with HP achieved an AUC of 0.60, $p<0.01$ for the subset of patients with thoracic cancers, comparable to the network without HP (AUC of 0.59 $p<0.01$). Cohen's kappa statistic was 0.53 between both networks, indicating moderate agreement between prognostic predictions. While both networks showed a decreased predictive performance compared to the network presented in chapter 1, the qualitative analysis showed that the model with HP managed to disentangle different factors of variations compared to the model without HP. The disentangled features were correlated with large deformations in specific anatomical locations, but no fine-grained clinically interpretable deformations were detected. Further research should incorporate disentanglement procedures in the original architecture, that improve prognostic performance and obtain features representing clinically significant deformations.

2.1 Introduction

A deep learning network, PAM, that can be used for prognostication, was presented in Chapter 1. However, PAM still has several limitations such as

a lack of explainability. It is not known what deformation is represented by which input feature, and therefore it is unclear how the predictions are exactly made. Legally, certain black-box models are allowed to be used in a clinical setting if they have an "appropriate level of transparency (clarity) of the output and the algorithm aimed at users"[30]. Besides legal arguments, explainability is essential for the further improvement of PAM. Explainable features can be used to assess which imaging characteristics are important for predicting survival. For example, these characteristics might represent currently unknown clinical information that is relevant for patient survival. Another reason for explainability is to study the limitations of PAM. Explainable features can also be monitored to identify artifacts, e.g. if input features are different from the ones the network was trained on, the predictions have a higher chance to be incorrect.

The problem that makes the features used for predicting survival not explainable is that they are entangled. An entangled representation is one where components of the representation do not independently capture the true underlying factors that explain the data[31]. A representation where the components independently describe the factors of variation is called disentangled. An ideal disentangled representation has at least these three properties: modularity, compactness, and explicitness[31]. Modularity means that a factor of variation in the data only affects a subset of the representation, and no other factors of variation affect this subset. The size of the subset is determined by the property of compactness, e.g. a factor of variation in the data only affects the smallest possible subset of the representation. One also wants the representation to completely describe the factors of interest, which is the property of explicitness. For survival prediction, a change of the representation space, which in our case are the features, should not influence different factors of variation (modularity) and should describe relevant factors of variation (explicitness). Among the methods proposed to obtain disentangled features, the Hessian Penalty (HP) is one of the few that can work with small-batch sizes[32].

The Hessian matrix represents the second partial derivative of a function. In a neural network, the Hessian matrix represents how changing two latent components changes the output. In an ideal case, where all components are independent of each other, the Hessian would be diagonal. An off-diagonal term represents how changing one latent component has an effect on the change of a different latent component to the output. For example, having a generator function G with as input latent components z, a non-diagonal term of the Hessian of this matrix would represent.

$$H_{ij} = \frac{\partial^2 G}{\partial z_i \partial z_j} = \frac{\partial}{\partial z_j}\left(\frac{\partial G}{\partial z_i}\right) \tag{2}$$

Therefore, minimizing off-diagonal terms would increase independence between latent components. The Hessian Penalty is a regularization function that tries to minimize the sum of the squared off-diagonal terms. The features PAM uses for predicting survival are obtained from the latent space of the U-Net and are responsible for the elastic deformation. Disentanglement in this case means that a change in one feature will lead to elastic deformation of a single component or aspect of the image, which is independent of a change in a different feature.

In this study, the Hessian Penalty will be implemented to achieve a disentangled representation. The effect of this disentanglement on the prognostic performance will be studied. The upsampling part of the U-Net can be seen as a generator, generating a deformation field from several input features. The effect of changing a feature in the latent space on the deformation field can therefore be directly visualized. This will be done in a qualitative manner, as no definite quantitative metric exists that can adequately capture disentanglement. Different quantitative metrics disagree in a large way about the degree of disentanglement[31, 33]. Our unique contributions are as follows:

- Implementation of the Hessian Penalty for image registration in PAM.
- Qualitative assessment of the degree of disentanglement.
- Study the effect of the Hessian Penalty on prognostic performance with PAM.

2.2 Methods

PAM is not suited for the implementation of HP as the U-Net responsible for creating the elastic deformation field has skip layers. In other words, the decoder, besides the features in the latent space, has other inputs originating from earlier layers of the encoder. To solve this problem, an autoencoder, without skip-layers, was added to a pre-trained instance of PAM, with the task of reconstructing the elastic deformation field generated by PAM. The pre-trained PAM was the same as the network trained in Chapter 1. The combination of PAM with the autoencoder will be named the 'combined network'. This autoencoder consists of 6 down- and upsampling layers, which can be seen in Figure 11.

This autoencoder was trained using four loss terms. The first is the correlation loss between the fixed image and the elastically deformed image, using the reconstructed deformation field. The second term is the mean square error between the original and reconstructed elastic deformation field. The third term is the Hessian Penalty. The last term is the

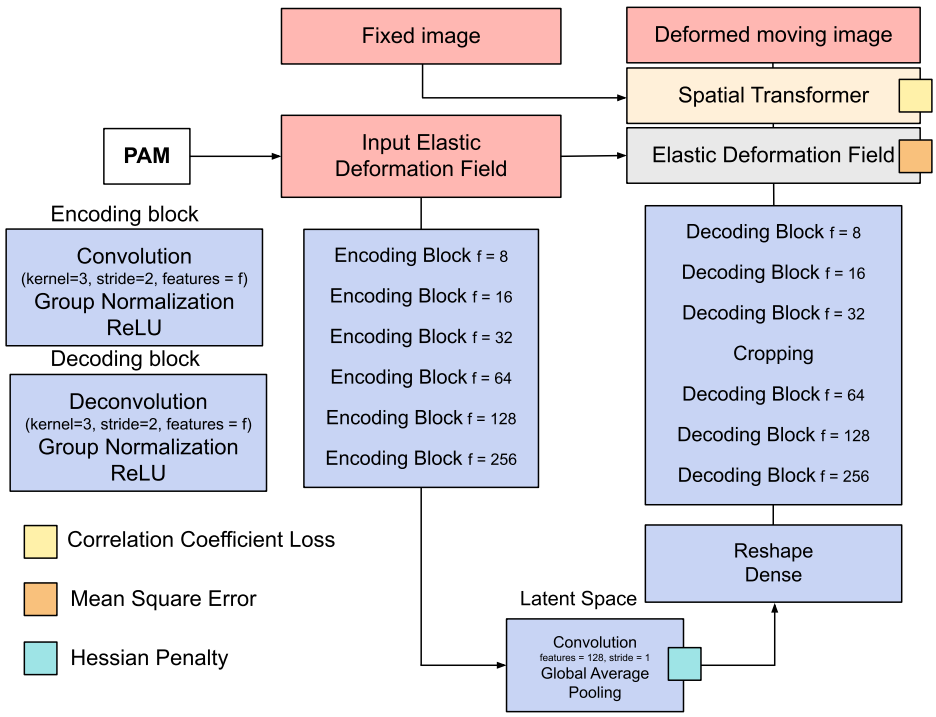


Figure 8: The architecture of the autoencoder that is added to the original network. The architecture of the original network can be found in figure 4.

total variation penalty on the reconstructed deformation field, stimulating smoothness. The loss function used was:

$$L = L_{CC} + L_{MSE} + 0.1L_{HP} + 0.001L_{TVP} \tag{3}$$

with CC being the correlation loss, TVP the total variation penalty and MSE the mean square error.

The processed TCIA dataset, the specifics of which are discussed in Chapter 1, was again used for pre-training the combined network, once with HP and once without. Other architectures and training regimes were also implemented, but those did not result in convergence.

Feature extraction from the latent space of both models and survival analysis was performed in a similar fashion as in Chapter 1. Wilcoxon signed-rank test was used to assess if the AUC were statistically different between both networks with and without HP and the original network. Cohen's kappa statistic was calculated to assess the concordance of the predictions of both networks with and without HP. Both networks were also compared to the original network. Shapley values were calculated and the top 5 most predictive features were chosen for qualitative analysis. Effects of changing individual features on the registration were visually inspected for both the network with and without the HP.

2.3 Results

The combined networks with and without HP converged to a correlation coefficient loss between the fixed image and the transformed moving image of 0.45. This is comparable to the correlation coefficient loss of the original network of 0.40. The combined network that was trained without HP achieved an AUC of 0.58 for the total test set, and 0.59 for the subset of patients with thoracic cancers. Results for the network trained with HP achieved an AUC of 0.59 for the total test set, and 0.60 for the subset of patients with thoracic cancers. There was no significant difference in prognostic performance between both networks (p=0.97). The original network achieved an AUC of 0.61 on the whole test-set and 0.64 on the subset of patients with thoracic cancers. No significant difference was found between the original network and the networks with and without HP (p=0.97 and p=0.94, respectively). Scan pairs in the test set were categorized in a low, medium, or high risk of survival based on the predictions of PAM using the 33rd and 67th percentile as thresholds. Cohen's kappa statistic was equal to 0.53 between the predictions of the network with and the network without HP. Figure 9 shows the results of the more fine-grained survival analysis for both networks.

Adjusting the features in the latent space of the registration network for an input scan pair will result in an accordingly modified deformation field. Figure 10 exemplifies these changes in elastic deformations by adjusting in opposite directions the top 5 predictive features according to the Shapley values. It can be seen that deformations are more localized and of a smaller magnitude for the network trained with Hessian Penalty. Some features change the position of the anterior thoracic wall, while a different feature deforms the lateral wall. In contrast, the results for the network trained without the Hessian Penalty show that changing a feature in the latent space results in deformations throughout the whole volume, not specific to any anatomical location. Similar results were obtained for different scan pairs.

2.4 Discussion

The results show that HP enforces a more disentangled representation. Deformations are more localized to certain anatomical regions and are also of a lesser magnitude. The deformations are mostly localized in the thoracic wall, where different features of the latent space change the shape of the thoracic wall in different directions. For example, features respectively deform the anterior, posterior, and lateral parts of the thoracic wall. No clear clinical significance can be given to these deformations.

To implement HP an autoencoder was added, which resulted in decreased prognostic performance. Not only that, but both combined networks converged to a higher correlation coefficient loss, which indicates worse registration performance. The cause can be the single bottleneck layer in the autoencoder, as the original U-Net network did not have a bottleneck layer. The skip-layers transferred information from the encoder to the decoder. State-of-the-art image registration architectures use U-Nets, which might explain why a simplified version of a U-Net, the autoencoder, produces worse image registrations. HP can be implemented such that multiple inputs can be used, making it suitable for integration with a U-Net. Further research that focuses on disentangling image registration models could implement this. Due to hardware constraints, it was not possible to implement it in this study.

The results also show a decrease in predictive performance in both the network with and without HP, compared to the results of the original network. However, this degradation is not statistically significant. There is no significant difference in prognostic performance between both combined networks, indicating HP does not limit the capacity of the model to capture morphological changes. As HP can be adjusted to incorporate multiple inputs, it would be interesting to implement HP into the original network.

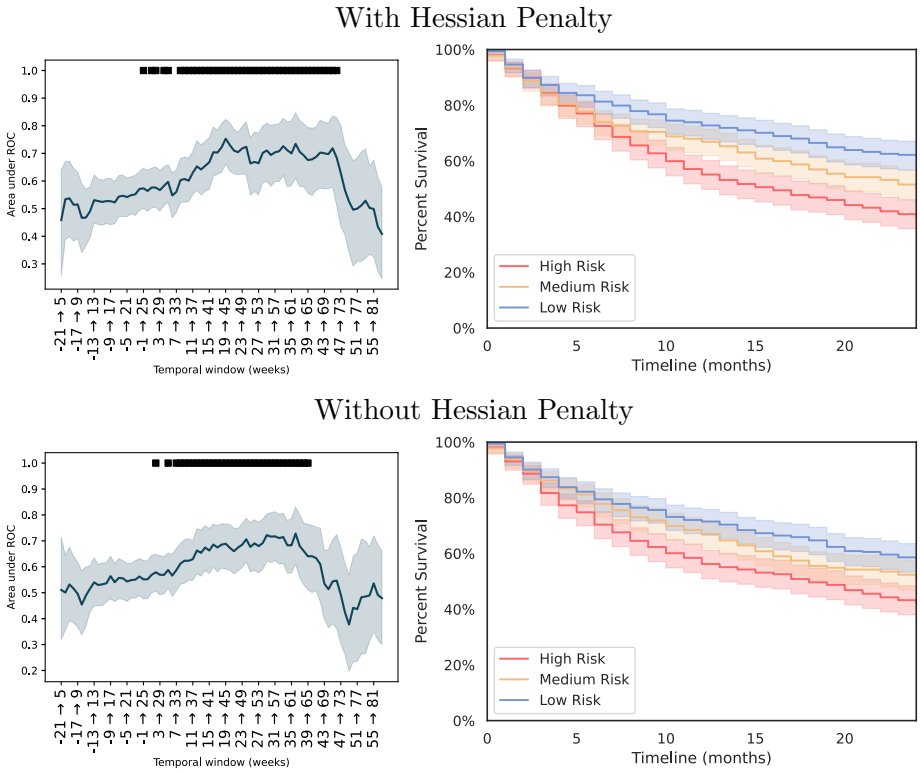


Figure 9: Left: ROC AUC with a sliding temporal window. The X-axis represents the temporal window in which both CT scans were performed with respect to the start of treatment. Black bars represent statistically significant results at that time-point. Right: Kaplan Meier curves with the timeline representing months after the second scan and the y-axis percent survivals. Patients were split based on the predictive probability of survival in three groups: high (blue), medium (yellow), and low (red) probability split on the 33rd and 67th percentile. All results were generated for the subset of patients with thoracic cancers. The upper row represents the network with HP and the lower row the network without HP.

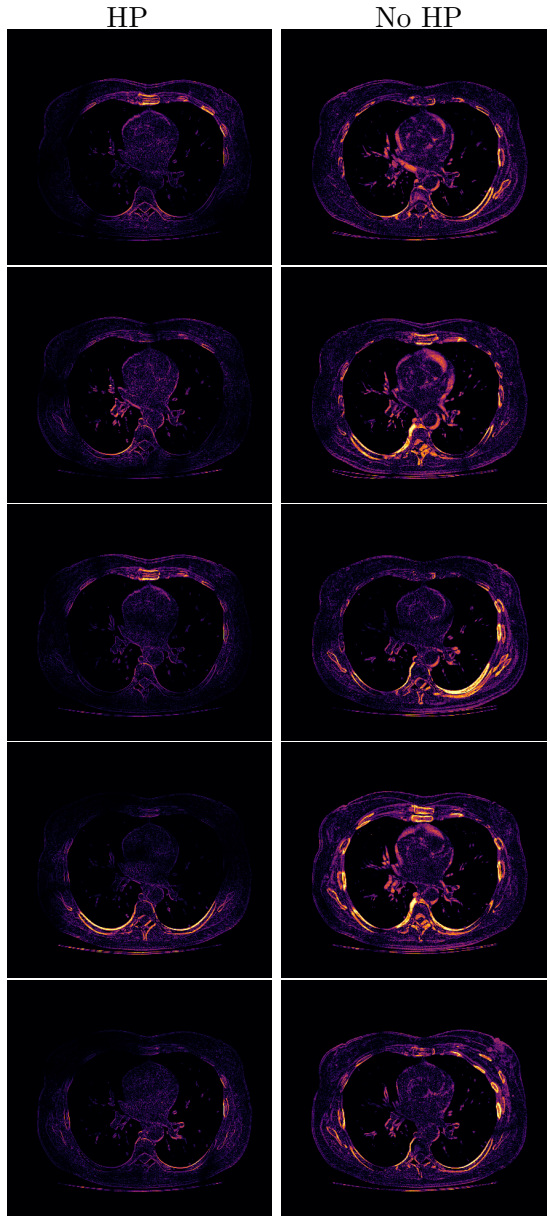


Figure 10: Plots represent the absolute difference between two elastically deformed images. These are created by changing one feature in a positive direction for one image and in a negative direction for the other image. Changing a feature results in a change in the elastically deformed image. Left column: network with Hessian Penalty. Right column: network without Hessian Penalty. The five rows represent the top 5 most predictive features with the top row representing the most predictive feature.

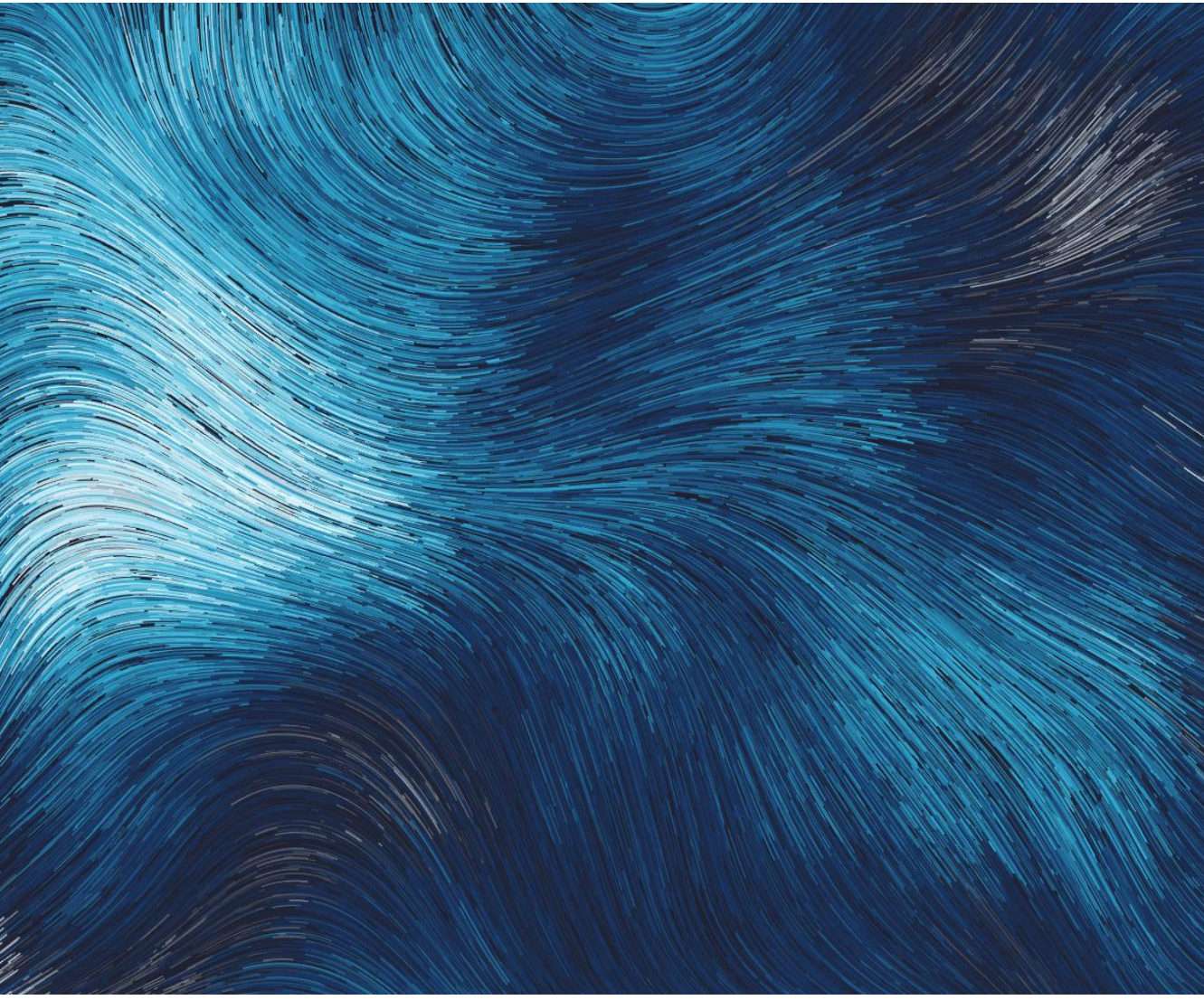
However, due to the presence of skip-layers between the encoder and decoder the explainability of the model is limited, as the features from the latent space will not fully represent the deformations. A choice has to be made between increased prognostic performance and explainability, which will guide the direction of further research.

HP enforces more anatomically localized deformations, and with a smaller magnitude, as we observed. For a disentangled representation, a smaller magnitude of deformation could be the result of minimizing the effect on the deformations caused by other features. The smaller magnitude of deformations can be seen However, the deformations obtained from the network trained with HP are not that fine-grained, and thus interpretable enough, such that clinical prognostic factors can be easily identified. As only the most prognostic features are changed, we hypothesized that these features would contain information about tumor growth, degree of inflammation, change in volume of atelectasis. However, the disentangled features seem to mainly influence relatively large deformation that seems to influence the general shape of the patient, e.g. change in muscle mass, and the volume of the lungs. We hypothesize that by implementing HP on a larger network, better capable of image registration, more fine-grained clinically relevant deformations might be obtained. Currently, however, the results indicate that disentanglement was partially successful as we obtained a modular representation but lack of explicitness.

We obtained a disentangled feature space but more measures are needed to create explainable predictions. Knowing what deformations a single feature represents does not in and of itself lead to explainability, as it is not clear how these features are combined to produce a final prediction. Also, using more than several features leads to a loss in explainability, as a very sparse feature space is necessary to obtain an explainable model.

2.5 Conclusion

Two deep learning models were trained to perform image registration, with and without the implementation of the Hessian Penalty, to study the effect of feature disentanglement on registration quality. Features from the disentangled model were extracted and linked to survival using a Random Forest. Effects of disentanglement on prognostic performance were studied. Results showed the model was capable of predicting 1-year survival to a statistically significant degree. Visual analysis of the image registration showed that a more disentangled latent space was achieved, as deformations were anatomically localized. However, the specific method for implementing the Hessian Penalty caused a drop in prognostic performance.



Chapter Three

Anatomical Fidelity and Realism in Morphological Changes via Adversarial Learning

Abstract

Chapter 1 showed that a deep learning model pre-trained for image registration is capable of predicting survival for cancer patients receiving immunotherapy. We hypothesize that increasing the capacity of the network to model morphological changes will increase the prognostic performance. The total variation penalty on the deformation field was lowered to accommodate for large morph ological changes. Furthermore, a discriminator was added that had to differentiate between a real and a fake image. The real image was the affinely transformed image and the fake image is the elastically deformed image. Two networks were trained to perform image registration with this reduced penalty, one with and one without the discriminator. Training with the discriminator was hypothesized to limit unrealistic deformations. Quantitative and visual analysis showed that training with the discriminator resulted in a deformation where less folding occurred. Reducing the penalty had a positive effect on prognostic performance; however, the difference was not statistically significant. No difference was found in prognostic performance between the network trained with and without the discriminator. Future research should study the impact of the discriminator on the realism of the deformation field.

3.1 Introduction

A deep learning network, PAM, was presented in Chapter 1 that can be used for prognostication. PAM is initially trained to perform image registration on CT scans, but features from the latent space of the network also are used to predict survival using a random forest classifier. It is hypothesized that

the features which are generated by the network are clinically significant, as the network is capable of capturing morphological changes between scans. However, there are opportunities for improvement in pre-training the network to perform image registration as in some cases the network cannot model large deformations between two scans. One approach to model large deformations is based on elastic registration which uses the total variation penalty [28] since it minimizes large unrealistic deformations. Despite the usefulness of the regularization penalty, it could lead to an inability of the network to model large clinically significant deformations(e.g. whole lung atelectasis). Reduction of this penalty might therefore lead to an increase in prognostic performance. However, reducing the penalty can lead to large discontinuities in the deformation field, losing trustworthiness in registration quality. Therefore, to improve the prognostic performance of the network it is essential to maximize the capacity of the network to model morphological changes while keeping the quality of the image registration realistic.

One of the most recent approaches to obtain realistic high-quality images in the deep learning domain is using generative adversarial networks[34]. A generative adversarial network consists of two neural networks known as generator and discriminator, which compete together. The generator is trained to generate images while the discriminator is trained to distinguish real from generated images. To improve the realism of the generated output, an adversarial loss is employed, which uses the capacity of the discriminator to detect fake images. A regularization term is added to the loss function of the generator that during training is minimized such that the generator produces images that can not be differentiated from real images by the discriminator. The generator and discriminator are trained in an alternating fashion, thereby improving both the generator and discriminator during training.

In the domain of medical image, registration adversarial loss has been implemented in both supervised and unsupervised settings [35][36][37]. However, the focus of these studies is to increase registration accuracy and not necessarily create realistic deformations. We hypothesize that training an image-registration network with adversarial loss will lead to an improvement of the capacity of the network to model morphological deformations while keeping the number of discontinuities to a minimum. In this study, an adversarial loss will be implemented to the original PAM network and a qualitative assessment will be performed to assess the realism of the registration. Quantitatively, both the smoothness of the deformation field as well as the effect of adversarial loss on prognostic performance will be assessed.

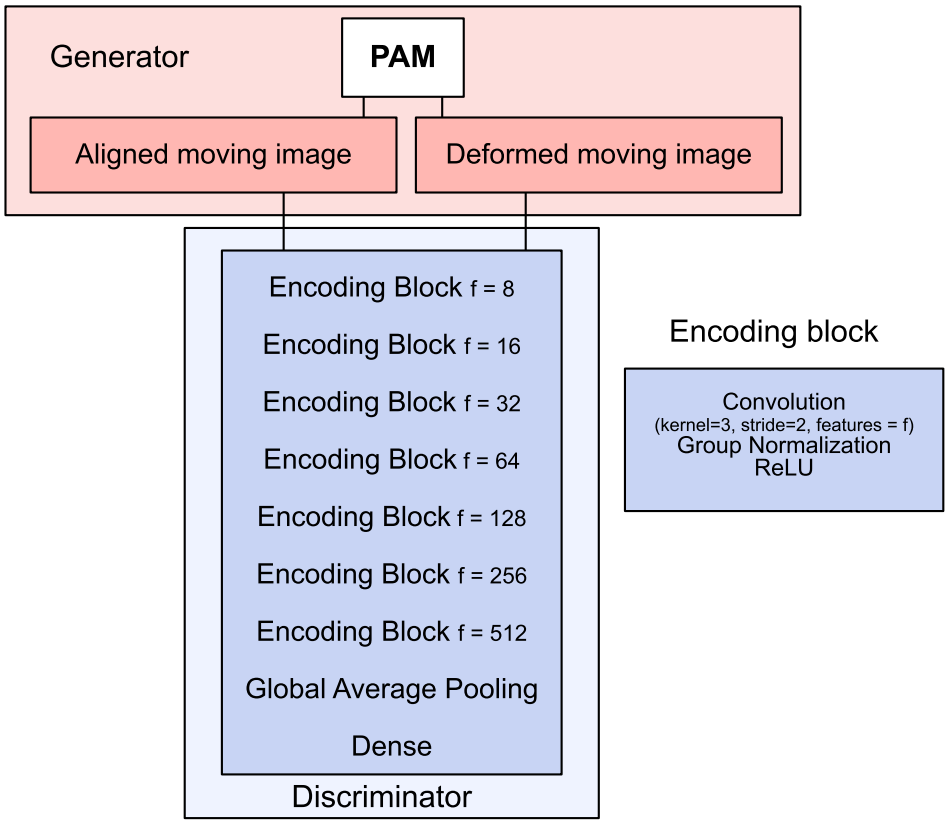


Figure 11: The architecture of the discriminator. The input to the discriminator are the aligned moving image (real image) and deformed moving image (fake image) from PAM. The location where these images are generated in PAM can be seen in Figure 4.

To summarize the contributions are as follows:

- Study the effects of adversarial loss on prognostic performance.
- Comparing quantitatively the smoothness of the deformation fields for the network with and without adversarial loss.

3.2 Methods

A discriminator was implemented, the architecture of which is depicted in Figure 11. The discriminator needs two inputs, a real and a generated image. The real image is the affinely transformed volume and the generated image is the elastically deformed image. The affinely transformed volume

was chosen instead of the fixed or moving image because the affine transformation creates boundaries around the volume that can be easily detected by the discriminator. Unrealistic deformations are in most cases caused by the elastic deformation, so the affinely transformed image can be used as the real image.

The loss function of the discriminator is:

$$L_{discriminator} = BCE^0_{real} + BCE^1_{fake} \tag{4}$$

with BCE representing the binary cross entropy, and the superscript denoting what the prediction should be compared to, with 0 representing a real image and 1 a fake image.

Only the elastic part of the generator is trainable. The loss function of the generator is:

$$L_{generator} = L^{elastic}_{cc} + 0.01L^{elastic}_{variation} + 0.1BCE^0_{fake} \tag{5}$$

The last term represents the loss from the discriminator, which was left out for the network trained without the discriminator. To minimize this term, the generator has to produce an image that can not be differentiated from a real image, in this case, the affinely transformed image.

Unsupervised training to perform image registration was performed on the same dataset as Chapter 1. To assess the regularity of the deformation field, ϕ, the Jacobian matrix around a voxel p is calculated using the voxelmorph package[24]: $J_\phi(\mathbf{p}) = \nabla\phi(\mathbf{p}) \in \mathcal{R}^{3\times3}$. For 20 input scan pairs all voxels with $|J_\phi(\mathbf{p})| \leq 0$ were counted, because these voxels represent areas where sharp deformations occur. These sharp discontinuities are called folding, as the vectors in the deformation field tend to overlap in these areas. This was performed for the same scan pairs for both the original and the adversarial network.

To perform a qualitative assessment of the registration quality, the deformation fields were visualized and compared to assess differences. To visualize the areas where the discriminator is focussing on, attention maps using the GradCAM method were plotted [38]. Feature extraction from the adversarial network was performed using the same protocol and dataset as also mentioned in Chapter 1. To study the impact of the adversarial loss on the prognostic capabilities of the network, results for the survival analysis of the adversarial network were compared to the results of the network trained without adversarial loss.

3.3 Results

The correlation loss reached for the network trained without the discriminator was 0.18, while the network trained with the discriminator reached

a cross-correlation loss of 0.26. To assess registration quality, the deformation fields were generated using both the network trained with and without the discriminator. Figure 12 shows the moving image and the deformation field. Figure 12 also shows the location of the voxels with a negative Jacobian determinant since these voxels represent areas where folding occurs. This figure also illustrates the areas where the discriminator focuses on to assess unrealistic deformations.

The mean number of negative Jacobian determinant voxels was calculated for both networks with 30 scan pairs used as input. $0.42\% \pm 0.26$ of all voxels had a negative Jacobian determinant for the network trained without the discriminator. In comparison, $0.13\% \pm 0.10$ of all voxels had a negative Jacobian determinant for the network trained with the discriminator.

The network trained without the discriminator achieved an AUC of 0.64 for the total test set, and 0.65 for the subset of patients with thoracic cancers. Results for the network trained with the discriminator achieved an AUC of 0.64 for the total test set, and 0.66 for the subset of patients with thoracic cancers. There was no significant difference in prognostic performance between both networks. Scan pairs in the test set were categorized in a low, medium, or high risk of survival based on the predictions of PAM using the 33rd and 67th percentile as thresholds. Cohen's kappa statistic was equal to 0.49 between the predictions of the network with and without the discriminator.

3.4 Discussion

The results show that the discriminator causes the network to converge to a higher correlation loss, indicating worse registration performance. However, as the original network with the larger penalty had an even higher correlation loss, both of the networks trained in this chapter had an increase in registration performance. The number of voxels with a negative Jacobian determinant was lower for the network with the discriminator, indicating less folding caused by the deformation field. There was a slight improvement visible in prognostic performance for both of the networks trained in this chapter compared to the original network; however, the difference was not statistically significant.

As we want to constrain the deformation field to provide a realistic registration it might seem counter-intuitive to then train a discriminator on the images instead of the deformation field. However, as we are training the registration algorithm in an unsupervised manner, it is not straightforward to get examples for the desired (and therefore real) deformation field to train a discriminator. One method could be a version of PAM trained with a higher penalty to generate deformation fields which will serve as the real

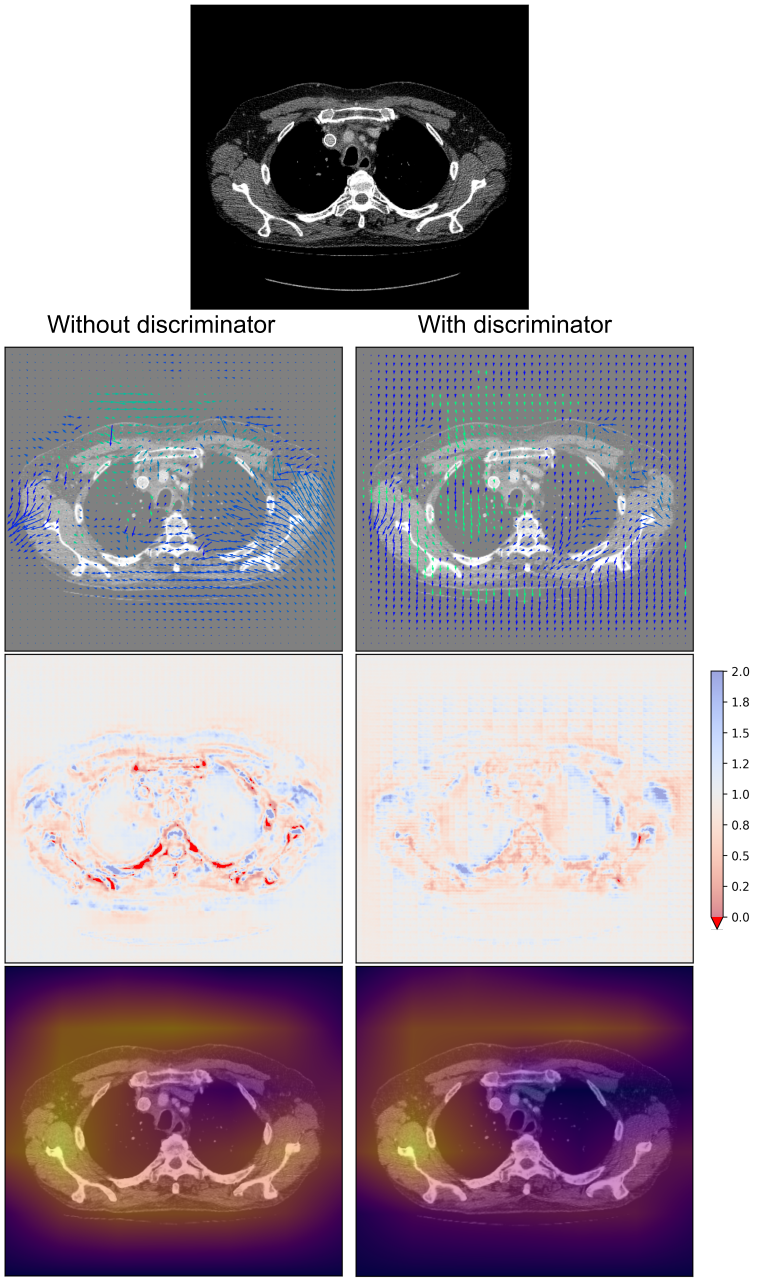


Figure 12: The first row is the target image for registration. The second row shows the deformation field overlaid on the affinely transformed moving image, where different colors represent the direction of the deformation. The third row shows the value of the determinant of the Jacobian at each voxel. The last row shows the GradCAM activation heatmap for the elastically deformed moving image.

image for the discriminator. The fake image will then be provided by the instance of PAM trained with a lowered penalty. The main problem with this approach is computational restraints, as two instances of PAM need to be active at the same time. Further research might try this approach to see if this increases performance.

Minimizing the occurrence of folding should not be a goal in and of itself. But having an excessive amount of folding indicates highly unlikely registration performance. Due to differences in the anatomy caused by the disease or the treatment, the deformation field might produce sharp deformations to accommodate these anatomical deformations. Folding might therefore be desired because these anatomical deformations might contain a lot of clinical information. Qualitative analysis showed that a large portion of the voxels where folding occurs is located in the bottom and top slices in the axial direction. These slices contain a registration artifact due to the affine transformation. Folding in these slices is more pronounced in the deformation field created by the network trained without the discriminator.

The network trained with the discriminator converges to a higher correlation loss compared to the other network. There are two possible explanations for this behavior. The first is that the discriminator keeps the network from deforming the volume in an unrealistic manner, thereby having a trade-off between realism and registration accuracy. This was the goal of including the discriminator. However, another possible explanation is that without the discriminator the network converges in a more stable and smoother way. This could indicate that training with the discriminator introduces too much noise to the loss making the network unable to converge to an optimal registration performance. The attention maps show that the discriminator is primarily focussing on areas with larger deformations, but no specific correlations can be seen with areas with a lot of voxels with a negative Jacobian determinant. No difference can be seen in attention maps in the results of both networks.

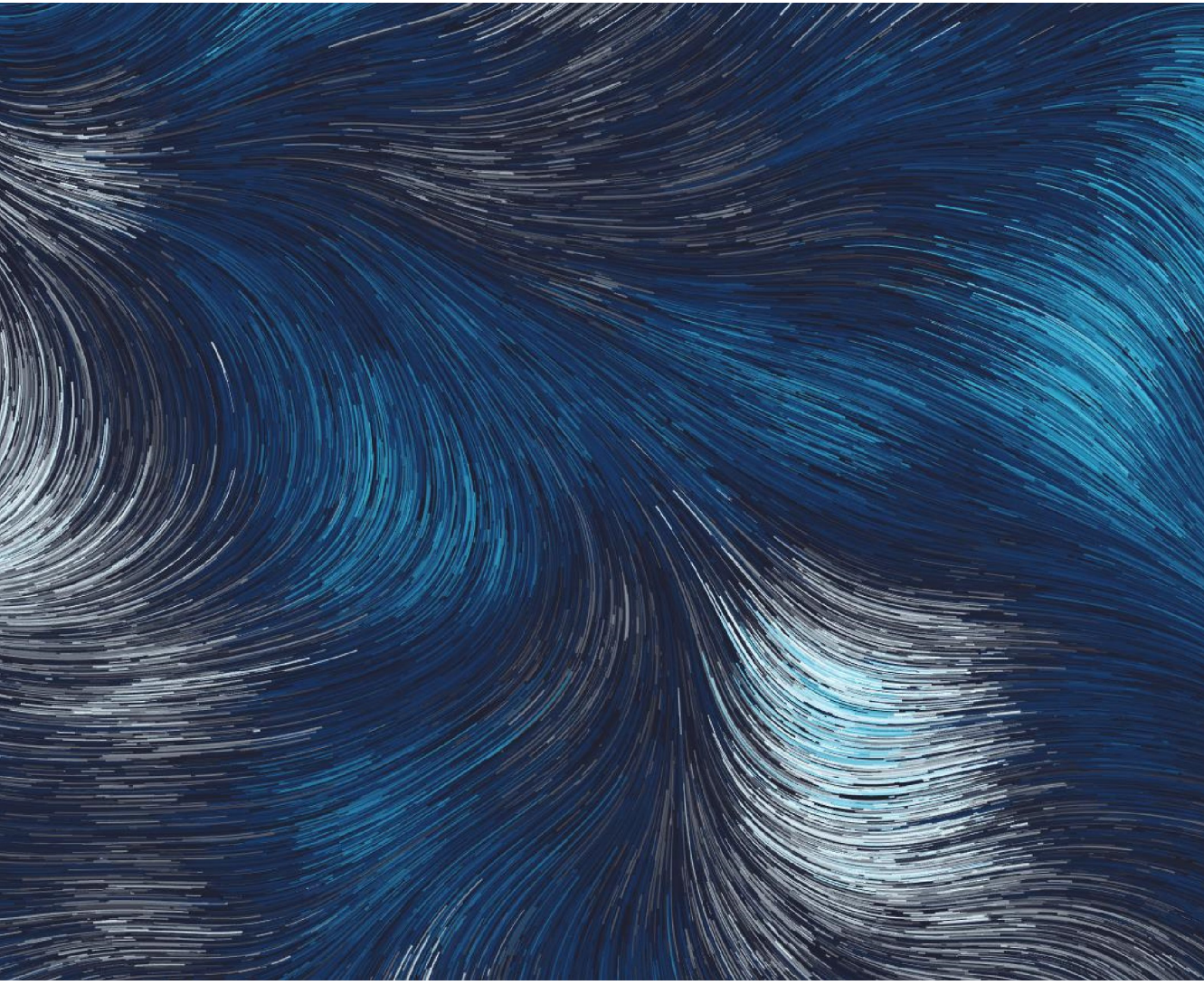
Survival analysis shows that decreasing the total variation improves the prognostic performance for both networks; however, this improvement is not statistically significant. The explanation for the increased prognostic performance for both networks is that because of the lowered total variation penalty the networks have a better capacity to model morphological changes.

Future work to improve registration performance should try to qualitatively evaluate which deformations are desired, and which ones are unrealistic. Improving the quality of the registration should focus on eliminating specific unrealistic deformations while increasing the registration accuracy. This might lead to increased prognostic performance. However, it is not

known how much further prognostic performance can be pushed by increasing the quality of the registration.

3.5 Conclusion

Training PAM with a discriminator allows for more realistic deformations because of the minimization of folding in the deformation field. Decreasing the total variation penalty increases the prognostic performance; however, no statistical difference was obtained.



General Conclusion and Outlook

General Conclusion

In this thesis, we have shown that PAM, a deep learning model trained to perform image registration, can extract prognostic features from thoracic CT scans of patients receiving immunotherapy. This supports the central hypothesis that morphological changes between scans can be captured using image registration, and that these changes contain prognostic information. We have shown that the quality of the image registration correlates with prognostic performance, further supporting this hypothesis. PAM was also compared to RECIST 1.1 and results showed that the prognostic performance of PAM is similar to RECIST.

We have studied two aspects of PAM's architecture. The first, presented in Chapter 2, is the disentangled feature space by implementing the Hessian Penalty. Disentangled features pave the way for increasing PAM's explainability, as PAM in the current state is a black-box model. As the disentangled features obtained in Chapter 2 are not yet fine-grained enough, they could not be directly linked to any clinically interpretable changes in the body. Further research should focus on extracting finer-grained morphological changes from prognostic features. The second aspect, presented in Chapter 3, is the improvement of the quality of the registration, by lowering the penalty on the elastic deformation field while keeping the deformation field realistic by training PAM with a discriminator. Reducing the penalty resulted in an increased prognostic performance compared to the network trained in Chapter 1. Also, training with the discriminator has been shown to reduce the number of unrealistic deformations in the deformation field compared to a similar network trained without the discriminator. However, the increased realism did not translate to an increase in prognostic performance. This indicates that only lowering the penalty affects prognostic performance. Because training with a discriminator is difficult, it should be considered if the added gain of a discriminator, the realism of the deformation field, outweighs the increased complexity.

Outlook

It is recommended that future research should first focus on implementing PAM for other anatomical regions, specifically the abdomen. The prognostic results from PAM can then be compared to RECIST, which should be done on a highly curated multi-center dataset. PAM was trained on data from the Netherlands Cancer Institute, which is a highly specialized cancer center. As this patient population is different from those at a general hospital, it is important to assess the prognostic performance of PAM on a

diverse patient population to ensure robust performance. The desired level of performance and accuracy needs to be quantified, which can be done by comparing with RECIST. PAM should be optimized to achieve this desired level of performance while reducing complexity as much as possible, maximizing the chances for clinical adoption. A quantified goal for the performance will give a clear indication in what way PAM needs to be further improved upon before clinical deployment.

PAM is currently trained to predict survival, but other outcomes might be more useful for clinicians. Survival was used as an easily obtainable proxy for treatment response, but survival is also a very noisy outcome metric. For example, patients can die from other causes not related to their disease or treatment, which is not accounted for in this study. It is therefore recommended to analyze if a model that can sufficiently predict survival will provide a benefit to the patient, or if other outcome metrics provide better value. An alternative outcome is pathological response, but in metastatic patients, this information is not available (patients cannot be followed-up by serial biopsies) nor can be considered representative (response of one lesion vs response of remaining lesions). An outcome that can also be used is the change in the experienced quality of life by patients. Information about the quality of life is however expensive to obtain, difficult to measure due to subjectivity, and not readily available in a retrospective cohort. As RECIST is currently used to assess radiological response, and the goal of PAM is to overcome the limitations of RECIST, it is not recommended to train PAM to predict radiological response based on RECIST. Similarly, total tumor volume (via segmentation) is expensive to obtain, and has intrinsic limitations that cannot be overcome (e.g. pseudo-progression). All alternatives to survival as an outcome have their drawbacks but might provide more clinical value.

If PAM can tackle RECIST's limitations, while increasing the prognostic performance, PAM can be highly beneficial to both patients and clinicians. However, a black-box model that can predict survival can be difficult to get accepted in a clinical setting. If clinical adoption proves difficult, PAM can still provide key insights into which morphological differences in longitudinal imaging are prognostic. A key prerequisite is that PAM can link and visualize prognostic features to fine-grained deformations, therefore providing explainability.

In a future scenario where PAM will be implemented in clinical practice, it is important to consider the specifics of the implementation. PAM should give quantitative prognostic information, where the contribution of single deformations to the prognosis can be visualized. For example, visualizing for a patient that PAM detected a decrease in tumor volume,

contributing to an improvement of 30% of the final outcome, and the onset of atelectasis, contributing to a decrease of 20% to the final outcome. The quantitative prognostic information output by PAM can then be taken into account when the patient is discussed at the tumor board. The AI system will serve as one of many tools used by doctors to find the best course of action for the patient. The doctors need to balance different objectives such as maximizing survival, maximizing the quality of life of the patient while also incorporating specific wishes of the patient. PAM could help to provide information that will lead to more objective clinical decision-making, resulting in better outcomes for the patient. To maximize the potential of PAM, clinical studies need to be performed to investigate in what context PAM provides a benefit, with direct comparisons to current standards such as RECIST. These results can then be translated into clinical guidelines, which will be needed to get PAM to be accepted by clinicians.

In conclusion, this work shows that AI can be used to obtain quantitative prognostic information from medical images. This work tries to contribute to the exploration of AI systems for possible implementation in clinical practice, to thereby improve healthcare and increase the quality of life for individual patients. The thesis serves as a base for further research, which is needed for eventual clinical adoption.

References

[1] Yiping Yang. "Cancer immunotherapy: harnessing the immune system to battle cancer". en. In: *The Journal of Clinical Investigation* 125.9 (Sept. 2015). Publisher: American Society for Clinical Investigation, pp. 3335–3337. ISSN: 0021-9738. DOI: 10.1172/JCI83871. URL: https://www.jci.org/articles/view/83871 (visited on 02/14/2022).

[2] K. Esfahani et al. "A Review of Cancer Immunotherapy: From the Past, to the Present, to the Future". en. In: *Current Oncology* 27.s2 (Apr. 2020). Number: s2 Publisher: Multidisciplinary Digital Publishing Institute, pp. 87–97. ISSN: 1718-7729. DOI: 10.3747/co.27.5223. URL: https://www.mdpi.com/1718-7729/27/12/5223 (visited on 02/14/2022).

[3] Mizuki Nishino, Hiroto Hatabu, and F. Stephen Hodi. "Imaging of Cancer Immunotherapy: Current Approaches and Future Directions". In: *Radiology* 290.1 (Jan. 2019), pp. 9–22. ISSN: 0033-8419. DOI: 10.1148/radiol.2018181349. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6312436/ (visited on 04/26/2021).

[4] Marco Calandri et al. "The role of radiology in the evaluation of the immunotherapy efficacy". eng. In: *Journal of Thoracic Disease* 10.Suppl 13 (May 2018), S1438–S1446. ISSN: 2072-1439. DOI: 10.21037/jtd.2018.05.130.

[5] Lawrence H. Schwartz et al. "RECIST 1.1-Update and clarification: From the RECIST committee". eng. In: *European Journal of Cancer (Oxford, England: 1990)* 62 (July 2016), pp. 132–137. ISSN: 1879-0852. DOI: 10.1016/j.ejca.2016.03.081.

[6] Lesley Seymour et al. "iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics". In: *The Lancet. Oncology* 18.3 (Mar. 2017), e143–e152. ISSN: 1470-2045. DOI: 10.1016/S1470-2045(17)30074-8. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5648544/ (visited on 01/20/2022).

[7] Liza C. Villaruz and Mark A. Socinski. "The clinical viewpoint: definitions, limitations of RECIST, practical considerations of measurement". eng. In: *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 19.10 (May 2013), pp. 2629–2636. ISSN: 1557-3265. DOI: 10.1158/1078-0432.CCR-12-2935.

[8] Christiane K. Kuhl et al. "Validity of RECIST Version 1.1 for Response Assessment in Metastatic Cancer: A Prospective, Multireader Study". eng. In: *Radiology* 290.2 (Feb. 2019), pp. 349–356. ISSN: 1527-1315. DOI: 10.1148/radiol.2018180648.

[9] Amylou C. Dueck et al. "Validity and Reliability of the U.S. National Cancer Institute's Patient-Reported Outcomes Version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE)". In: *JAMA oncology* 1.8 (Nov. 2015), pp. 1051–1059. ISSN: 2374-2437. DOI: 10.1001/jamaoncol.2015.2639. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4857599/ (visited on 01/20/2022).

[10] Ethan Basch et al. "Development of the national cancer institute's patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE)". In: *Journal of the National Cancer Institute* 106.9 (Sept. 2014). ISSN: 0027-8874. DOI: 10.1093/jnci/dju244. URL: http://www.scopus.com/inward/record.url?scp=84909994482&partnerID=8YFLogxK (visited on 01/20/2022).

[11] Kathleen E. Fenerty et al. "Predicting clinical outcomes in chordoma patients receiving immunotherapy: a comparison between volumetric segmentation and RECIST". In: *BMC Cancer* 16.1 (Aug. 2016), p. 672. ISSN: 1471-2407. DOI: 10.1186/s12885-016-2699-x. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4995658/ (visited on 01/20/2022).

[12] Cheng Jin et al. "Predicting treatment response from longitudinal images using multi-task deep learning". en. In: *Nature Communications* 12.1 (Dec. 2021), p. 1851. ISSN: 2041-1723. DOI: 10.1038/s41467-021-22188-y. URL: http://www.nature.com/articles/s41467-021-22188-y (visited on 01/20/2022).

[13] Lin Lu et al. "Deep learning for the prediction of early on-treatment response in metastatic colorectal cancer from serial medical imaging". en. In: *Nature Communications* 12.1 (Nov. 2021), p. 6654. ISSN: 2041-1723. DOI: 10.1038/s41467-021-26990-6. URL: https://www.nature.com/articles/s41467-021-26990-6 (visited on 01/20/2022).

[14] Alessa Hering et al. "Whole-Body Soft-Tissue Lesion Tracking and Segmentation in Longitudinal CT Imaging Studies". In: *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning*. Ed. by Mattias Heinrich et al. Vol. 143. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 312–326. URL: https://proceedings.mlr.press/v143/hering21a.html.

[15] Anirudh Joshi et al. "OncoNet: Weakly Supervised Siamese Network to automate cancer treatment response assessment between longitudinal FDG PET/CT examinations". In: *arXiv:2108.02016 [cs, eess]* (Aug. 2021). arXiv: 2108.02016 version: 1. URL: http://arxiv.org/abs/2108.02016 (visited on 01/20/2022).

[16] Stefano Trebeschi et al. "Prognostic Value of Deep Learning-Mediated Treatment Monitoring in Lung Cancer Patients Receiving Immunotherapy". eng. In: *Frontiers in Oncology* 11 (2021), p. 609054. ISSN: 2234-943X. DOI: 10.3389/fonc.2021.609054.

[17] Stefano Trebeschi et al. "Development of a Prognostic AI-Monitor for Metastatic Urothelial Cancer Patients Receiving Immunotherapy". eng. In: *Frontiers in Oncology* 11 (2021), p. 637804. ISSN: 2234-943X. DOI: 10.3389/fonc.2021.637804.

[18] *Pattern Recognition and Machine Learning*. en. URL: https://link.springer.com/book/9780387310732 (visited on 01/20/2022).

[19] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. en. Ed. by Francis Bach. Adaptive Computation and Machine Learning series. Cambridge, MA, USA: MIT Press, Nov. 2016. ISBN: 978-0-262-03561-3.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *arXiv:1505.04597 [cs]* (May 2015). URL: http://arxiv.org/abs/1505.04597 (visited on 12/03/2021).

[21] A. Sotiras, C. Davatzikos, and N. Paragios. "Deformable Medical Image Registration: A Survey". en. In: *IEEE Transactions on Medical Imaging* 32.7 (July 2013), pp. 1153–1190. ISSN: 0278-0062, 1558-254X. DOI: 10.1109/TMI.2013.2265603. URL: http://ieeexplore.ieee.org/document/6522524/ (visited on 02/14/2022).

[22] Hamid Reza Boveiri et al. "Medical image registration using deep neural networks: A comprehensive review". en. In: *Computers & Electrical Engineering* 87 (Oct. 2020), p. 106767. ISSN: 0045-7906. DOI: 10.1016/j.compeleceng.2020.106767. URL: https://www.sciencedirect.com/science/article/pii/S0045790620306224 (visited on 02/17/2022).

[23] Francisco P. M. Oliveira and João Manuel R. S. Tavares. "Medical image registration: a review". eng. In: *Computer Methods in Biomechanics and Biomedical Engineering* 17.2 (2014), pp. 73–93. ISSN: 1476-8259. DOI: 10.1080/10255842.2012.670855.

[24] Guha Balakrishnan et al. “VoxelMorph: A Learning Framework for Deformable Medical Image Registration”. eng. In: *IEEE transactions on medical imaging* (Feb. 2019). ISSN: 1558-254X. DOI: 10.1109/TMI.2019.2897538.

[25] E. A. Eisenhauer et al. “New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)”. eng. In: *European Journal of Cancer (Oxford, England: 1990)* 45.2 (Jan. 2009), pp. 228–247. ISSN: 1879-0852. DOI: 10.1016/j.ejca.2008.10.026.

[26] 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)*. Apr. 2017, pp. 578–582. DOI: 10.1109/ISBI.2017.7950587.

[27] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *arXiv:1412.6980 [cs]* (Jan. 2017). URL: http://arxiv.org/abs/1412.6980 (visited on 12/03/2021).

[28] Shengyu Zhao et al. “Unsupervised 3D End-to-End Medical Image Registration with Volume Tweening Network”. In: *IEEE Journal of Biomedical and Health Informatics* 24.5 (May 2020), pp. 1394–1404. ISSN: 2168-2194, 2168-2208. DOI: 10.1109/JBHI.2019.2951024. URL: http://arxiv.org/abs/1902.05020 (visited on 12/03/2021).

[29] Ryan Datteri et al. “Applying the algorithm ”assessing quality using image registration circuits” (AQUIRC) to multi-atlas segmentation”. In: *Medical Imaging 2014: Image Processing*. Vol. 9034. SPIE, Mar. 2014, pp. 355–361. DOI: 10.1117/12.2043756. URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/9034/90341F/Applying-the-algorithm-assessing-quality-using-image-registration-circuits-AQUIRC/10.1117/12.2043756.full (visited on 12/03/2021).

[30] Julia Amann et al. “Explainability for artificial intelligence in healthcare: a multidisciplinary perspective”. In: *BMC Medical Informatics and Decision Making* 20.1 (Nov. 2020), p. 310. ISSN: 1472-6947. DOI: 10.1186/s12911-020-01332-6. URL: https://doi.org/10.1186/s12911-020-01332-6 (visited on 12/14/2021).

[31] Julian Zaidi et al. “Measuring Disentanglement: A Review of Metrics”. In: *arXiv:2012.09276 [cs]* (Jan. 2021). URL: http://arxiv.org/abs/2012.09276 (visited on 12/14/2021).

[32] William Peebles et al. “The Hessian Penalty: A Weak Prior for Unsupervised Disentanglement”. In: *arXiv:2008.10599 [cs]* (Aug. 2020). URL: http://arxiv.org/abs/2008.10599 (visited on 12/14/2021).

[33] Francesco Locatello et al. “Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations”. In: *arXiv: 1811.12359 [cs, stat]* (June 2019). URL: http://arxiv.org/abs/1811.12359 (visited on 12/14/2021).

[34] Ian Goodfellow et al. “Generative Adversarial Nets”. In: *Advances in Neural Information Processing Systems*. Vol. 27. Curran Associates, Inc., 2014. URL: https://arxiv.org/abs/1406.2661 (visited on 01/19/2022).

[35] Mohamed S. Elmahdy et al. “Adversarial Optimization for Joint Registration and Segmentation in Prostate CT Radiotherapy”. en. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by Dinggang Shen et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 366–374. ISBN: 978-3-030-32226-7. DOI: 10.1007/978-3-030-32226-7_41.

[36] Yipeng Hu et al. “Adversarial Deformation Regularization for Training Image Registration Neural Networks”. en. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Ed. by Alejandro F. Frangi et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 774–782. ISBN: 978-3-030-00928-1. DOI: 10.1007/978-3-030-00928-1_87.

[37] Jingfan Fan et al. “Adversarial Similarity Network for Evaluating Image Alignment in Deep Learning Based Registration”. en. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Ed. by Alejandro F. Frangi et al. Vol. 11070. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 739–746. ISBN: 978-3-030-00927-4 978-3-030-00928-1. DOI: 10.1007/978-3-030-00928-1_83. URL: http://link.springer.com/10.1007/978-3-030-00928-1_83 (visited on 01/19/2022).

[38] Ramprasaath R. Selvaraju et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”. In: *International Journal of Computer Vision* 128.2 (Feb. 2020). arXiv: 1610.02391, pp. 336–359. ISSN: 0920-5691, 1573-1405. DOI: 10.1007/s11263-019-01228-7. URL: http://arxiv.org/abs/1610.02391 (visited on 01/31/2022).