VNIVERSITAT DE VALÈNCIA

PhD Thesis

Contributions of biomechanical modeling and machine learning to the automatic registration of Multiparametric Magnetic Resonance and Transrectal Echography for prostate brachytherapy

Oscar José Pellicer Valero

Intelligent Data Analysis Laboratory, Department of Electronic Engineering,
ETSE (Engineering School), Universitat de València (UV),
Av. Universitat, sn, Bujassot, València 46100, Spain

València, July 2022

José David Martín Guerrero
Intelligent Data Analysis Laboratory,
Department of Electronic Engineering,
ETSE (Engineering School),
Universitat de València (UV),
Av. Universitat, sn, Bujassot,
València 46100, Spain

María José Rupérez Moreno
Instituto de Ingeniería Mecánica y
Biomecánica, Departamento de
Ingeniería Mecánica y de Materiales,
Universitat Politècnica de València (UPV),
Camino de Vera, sn,
València 46022, Spain

Doctoral Program in Electronic Engineering
Universitat de València

Contributions of biomechanical modeling and machine learning to the automatic registration of Multiparametric Magnetic Resonance and Transrectal Echography for prostate brachytherapy

Oscar José Pellicer Valero

València, July 2022

Universitat de València Escola Tècnica Superior d'Enginyeria

José David Martin Guerrero, PhD by Universitat de València, Full Professor at Department of Electronic Engineering, ETSE (Engineering School), Universitat de València (UV),

and María Jose Rupérez Moreno, PhD by Universitat Politècnica de València, Associate Professor at Departament of Mechanical and Material Engineering (DIMM), Universitat Politècnica de València (UPV),

STATE THAT:

Oscar José Pellicer Valero, BSc in Engineering in Industrial Technologies and MSc in Industrial Engineering (Specialty in Automatics and Industrial Robotics) by Universitat Politècnica de València, and BSc in Computer Science by Universidad Nacional de Educación a Distancia has undertaken under our supervision the PhD Thesis titled *Contributions of biomechanical modeling and machine learning to the automatic registration of Multiparametric Magnetic Resonance and Transrectal Echography for prostate brachytherapy*, and he submits it to the current call to apply for the Doctoral Degree,

and for the record, we sign this certificate in València on July 2022:

José David Martín Guerrero	María Jose Rupérez Moreno	Vicente González Millán (Department director)

PhD Thesis	*Contributions of biomechanical modelling and machine learning to the automatic registration of Multiparametric Magnetic Resonance and Transrectal Echography for prostate brachytherapy*
Author	MSc Oscar José Pellicer Valero
Supervisors	PhD José David Martín Guerrero
	PhD María José Rupérez Moreno

The tribunal named for evaluating the PhD Thesis cited above, formed by Doctors:

President ______________________________

Vocal ______________________________

Secretary ______________________________

They agree to concede the grade of: ______________________________

And for the record, we sign this certificate:

Acknowledgements

Desarrollar una tesis doctoral nunca es sencillo pero, sin la ayuda de ciertas personas, es sencillamente imposible:

Gracias, José David Martín, por haberme acompañado durante todo este camino como director y como amigo, asesorándome sobre los entresijos de la ciencia, estando siempre atento y dispuesto a echar una mano, haciéndome partícipe de proyectos que me han enriquecido, involucrándote personalmente en mi formación y, en definitiva, siendo un director de tesis ejemplar. Extiendo este agradecimiento a María José Rupérez, quién también me ha apoyado en todo momento, preocupada por que mi proyecto resultara un éxito incluso antes de conocerme, desde mi reunión en septiembre de 2017 con Enrique Nadal, a quien también agradezco su consejo en aquellos momentos de duda.

Gracias a los miembros del IDAL por la total acogida que he recibido desde el primer día. Son pocos y afortunados los que trabajan entre amigos. Sin orden específico, gracias Marcelino Martínez, Rafael Magdalena, Antonio Serrano, Joan Vila, Emilio Soria, Juan Gómez, Juanjo Carrasco y Fernando Mateo.

Gracias a Juan Casanova por haberme acogido en el IVO como si fuera mi casa, por facilitarme mi trabajo, y por haberme involucrado (y haberte involucrado tú también) en proyectos de gran interés científico y personal para ambos. Creo que esta colaboración nos ha ensanchado a ambos las miras, y espero que en un futuro recojamos aún más frutos. Del mismo modo, gracias a Víctor González por haber estado desde el principio a mi lado, asesorándome científicamente, y ayudándome en tareas tan fundamentales como poco recompensadas como es la recogida de datos. Quiero agradecer también a José Luís Marenco y al resto del equipo del IVO que ha contribuido en el éxito de este trabajo de cuatro años: Isabel Martín, María Barrios, Paula Pelechano y José Rubio.

I also want to express my gratitude towards the DKFZ in Heidelberg, which hosted me during a three-month stay under the supervison of Klaus H. Maier-Hein, offering me the privilege to experience working at one of Europe's leading institutions in medical image analysis. Thanks to my colleagues at MIC, and specially to Dimitrios Bounias and Michael Baumgartner, for their patience and warm welcome.

Agradezco también al Ministerio de Ciencia e Innovación y a la Universitat de València por financiar este proyecto y permitirme dedicarme a la investigación por completo, con total libertad y responsabilidad durante estos últimos cuatro años. Gracias también a FME y a Carlo Barbieri por su financiación y confianza durante mis primeros pasos en el mundo de la investigación en analítica de datos médicos.

Finalmente, gracias a mi padre y a mi madre, por creer en mí más que yo mismo, por hacer posible el inicio de este proyecto, y por el apoyo para desarrollarlo y divulgarlo,

evitando (esperemos) que se quede en un cajón. Gracias también a mis amigos, que siempre me han apoyado en esta aventura, y deseo suerte en especial a los que están en mi mismo barco, y aún no tienen la fortuna de poder dedicarse a escribir sus propios agradecimientos; aunque pronto la tendrán.

Gracias a quién no he nombrado, pero que también ha contribuido a que pueda emerger de esta larga batalla con dignidad.

Oscar José Pellicer Valero
València, July 2022.

Contents

List of Figures

List of Tables

Abstract

Prostate cancer (PCa) is the most common malignancy in western males, and third by mortality. After detecting elevated Prostate Specific Antigen (PSA) blood levels or after a suspicious rectal examination, a Magnetic Resonance (MR) image of the prostate is acquired and assessed by radiologists to locate suspicious regions. These are then biopsied, i.e. living tissue samples are collected and analyzed histopathologically to confirm the presence of cancer and establish its degree of aggressiveness.

During the biopsy procedure, Ultrasound (US) is typically used for guidance and lesion localization. However, lesions are not directly visible in US, and the urologist needs to use fusion software to performs MR-US registration, so that the MR-marked locations can be transferred to the US image. This is essential to ensure that the collected samples truly come from the suspicious area.

This work compiles five publications employing several Artificial Intelligence (AI) algorithms to analyze prostate images (MR and US) and thereby improve the efficiency and accuracy in diagnosis, biopsy and treatment of PCa:

1. **Automatic prostate segmentation in MR and US:** Prostate segmentation consists in delimiting or marking the prostate in a medical image, separating it from the rest of the organs or structures. Automating this task fully, which is required for any subsequent analysis, saves significant time for radiologists and urologists, while also improving accuracy and repeatability.

2. **Segmentation resolution enhancement:** A methodology for improving the resolution of the previously obtained automatic segmentations is presented.

3. **Automatic detection and classification of MR lesions:** An AI model is trained to detect lesions as a radiologist would and to estimate their risk. The model achieves improved diagnostic accuracy, resulting in a fully automatic system that could be used as a second clinical opinion or as a criterion for patient prioritization.

4. **Simulation of biomechanical behavior in real time:** It is proposed to accelerate the simulation of biomechanical behavior of soft organs using AI.

5. **Automatic MR-US registration:** Registration allows localization of MR-marked lesions on US. High accuracy in this task is essential for the correctness of the biopsy and/or focal treatment procedures (such as high-rate brachytherapy). Here, AI is used to solve the registration problem in near-real time, while exploiting underlying biomechanically-compatible models.

Resumen

El cáncer de próstata (CaP) es el primer cáncer por incidencia en hombres en países occidentales, y el tercero en mortalidad. Tras detectar en sangre una elevación del Antígeno Prostático Específico (PSA) o tras tacto rectal sospechoso se realiza una Resonancia Magnética (RM) de la próstata, que los radiólogos analizan para localizar las regiones sospechosas. A continuación, estas se biopsian, es decir, se toman muestras vivas que posteriormente serán analizadas histopatológicamente para confirmar la presencia de cáncer y establecer su grado de agresividad.

Durante la biopsia se emplea típicamente Ultrasonidos (US) para el guiado y la localización de las lesiones. Sin embargo, estas no son directamente visibles en US, y el urólogo necesita usar software de fusión que realice un registro RM-US que transfiera la localizaciones marcadas en MR al US. Esto es fundamental para asegurar que las muestras tomadas provienen verdaderamente de la zona sospechosa.

En este trabajo se compendian cinco publicaciones que emplean diversos algoritmos de Inteligencia Artificial (IA) para analizar las imágenes de próstata (RM y US) y con ello mejorar la eficiencia y precisión en el diagnóstico, biopsia y tratamiento del CaP:

1. **Segmentación automática de próstata en RM y US:** Segmentar la próstata consiste en delimitar o marcar la próstata en una imagen médica, separándola del resto de órganos o estructuras. Automatizar por completo esta tarea, que es previa a todo análisis posterior, permite ahorrar un tiempo significativo a radiólogos y urólogos, mejorando también la precisión y repetibilidad.

2. **Mejora de la resolución de segmentación:** Se presenta una metodología para mejorar la resolución de las segmentaciones automáticas anteriores.

3. **Detección y clasificación automática de lesiones en RM:** Se entrena un modelo basado en IA para detectar las lesiones como lo haría un radiólogo, asignándoles también una estimación del riesgo. Se logra mejorar la precisión diagnóstica, dando lugar a un sistema totalmente automático que podría implantarse para segunda opinión clínica o como criterio para priorización.

4. **Simulación del comportamiento biomecánico en tiempo real:** Se propone acelerar la simulación del comportamiento biomecánico de órganos blandos mediante el uso de IA.

5. **Registro automático RM-US:** El registro permite localizar en US las lesiones marcadas en RM. Una alta precisión en esta tarea es esencial para la corrección de la biopsia y/o del tratamiento focal del paciente (como braquiterapia de alta tasa). Se plantea el uso de la IA para resolver el problema de registro en tiempo casi real, utilizando modelos biomecánicos subyacentes.

Resum

El càncer de pròstata (CaP) és el primer càncer per incidència en homes en països occidentals, i el tercer en mortalitat. Després de detectar en sang una elevació de l'Antigen Prostàtic Específic (PSA) o després de tacte rectal sospitós es realitza una Ressonància Magnètica (RM) de la pròstata, que els radiòlegs analitzen per a localitzar les regions sospitoses. A continuació, aquestes es biopsien, és a dir, es prenen mostres vives que posteriorment seran analitzades histopatològicament per a confirmar la presència de càncer i establir el seu grau d'agressivitat.

Durant la biòpsia s'utilitza típicament Ultrasons (US) per al guiat i la localització de les lesions. No obstant això, aquestes no són directament visibles en US, i l'uròleg necessita usar software de fusió que realitze un registre RM-US que transferisca les localitzacions marcades en MR a l'US. Això és fonamental per a assegurar que les mostres preses provenen veritablement de la zona sospitosa.

En aquest treball es compendien cinc publicacions que utilitzen diversos algorismes d'Intel·ligència Artificial (IA) per a analitzar les imatges de pròstata (RM i US) i amb això millorar l'eficiència i precisió en el diagnòstic, biòpsia i tractament del CaP:

1. **Segmentació automàtica de pròstata en RM i US:** Segmentar la pròstata consisteix a delimitar o marcar la pròstata en una imatge mèdica, separant-la de la resta d'òrgans o estructures. Automatitzar per complet aquesta tasca, que és prèvia a tota anàlisi posterior, permet estalviar un temps significatiu a radiòlegs i uròlegs, millorant també la precisió i repetibilitat.

2. **Millora de la resolució de segmentació:** Es presenta una metodologia per a millorar la resolució de les segmentacions automàtiques anteriors.

3. **Detecció i classificació automàtica de lesions en RM:** S'entrena un model basat en IA per a detectar les lesions com ho faria un radiòleg, assignant-les també una estimació del risc. S'aconsegueix millorar la precisió diagnòstica, donant lloc a un sistema totalment automàtic que podria implantar-se per a segona opinió clínica o com a criteri per a priorització.

4. **Simulació del comportament biomecànic en temps real:** Es proposa accelerar la simulació del comportament biomecànic d'òrgans blans mitjançant l'ús d'IA.

5. **Registre automàtic RM-US:** El registre permet localitzar en US les lesions marcades en RM. Una alta precisió en aquesta tasca és essencial per a la correcció de la biòpsia i/o del tractament focal del pacient (com braquiteràpia d'alta taxa). Es planteja l'ús de la IA per a resoldre el problema de registre en temps quasi real, utilitzant models biomecànics subjacents.

Objectives

In this work, artificial intelligence (AI) algorithms will be used to improve the quality of life of prostate cancer (PCa) suspect patients, and of the clinicians caring for them, by enhancing the efficiency and precision in diagnosis, biopsy, and treatment of this condition (such as high-rate brachytherapy). All this will be achieved trough two secondary objectives:

1. Development of a system for PCa lesion detection, segmentation, and classification in multi-parametric Magnetic Resonance (MR) Imaging (mpMRI). This will require the creation of a model for MR prostate segmentation in advance.

2. Development of a system for MR - Ultrasound (US) prostate registration. This will require MR and US prostate segmentation models that are able to produce high resolution segmentation masks.

All previous systems will need to be fully automatic (require no intervention), perform on par or better than expert radiologists, and be very quick to use (real-time or near-real-time) in order to enable their use in daily practice.

Chapter 1

Introduction

1.1 Medical perspective

In 2020, prostate cancer (PCa) was the first malignancy by incidence in the male population in Europe (Gandaglia et al., 2021), with a cumulative risk of 8.2% of being diagnosed before the age of 75, and a cumulative death risk of 1% (Dyba et al., 2021). Despite being the most common cancer in males, it is only the third by death count (10% from all cancer-related deaths in men), following colorectal cancer (12.3% of the male total) and lung cancer (24.2%) (Dyba et al., 2021). In fact, 59% of men over the age of 79 who died from unrelated causes were found to have incidental PCa upon necropsy (Bell et al., 2015). This disagreement between incidence and mortality is due to the heterogeneous aggressiveness of PCa lesions, generally being of slow evolution, as well as the positive outcomes of current treatments. In any case, PCa is a major socio-economical and healthcare burden, and any improvements to its diagnosis, handling or treatment will certainly result in a significant positive impact over the lives of millions of people.

Standard clinical pathway for PCa diagnosis typically consists in periodical measurements of Prostate Specific Antigen (PSA), a protein produced by prostate cells and measured in plasma, along with digital rectal examinations (DRE). PSA is usually produced in higher quantities by malignant prostate cells, hence an elevation in its concentration (e.g. above 4ng/mL) may be indicative of PCa. However, many other factors, such as benign prostate hyperplasia or an enlarged prostate may also raise PSA levels. Therefore, although highly sensitive, PSA remains a very unspecific test for PCa, with a positive predictive value of only 24% (i.e., only one out of four men with high PSA levels actually have PCa) (Hugosson et al., 2019).

Traditionally, high-PSA or DRE-positive patients undergo Ultrasound (US)-guided confirmation biopsy directly, wherein 20-30 tissue samples are collected from the patient's prostate with the use of needles, and their aggressiveness is then assessed by careful histopathological evaluation. Each sample is assigned a Gleason score (GS, from 3 to 5) (Epstein et al., 2005) depending on the appearance of the cells, e.g.: normal-looking cells are assigned a lower GS, are likely to grow slowly, and are not very aggressive, while very abnormal-looking cells receive a higher GS and can be extremely aggressive. Depending on the two most common GSs detected within a sample, these are further classified into a 1-5 grading system known as International Society of Urological Pathology (ISUP) grade (also known as Gleason

Grade Group, GGG) (Epstein et al., 2016). Figure 1.1 shows an example of several prostate histological samples along with their corresponding GS and GGG. Also, rightmost image in Figure 1.2 corresponds to a slice of a transrectal US (TRUS) for prostate biopsy guidance.

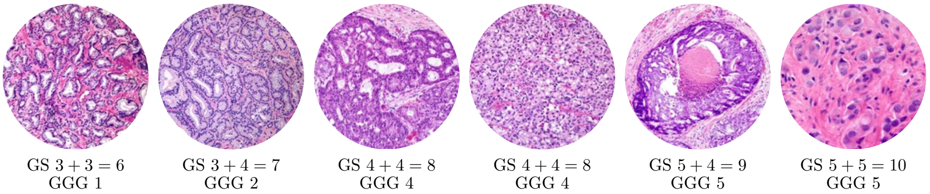


Figure 1.1: GSs and corresponding GGGs of several prostate tissue samples. GS: Gleason score, GGG: Gleason Grade Group. Images taken from Epstein et al. (2016)

In recent years, the introduction of Magnetic Resonance Imaging (MRI) and, in particular, pre-biopsy multi-parametric MRI (mpMRI) has drastically shifted this paradigm. MRI is a non-invasive non-ionizing medical imaging technique that employs very powerful magnetic fields (1.5 to 3T typically) to obtain a three-dimensional (3D) image of the internal structures of the body. Depending on the acquisition protocol (i.e., how, when, where, and for how long magnetic fields are activated), different MRI sequences can be obtained, highlighting distinct properties of the same underlying tissue; the combination of several of these sequences produces an mpMRI. Most common prostate mpMRI sequences can be seen in Figure 1.2.

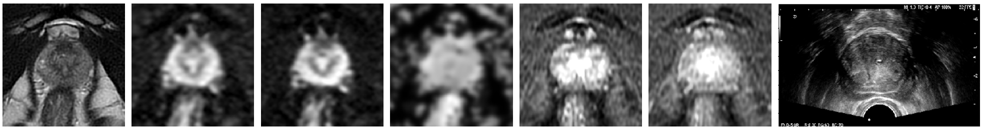


Figure 1.2: From left to right: most common prostate mpMRI sequences (T2, b500, b1000, ADC, DCE t = 10, DCE t = 30) and a prostate US image (not from the same patient). mpMRI: multi-parametric Magnetic Resonance Imaging, US: Ultrasound. For further explanations on the sequences refer to Section 4.5.1. Images taken from Pellicer-Valero et al. (2022, 2021)

In particular, a trained radiologist is able to identify PCa lesions by visual assessment of mpMRIs, enabling for better selection of patients for prostate biopsy (Mehralivand et al., 2018), significantly reducing the number of unneeded biopsies, increasing the diagnostic yield of the procedure (Ahmed et al., 2017), and allowing for more precise fusion-guided biopsy examinations and focal therapies as compared with cognitive fusion approaches (Marra et al., 2019), as will be later elaborated. A 2019 systematic review (Zhen et al., 2019) including 29 publications and 8503 patients found mpMRI to have a sensitivity and specificity of 0.87 [95% confidence interval -CI-, 0.81–0.91] and 0.68 [95% CI,0.56–0.79], respectively, and an area under the ROC curve (AUC-ROC) of 0.87 [95% CI,0.84–0.90], which helps explain its current widespread acceptance as a standard PCa diagnosis tool. This has been further pushed forward by global standardization efforts in the interpretation of mpMRI examinations, such as the Prostate Imaging Reporting and Data System (PI-RADS) which, in its latest 2.1 version, combines available evidence to assign scores to objective findings in each sequence (Turkbey and Choyke, 2019), or the Likert scale (Khoo et al., 2020).

Despite the positive aspects of mpMRI, it does come with its own set of problems. Firstly, mpMRI interpretation is time-consuming, expertise dependent (Gaziev et al., 2016), and is usually accompanied by a non-negligible inter-observer variability (Sonn et al., 2019). This is particularly relevant outside of expert high-volume centers (Kohestani et al., 2019). Secondly, mpMRI acquisitions are costly, and so is hiring the radiologists needed for analyzing an ever-increasing number of mpMRIs, as periodical PCa screenings are becoming widespread. Thirdly, although MRI technology is in itself non-ionizing and safe, contrast agents such as gadolinium, typically employed in dynamic contrast enhanced (DCE) MRI sequences, are increasingly controversial. Recent studies (Le Fur and Caravan, 2019) show that gadolinium is partially retained in the brain, bone, skin, and other tissues for months to years, although the health implications, if any, are not yet fully understood. As an alternative to standard mpMRI protocols containing DCE sequences, bi-parametric MRI (bpMRI) gets rid of contrast-enhanced sequences, relying solely on T2 and diffusion sequences for PCa diagnosis, with the added advantage of faster acquisition times. A recent meta-analysis (ke Niu et al., 2018) including 33 studies found mpMRI to have a pooled sensitivity / specificity of 0.85 [95% CI, 0.78–0.93] / 0.77 [95% CI, 0.58–0.95], and bpMRI having 0.80 [95% CI, 0.71–0.90] / 0.80 [95% CI, 0.64–0.96], with mpMRI yielding a significantly higher sensitivity than bpMRI and no statistical difference in terms of specificity. As of now, the relative advantages of either protocol remain largely under discussion.

mpMRI has not only marked a turning point in PCa diagnosis, but also in biopsy and focal PCa treatment interventions. While classical systematic biopsies required the collection and analysis of 20-30 samples, MR-guided biopsies can directly target the radiologist-marked lesions, hence needing much fewer samples and improving the detection of clinically significant PCa as compared to systematic biopsies alone (Marra et al., 2019). Commonly, MR-guided biopsies are performed using TRUS for guidance during the operations, since using intraoperative MRI would be prohibitively expensive for most medical institutions (Hambrock et al., 2010). In contrast, mpMR-guided TRUS biopsies are much more accessible, but require locating the exact lesion positions (marked by radiologists in the pre-acquired mpMRI) within the intraoperative US image, where unfortunately lesions lack contrast with respect to surrounding tissue and cannot therefore be visually identified (Kaplan et al., 2002). The problem of finding the full correspondences between mpMRI and TRUS prostate positions is known as registration, and it can either be performed mentally by an expert urologist during the biopsy procedure (which is known as cognitive fusion, and has shown some contradictory results, Puech et al. (2013)), or computationally, which offers increased accuracy and reproducibility, and is the focus of much active research. Similarly, accurate MR-TRUS registration techniques are needed for the increasingly popular focal therapies. As opposed to radical prostatectomy (i.e., removing the whole prostate), focal therapies (such as cryoablation or high dose-rate brachiterapy) target the lesions exclusively and leave the surrounding healthy tissue unaffected, hence boasting a much lower complication profile (Ahdoot et al., 2019).

In summary, a major socio-economical problem ensues: while PSA screening has been shown highly effective for PCa diagnosis, reducing PCa mortality over 20% (Schröder et al., 2009, 2012), it comes at a high risk of overdiagnosis, entailing an economic burden that health systems are unable to assume. mpMRI has significantly improved the situation, reducing unneeded biopsies and improving the yield of the procedure; however, mpMRI has its own set of problems, such as a high cost of acquisition, lack of sufficient radiologists for analyzing an ever-increasing number of mpMRIs, and the complexity, expert-dependence, and variability in its interpretation, which ultimately may result in missing clinically significant lesions and putting the patient at risk. Lastly, MR-guided US interventions have the potential to further reduce costs and further improve the yield by accurately sampling the suspect lesion directly, but they require solving the complex problem of MR-US image registration.

This work is comprised of five peer-reviewed publications attempting to tackle the aforementioned problems with the help of convolutional neural networks (CNNs, Section 1.2) an artificial intelligence (AI) algorithm specialized in image processing (medical and otherwise), as well as other useful techniques, such as the finite element method (FEM, Section 1.3) for simulating the biomechanical behavior of the prostate, and coherent point drift (CPD, Section 1.4), a point set registration method used for tackling the MR-US registration problem. These will be introduced in the following Sections 1.2-1.4, while Section 1.5 will overview the distribution of the rest of the document, including an outline of each of the publications (Chapters 2-6).

1.2 Artificial intelligence for medical image analysis

1.2.1 Historical perspective

AI is a very loosely used term, which encompasses many different fields with a shared purpose of developing systems able to manifest intelligent behaviors. Frequently, however, AI is used exclusively to refer to Machine Learning (ML), which is a sub-field of AI that studies algorithms able to learn from experience (Benet-Ferrus et al., 2022). While the field of ML started more than half a century ago, it was not until the early 2010s when the true revolution ensued, due to the confluence of several factors, namely: theoretical developments (new architectures, better wight initialization, autograd frameworks, etc.) allowed for training deeper and more powerful neural networks (NNs), hence the emergence of the field known as deep learning (DL); data, which is the fuel needed to train them, became ubiquitous thanks to the growth of the Internet and the digitalization; and Graphical Computing Units (GPUs) started to be employed for general computing (beyond computer gaming), accelerating computations required by NNs by several orders of magnitude.

At the turning point was AlexNet (Krizhevsky et al., 2012), a CNN that won the 2012 ImageNet (Russakovsky et al., 2015) image classification competition by a large margin, sparking an interest in DL that has been growing exponentially ever since. Pivotal papers of recent years include DeepMind's Alpha Go (Silver et al., 2016), a model trained by playing against itself that beat the world's champion in Go, a complex game involving very-long-term planning; OpenAI's GPT2 (Brown et al., 2020), a very large language model trained by predicting the next word in a huge corpus of web-scrapped data, that was able to generate text and answer questions in a way that was indistinguishable from humans; DeepMind's AlphaFold 2 (Jumper et al., 2020) which, trained on experimentally-obtained protein structures, was able to achieve unparalleled accuracy in folding unknown structures, and that was labeled by many as the biggest discovery in computation biology in decades (Callaway, 2020); and lastly, OpenAI's DALL-E 2 (OpenAI, 2022), a multi-modal (i.e. using both text and images) image generation model, that was able to create surprisingly realistic images given a description in natural language, revolutionizing the field of image synthesis.

In the context of medical image analysis, CNNs, such as AlexNet, have been the driving force behind most current developments. Esteva et al. (2017) trained a classification CNN on $\sim 130,000$ images of skin lesions, achieving a performance on par with experts; De Fauw et al. (2018) employed $\sim 15,000$ optical coherence tomography (OCT) images to train a set of two CNNs able to detect a wide range of retinal diseases, with a performance matching or exceeding that of experts; Balakrishnan et al. (2019) proposed the CNN-based VoxelMorph framework, a learning-based method for medical image registration, achieving registration speeds several orders of magnitudes faster than classical pair-wise optimization-based alternatives, and enabling the inclusion of auxiliary segmentation information to further improve accuracy; finally, Minaee et al. (2020) trained a CNN on 5,000 chest X-rays to detect the presence of COVID-19, achieving a sensitivity/specificity of 0.98/0.90. As can be seen, the single algorithm behind all these contributions is the CNN; due to its importance, Section 1.2.2 will introduce the main building blocks and ideas behind it. Also, these papers show that CNNs need a lot of data for learning but, once collected, the trained systems are able to perform on par (and sometimes above) the experts.

1.2.2 Technical overview of convolutional neural networks

Simply put, CNNs can be seen as stack of learnable convolutional filters along with other non-linearities, wherein filter parameters are learned by gradient descent. In this context, images must be understood as arrays of numbers; for instance, the two-dimensional gray-scale image I to the left of Figure 1.3 can also be seen as an array (of dimensions 5×5 in this case), where every element I_{ij} represents the intensity of the pixel at that location. Similarly, the convolutional kernel θ is just another array of numbers (dimensions 3×3 in this case); by applying convolution θ to I, the output activation map O (also known as feature map) is generated, i.e. $O = I * \theta$.

To the right of Figure 1.3, θ is being applied to I at position I_{31} to produce the output activation O_{31}. This operation is simply the dot product between image intensities I_{ij} "under the kernel" and the parameters of the kernel θ_{kl}, i.e.: $O_{31} = I_{20} \cdot \theta_{00} + I_{21} \cdot \theta_{01} + I_{22} \cdot \theta_{02} + I_{30} \cdot \theta_{10} + I_{31} \cdot \theta_{11} + I_{32} \cdot \theta_{12} + I_{40} \cdot \theta_{20} + I_{41} \cdot \theta_{21} + I_{42} \cdot \theta_{22}$.

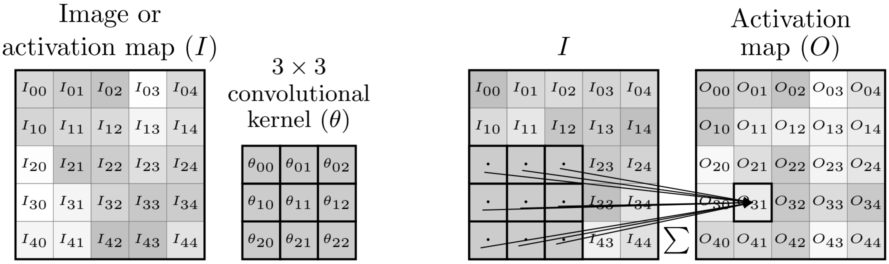


Image or activation map (I)
I_{00} I_{01} I_{02} I_{03} I_{04}
I_{10} I_{11} I_{12} I_{13} I_{14}
I_{20} I_{21} I_{22} I_{23} I_{24}
I_{30} I_{31} I_{32} I_{33} I_{34}
I_{40} I_{41} I_{42} I_{43} I_{44}
3×3 convolutional kernel (θ)
θ_{00} θ_{01} θ_{02}
θ_{10} θ_{11} θ_{12}
θ_{20} θ_{21} θ_{22}
I
$\sum$
Activation map (O)
O_{00} O_{01} O_{02} O_{03} O_{04}
O_{10} O_{11} O_{12} O_{13} O_{14}
O_{20} O_{21} O_{22} O_{23} O_{24}
O_{30} O_{31} O_{32} O_{33} O_{34}
O_{40} O_{41} O_{42} O_{43} O_{44}

Figure 1.3: Convolution operation on a simple 2D gray-scale image. 2D: Two-dimensional

By sliding the kernel over the whole image, all activations O_{ij} can be computed, and hence O can be obtained. As an example, if all the parameters of the kernel were $\theta_{kl} = 1/9$, every activation O_{ij} would just be the mean value of I around I_{ij}, and the resulting activation map would just be a blurred version of the input image. Depending on the parameters, convolutional filters can detect features such lines, dots, etc. Furthermore, convolutions can also be applied to activation maps, since both images I and activations O are just arrays of numbers, and by stacking convolutions more complex patterns can be detected.

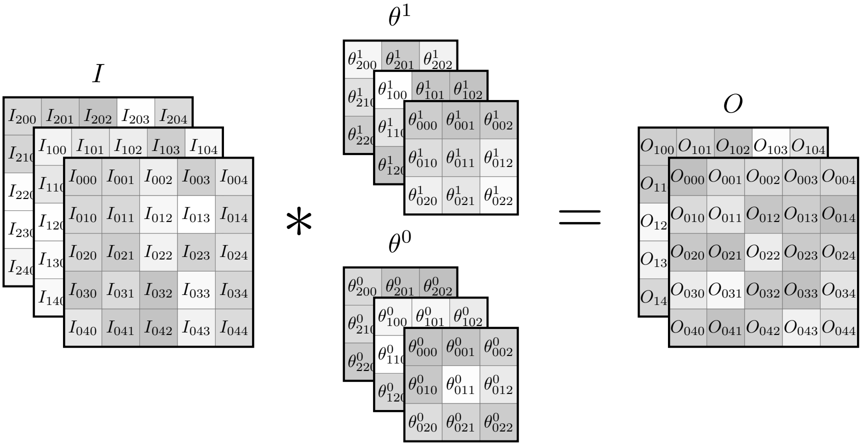

I
I_{000} I_{001} I_{002} I_{003} I_{004}
I_{010} I_{011} I_{012} I_{013} I_{014}
I_{020} I_{021} I_{022} I_{023} I_{024}
I_{030} I_{031} I_{032} I_{033} I_{034}
I_{040} I_{041} I_{042} I_{043} I_{044}
$*$
θ^1
θ^1_{000} θ^1_{001} θ^1_{002}
θ^1_{010} θ^1_{011} θ^1_{012}
θ^1_{020} θ^1_{021} θ^1_{022}
θ^0
θ^0_{000} θ^0_{001} θ^0_{002}
θ^0_{010} θ^0_{011} θ^0_{012}
θ^0_{020} θ^0_{021} θ^0_{022}
$=$
O
O_{000} O_{001} O_{002} O_{003} O_{004}
O_{010} O_{011} O_{012} O_{013} O_{014}
O_{020} O_{021} O_{022} O_{023} O_{024}
O_{030} O_{031} O_{032} O_{033} O_{034}
O_{040} O_{041} O_{042} O_{043} O_{044}

Figure 1.4: Convolution operation with more than one channel: I is a three-channel image being convolved with a $3 \times 3 \times 3 \times 2$ convolutional kernel θ to obtain the output activation map O, which has two channels

Now, natural images typically have several channels (red, green and blue); so have medical images such as mpMRI, in which each modality can be stored in a different channel. Each channel can be seen as an individual 2D gray-scale image, like I in Figure 1.4. Furthermore, in this case I is being convolved with two $3 \times 3 \times 3$ convolutional filters θ^0 and θ^1 which, applied to I, will produce an output activation map O with two channels. In general, activation maps can have any number of channels, each one reacting to a different feature (e.g. vertical lines, transitions from red to green, etc.) of the input image or activation map.

Another very common operation in CNNs is downsampling and upsampling the activation maps (i.e. reducing and increasing their resolution, respectively). A very common way of performing downsampling is by changing the stride of the convolution, i.e, the step with which it slides over the image. In the example in Figure 1.3 the stride was one, hence the output map had the same resolution as the input image (assuming necessary padding was added to the borders); if the stride were two (and assuming no padding this time), the convolutional kernel would skip all even positions, and the output map would have a resolution of only 2×2. More precisely, the center of the kernel would only visit positions O_{11}, O_{13}, O_{31} and O_{33}. Another way of performing downsampling is using the maxpooling operator, which is just like a convolution, but instead of applying a dot product, it just applies the maximum operator to the elements of I within the window of the kernel. For upsampling, transpose convolutions operate with a stride > 1 over the output, rather than the input, hence achieving this effect; still, simple linear or cubic interpolation is also commonly used.

Lastly, CNNs employ non-linearities to further enhance their ability to recognize complex patterns. The simplest non-linearity (or activation function in this context) is the ReLU (Rectified Linear Unit) activation function (Equation 1.1), which was proposed in the late 1960s and has both biological and mathematical motivations. Despite its simplicity, it is currently the most commonly used activation function in DL, along with its many variants, such as Leaky ReLU (which has a small slope for $x \leq 0$), PReLU (which makes the slope a learnable parameter), or GELU (which is smooth approximation to the ReLU).

$$\mathrm{ReLU}(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{otherwise} \end{cases} \tag{1.1}$$

By combining all the previous elements, different CNN architectures can be obtained. For instance, Figure 1.5 shows the AlexNet CNN architecture (Krizhevsky et al., 2012), which is just a stack of convolutions followed by ReLU activations, with some max-pooling operations in between, gradually reducing the input image resolution from 224×224 to 13×3 while increasing the number of channels from 3 to 256; then, the last feature map is then flattened and passed through a standard NN with a final softmax activation function predicting the probability for each of the 1000 ImageNet's classes (Russakovsky et al., 2015) when given an input image. Most CNNs share a similar structure: as the network grows deeper, the spatial resolution is reduced while the number of channels is increased, hence transforming spatially-dependent information (lines at a given position, a color, etc.), into progressively more semantically-rich information (a shape, a combination of colors, etc.), and eventually into highly-informative features (a face, a wheel, a flower, etc.) that can then be used for predicting the output.

For training a CNN (and all DL algorithms, in fact) the gradient descent (GD) algorithm is employed. After initializing the weights (or parameters θ) of the CNN to

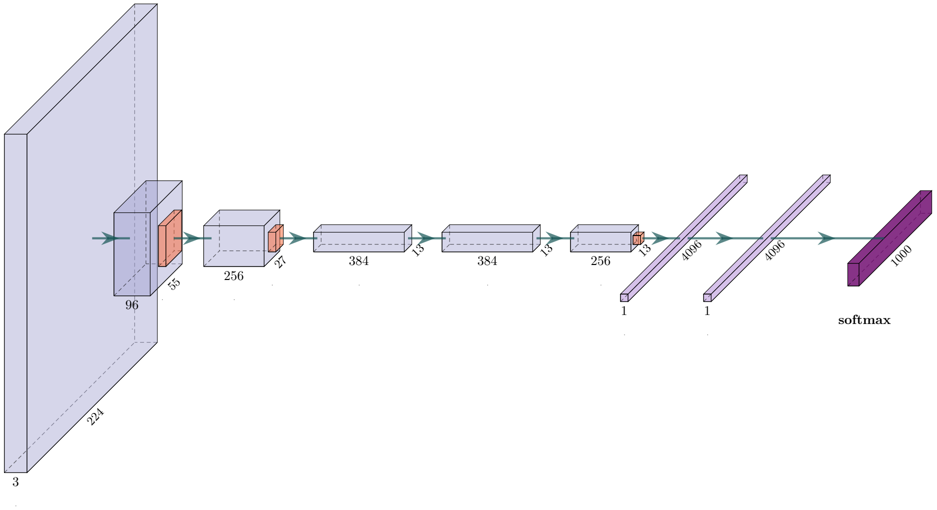


Figure 1.5: AlexNet architecture. ReLU activation function is applied after each convolution and / or transformation. LEGEND: Blue boxes: activation maps; orange boxes: max-pooling; purple boxes: flattened feature vectors

a small random value, GD works by iteratively pushing them a small amount μ in the direction opposite to the gradient of the loss function J (some measure of prediction error) with respect to them (Equation 1.2), hence enforcing a minimization of the error over time. Thanks to GD, CNNs manage to learn convolutional filters that are useful for a given problem, without being explicitly programmed to do so; this is in contrast to classical computer vision, which employed manually defined filter parameters, and performed much worse in complex perception tasks.

$$\text{GD}: \ \ \theta \leftarrow \theta - \mu \frac{\delta J(\theta, x, y)}{\delta \theta} \tag{1.2}$$

Several CNN architectures have been introduced over the years, many of which are still in active use. Simonyan and Zisserman (2015) proposed the VGG16 and VGG19 architectures, very similar to AlexNet, but increasing both depth (number of convolutional layers, which were increased to 16 and 19 respectively) and width (number of channels per layer, increased up to 512 channels), hence attaining a more powerful model. However, with greater depth, the flow of information through the network deteriorated; to solve this issue, He et al. (2016) incorporated residual connections to their ResNet architecture, connecting previous stages of the CNN to later ones by means of feature map addition, hence creating a low-resistance pathway for the information, and allowing for much deeper architectures, such as the 152-layer ResNet-152 model. This same idea was further extended by the DenseNet architecture (Huang et al., 2017). Recently Tan and Le (2019) employed neural architectural search (Zoph and Le, 2017), a reinforcement learning-based technique for finding an optimal base CNN architecture, together with optimal simultaneous scaling of depth/width/resolution, to find the family of architectures known as EfficientNet.

Beyond classification, for problems such as segmentation, both input and output are images. In the context of prostate imaging, segmenting consists in delineating or marking the prostate within a medical image, separating it from the rest of organs or structures. Similar to images, segmentation masks can be seen as a binary image (i.e., containing just 1s and 0s) with a one in all positions within the region of interest (e.g., the prostate) and zero elsewhere. For these sorts of tasks, the U-Net architecture (Ronneberger et al., 2015) (or one of its many variants) is predominantly employed. The input image is first processed through an encoder CNN, which is typically identical to a classification CNN, without the final classification layers. The output of the encoder is then connected to the input of the decoder CNN, which is an inverted version of the encoder where the downscaling operators have been exchanged by upscaling operators. Additionally, skip connections transfer information from the encoder to the decoder at several stages other than the output (Figure 1.6). This idea, in combination with residual connections (He et al., 2016) and a cost function based on Dice Similarity Coefficient (DSC), was quickly extended from 2D convolutions to 3D convolutions by the V-net (Milletari et al., 2016), in order to better deal with 3D medical images. DSC is a very commonly used metric both for training (as a loss) and evaluating segmentation CNNs, that measures the overlap between two surfaces (in 2D) or volumes (in 3D). It is defined in Equation 1.3 for two arbitrary binary masks m^1 and m^2; it can have a value between 0 and 1, where 0 means no overlap, and 1 denotes perfect overlap. Although many of the previously discussed architectural improvements (such as residual connections) have been ported to the U-Net architecture, the vanilla version remains still widely used.

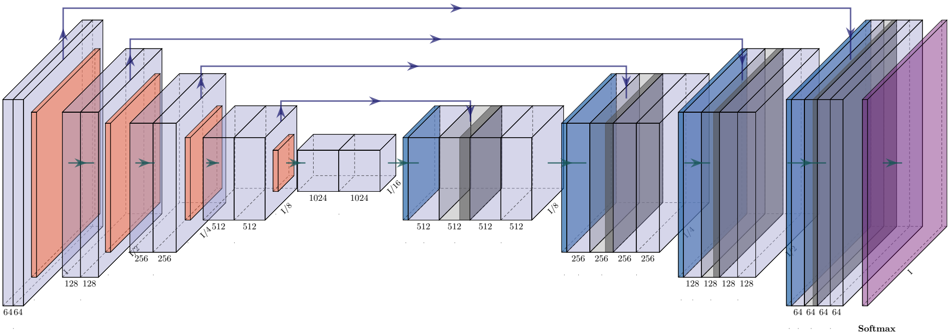


Figure 1.6: U-Net architecture. ReLU activation function is applied after each convolution and / or transformation. LEGEND: Blue boxes: activation maps; orange boxes: max-pooling; dark-blue boxes: upscaling operations

$$DSC(m^1, m^2) = \frac{2 \cdot \sum_i^N m_i^1 \cdot m_i^2 + \epsilon}{\sum_i^N m_i^1 + \sum_i^N m_i^2} \tag{1.3}$$

This work employs CNNs extensively: Chapter 2 proposes a specific U-Net-like architecture, combining both dense and residual connections, as well as many other techniques, to solve the problem of MRI and TRUS prostate segmentation; Chapter 3 develops a technique for improving the output resolution of any segmentation CNN; Chapter 4 makes use of the Retina-U-Net (Jaeger et al., 2020), an architecture combining a U-Net for PCa lesion segmentation with the Retina Net (Lin et al., 2017b) for lesion detection, classification and bounding box refinement; lastly, Chapter 6 uses the same ideas of the VoxelMorph (Balakrishnan et al., 2019) framework to train a CNN that, given a corresponding pair of prostate MRI and TRUS, it is able to directly perform near-real-time MR-US registration; with near-real time meaning that the system can be used without having to wait for the answer (e.g., less than half a second). In fact, although slow in training, DL models generally have the advantage of being extremely quick in inference (i.e. predicting on a new sample).

1.3 Simulation of biomechanical behavior through finite element method

FEM is a numerical method used for finding approximate solutions to engineering and mathematical physics problems that cannot be solved analytically due to the complexity of their constitutive equations, the geometry of the problem, and/or its boundary conditions. The method was originally created in the middle of the 20^{th} century for structural analysis and has experienced a continued growth from the 70s with the increased availability of both closed and open source FEM code, and its mathematical formalization. Nowadays, it is widely used, both in industry and academia, for simulating a variety of real-world problems in the fields of structural mechanics, fluid dynamics, electromagnetism, acoustics, and heat transfer, among others. From metal casting simulations (Lewis and Ravindran, 2000) to computer fluid dynamics simulations for heat assessment on NASA's Mars Rovers due to martian air resistance (Bhandari and Anderson, 2013), to simulating the electromagnetic forces within an electrical transformer and analyzing the mechanical stresses due to short-circuit force (Ahn et al., 2012), FEM is a standard tool in a variety of industries, helping with product validation and design, specially in instances where physical testing would be very costly or even unfeasible (such as simulating martian atmosphere and gravity, or simulating the mechanical behavior of living tissue).

In the context of this work, FEM is used to obtain the displacement field within a liver (Chapter 5), or a prostate (Chapter 6), given some boundary conditions, such as external forces, or a surface displacement field, respectively. Roughly, solving a biomechanical problem (as is the case here) with the FEM requires meshing the volumes of interest (i.e., discretizing them into a mesh of finite elements), providing some constitutive equations and parameters for their mechanical behavior, and setting up adequate boundary conditions. The FEM solver will then try to find the displacement field that minimizes the potential energy of the system, in virtue of the minimum total potential energy theorem.

Figure 1.7 shows the process for obtaining a prostate mesh automatically and simulating its mechanical behavior using FEM: first, a segmentation CNN is employed to obtain the segmentation mask, then the mask is meshed using TetGen (Si, 2010), and finally, mesh, material properties and boundary conditions are input to the mesh solver to obtain the displacement field within the gland. Notice that the mesh surface has been divided in many triangular elements, while the interior (not shown) uses volumetric tetrahedral elements. FEM makes good use of this discretization, by finding the displacements only in the vertices of the mesh (also called nodes), and then interpolating to the rest of the volume, hence reducing to degrees of freedom from an infinite number to as many as vertices. Obviously, the finer the mesh, the more accurate the solution will be, but also the slower it will be to compute.

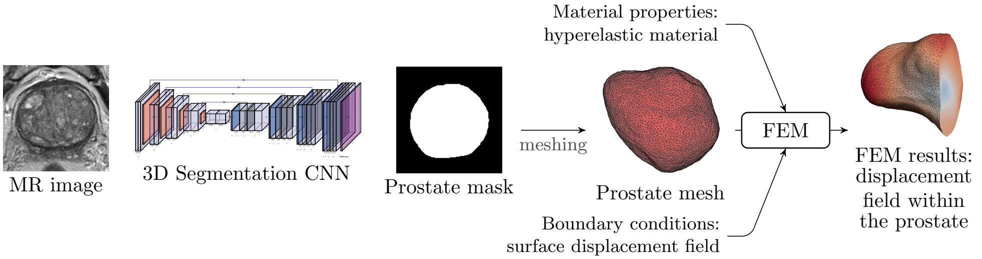


Figure 1.7: From left to right: a prostate MR is automatically segmented into a binary mask, the prostate mask is meshed, and its mechanical behavior is simulated through FEM, eventually obtaining the displacement field within the gland (with colors representing the magnitude of the displacements). Note that only a slice from the MR image and the binary mask is shown; MR: magnetic resonance, FEM: finite element method

For a mechanical problem (such as in Figure 1.7), Equation 1.4 states that the total potential energy Π_p equals the strain energy of the system W_s minus the work potential W_p. In particular, for linear elastic materials (i.e. with a linear stress-strain relationship), $W_s = \frac{1}{2}U^T KU$, where U is a matrix with the displacement of the nodes and K is the global stiffness matrix that has been built by assembling the stiffness matrices from each element; and $W_p = U^T F$, where F represents the nodal forces. Minimization of the total potential energy can be achieved by taking the derivative of Π_p with respect to the nodal displacement field U and equating it to zero, U simply being the solution to a linear system (Equation 1.5). Further explanation on FEM can be found in Section 6.4.4.

$$\Pi_p = W_s - W_p \tag{1.4}$$

$$\frac{\delta \Pi_p(U)}{\delta(U)} = 0 \rightarrow KU = F \tag{1.5}$$

In practice, soft tissue (such as that in the prostate or the liver) tends to present a non-linear mechanical behavior, and more complex formulations for W_s must be employed. Furthermore, constitutive equations for these materials are very hard to parameterize, since direct mechanical measurements within the body are not usually feasible, and properties change significantly when determined ex-vivo. In addition, proper boundary conditions (i.e., how an organ interacts with the surrounding tissue) are often even harder to obtain. Usually, strong assumptions and simplifications must be made, as is discussed in Sections 5.3.3.1 and 5.3.3.2 for the case of the liver.

In this work, FEM is employed in Chapter 5 for simulating the biomechanical behavior of the liver, and then using those simulations to train a DL model that achieves real-time inference speeds while maintaining high accuracy; both high speed and accuracy are required in the context of surgical simulators, computed-assisted surgery, and guided tumor irradiation, among other applications. In Chapter 6, FEM is used to obtain the displacement field experienced by the prostate during a biopsy (or any targeted procedure, such as brachitherapy) with respect to the resting prostate (from MRI), hence solving the MR-US registration problem; similarly as with the previous chapter, a CNN is eventually trained to imitate those simulations so as to perform MR-US registration in near-real-time, which opens the door to continuous registration and increased accuracy due to this adaptability, which current methods lack.

1.4 Point set registration with coherent point drift

Point set registration consists in finding the correspondence between two sets of points and/or recovering the transformation that maps the moving point set $X_{N \times d}$ to the fixed point set $Y_{M \times d}$, where N and M is the number of samples of X and Y, respectively, and d are the dimensionalities of the point sets (typically 2 or 3) (Figure 1.8).

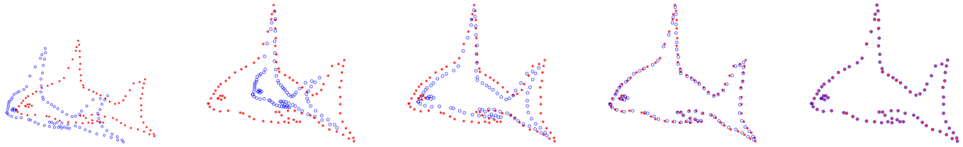


Figure 1.8: Non-rigid point set registration ($d = 2$) between moving set X (in blue) and fixed point set Y (in red) at iterations (from left to right): 0, 10, 20, 40, and 50. Image from Myronenko and Song (2009)

Point clouds can be obtained either from LiDARs (Light Detection and Ranging, wherein a laser is moved over a surface and distance is computed based on time of flight between emission and reception), range imaging techniques (stereo triangulation, structured light, etc.), or by means of post-processing, such as monocular depth estimation using DL, or feature extraction in images. For the purposes of this work, point clouds will be obtained from the vertices of liver (Chapter 5) and prostate (Chapter 6) surface meshes (see Section 1.3). Point set registration is widely used in the context of Simultaneous Localization and Mapping (SLAM), which consists in the concurrent construction of a model of the environment (the map), and the estimation of the state of the agent moving within it (Cadena et al., 2016); this problem appears in fields such as autonomous driving, 3D reconstruction, and object detection.

The sought after transformation can be either rigid (consisting only in rotation and/or translation), or non-rigid. For the simple rigid case, a more formal objective would be to find the transformation parameters $\theta = \{R, t\}$ (where R is a rotation matrix, and t is translation vector) that most closely map X to Y. This can be accomplished by optimizing cost function J with respect to θ (Equation 1.7), subject to R being a rotation matrix, where $\|\cdot\|_F$ denotes the Frobenius norm and $\hat{X}$ is X transformed (Equation 1.6). Note that this optimization problem can still yield solutions where X is mirrored, which happens when $\det(X) = -1$.

$$\hat{X} = R \cdot X + t \tag{1.6}$$

$$J = \|\hat{X}(\theta) - Y\|_F \quad \text{s.t. } R^T \cdot R = \mathbb{I} \tag{1.7}$$

If the number of points in both X and Y is the same (i.e., points are paired), this is known as the Procrustes problem, which has a closed form solution: the optimal translation vector t^* is just the vector joining the centroids of both point sets $t^* = \bar{Y} - \bar{X}$, while the optimal rotation matrix R^* can be obtained as $R^* = UV^T$, with U, V coming from the singular value decomposition of $(Y - \bar{Y})(X - \bar{X}) = U\Sigma V^T$, where both Y, X have been detrended by subtracting their respective centroids.

If the number of points in X and Y is different, the simplest point set registration method is the Iterative Closest Point (ICP) algorithm, developed by (Besl and McKay, 1992), which works in the following way: first, parameters are initialized (for instance: $R = \mathbb{I}, t = [0, ..., 0]^T$); second, each point in X is matched to its closest point in Y, thus obtaining a subset of Y which we call Y_s that is now paired with X; third, solve the Procrustes problem between X and Y_s. Finally, X is transformed according to the parameters found by solving Procrustes, and steps two and three are repeated until X does not change between iterations.

Unfortunately, basic ICP algorithm presents several important flaws, namely: X and Y must be sufficiently close together for ICP to converge to the optimal solution, it is not robust against noise or the presence of outliers, and it does not perform non-rigid registration. Many methods have been proposed to overcome these limitations, such as probabilistic methods, which assign point correspondences that are not binary, but rather probabilistic, i.e. there is no longer a “closest” point, but rather all points in Y are somewhat close to every point in X according to some probabilistic weighting.

CPD (Myronenko and Song, 2009) is a probabilistic point set registration method formulated as a probability density estimation problem, where the set of moving points Y make up the centroids of a gaussian mixture model (GMM, a probabilistic distribution defined as a sum of gaussians), and the set of fixed points X represent the observations of the GMM, which have some uniform noise (note that fixed and moving point sets now have switched names to be consistent with the notation in Myronenko and Song (2009)). In CPD, the objective is to maximize the likelihood of observations

X belonging to the GMM defined by the points in Y. For more information, refer to Section 6.4.3. By virtue of its probabilistic nature, CPD performs better than ICP, especially in presence of noise and outliers, while including both rigid and non-rigid variants of the algorithm. An example of CPD applied to non-rigid surface point set registration in the liver can be seen in Figure 5.8 and, for prostate, in Figure 6.4.

In the present work, CPD is employed in two occasions. In Chapter 5, it is employed to match a common reference mesh (whose vertices are the moving point set) to a set of liver meshes (whose vertices are the fixed point set); the first few principal components of the transformation that the common liver mesh must undergo can then be used to efficiently parameterize the shape of any liver mesh. In Chapter 6, CPD is used to perform registration between prostate meshes obtained from segmenting an MRI and a TRUS from the same patient.

1.5 Document overview

The rest of the document is organized as follows:

Chapter 2: *Robust Resolution-Enhanced Prostate Segmentation in Magnetic Resonance and Ultrasound Images through Convolutional Neural Networks*, published in Applied Sciences, 2021 (2021 Journal Impact Factor -JIF- 2.84, Q2, percentile 58.15 in Engineering, Multidisciplinary). A fast, robust, accurate and generalizable model for MR and TRUS prostate segmentation employing CNNs is proposed. It achieves a consistently strong performance and even outperforms the inter-expert variability in MR segmentation. Prostate segmentations are routinely done in MR and TRUS, as they are needed for analyzing the mpMRI as well as performing MR-US registration. More accurate segmentations may lead to better registration and prognosis, while almost instant results are of particular interest to urologists, who currently have to spend around ten minutes manually performing segmentation the middle of the biopsy or tumor ablation operation. See Figure 2.6 for some examples of automatic segmentation.

Chapter 3: *Cost-free Resolution Enhancement in Convolutional Neural Networks for Medical Image Segmentation*, published in ESANN proceedings, 2021 (2021 Computer Research and Education -CORE- Rank B). This publication proposes a simple yet effective method for improving the output resolution of any already trained segmentation CNN (such as those developed in Chapter 2), even beyond that of the original image. High-resolution prostate segmentations may lead to improved registration and/or more accurate simulations of its biomechanical behavior. See Figure 3.3 for an example of this technique.

Chapter 4: *Deep Learning for Fully Automatic Detection, Segmentation, and Gleason Grade Estimation of Prostate Cancer in Multiparametric Magnetic Resonance Images* published in Scientific Reports, 2022 (2021 JIF 5.00, Q2, percentile 74.66 in multidisciplinary sciences). This paper presents a CNN-based model for automatic mpMRI analysis, achieving an excellent lesion-level AUC-ROC/sensitivity/specificity of 0.95/1.00/0.80 for the GGG $\geq$ 2 PCa significance criterion, outperforming the expert radiologists. It uses the model in Chapter 2 for performing automatic prostate segmentation. The clinical applications of this model are countless: it could be used a second clinical opinion -a safety net to reduce the likelihood of missing PCa lesions-, for mpMRI analysis prioritization and/or biopsy prioritization, or even as a fully automatic referral suggestion tool in the context of future population-wide PCa screening programs. See Figure 4.2 for some examples of this model in action.

Chapter 5: *Real-time Biomechanical Modeling of the Liver using Machine Learning Models trained on Finite Element Method Simulations* published in Expert Systems with Applications, 2020 (2020 JIF 6.95, Q1, percentile 91.39 in Engineering, Electrical and Electronic). Living tissue and organs present a complex biomechanical behavior, whose simulation is nonetheless of great interest in the context of surgical planning, computed-assisted surgery, or mechanically-constrained registration. While FEM is typically employed to this end, it is generally too slow for real-time use. Here, it is proposed to use machine learning (ML) for accelerating the FEM simulations, retaining a high accuracy while improving simulation speed by several orders of magnitude. See Figure 5.21 for a proof-of-concept simulation of the mechanical behavior of the liver running in real time.

Chapter 6: *Deep Learning Contributions for Reducing the Complexity of Prostate Biomechanical Models* accepted for publication in Reduced Order Models for the Biomechanics of Living Organs, Elsevier, 2022 (Recognized Publisher, Book Citation Index, Thomson Reuters). This last publication tackles the complex problem of MR-US registration by using a CNN to learn the registration task, similar as it was done in Chapter 5. As a ground truth for training, CPD was used to first match MR-US prostate surfaces (automatically generated by the segmentation models from Chapters 2 and 3), followed by a FEM simulation to obtain mechanically plausible internal deformations. The trained CNN achieved an almost perfect approximation, hence significantly reducing registration error, while reaching near-real-time speeds.

Finally, Chapter 7, Main results and conclusion, analyzes whether the original objectives were eventually met, summarizes the contributions of each of the papers both from a patient and a technical point of view (Section 7.1), and concludes by discussing some general limitations of medical AIs, as well as further work (Section 7.2).

References from all chapters have been collected at the end of the document to avoid duplication. The code for this work has been made publicly available at https://github.com/OscarPellicer/Deep-Learning-in-Prostate-PhD.

Chapter 2

Robust Resolution-Enhanced Prostate Segmentation in Magnetic Resonance and Ultrasound Images through Convolutional Neural Networks

Reference: Oscar J. Pellicer-Valero, Victor Gonzalez-Perez, Juan Luis Casanova Ramón-Borja, Isabel Martín García, María Barrios Benito, Paula Pelechano Gómez, José Rubio-Briones, María José Rupérez, and José D. Martín-Guerrero. Robust Resolution-Enhanced Prostate Segmentation in Magnetic Resonance and Ultrasound Images through Convolutional Neural Networks. *Appl. Sci.*, 11(2):844, jan 2021. ISSN 2076-3417. doi: 10.3390/app11020844. URL https://www.mdpi.com/2076-3417/11/2/844

Oscar J. Pellicer-Valero*[1], Victor Gonzalez-Perez[2], Juan Luis Casanova Ramón-Borja[3], Isabel Martín García[2], María Barrios Benito[2], Paula Pelechano Gómez[2], José Rubio-Briones[3], María José Rupérez[4], José D. Martín-Guerrero[1]

[1] Intelligent Data Analysis Laboratory, Department of Electronic Engineering, ETSE (Engineering School), Universitat de València (UV), Av. Universitat, sn, 46100 Bujassot, València, Spain
jose.d.martin@uv.es

[2] Department of Radiodiagnosis, Fundación Instituto Valenciano de Oncología (FIVO), Beltrán Báguena, 8, 46009 València, Spain
vgonzalezper@hotmail.com, mismaga99@gmail.com, mar7esc@gmail.com, ppelechano@hotmail.com

[3] Department of Urology, Fundación Instituto Valenciano de Oncología (FIVO), Beltrán Báguena, 8, 46009 València, Spain
jcasanova@fivo.org, jrubio@fivo.org

[4] Centro de Investigación en Ingeniería Mecánica (CIIM), Universitat Politècnica de València (UPV), Camino de Vera, sn, 46022 València, Spain
mjrupere@upvnet.upv.es

2.1 Abstract

Prostate segmentations are required for an ever-increasing number of medical applications, such as image-based lesion detection, fusion-guided biopsy and focal therapies. However, obtaining accurate segmentations is laborious, requires expertise and, even then, the inter-observer variability remains high. In this paper, a robust, accurate and generalizable model for Magnetic Resonance (MR) and three-dimensional (3D) Ultrasound (US) prostate image segmentation is proposed. It uses a densenet-resnet-based Convolutional Neural Network (CNN) combined with techniques such as deep supervision, checkpoint ensembling and Neural Resolution Enhancement. The MR prostate segmentation model was trained with five challenging and heterogeneous MR prostate datasets (and two US datasets), with segmentations from many different experts with varying segmentation criteria. The model achieves a consistently strong performance in all datasets independently (mean Dice Similarity Coefficient -DSC-above 0.91 for all datasets except for one), outperforming the inter-expert variability significantly in MR (mean DSC of 0.9099 vs. 0.8794). When evaluated on the publicly available Promise12 challenge dataset, it attains a similar performance to the best entries. In summary, the model has the potential of having a significant impact on current prostate procedures, undercutting, and even eliminating, the need of manual segmentations through improvements in terms of robustness, generalizability and output resolution.

Keywords: MR prostate imaging; US prostate imaging; convolutional neural network; prostate segmentation; neural resolution enhancement

2.2 Introduction

In the field of medical imaging, segmentations are extremely useful for a plethora of tasks, including image-based diagnosis, lesion detection, image fusion, surgical planning or computer-aided surgery. For the prostate, in particular, fusion-guided biopsy and focal therapies are quickly gaining popularity due to the improved sensitivity and specificity for lesion detection (Marra et al., 2019), and the low complication profile (Ahdoot et al., 2019), respectively, although they are still not fully accepted in clinical guidelines.

Nevertheless, accurate prostate segmentations are still hard and laborious to obtain, since they have to be manually annotated by expert radiologists and, even then, the inter- and intra-observer variability may be significant due to factors such as the lack of clear boundaries between neighboring tissues or the huge size and texture variation of this gland among patients. In our experiments, 14 images were segmented by a second expert, and the inter-expert agreement was found to be of 0.8794 in terms of Sørensen-Dice Similarity Coefficient (DSC) and 1.5619 mm in terms of the Average Boundary Distance (ABD). Very similar results were obtained in Shahedi et al. (2019), with experts achieving a DSC and an ABD of 0.83 and 1.5 mm, respectively. Because of this, automatic segmentation algorithms for the prostate are increasingly sought-after.

Before the rise of Deep Learning (DL) around 2012 (Krizhevsky et al., 2012), many different techniques for automatic prostate segmentation coexisted. For instance, in Allen et al. (2006), MR images of the prostate were segmented by using voxel threshold-based classification followed by 3D statistical shape modeling. An alternative approach suggested by Freedman et al. (2005) attempted to match the probability distributions of the photometric variables inside the object of interest with an appearance model, and then evolved the shape of the object until both distributions matched best. Another technique that has been widely used in the literature for medical image segmentation is atlas matching, which consists in non-rigidly registering a set of labeled atlas images to the image of interest, and then somehow combining all resulting segmentations into a single one (Klein et al., 2008).

Despite the strengths of these methods, the true revolution in this field came with the advent of CNNs, which are a kind of DL algorithm formed by a stack of convolutional filters and non-linear activation functions, wherein the filter parameters are learned by stochastic gradient descent. New CNN architectures have been steadily raising state-of-the-art performance in computer vision tasks, such as image classification (Tan and Le, 2019), image segmentation (Ronneberger et al., 2015) or object detection (He et al., 2016). Similarly, this trend has carried over to medical imaging, and prostate segmentation in particular.

One of the first approaches to CNN-based segmentation consisted in sliding a classification CNN over a whole image to provide pixel-wise classifications, which then were combined into a single segmentation mask (Cireşan et al., 2012). Shortly after, fully convolutional neural networks for semantic segmentation were proposed (Long et al., 2014); they allowed for much faster training and inference, as the whole image was processed at once, and also made a better use of spatial information by utilizing activation maps from different layers. Later, the U-net architecture (Ronneberger et al., 2015) introduced the encoder-decoder design with skip connections that is still predominantly used. In a U-net, the image is first processed through an encoder CNN, which is similar to a classification CNN. The output of the encoder is then connected to the input of the decoder CNN, which is an inverted version of the encoder where the pooling operators have been exchanged by up-scaling convolutions. Additionally, skip connections transfer information from the encoder to the decoder at several stages other than the output. This idea, in combination with residual connections (He et al., 2016) and a cost function based on DSC, was quickly extended from two-dimensional (2D) convolutions to 3D convolutions by the V-net (Milletari et al., 2016), in order to better deal with 3D medical images. Similar to the transition from pixel-wise classification to fully convolutional CNNs, 3D-CNNs are better able to use the context of the whole image and provide faster speeds in comparison with a per-slice 2D segmentation.

In the field of prostate segmentation in MR imaging, many different CNN architectures building on top of the V-net or the U-net have been proposed. For instance, Zhu et al. (2017) proposed the addition of deep supervision, To et al. (2018) used a more recent densenet-resenet architecture (Huang et al., 2017; He et al., 2016) and Zhu et al. (2019b) introduced a boundary-aware cost function. Prostate segmentation in 3D Ultrasound (US) imaging, although much less prevalent, has experienced a similar development, with a recent paper employing the attention mechanism to exploit the information from several layers (Wang et al., 2019). Some of these architectural choices, and several others, will be further elaborated in Section 2.3.

Despite the high performances reported by many of the aforementioned papers, it could be argued that they all incur in a common pitfall: they are designed to perform well on one single prostate dataset. Therefore, it is unknown how robust the model would be when applied to any other dataset. This kind of robustness is paramount if the model is to be applied in a real-life scenario, where the images may come from many different scanners, and may be analyzed by many different experts. Furthermore, robustness is also desirable in the sense that the produced segmentations should be accepted by different experts, despite their possibly varying criteria for segmentation; in other words, the produced segmentations should ideally behave like an average prediction from several experts.

In this paper, a robust algorithm based on CNNs for MR and US prostate image segmentation is proposed. It leverages both common and not-so-common design choices, such as a hybrid densenet-resnet architecture (Section 2.3.3), deep supervision (Section 2.3.4), 3D data augmentation (Section 2.3.5), a cyclic learning rate (Section 2.3.7), checkpoint ensembling (Section 2.3.8) and a simple yet effective post-processing technique to increase the resolution of the segmentations known as Neural Resolution Enhancement (Section 2.3.9). This technique, besides improving the segmentation performance, allows the CNN to produce segmentations with resolutions beyond that of the original image. Furthermore, the model is trained on five different datasets simultaneously (Section 2.3.1), achieving an excellent performance on all of them. Finally, the weights obtained from this model are used to train (through transfer learning) an US segmentation model on two different datasets, achieving also an excellent performance on them both. Results are presented both quantitatively (Section 2.4.1), by presenting the metrics (Section 2.3.10) achieved on every dataset, and qualitatively (Section 2.4.2), by showing images of the predicted and Ground Truth (GT) segmentations on several patients. The paper is closed by a discussion, (Section2.5), about the clinical impact of the proposed model, and a brief conclusion (Section 2.6).

2.3 Materials and Methods

2.3.1 Description of the Datasets

One of the main strengths of this study is the use of five different prostate T2-weighted MR datasets. As shown in Table 2.1, there is a significant variability in scanner manufacturers, resolutions and magnetic field strengths, among other factors. Datasets “Girona” (Lemaître et al., 2015), “Promise12” (Litjens et al., 2014b) and “Prostate-3T” (Litjens et al., 2015) are all freely available for download on the Internet, while

Table 2.1: Details of the MR datasets.

Dataset	*N*	*Scanner Manufacturer (%)*	*Endorectal Coil*	*Pixel Spacing (mm)*	*Slice Spacing (mm)*	*Field Strength (T)*
Girona	34	G. Elec. (59%)	No	0.27–0.55	1.00	1.5
		Siemens (41%)	No	0.68–0.79	1.00	3.0
Promise12	48	Siemens (25%)	Yes	0.63	3.6	1.5
		G. Elec. (25%)	Yes	0.25	2.20–3.00	3.0
		Siemens (25%)	No	0.33–0.63	3.00–3.60	1.5 & 3.0
		Siemens (25%)	No	0.50–0.75	3.60–4.00	3.0
Promise12_test	30	Unknown	Yes & No	0.27–0.63	2.2–3.6	1.5 & 3.0
Prostate-3T	12	Siemens	No	0.60–0.62	3.60–4.00	3.0
IVO	280	G. Elec. (96%)	No	0.35–0.74	0.60–7.00	1.5
		Philips (3%)	No	0.28–0.49	3.00	1.5 & 3.0
		Siemens (1%)	No	0.62–0.69	3.00–3.50	1.5
Private	90	Philips (81%)	No	0.30–0.62	2.91–5.00	1.5 & 3.0
		Siemens (11%)	No	0.52–0.69	3.00–3.60	1.5 & 3.0
		G. Elec. (8%)	No	0.37–0.43	3.40–6.00	1.5 & 3.0

“IVO” comes from the Valencian Institute of Oncology, and “Private” comes from a private institution which has decided to remain anonymous. Furthermore, Promise12 is an ongoing prostate segmentation challenge, wherein 50 MR prostate images are provided along with their segmentation masks (dataset “Promise12”), and 30 additional images are provided without segmentations as a test set (dataset “Promise12_test”). The participants must submit their predictions to the challenge server, where they are evaluated. Hence, “Promise12_test” will only be used for testing.

In addition to that, the prostate segmentations follow varying criteria depending on the expert who segmented them. In “IVO” dataset, three different radiologists with two, five and seven years of experience in prostate cancer imaging took turns to perform the segmentations. In “Private” dataset, a single medical physicist with two years of experience in MR prostate imaging segmented all the images. In “Promise12”, each of the four rows in Table 2.1 corresponds to a different medical center and, by extension, were also segmented by at least one different expert each, although an expert from the Promise12 challenge (Litjens et al., 2014b) corrected some of them. For the other datasets, no further information about the segmentations is known.

Regarding exclusion criteria, before separating the images into any subsets, all segmentations were examined and those with obvious errors were directly excluded. Therefore, no corrections were made, so as to better preserve the particular criteria from each expert (except for the “Private” dataset, in which all segmentations were revised). The number of samples (N) in Table 2.1 is computed after this filtering. As a special mention, 18 images from “Prostate-3T”, which were also present in “Promise12” or “Promise12_test” (although with different GT segmentations), were also discarded; and other 30 images from “Prostate-3T” (half of the original dataset), which systematically left many slices in the base and apex unsegmented, had to be discarded as well. Figure 2.1 shows the center slice of a sample from each of the datasets.

For the 3D-US segmentation model, two different datasets were employed: “IVO” and “Private”, both coming from the same institutions as their homonymous MR datasets. For “IVO” ($N = 160$ images), five different urologists with six to thirty years of experience segmented the images, while for “Private” ($N = 82$ images), it was two urologists with more than ten years of experience; no exclusion criteria were applied. Images from both datasets were captured using Hitachi scanners at spacings of 0.20 mm to 0.41 mm in any axis. Figure 2.1 shows the center slice of a sample from each axis. Unfortunately, no further segmented datasets were found on the Internet for this image modality.

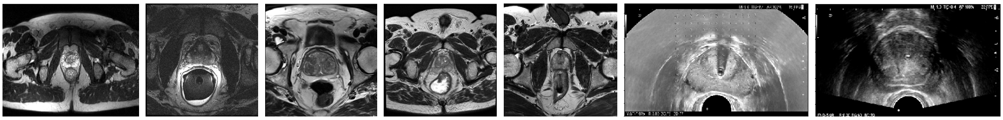


Figure 2.1: Center slice of a sample from MR and US datasets (from left to right): “Girona”, “Promise12”, “Prostate-3T”, “IVO (MR)”, “Private (MR)”, “IVO (US)” and “Private (US)”.

2.3.2 Image Pre-Processing

Before using the images to train the CNN, they all had to be pre-processed to alleviate their heterogeneity. First, their intensity was normalized by applying Equation (2.1) to every image I, such that 98% of the voxels in I_{new} fall within the range $[0, 1]$.

$$I_{new} = \frac{I - percentile(I, 1)}{percentile(I, 99) - percentile(I, 1)} \tag{2.1}$$

Then, the center crop of each image (and its respective segmentation mask) was taken, using a size of $112 \times 112 \times 32$ and a spacing of (1, 1, 3) mm for the MR images, and a size of $160 \times 112 \times 80$ and a spacing of (0.75, 0.75, 0.75) mm for the US images. B-Spline interpolation of third order was employed for all image interpolation tasks, while Gaussian label interpolation was used for the masks.

2.3.3 Hybrid Densenet-Resnet Architecture

The proposed CNN architecture (Figure 2.2) is based on the V-Net and, more precisely, on the architecture proposed by To et al. (2018), which combines a densenet (Huang et al., 2017) encoder with a resnet (He et al., 2016) decoder. All design decisions were guided by validation results.

The full architecture is sufficiently described in Figure 2.2. Therefore, only a few interesting design choices will be discussed here. Firstly, the proposed Dense block

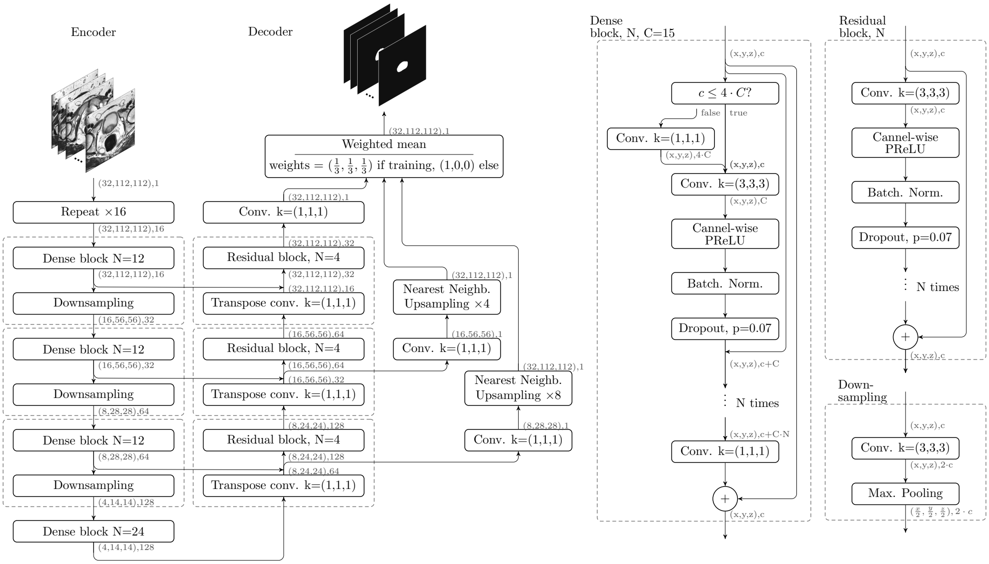


Figure 2.2: Architecture of the CNN. The encoder is composed of four Dense Blocks connected by Downsampling blocks. The decoder uses three Residual Blocks connected by transpose convolutions. Several skip connections transfer information from the encoder to the decoder. Furthermore, to the right of the decoder, intermediate outputs are used to perform Deep Supervision.

includes a residual connection, which empirically helped the CNN converge faster. Secondly, every Dense block contains between 12 and 24 “standard” convolutions (kernel of size (3, 3, 3)), as well as several “bottleneck” convolutions (kernel of size (1, 1, 1)), for a total of 72 “standard” convolutions, which is a huge number compared to similar architectures such as V-net (with only 12 convolutions in its resnet-based encoder) or BOWDA-net (Zhu et al., 2019a) (with 28 convolutions in its densenet-based encoder). This makes the encoder better capable of learning more complex representations of the input data. Comparatively, the decoder can have a simpler resnet architecture, since the heavy lifting (which is feature extraction) has already been done by the encoder. Thirdly, channel-wise PReLU was employed as activation function (He et al., 2015), as it provides a slightly better performance at a negligible additional computational cost. A channel-wise PReLU function is similar to a ReLU (Rectified Linear Unit) (Nair and Hinton, 2010) function, but with a learnable slope α for the negative inputs (instead of being just zero); α is shared among all activations in a channel, but is different for every channel. Fourthly, transpose convolutions were used in the decoder, since they were found to provide a better performance when compared to upsampling followed by a convolution.

Due to the huge memory requirements intrinsic to the densenet architecture, very small batch sizes had to be employed (4 for the MR dataset, and 2 for the US dataset), as well as a technique known as Gradient Checkpointing (Chen et al., 2016), which allows to reduce the Graphics Processing Unit (GPU) memory requirements at the cost of increased computation times. It works by keeping a fraction of the CNN activations in memory at any given time (instead of all of them), and recomputing the rest when they are needed.

2.3.4 Deep Supervision

To further improve the performance of the CNN, a simple implementation of Deep Supervision (Lee et al., 2014) is used. Unlike regular CNNs, which predict the segmentation mask from the last layer only, deeply supervised CNNs attempt to predict it from several intermediate layers as well. In Figure 2.2 this is implemented by the branches to the right of the decoder, which take the activation maps at two points along the decoder, reduce the number of channels to one by means of a "bottleneck" convolution, and then upsample them to the CNN output resolution using Nearest Neighbors interpolation. During training, the final output of the CNN is averaged with these intermediate predictions while, during inference, only the final output is considered. A similar implementation for this technique is also successfully used by Zhu et al. (2017). Figure 2.3 shows the GT mask of a prostate MR image, as well as the final and intermediate predictions, which are used for Deep Supervision. As it can be seen, intermediate predictions resemble a downscaled version of the final mask.

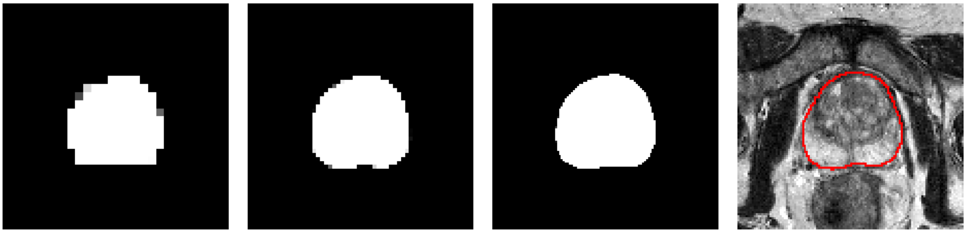


Figure 2.3: From left to right: first intermediate prediction, second intermediate prediction, third (and final) prediction, and original MR prostate image with GT label.

As demonstrated in Lee et al. (2014), Deep Supervision serves a twofold purpose: on one hand, it forces all the layers throughout the network to learn features which are directly useful for the task of image segmentation; on the other hand, the gradients are better able to flow towards the deeper layers, which accelerates training, and helps prevent problems related to gradient vanishing.

2.3.5 Online Data Augmentation

Online data augmentation was used to artificially increase the amount and variability of the training images, thus improving the generalization capabilities of the model and, ultimately, its performance. Before feeding an image to the CNN during training, the following transformations were sequentially applied to it:

1. 3D rotation along a random axis with random magnitude in the range $[0, \pi/20]$ radians.

2. 3D shift of random magnitude in the range $[0, 15]$ mm along every axis.

3. 3D homogeneous scaling of random magnitude in the range $[1/1.15, 1.15]$ times.

4. Flipping along x-axis with probability 1/2.

5. Adding Normally distributed noise with a random magnitude in the range $[0, 0.05]$ relative to the normalized image.

When required, a random number would be sampled from the uniform distribution and then scaled and shifted to the appropriate range.

2.3.6 Model Training

For both the MR and US segmentation models, images were split into three subsets: training (70% of the images), validation (15%) and test (15%). These proportions were computed dataset-wise, such that the relative representation of each datset on every subset was the same. The MR training set was used to update the weights of the CNN through stochastic gradient descent (Adam optimizer with default parameters), while the MR validation set was used to choose the best set of hyper-parameters (such as learning rate schedule, CNN depth, CNN width, input resolution or even internal CNN architecture). Once the MR segmentation model was considered final, the CNN was retrained one last time using both MR training and validation subsets, and the results were evaluated in the MR test subset.

For the US segmentation model no hyper-parameters were changed, except the input size and spacing (to better fit the prostate in the image, as discussed in Section 2.3.2) and the batch size (due to GPU memory limitations, as discussed in Section 2.3.3). Furthermore, transfer learning was employed (Pan and Yang, 2010): the weights from the MR segmentation model were used as an initialization to the US segmentation model, thus leveraging the feature extraction capabilities of the pre-trained model. The US model was directly trained using both US training and validation subsets (as no validation subset was actually required), and the results were evaluated in the US test subset.

2.3.7 Cyclic Learning Rate

A cyclic learning rate (Smith, 2017) was chosen for training, as it presents several advantages with respect to a fixed schedule. Firstly, an optimal fixed schedule must be learned from the data, which is cumbersome and requires extensive trial and error; secondly, this cyclic schedule will allow us to use a technique known as Checkpoint Ensembling, which will be explained in Section 2.3.8. Thirdly, a cyclic learning rate supposedly helps the optimizer escape saddle point plateaus, which is desirable. The chosen cyclic schedule is a decaying triangular wave (schedule known as "triangular2" in Smith (2017)) of period 48 epochs (at 180 batches per epoch), with a minimum learning rate of $5.5 \cdot 10^{-5}$, a maximum of $7.5 \cdot 10^{-4}$, and a decay such that the maximum value of the wave is halved every period. The CNN was trained for six periods (a total of 288 epochs).

2.3.8 Checkpoint Ensembling

Checkpoint Ensembling (Chen et al., 2017) is a strategy that allows to capture the effects of traditional ensembling methods within a single training process. It works by collecting checkpoints of the best k weights (those that lead to the best validation scores of the CNN during its training process). Then, during inference, for each input to the CNN, k predictions are obtained and combined into a single one (by averaging, for instance). In theory, this method makes a compound prediction from weights which may have settled into different local minima, thus simulating the compound segmentation proposal from several experts.

As for our model, using a cyclic learning rate opens up the possibility of using weight checkpoints that coincide with the minima of the learning rate schedule, as it is at these points where the gradient stabilizes most, and local minima are supposedly reached. Therefore, in our particular case, six checkpoints will be used for Checkpoint Ensembling. As a bonus, this technique incurs in no additional costs, other than inference costs, which are obviously increased by a factor of six. Traditional ensembling was also tested, although finally discarded, as it did not provide any performance improvements and incurred in much higher training costs.

2.3.9 Neural Resolution Enhancement

The last technique that will be discussed is Neural Resolution Enhancement (Pellicer-Valero et al., 2020a), which leverages the properties of any already trained image segmentation CNN to intelligently increase the resolution of the output mask at no cost, even beyond the resolution of the original image.

To understand how this method operates, let us describe how a threefold increase in the resolution along the z-axis would be performed (refer to Figure 2.4). First, the resolution of the input image X is triplicated along the z-axis (by using bicubic interpolation, for instance), therefore becoming $\hat{X}$. Then, three new images (X_0, X_1, X_2) are built by taking z-slices from $\hat{X}$ in such a way that they have the same size (dimensions) and spacing (voxel size) as X, but are offset by different sub-voxel amounts along z-axis (in fact, note that $X_0 \equiv X$). Then, X_0, X_1 and X_2 are fed through the CNN, and three segmentation masks are obtained (Y_0, Y_1, Y_2). Finally, all three predictions are combined by stacking them in the correct order, hence obtaining $\hat{Y}$, which is a predicted mask with three times the resolution of X along the z-axis. This same procedure could be applied to any number of dimensions simultaneously, although the inference cost would scale abruptly, as all the possible sub-voxel displacement combinations would have to be computed.

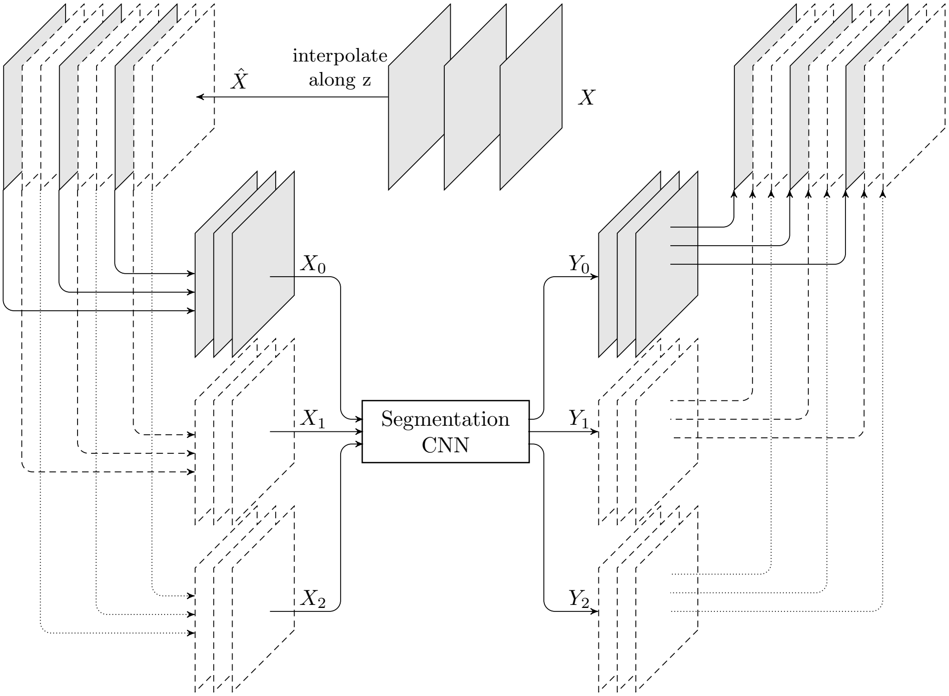


Figure 2.4: Visual representation of the Neural Resolution Enhancement method to triplicate the resolution along the z-axis.

This method, albeit simple, is extremely powerful, as it allows to predict (rather than to interpolate) segmentation masks beyond the resolution of the original image. The problem of interpolation is therefore shifted from the mask domain to the image domain, where the conveyed information is still complete and not yet binarized. Furthermore, it can be applied to any already trained segmentation CNN, as a simple post-processing step. Figure 2.5 shows an example application.

In the context of our problem, z-axis resolution is triplicated to reduce the impact of the final mask interpolation. This is: once the CNN outputs a segmentation mask, it must be transformed back to the space of the original input image (same resolution, spacing, physical orientation, position, etc.). Although this is necessarily a lossy process, by leveraging this technique, the predicted mask can have a higher resolution, which significantly mitigates the issue.

2.3.10 Evaluation Metrics and Loss

As it is customary in semantic segmentation problems, DSC was employed as the main evaluation metric, which guided most design decisions. DSC is defined in Equation (2.2), where N denotes the total number of voxels in an image, $\hat{y}_i \in [0, 1]$ represents the prediction of the CNN at voxel i, $y_i \in \{0, 1\}$ is the GT label at voxel i, and $\epsilon = 1$ is a

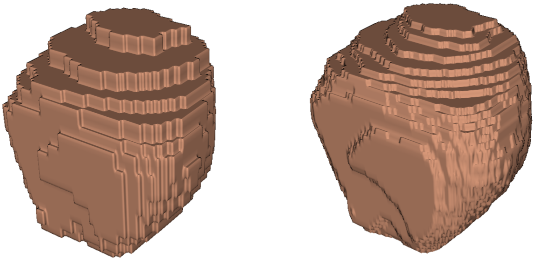


Figure 2.5: Mask predicted by a prostate segmentation CNN and upscaled along the z-axis three times using: nearest-neighbor interpolation (left), and Neural Resolution Enhancement (right).

small arbitrary value that prevents division by zero.

$$DSC(y,\hat{y}) = \frac{2 \cdot \sum_i^N \hat{y}_i \cdot y_i + \epsilon}{\sum_i^N \hat{y}_i + \sum_i^N y_i + \epsilon} \tag{2.2}$$

As a loss function, DSC is much better able to deal with unbalanced segmentation masks in comparison with binary cross-entropy. However, several studies acknowledge its deficiencies along the boundaries of the mask (Zhu et al., 2019a), or when the target is very small (Abraham and Khan, 2019) (as in lesion segmentation). Zhu et al. (2019a), for instance, utilizes a composite loss which penalizes wrong segmentations proportionally to the distance to the boundary of the GT. Despite multiple attempts at incorporating a similar loss to our model, we finally decided against it, since it did not provide any performance advantages during validation. Therefore, the finally used loss function $\mathcal{L}$ is directly derived from DSC, as illustrated in Equation (2.3).

$$\mathcal{L} = 1 - DSC \tag{2.3}$$

In addition to DSC, two distance-based metrics were also employed: Average Boundary Distance (ABD) and 95th percentile Hausdorff Distance (HD95). These metrics were computed as described in the Promise12 challenge (Litjens et al., 2014b) and represent the average and the 95th percentile largest distance (in mm) between the surface of the predicted mask and the GT mask, respectively.

When comparing these metrics among groups, the Wilcoxon signed-rank test was employed, which is the non-parametric equivalent of the paired t-test. The Wilcoxon test was needed due to the distribution of the metrics in the test set not being normal (p-value ≤ 0.001 using D'Agostino and Pearson's normality test for DSC, ABD and HD95 results).

Table 2.2: Quantitative results for all datasets and models.

	Dataset	*N*	*DSC*			*HD95 (mm)*			*ABD (mm)*		
			Mean	*Median*	*Min.*	*Mean*	*Median*	*Max.*	*Mean*	*Median*	*Max.*
MR	Girona	5	0.8980	0.9113	0.8467	3.7240	3.7873	4.2054	1.3305	1.3187	2.0323
	Promise12	7	0.9148	0.9118	0.8919	4.2876	3.6000	7.2000	1.0135	0.9680	1.2757
	Prostate-3T	2	0.9222	0.9222	0.9099	3.6000	3.6000	3.6000	0.9190	0.9190	1.0719
	IVO	42	0.9136	0.9182	0.8094	3.9947	4.0002	6.9347	0.9569	0.9013	2.0356
	Private	13	0.9251	0.9228	0.8993	3.3363	3.1736	4.5122	0.9190	0.8525	1.2815
	All	69	0.9150	0.9179	0.8094	3.8693	3.9995	7.2000	0.9815	0.9311	2.0356
US	IVO	24	0.9215	0.9256	0.8456	3.4295	3.1210	9.9997	1.1825	1.0539	2.8573
	Private	12	0.9131	0.9133	0.8960	3.6317	3.7025	6.1216	1.1872	1.2008	1.7809
	All	36	0.9187	0.9235	0.8456	3.4969	3.2863	9.9997	1.1840	1.1102	2.8573

2.4 Results

2.4.1 Quantitative Results

The quantitative test results (in terms of DSC, HD95 and ABD metrics) for both MR and US segmentation models (globally, and by dataset) are shown in Table 2.2. Both models achieve a mean and median DSC above the 0.91 threshold for all datasets, meaning that they are very strong performers and, more interestingly, that they are robust to the heterogeneity of the various datasets. As an exception, the mean DSC on the "Girona" dataset falls to around 0.90 due to a single relatively weak prediction (DSC of 0.8467) dragging down the mean of this extremely small set, as evidenced by the otherwise inexplicably high median value. Also, the mean DSC for the MR segmentation model on the "Private" dataset is exceptionally high, probably due to it being the only dataset where GT segmentations were revised.

These observations are further supported by the HD95 metric. For all MR and US datasets, it sits mostly just below 4 mm in average. Since the slices in a typical prostate MR image are about 3 mm apart, achieving an HD95 below 3 mm is extremely unlikely due to the different criteria regarding how far the base and the apex should extend. Thus, an average HD95 below 4 mm is a very good result. Finally, the ABD metric lies mainly below 1 mm in average for all MR datasets, and below 1.2 mm for the US datasets.

For comparison purposes, a second segmentation (GT2) was created for the first three datasets by one of the IVO experts. Table 2.3 shows the mean DSC of the predictions of the model against each of the GTs (GT and GT2), as well as the mean DSC of the GTs against themselves (the inter-expert agreement). As it can be seen, the DSC of the model against both GT and GT2 surpasses by a large margin the inter-expert agreement (except for one case), suggesting that the model is more robust and reliable than any given expert by itself. Two Wilcoxon tests confirm that these

Table 2.3: Evaluation of the predictions against GT and GT2, as well as GT against GT2 (inter-expert performance).

		Mean DSC		
Dataset (MR)	*N*	*Predicted & GT*	*Predicted & GT2*	*GT & GT2*
Girona	5	0.8980	0.9057	0.8657
Promise12	7	0.9148	0.9032	0.8825
Prostate-3T	2	0.9222	0.8995	0.9026
All above	14	0.9099	0.9035	0.8794
		p-*value against last column*		
		0.0035	0.0258	-

differences in DSC are statistically significant (at a significance threshold of 0.05).

Since most authors focus on performing well in one single dataset, it is difficult to compare these results against other published models. As an exception To et al. (2018) used a private dataset (with diffusion-weighted MR images and ADC-maps in addition to the T2-weighted MR images) in conjunction with "Promise12", achieving an impressive mean DSC of 0.9511 on their own dataset, but only a mean DSC of 0.8901 on "Promise12_test".

Regarding 3D-US image segmentation, few publications were found (most use 2D-US), and none employed more than one dataset. As for recent 3D-US papers, Wang et al. (2019) achieved a mean DSC of 0.90 by leveraging an attention mechanism. Lei et al. (2019) obtained a DSC 0.919 by using a contour-refinement post-processing step, however, the results are not reported on a proper test set, but rather, using leave-one-out cross-validation. More recently, Orlando et al. (2020) achieved an excellent 0.941 mean DSC by applying a 2D U-Net on radially sampled slices of the 3D-US and then reconstructing the full 3D volume. As an example on the problem of 2D-US segmentation, Karimi et al. (2019) achieved a mean DSC of 93.9 by using an ensemble of five CNNs. This last result, however, is not directly comparable, as in 2D-US segmentation the DSC is evaluated on a per-slice basis, instead of the prostate as a whole.

Promise12 is an ongoing prostate segmentation challenge, wherein 50 MR prostate images are provided along with their segmentation masks (dataset "Promise12"), and 30 additional images are provided without segmentations as a test set (dataset "Promise12_test"). Table 2.4 shows the performance of the model on "Promise12_test" along with the five best entries to the Promise12 challenge. For this specific dataset, the predicted segmentation masks are uploaded and evaluated in the servers of the challenge, and the results are publicly posted online thereafter (Litjens et al., 2020).

As it should be expected, the mean and median for our model are similar to the

Table 2.4: MR model performance on "Promise12_test" along with the five best entries as of December 2020.

Challenge Score	*Name*	*DSC*			*HD95 (mm)*			*ABD (mm)*		
		Mean	*Median*	*Min.*	*Mean*	*Median*	*Max.*	*Mean*	*Median*	*Max.*
91.9072	MSD-Net	0.9299	0.9323	0.8890	3.5512	3.3454	7.3344	1.1160	1.0968	1.6777
90.7993	Edge Att.	0.9118	0.9136	0.8672	4.3095	3.9362	7.6217	1.4264	1.4004	2.3584
90.3441	HD_Net	0.9135	0.9129	0.8398	3.9331	3.7134	5.9674	1.3614	1.3090	2.2662
89.6507	nnU-Net	0.9194	0.9272	0.8406	3.9509	3.7276	6.8301	1.2431	1.1771	2.1693
89.5858	Bowda-Net	0.9141	0.9222	0.8367	4.2654	3.8969	7.7235	1.3451	1.2763	2.2920
88.5397	Ours	0.9137	0.9168	0.8741	4.1176	3.8449	7.8605	1.3197	1.2864	1.8129

results obtained for the other test sets (Table 2.2). Also, comparing it to the other entries (Table 2.4), our model achieves very similar results for all metrics; yet, its Challenge Score falls behind. That said, this is also to be expected since, unlike the other contestants, no fine-tuning was performed to improve the results for this dataset in particular. BOWDA-net (Zhu et al., 2019a), for example, uses an adversarial domain adaptation strategy to transform the images from a second training dataset to the domain of the "Promise12" dataset, therefore improving the performance only on "Promise12_test". Lastly, our model used just 41 out of the 50 images provided in the "Promise12" dataset for training, as two were discarded and seven were used for testing. When comparing our model against each of the others with a Wilcoxon test, only the first contender (MSD-Net) was found to be significantly (p-value ≤ 0.01) better in all metrics, while the fourth contender (nnU-Net) was better in terms of DSC (p-value $= 0.037$) and ABD (p-value $= 0.030$), but not HD95 (p-value $= 0.439$). The nnU-Net (Isensee et al., 2020) is a very recent and interesting method that tries to automate the process of adapting a CNN architecture to a new dataset by making use of a sensible set of heuristics. Regarding the MSD-Net, unfortunately, its specifics are yet to be published as of the writing of this paper.

Ultimately, beating this challenge was never the focus of this paper. No other single model (to the author's knowledge) is able to perform as consistently as ours in so many different datasets simultaneously. This is of utmost importance if such a model is to be used in a real-life scenario, where the MR images may come from many different scanners, and may be analyzed by many different experts.

2.4.2 Qualitative Results

To asses these results qualitatively, in Figure 2.6a–n, the center 100 mm×100 mm crop (85 mm×85 mm in the case of US images) of three slices from the worst and best performing images (in terms of DSC) from each dataset have been represented, along with the GT (in red), the GT2 (in blue, when available) and the predicted segmentations (in green). Figures were generated using Python library plot_lib (Pellicer-Valero, 2020).

Regarding the worst cases, despite being the poorest performers, the differences are relatively small and often the model proposal is arguably superior to the GT. Furthermore, the central slices are almost identical in all instances, and it is only towards base and apex where the differences emerge. One of such discrepancies is the point at which the apex and the base begin, which oftentimes depends on the segmentation criteria, as it can be seen, for instace, in Figure 2.6a, where the CNN indicates the presence of prostate in the rightmost slice (at the base), while the GT label does not (although GT2 does). Finally, at several ambiguous instances (such as in the middle slice of Figure 2.6a and the rightmost slice of Figures 2.6b,j), the predicted mask (in green) behaves as an average between both experts. As discussed, this is a very desirable property for the model to have, and this is what allows it to outperform single experts on their own (as demonstrated in Table 2.3).

As for the best cases, it can be seen that they are mostly represented by larger prostates, as they are comparatively easier to segment, and also the DSC metric is biased towards them. As a curiosity, the rightmost slice in Figure 2.6n shows how the model has learned to avoid segmenting the catheter balloon that is used in prostate biopsies, the procedure during which the US images were acquired.

In terms of HD95, the worst MR case, which corresponds to an HD95 of 7.2 mm, is shown in Figure 2.7. As it can be seen, two slices from the apex are missed by the algorithm, hence amounting to a minimum of 2 × 3 mm of error, plus some extra mm. The worst performing MR case in terms of ABD coincides with the worst performing prostate in terms of DSC, which can be found in Figure 2.6d.

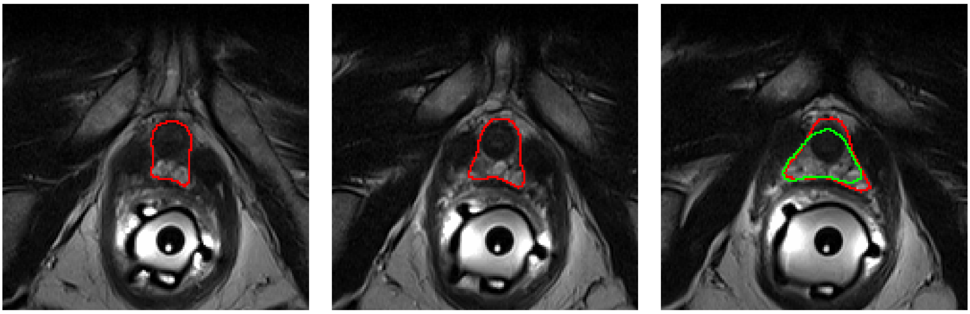


Figure 2.7: Worst HD95 (7.2 mm) of all MR test datasets. Two slices from the apex (left and center) are missed by the algorithm, hence amounting to a minimum of 2 × 3 mm of error, plus some extra mm from the segmentation errors commited in the third slice (right).

2.4.3 Ablation Studies

Table 2.5 contains the results of the ablation studies, which were performed by changing one single aspect of the baseline MR model at a time. Wilcoxon tests were performed against the baseline to check for significance (p-value < 0.05). Also, a single experiment was performed on the US model by retraining it without the use of transfer learning.

Firstly, the two post-processing techniques discussed in this paper (Checkpoint Ensembling and Neural Resolution Enhancement) are analyzed. Both show high statistical significance (p-values < 0.01) in terms of DSC and ABD. In fact, out of all the experiments conducted in this Section, only these two were found to make a statistically significant difference, probably since the worsening of the metrics, even if minor, is sustained for all images. These post-processing methods affect in no way the training process or the model, as they are applied at a later stage; therefore they are a simple and free bonus in performance, only at the cost of increased inference time.

Secondly, a battery of tests involving architectural changes (which require retraining) is presented. Even if none of these experiments showed statistical significance, several conclusions can still be extracted cautiously.

The first two experiments are an attempt to lower the complexity of the baseline model by either reducing the number of resolution levels of the network, or the amount of layers (this is: “standard” convolutions) per level. In both cases, even if the differences with respect to the baseline were small, a decrease in performance can be observed for the majority of the metrics, which justifies the use of the more complex baseline architecture if possible.

In the next two experiments, models based exclusively on the resnet architecture (with residual connections applied every four consecutive convolutions) were employed. Despite having as much as four times the amount of parameters as compared to the baseline, these models were the worst performing out of all analyzed in this Section, hence showing the power of the densenet architecture.

The next test consisted in replacing the PReLU activations with ReLU activations. Despite this having a very small influence in performance, the metrics are overall better in the baseline, and PReLU is therefore preferred given its negligible impact on model complexity.

For the following test, Deep Supervision was deactivated. In general, most of the metrics show a small improvement with this architectural modification. In our internal validation tests, this technique seemed to provide a small boost in performance and, as such, it was added to the final model. Furthermore, it stabilized the initial steps of the training procedure. However, in light of these results, its usefulness remains now in question.

For the last test, the US segmentation model was retrained using random weight initialization, instead of using the weights from the MR model for transfer learning. Although the fine-tuned model converged faster, the results suggest that training from scratch might be preferable in situations like this one, where the amount of training data is sufficient. In any event, the generality of the architecture and pre-/post-processing methodology still holds, even if the weights of the MR prostate segmentation model are not particularly useful for the problem of US prostate segmentation.

Table 2.5: Ablation studies.

	Experiment Description	*DSC*			*HD95 (mm)*			*ABD (mm)*		
		Mean	*Med.*	*Min.*	*Mean*	*Med.*	*Max.*	*Mean*	*Med.*	*Max.*
MR	Baseline (4M params.)	0.9150	0.9179	0.8094	3.8693	3.9995	7.2000	0.9815	0.9311	2.0356
	No neural Resolution Enhancement	0.9127	0.9153	0.8064	3.8991	3.9994	7.2000	1.0344	1.0090	2.1380
	No Checkpoint Ensembling	0.9128	0.9178	0.8009	4.0071	3.9997	8.0018	1.0269	0.9674	2.2703
	One less resolution reduction level	0.9145	0.9163	0.8404	3.9413	3.8665	11.1732	0.9952	0.9468	2.6608
	Half the amount of layers per level	0.9144	0.9161	0.7556	3.9394	3.8177	9.2372	0.9912	0.9316	2.3902
	Resnet with 20 layers (7M params.)	0.9063	0.9112	0.4647	4.2393	3.9995	15.1950	1.1174	0.9842	6.4507
	Resnet with 72 layers (16M params.)	0.9033	0.9150	0.1525	4.1926	3.9997	17.4923	1.1348	0.9207	9.9787
	ReLU instead of PReLU activations	0.9146	0.9171	0.8135	3.9631	3.7499	13.8931	1.0052	0.9561	3.1201
	No Deeep Supervision	0.9152	0.9143	0.8257	3.8033	3.7499	7.0495	0.9732	0.9802	1.9869
US	Baseline	0.9187	0.9235	0.8456	3.4969	3.2863	9.9997	1.1840	1.1102	2.8573
	No transfer learning (fine-tuning)	0.9207	0.9236	0.8576	3.4166	3.2372	8.3040	1.1536	1.1096	2.2470

2.5 Discussion

In this paper, a robust and generalizable model for accurate prostate segmentation has been proposed.

To achieve robustness, the model was trained with five very challenging and heterogeneous MR prostate datasets (and two US prostate datasets) with GTs originating from many different experts with varying segmentation criteria. Additionally, several key design choices, such as the use of Checkpoint Ensembling and a relatively heavy data augmentation regime, were explicitly made.

In clinical practice, the MR and US images may originate from many different scanners, with widely different characteristics (field intensities, scanner manufacturers, use of endorectal coil, etc.), to which any segmentation algorithm should be robust by design. As we have seen in Table 2.2, the proposed model has a similarly good performance for all images, no matter the dataset nor its specific characteristics.

Furthermore, such an algorithm may be used by different experts with varying criteria for segmenting the prostate. Even if it is impossible to please every criterion simultaneously, the proposed model is shown to behave as an average prediction among the different experts, as seen for instance in the rightmost slice of Figure 2.6j. This is corroborated by Table 2.3, where it shows a significantly higher overlap with any given expert (it tries to please all criteria), than the experts between themselves.

Concerning generalizability, the proposed architecture can be very easily applied to different tasks by means of transfer learning. In this paper, the MR segmentation model is simply retrained, with no hyperparameter tuning or image pre- or post-processing changes (other than the change of input resolution), on the problem of US prostate segmentation, achieving equally good performances despite the smaller dataset sizes, as seen in Table 2.2.

The main clinical applications of the proposed model lie in the context of fusion-guided biopsy and focal therapies on the prostate, which require an accurate segmentation of both MR and US prostate images. These segmentations are employed to perform registration between both modalities, which is needed to transfer the prostate lesions detected by the radiologists in the preoperative MR to the intraoperative US image in order to guide the procedure.

The proposed model can undercut, and even eliminate, the need of manual segmentations, which require expertise, are very time-consuming, and are prone to high inter- and intra-expert variability. Hence, more accurate segmentations may lead to better inter-modal prostate image registration and better prognosis in the aforementioned procedures; while almost instant results can be of particular interest for the segmentation of intraoperative US images, where the urologist currently has to spend around ten minutes manually performing this task next to the sedated patient.

Finally, a technique known as Neural Resolution Enhancement was employed as a post-processing step to reduce the impact of the lossy CNN output interpolation. This method, which leverages any already trained segmentation CNN, can also be used to improve the resolution of the output mask even beyond that of the original input image, as discussed in Figure 2.5.

This technique could be especially appealing for simulating the biomechanical behavior of the prostate, which is required by many registration algorithms and surgical simulators. To function properly, such simulations demand very high resolution meshes of the prostate geometry, which are inherently impossible to obtain due to the reduced resolution of the original MR and US images. However by using Neural Resolution Enhancement, a much higher resolution mask is obtained, which is not the result of mere interpolation, but rather a prediction of the missing geometry by combining the contextual information contained in the original image with the knowledge that the CNN has acquired about the general shape of this gland.

2.6 Conclusions

In conclusion, this paper proposes a prostate segmentation model with the potential of having a significant impact on the efficacy and efficiency of current guided prostate procedures, through improvements in terms of performance, robustness, generalizability and output resolution.

In our future work, the proposed model will be applied to different organs and tasks, such as lesion segmentation. Furthermore, different backbone architectures, such as those based on Neural Architecture Search, will be tested.

Author contributions

Conceptualization, J.D.M.-G.; Data curation, O.J.P.-V. and V.G.-P.; Methodology, O.J.P.-V.; Project administration, J.D.M.-G.; Resources, V.G.-P., I.M.G. and M.B.B.; Software, O.J.P.-V.; Supervision, J.L.C.R.-B., J.R.-B., M.J.R. and J.D.M.-G.; Validation, I.M.G., M.B.B. and P.P.G.; Visualization, O.J.P.-V.; Writing—original draft, O.J.P.-V.; Writing—review & editing, V.G.-P., M.J.R. and J.D.M.-G. All authors have read and agreed to the published version of the manuscript.

Funding

This work has been partially supported by a doctoral grant of the Spanish Ministry of Innovation and Science, with reference FPU17/01993.

Institutional review

The study was approved by the Ethical Committee of the València Institute of Oncology (CEIm-FIVO), protocol code PROSTATEDL (2019-12) and date 17th of July, 2019.

Informed consent

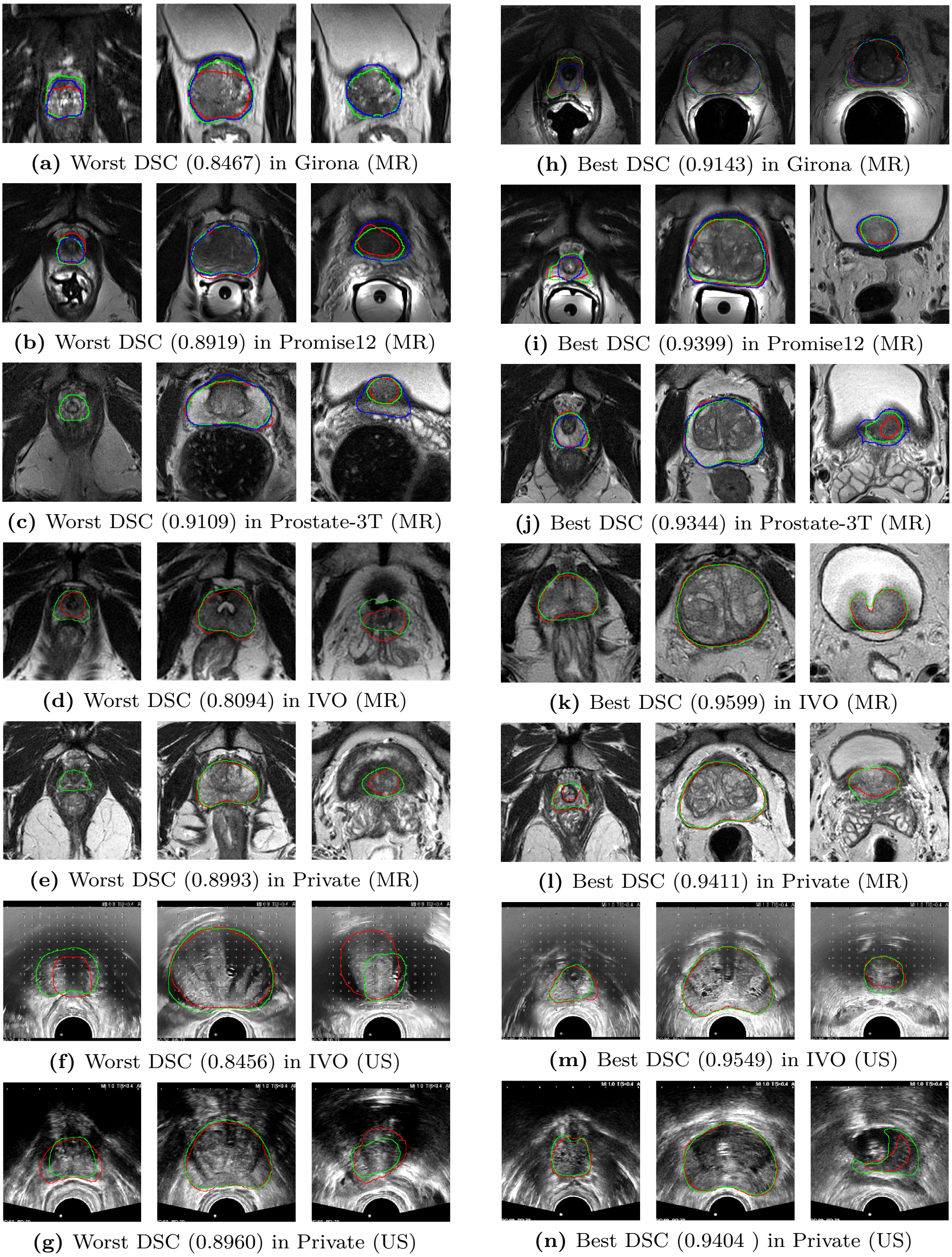
Informed consent was obtained from all subjects involved in the study.

Data availability

Dataset Girona is available at Zenodo (https://doi.org/10.5281/zenodo.162231), Promise12 at Grand Challenge (https://promise12.grand-challenge.org/Download) and Prostate-3T at the Cancer Imaging Archive (https://wiki.cancerimagingarchive.net/display/Public/Prostate-3T). Dataset IVO, from the Valencian Institute of Oncology, is not publicly available, since the ethical committee (CEIm-FIVO) only approved its use for the current study. Dataset Private comes from a private institution which retains all rights of usage for the images, and hence it is not publicly available either.

Conflicts of interest

The authors declare no conflict of interest.



(a) Worst DSC (0.8467) in Girona (MR)

(h) Best DSC (0.9143) in Girona (MR)

(b) Worst DSC (0.8919) in Promise12 (MR)

(i) Best DSC (0.9399) in Promise12 (MR)

(c) Worst DSC (0.9109) in Prostate-3T (MR)

(j) Best DSC (0.9344) in Prostate-3T (MR)

(d) Worst DSC (0.8094) in IVO (MR)

(k) Best DSC (0.9599) in IVO (MR)

(e) Worst DSC (0.8993) in Private (MR)

(l) Best DSC (0.9411) in Private (MR)

(f) Worst DSC (0.8456) in IVO (US)

(m) Best DSC (0.9549) in IVO (US)

(g) Worst DSC (0.8960) in Private (US)

(n) Best DSC (0.9404) in Private (US)

Figure 2.6: Worst (left) and best (right) segmentations in terms of DSC for each dataset (green: model, red: GT, blue: GT2).

Chapter 3

Cost-free Resolution Enhancement in Convolutional Neural Networks for Medical Image Segmentation

Reference: Oscar J. Pellicer-Valero, José D. Martín-Guerrero, and M.J. Rupérez. Cost-free resolution enhancement in Convolutional Neural Networks for medical image segmentation. In *Proc. ESANN*, pages 145–150, 2020a. URL https://www.esann.org/sites/default/files/proceedings/2020/ES2020-51.pdf

Oscar J. Pellicer-Valero*[1], María J. Rupérez-Moreno[2], José D. Martín-Guerrero[1]

[1] Intelligent Data Analysis Laboratory, Department of Electronic Engineering, ETSE (Engineering School), Universitat de València (UV), Spain.
Oscar.Pellicer@uv.es, jose.d.martin@uv.es

[2] Centro de Investigación en Ingeniería Mecánica (CIIM), Universitat Politècnica de València (UPV), Spain.
mjrupere@upvnet.upv.es

3.1 Abstract

High-resolution segmentations of medical images are imperative for applications such as treatment planning, image fusion or computer-aided surgery. Nevertheless, these are often hard and time-consuming to produce. This paper presents a method for improving the output resolution of Convolutional Neural Networks (CNNs) for medical image segmentation. It is straightforward to implement and works with any already trained CNN with no modification nor retraining required. It is able to produce better results than binary interpolation methods since it exploits all the contextual information to predict the sought values.

3.2 Introduction

Segmentation in medical images is a voxel-level classification task such that all voxels corresponding to a particular class represent a single semantical entity in the body: an organ, a bone, a tissue, a lesion, etc. Segmentation algorithms take an image as an input (e.g., a chest radiography), and one or several masks (e.g., lungs and lesions) are obtained as an output (Figure 3.1). These algorithms are commonly applied to medical imaging techniques such as Magnetic Resonance (MR), Computerized Tomography (CT) and Ultrasound (US).

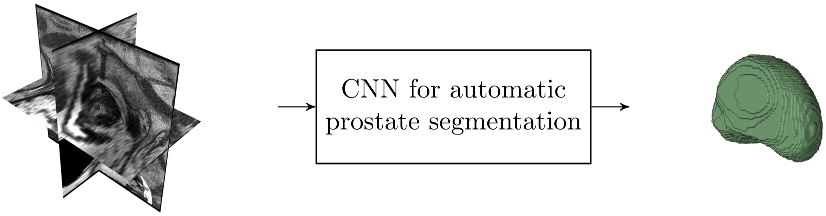


Figure 3.1: A CNN trained on the task of prostate segmentation takes a 3D prostate MR image as an input and obtains a 3D prostate mask as an output.

Obtaining accurate segmentations is a very valuable yet difficult endeavor. On one hand, segmentations are valuable as they are mandatory inputs for image-based diagnosis, lesion detection and treatment planning; furthermore, for three-dimensional (3D) images, the obtained geometries can be used to perform simulations of the biomechanical behavior of a body, which can then be used in image fusion, surgical planning, computer-aided surgery or bone-strength simulations, to cite a few applications.

On the other hand though, accurate segmentations are hard and laborious to obtain, since they have to be manually annotated by expert radiologists, and even then, the inter- and intra-observer variations may be significant (Ibragimov and Xing, 2017). Because of this, automatic segmentation algorithms for medical images have become increasingly prevailing.

Although several different automatic segmentation frameworks have been suggested in the past, current state-of-the-art techniques usually employ Convolutional Neural Networks (CNNs) and, more specifically, those based on the U-Net architecture (Ronneberger et al., 2015), which has lead to segmentation accuracies above the inter-observer threshold in increasingly more scenarios (Ibragimov and Xing, 2017).

High-resolution segmentations are often essential in the aforementioned applications. However, automatic segmentation techniques tend to present two closely related problems. Firstly, CNNs usually require the input image to be downscaled before processing it to alleviate the Graphics Processing Unit (GPU) memory costs associated with 3D convolutions. Secondly, so-called 3D medical imaging techniques are often actually two-dimensional (2D) multi-slice images instead, which are then stacked to form the final 3D geometry. These images, however, usually have a dimension (perpendicular to all the individual 2D slices) along which the resolution is much coarser than the rest. For instance, the MR image in Figure 3.1 suffers from this inconvenience.

Even if the first problem could be solved by using a finer resolution image as input to the CNN, the second problem remains still a challenge, since no inter-slice information can be extracted from 2D multi-slice images in order to perform a finer segmentation.

One possible solution would be to improve the resolution of the input image in an intelligent manner, which is a problem known as super-resolution (Ledig et al., 2017). These upscaled images could then be used to train a segmentation CNN, thus obtaining higher resolution output masks. Some works have already studied this in the medical domain; for instance, Oktay et al. (2016) proposes a CNN to upscale 2D multi-slice images of the heart along the axis perpendicular to the slices, achieving perceptible improvements. However, with this approach, the problem of GPU memory limitations still remains. Furthermore, in order to train the CNN, many images should be manually segmented at the new increased resolution, thus making the process even more time-consuming.

Another possible solution is to employ binary interpolation techniques to produce a high-resolution mask from a lower resolution one. The simplest approach is nearest-neighbor interpolation, which simply takes the value of the closest neighbor for any given point. Even if this procedure produces very “blocky” low-quality interpolations, it is still widely used due to its simplicity and speed. A better approach consists in taking any kind of interpolator for real numbers and using it to interpolate the binary masks for each class independently; then, the class with the highest value at any given point is used as the final label for that point. This approach provides much smoother results and, in combination with linear interpolation, it is also very quick. Finally, some more complex algorithms have been proposed to deal specifically with the problem of inter-slice interpolation in 2D multi-slice images, such as in Albu et al. (2008). However, all these methods present one important pitfall: they completely disregard the contextual information contained either in the original image or in the domain knowledge of the problem. The single notable exception seems to be Liao et al. (2011), where the authors combine both the binary morphology and the local intensities to perform the interpolation.

In this paper, a method for intelligent upscaling of the output mask of a medical image segmentation CNN is proposed. This method takes into account all the available contextual information and it is cost-free in the sense that it can be applied to any already trained CNN with no modifications to its architecture or any retraining required.

3.3 Materials and methods

The proposed method exploits a very simple yet effective idea for performing intelligent output upscaling on already trained segmentation CNNs. It consists in shifting the input image by several different sub-voxel amounts, feeding these transformed images to the CNN in order to obtain the segmentation masks, and then combining them into a single final high resolution mask. Despite its simplicity, this procedure achieves high resolution segmentation masks which outperform other discussed approaches (as it will be discussed in Section 3.4) from already trained (an possibly low resolution) segmentation CNNs. Thus, the problem of interpolation is shifted from the mask domain to the image domain, where the conveyed information is still complete and not yet binarized.

For a more detailed description of the method, consider a CNN with input and output dimensionality (or resolution) of $(d_1 \times \cdots \times d_N)$, where N is the number of dimensions (e.g., $N = 3$ for a 3D image). Suppose we wanted to increase the output resolution along a single dimension i by an integer factor of k_i, such that the output resolution were: $(d_1 \times \cdots \times d_i \cdot k_i \times \cdots \times d_N)$.

First, we would need to generate $k_i - 1$ images $[I_1, \ldots, I_{k_i-1}]$ from the original input image I_0, each one shifted $+\frac{1}{k_i}$ voxels along dimension i with respect to the previous one. Therefore, in order to obtain $[I_1, \ldots, I_{k_i-1}]$, I_0 must be interpolated and evaluated at the positions given by the translation transforms $\left[(0, \ldots, \frac{1}{k_i}, \ldots, 0), \ldots (0, \ldots, \frac{k_i-1}{k_i}, \ldots, 0)\right]$ applied to I_0, where the translation has a value of zero for all dimensions except for i. Second, all the k_i images $[I_0] \cup [I_1, \ldots, I_{k_i-1}]$ are fed to the CNN one by one, and k_i outputs masks $[O_0, O_1, \ldots, O_{k_i-1}]$ are obtained in return. Finally, $[O_0, O_1, \ldots, O_{k_i-1}]$ are combined by interleaving them voxel-wise along i in order to obtain a single output mask $O_{combined}$, which will have a k_i-times higher resolution along axis i. In this context, interleaving can be defined as stacking $[O_0, O_1, \ldots, O_{k_i-1}]$ to produce $O_{combined}$ in such a way that the n^{th} slice along dimension i in $O_{combined}$ corresponds to the $\lfloor \frac{n}{k_i} \rfloor^{th}$ slice along dimension i of $O_{n\%k_i}$. Figure 3.2 provides a visual representation of the described methodology.

The proposed method can be extended to simultaneously improve the resolution of the output mask along any number of dimensions. As an overview, a new set of transformations T must be computed and applied to I_0, by combining in all possible ways the transformations which would be required to increase the resolution by a factor k_m along a single dimension $m \in (1, N)$ independently:

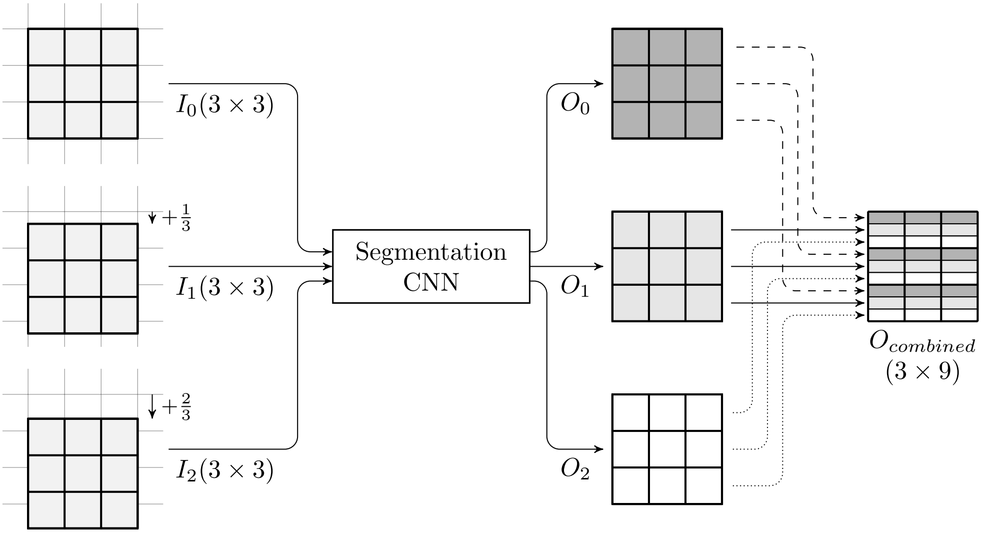


Figure 3.2: Visual representation of the method for an image of dimensions $(d_1, d_2) = (3, 3)$, for $k_2 = 3$.

$$T = \left[(\frac{1}{k_1}, 0, \ldots, 0), \ldots (\frac{k_1 - 1}{k_1}, 0, \ldots, 0) \right] \times \cdots$$
$$\times \left[(0, \ldots, 0, \frac{1}{k_N}), \ldots (0, \ldots, 0, \frac{k_N - 1}{k_N}) \right] \tag{3.1}$$

Finally, the resulting masks are combined to form the final mask. It must be however noted that, for some positions in $O_{combined}$ there will be several possible values due to overlap among masks. In those instances, a binary fusion function (such as the majority vote) must be used to produce a final value.

The computational cost of the method is approximately $c_0 \cdot \prod_{i=1}^{N} k_i$, where c_0 is the cost of interpolating an image and passing it through the CNN. This cost comes from computing the size of the set of transformations shown in Eq. (3.1). As an example, if we wanted to increase the resolution along a single axis i by a factor of k_i, the cost would be k_i times the cost of obtaining a single mask in the native CNN resolution.

3.4 Results and discussion

This method was applied to a segmentation CNN trained on several datasets (Litjens et al., 2014b, 2015) of prostate MR 2D multi-slice images, where the approximate physical voxel spacings are $(s_x, s_y, s_z) \approx (0.5, 0.5, 3)mm$, and s_z is to be improved by a factor $k_z = 3$ ($s_z = 1mm$) and $k_z = 6$ ($s_z = 0.5mm$). Figure 3.3 shows a comparison between two of the most common binary interpolation methods and the proposed method.

The proposed method produces the smoothest results of the three, as it can be noticed by comparing the upper slices of the prostate masks in Figure 3.3. Furthermore, it does not just interpolate between slices, but rather predicts the mask at several inter-slice levels. Therefore, it is able to obtain more accurate results by incorporating the underlying image as input, as well as all the contextual information that the CNN has learned about the problem of segmenting a particular part of the body. Unfortunately, no numerical results can be provided, as no ground truth is available, since all the methods are interpolating beyond the resolution of the original image.

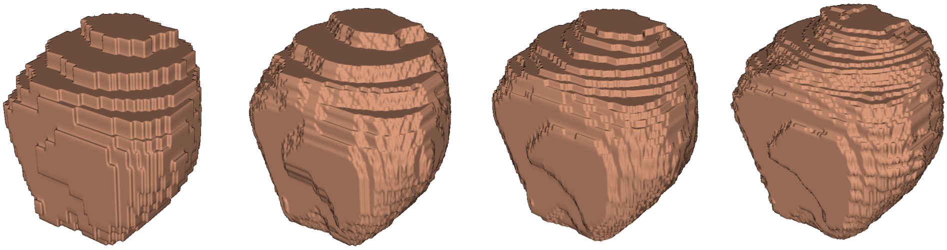


Figure 3.3: Mask predicted by a prostate segmentation CNN and upscaled along the z-axis using (from left to right): nearest-neighbor interpolation, Gaussian interpolation ($k_z = 6$), the proposed method with $k_z = 3$ and the proposed method with $k_z = 6$.

This approach seems to be closely related to a technique known as Test-Time Augmentation, wherein an already trained CNN is provided with several randomly augmented (translated, rotated, shifted, etc.) versions of the same input image, and the outputs are combined into a single output prediction, which is oftentimes more accurate than any individual prediction. Similarly, the proposed method feeds the CNN several transformed versions of the same input and then combines all the outputs. However, by contrast, the transformations are not random and follow instead a very precise structure which must also be taken into account in the output combination process.

3.5 Conclusion and further work

This paper presents a method for intelligently improving the output resolution of CNNs for medical image segmentation. It is better than other uspcaling methods since it does not perform interpolation, but rather it predicts the sub-voxel values using the image and the context information that the CNN has encoded about the particular problem. It can be used to improve the resolution of 2D multi-slice images beyond the original resolution of the image, thus providing an accurate 3D segmentation for methods that require it, such as in the simulation of the biomechanical behavior of a body. Finally, it is a very simple to implement post-processing step that can make use of any already existing CNN with no modifications required whatsoever.

As a main downside, the method can only improve the resolution of the predictions of a CNN, unlike general binary interpolation algorithms, which can upscale any binary image. Also, the computational cost of this procedure can be high if the resolution is increased along many different axes simultaneously. Lastly, it is not proven in any way that the results should be smooth and/or correct, and it is instead trusted in the empirical results and the robustness intuitions about CNN architectures.

From the ideas here presented, two main research lines arise. First, it should be explored how well this technique generalizes to natural image segmentation CNNs, and how useful it would be in this context. Second, and more interestingly, a niche for improvement has been discovered in binary interpolation algorithms for segmentations. Namely, almost all current binary interpolators disregard the precious information contained in the original image to perform the interpolation. A clever exploitation of this information could yield improved interpolations for segmentation masks.

Chapter 4

Deep Learning for Fully Automatic Detection, Segmentation, and Gleason Grade Estimation of Prostate Cancer in Multiparametric Magnetic Resonance Images

Reference: Oscar J. Pellicer-Valero, José L. Marenco Jiménez, Victor Gonzalez-Perez, Juan Luis Casanova Ramón-Borja, Isabel Martín García, María Barrios Benito, Paula Pelechano Gómez, José Rubio-Briones, María José Rupérez, and José D. Martín-Guerrero. Deep learning for fully automatic detection, segmentation, and Gleason grade estimation of prostate cancer in multiparametric magnetic resonance images. *Sci. Rep.*, 12(1):2975, dec 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-06730-6. URL https://www.nature.com/articles/s41598-022-06730-6

Oscar J. Pellicer-Valero*[1], José L. Marenco Jiménez[2], Victor Gonzalez-Perez[3], Juan Luis Casanova Ramón-Borja[2], Isabel Martín García[4], María Barrios Benito[4], Paula Pelechano Gómez[4], José Rubio-Briones[2], María José Rupérez[5], José D. Martín-Guerrero[1]

[1] Intelligent Data Analysis Laboratory, Department of Electronic Engineering, ETSE (Engineering School), Universitat de València (UV), Av. Universitat, sn, 46100 Bujassot, València, Spain.
Oscar.Pellicer@uv.es (+34 9635 44022), jose.d.martin@uv.es

[2] Department of Urology, Fundación Instituto Valenciano de Oncología (FIVO), Beltrán Báguena, 8, 46009 València, Spain.
jlmarencoj@gmail.com, jcasanova@fivo.org, jrubio@fivo.org

[3] Department of Medical Physics, Fundación Instituto Valenciano de Oncología (FIVO), Beltrán Báguena, 8, 46009 València, Spain.
vgonzalezper@hotmail.com

[4] Department of Radiodiagnosis, Fundación Instituto Valenciano de Oncología (FIVO), Beltrán Báguena, 8, 46009 València, Spain.
mismaga99@gmail.com, mar7esc@gmail.com, ppelechano@hotmail.com

[5] Instituto de Ingeniería Mecánica y Biomecánica, Universitat Politècnica de València (UPV), Camino de Vera, sn, 46022, València, Spain.
mjrupere@upvnet.upv.es

4.1 Abstract

Although the emergence of multi-parametric magnetic resonance imaging (mpMRI) has had a profound impact on the diagnosis of prostate cancers (PCa), analyzing these images remains still complex even for experts. This paper proposes a fully automatic system based on Deep Learning that performs localization, segmentation and Gleason grade group (GGG) estimation of PCa lesions from prostate mpMRIs. It uses 490 mpMRIs for training/validation and 75 for testing from two different datasets: ProstateX and Valencian Oncology Institute Foundation. In the test set, it achieves an excellent lesion-level AUC/sensitivity/specificity for the GGG≥2 significance criterion of 0.96/1.00/0.79 for the ProstateX dataset, and 0.95/1.00/0.80 for the IVO dataset. At a patient level, the results are 0.87/1.00/0.375 in ProstateX, and 0.91/1.00/0.762 in IVO. Furthermore, on the online ProstateX grand challenge, the model obtained an AUC of 0.85 (0.87 when trained only on the ProstateX data, tying up with the original winner of the challenge). For expert comparison, IVO radiologist's PI-RADS 4 sensitivity/specificity were 0.88/0.56 at a lesion level, and 0.85/0.58 at a patient level. The full code for the ProstateX-trained model is openly available at https://github.com/OscarPellicer/prostate_lesion_detection. We hope that this will represent a landmark for future research to use, compare and improve upon.

Keywords: multi-parametric magnetic resonance imaging; prostate cancer; deep learning; convolutional neural network; cancer detection; lesion segmentation; computer-aided diagnosis; prostate zonal segmentation

4.2 Introduction

Prostate cancer (PCa) is the most frequently diagnosed malignancy in males in Europe and the USA and the second in the number of deaths (Bray et al., 2018). Magnetic resonance imaging (MRI) is a medical imaging technique that employs very strong magnetic fields (typically 1.5-3T) to obtain three-dimensional (3D) images of the body; multi-parametric MRI (mpMRI) extends MRI by combining several MRI sequences into a multi-channel 3D image, each sequence providing different information on the imaged tissue. mpMRI has drastically changed the diagnostic approach of PCa: The traditional pathway includes screening based on the determination of prostate serum antigen (PSA) levels and digital rectal examination followed by a systematic random transrectal biopsy (Mottet et al., 2017). However, in recent years, the introduction of pre-biopsy mpMRI has enabled better selection of patients for prostate biopsy (Mehralivand et al., 2018), increasing the diagnostic yield of the procedurep (Ahmed et al., 2017) and allowing for more precise fusion-guided biopsy examinations and focal therapies as compared with cognitive fusion approaches (Marra et al., 2019). Additionally, mpMRI-derived parameters, such as tumor volume or PSA density (PSA divided by prostate volume) have proven helpful prognosis and stratification tools (Cellini et al., 2002).

To promote global standardization in the interpretation of prostate mpMRI examinations, the Prostate Imaging Reporting and Data System (PI-RADS) in its latest 2.1 version combines available evidence to assign scores to objective findings in each sequence (Turkbey and Choyke, 2019). However, mpMRI interpretation is time-consuming, expertise dependent (Gaziev et al., 2016), and is usually accompanied by a non-negligible inter-observer variability (Sonn et al., 2019). This is particularly the case outside of expert high-volume centers (Kohestani et al., 2019). Although promising alternative mpMRI scoring criteria are being developed, such as Likert (Khoo et al., 2020), PI-RADS remains still the most widely used criterion for both clinical and academic purposes.

Computer-aided diagnosis (CAD) systems have been broadly defined as "the use of computer algorithms to aid the image interpretation process" (Giger and Suzuki, 2008). In this sense, CAD is one of the most exciting lines of research in medical imaging and has been successfully applied to interpret images in different medical scenarios (Morton et al., 2006). CAD poses several theoretical advantages, namely speeding up the diagnosis, reducing diagnostic errors, and improving quantitative evaluation (Van Ginneken et al., 2011). On the topic of mpMRI-based PCa CAD, different methods have been proposed since the early 2000s (Chan et al., 2003). These pioneered the field but were nonetheless limited in some important aspects (e.g. they lacked proper evaluation, expert comparison, and large enough datasets). In 2014, Litjens et al. (2014a) proposed the first CAD system able to provide candidate regions for lesions along with their likelihood for malignancy using pharmacokinetic and appearance-derived features from several MRI sequences using classical (non-Deep Learning) voxel-based classification algorithms and evaluated the results on a large cohort of 347 patients.

Since the advent of Deep Learning (Krizhevsky et al., 2012), however, Deep Convolutional Neural Networks (CNNs) have quickly dominated all kinds of image analysis applications (medical and otherwise), phasing out classical classification techniques. In the context of the prostate, the turning point can be traced back to the ProstateX challenge in 2016 (Litjens et al., 2014a, 2017; Armato et al., 2018). The challenge consisted in the classification of clinically significant PCa (csPCa) given some tentative locations on mpMRI. More importantly, a training set of 204 mpMRIs (330 lesions) was provided openly for training the models, hence enabling many researchers to venture into the problem (further details of this dataset can be found in Section 4.5.1). At the time, half of the contestants employed classical classification methods (Kitchen and Seah, 2017) and the other half CNNs (Liu et al., 2017). In all cases, a patch (or region of interest, ROI) of the mpMRI around the lesion was extracted, and a machine learning algorithm was trained to classify it as either csPCa or not. The second-highest-scoring method (Liu et al., 2017), with a receiver operating characteristic -ROC- curve (AUC) of 0.84, used a simple VGG-like (Simonyan and Zisserman, 2015) CNN architecture trained over the mpMRI ROIs to perform classification. The main limitation of all these approaches is that ROIs have to be manually located beforehand (even after the model has been trained), hence limiting their interest and applicability to clinical practice.

In 2019, Cao et al. (2019) employed a slice-wise segmentation CNN, FocalNet, not only to predict csPCa but also a to obtain a map of the Gleason grade group (GGG) (Epstein et al., 2005, 2016) of the prostate. Very briefly, GGG is a standard 1-5 grading system for PCa, where GGG1 cancer cells look normal and are likely to grow slowly (if at all), while GGG5 cells look very abnormal and are likely to grow very quickly. Segmentation-based models are a step up from previous patch classification approaches because they provide a csPCa map of the prostate; however, they cannot directly identify lesions as individual entities and assign a score to each one, as is common procedure in clinical practice. This is natively solved in an instance detection+segmentation framework, which is very common in natural image detection tasks (He et al., 2017); but has never been applied to csPCa detection. Additionally, two-dimensional (2D) slice-wise CNNs are known to generally underperform as compared with actual 3D CNNs in lesion detection tasks (Jaeger et al., 2020). Indeed, in 2020 several authors turned to 3D CNNs, such as Arif et al. (2020) or Aldoj et al. (2020b).

To the best of our knowledge, the model we propose is the first to leverage a proper instance detection and segmentation network, the 3D Retina U-Net (Jaeger et al., 2020), to simultaneously perform detection, segmentation, and Gleason Grade estimation from mpMRIs to a state-of-the-art performance level. It is also one of the few works that combines two very different mpMRI datasets into a single model: the ProstateX dataset and the IVO (Valencian Institute of Oncology Foundation) dataset (view Section 4.5.1), achieving similarly excellent results in both. It uses prior prostate zonal segmentation information, which is provided by an automatic segmentation model, and leverages an automatic non-rigid MRI sequence registration algorithm, among other subsystems, allowing for a fully automatic system that requires no intervention. The code of this project has been made available online at `https://github.com/OscarPellicer/prostate_lesion_detection`.

4.3 Results

4.3.1 Lesion detection, segmentation, and classification

4.3.1.1 Quantitative results

A comprehensive quantitative evaluation of the trained model on the ProstateX and IVO test sets has been compiled in Table 4.1 (showing sensitivity and specificity) and in Table 4.2 (showing positive predictive value and negative predictive value). The computation procedure for patient- and lesion-level metrics is explained in Section 4.5.4. For the evaluation of sensitivity and specificity, the model-predicted scores were thresholded at two working points (computed a posteriori on the test data): maximum sensitivity and balanced (similar sensitivity and specificity). Furthermore, radiologist-assigned pre-biopsy PI-RADS scores for all IVO patients with no missing sequences and with PI-RADS information available (N=106 patients, 111 lesions) has

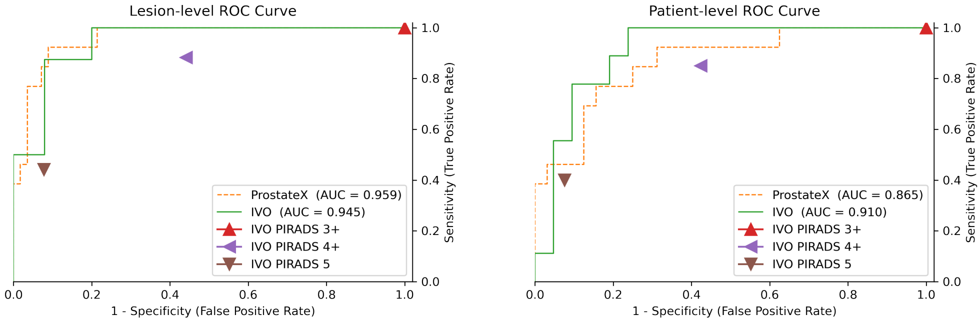


Figure 4.1: ROC curve of the model for significance criterion Gleason Grade Group ≥2, evaluated at the lesion level (left) and the patient level (right). For comparison, triangular marks represent the radiologist-assigned pre-biopsy PI-RADS. AUC: area under the ROC curve.

also been included in Table 3 for comparison. Please notice that PI-RADS≥ 3 is omitted since all IVO lesions were assigned at least a PI-RADS 3 score, and hence PI-RADS≥ 3 acts just as a naïve classifier that considers all samples as positive (sensitivity 1 and specificity 0). A graphical representation of the area under the receiver operating characteristic (ROC) curve for the main significance criterion (GGG≥2) can be found in Figure 4.1. Also, Table 4.3 uses a single threshold for all tests (but different for IVO and ProstateX datasets), computed a priori from the training data; this table might be a better proxy for the prospective performance of the model.

Focusing on the results for the GGG≥2 significance criterion, at the highest sensitivity working point, the model achieves a perfect lesion-level sensitivity of 1 (no csPCa is missed) and a specificity of 0.786 and 0.875 for ProstateX and IVO, respectively (AUCs: 0.959 and 0.945). At the patient level, the specificity falls to 0.375 and 0.762 for each dataset (AUCs: 0.865 and 0.910).

For the GGG≥1 significance criterion, the model achieves a lesion-/patient-level maximum sensitivity of 0.941 (spec. 0.788) / 1 (spec. 0.138) in the ProstateX dataset, and a maximum sensitivity of 1 (spec. 0.350) / 1 (spec. 0.667) in the IVO dataset. In summary, no GGG≥1 patient was missed, although at a cost of low specificity. Using the GGG≥3 significance criterion the model reaches a lesion- and patient-level sensitivity of 0.714 (spec. 0.887) / 1 (spec.: 0.395) in the ProstateX dataset, and a maximum sensitivity of 1 (spec. 0.800) / 1 (spec. 0.778) in the IVO dataset.

Regarding lesion segmentation performance, the mean DSC across all patients for segmenting any type of lesion irrespective of their GGG (including GGG0 benign lesions), was 0.276/0.255 for the IVO/ProstateX dataset when evaluated at the 0.25 segmentation threshold, and 0.245/0.244 when evaluated at 0.5.

Table 4.1: Quantitative results for IVO (top) and ProstateX (bottom) test data evaluated with different Gleason Grade Group (GGG) significance criteria (e.g. lesions with GGG≥1, 2, or 3 are considered positive), at lesion- and patient-level ($N_{positives}/N_{total}$), and at two thresholds (t): maximum sensitivity and balanced. For IVO data, results are compared with radiologist-assigned pre-biopsy PI-RADS scores for all IVO patients with no missing sequences and with PI-RADS information available (N=106 patients, 111 lesions). AUC: Area under the ROC curve.

(Dataset) & Significance criterion	*Level*	*AUC*	*Max. sensitivity*			*Balanced*			*PI-RADS≥4*		*PI-RADS=5*	
			t	*Sens.*	*Spec.*	*t*	*Sens.*	*Spec.*	*Sens.*	*Spec.*	*Sens.*	*Spec.*
(IVO) GGG≥1	Lesion (13/33)	0.892	0.027	1.000	0.350	0.105	0.923	0.700	0.741	0.604	0.328	0.962
	Patient (15/30)	0.920	0.253	1.000	0.667	0.301	0.867	0.800	0.710	0.649	0.290	0.973
GGG≥2	Lesion (8/33)	0.945	0.173	1.000	0.800	0.301	0.875	0.920	0.882	0.558	0.441	0.922
	Patient (9/30)	0.910	0.219	1.000	0.762	0.262	0.889	0.810	0.850	0.576	0.400	0.924
GGG≥3	Lesion (3/33)	0.856	0.301	1.000	0.800	0.315	0.667	0.867	0.727	0.440	0.455	0.840
	Patient (3/30)	0.840	0.301	1.000	0.778	0.315	0.667	0.852	0.727	0.432	0.455	0.832
(ProstateX) GGG≥1	Lesion (17/69)	0.898	0.028	0.941	0.788	0.053	0.824	0.865	-	-	-	-
	Patient (16/45)	0.866	0.108	1.000	0.138	0.104	0.938	0.655	-	-	-	-
GGG≥2	Lesion (13/69)	0.959	0.028	1.000	0.786	0.108	0.923	0.911	-	-	-	-
	Patient (13/45)	0.865	0.028	1.000	0.375	0.108	0.923	0.688	-	-	-	-
GGG≥3	Lesion (7/69)	0.751	0.195	0.714	0.887	0.195	0.714	0.887	-	-	-	-
	Patient (7/45)	0.767	0.016	1.000	0.395	0.026	0.857	0.500	-	-	-	-

4.3.1.2 Qualitative results

Figure 4.2 shows the output of the model evaluated on two IVO test patients and three ProstateX test patients. For the sake of clarity, GGG0 (benign) bounding boxes (BBs) are not shown and, for highly overlapped detections (Intersection over Union, -IoU- > 0.25), only the highest-scoring BB is drawn. Detections with confidence below the GGG≥2 lesion-wise maximum sensitivity threshold (0.173 for IVO, and 0.028 for ProstateX) are not shown either. The first IVO patient (Figure 4.2, row 1) is of special interest, as it is one of the relatively few IVO cases where the targeted biopsy did not find csPCa (as evidenced by the GGG0 BB in the GT image to the left), but the massive biopsy (20-30 cylinders) detected GGG2 csPCa. As can be seen, the model was able to detect this GGG2 lesion while ignoring the benign GGG0 one, hence outperforming the radiologists for this particular instance. For the second IVO patient (Figure 4.2, row 2) a GGG3+ GT lesion (GGG4 specifically) was properly detected by the model with very high confidence.

The first ProstateX patient (Figure 4.2, row 3) is a case of failure, where the model detects a non-existent GGG2 lesion, albeit with relatively low confidence; in fact, it would have been ignored at the balanced sensitivity setting ($t = 0.108$). For the next patient (Figure 4.2, row 4), the model has been able to segment both GT lesions; however, only the csPCa lesion is detected, while the other is ignored (actually, the model correctly detected the other lesion as a GGG0, but BBs for those lesions are not shown). For the third patient (Figure 4.2, row 5), the model could correctly identify

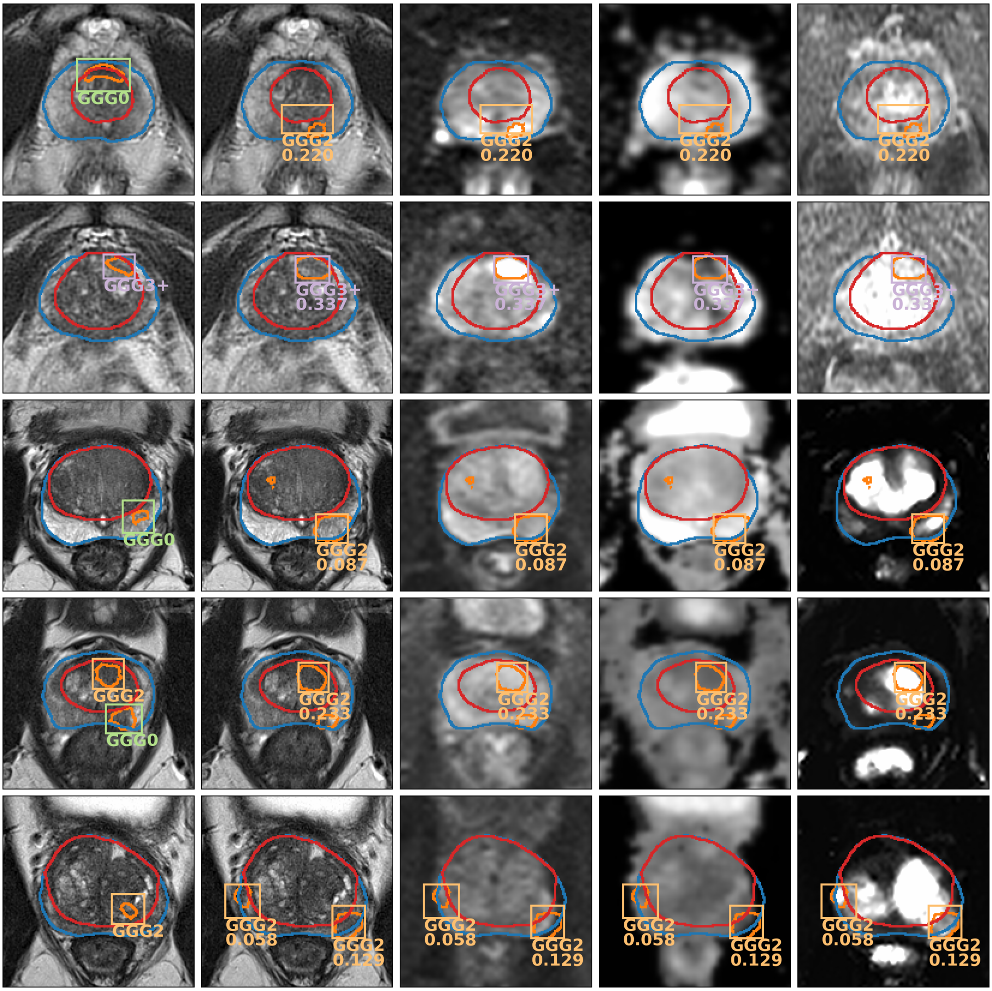


Figure 4.2: Output of the model (every row corresponds to a different patient) evaluated on two IVO test patients (first two rows) and three ProstateX test patients (last three rows). For each patient, first image from the left shows the ground truth on the T2 sequence; the rest show the output predictions of the model on different sequences (from left to right: T2, b800, ADC, K^{trans} -IVO- / DCE $t = 30$ -ProstateX-). Gleason Grade Group (GGG) 0 -benign- bounding boxes (BBs) are not shown and only the highest-scoring BB is shown for sets of highly overlapped detections (intersection over union > 0.25). Detections with confidence below the GGG$\geq$2 lesion-wise maximum sensitivity threshold (0.173 for IVO, and 0.028 for ProstateX) are not shown either.

Table 4.2: Quantitative results for IVO (top) and ProstateX (bottom) test data evaluated with different Gleason Grade Group (GGG) significance criteria (e.g.: lesions with GGG≥1, 2, or 3 are considered positive), at lesion- and patient-level ($N_{positives}/N_{total}$), and at two thresholds (t): maximum sensitivity and balanced (same thresholds as Table 1 in the main text). Results here are expressed in terms of positive predictive value (PPV) and negative predictive value (NPV). For IVO data, results are compared with radiologist-assigned pre-biopsy PI-RADS scores for all IVO patients with no missing sequences and with PI-RADS information available (N=106 patients, 111 lesions). AUC: Area under the ROC curve.

(Dataset) & Significance criterion	*Level*	*AUC*	*Max. sensitivity*			*Balanced*			*PI-RADS≥4*		*PI-RADS=5*	
			t	*PPV*	*NPV*	*t*	*PPV*	*NPV*	*PPV*	*NPV*	*PPV*	*NPV*
(IVO) GGG≥1	Lesion (13/33)	0.892	0.027	0.500	1.000	0.105	0.667	0.933	0.672	0.681	0.905	0.567
	Patient (15/30)	0.920	0.253	0.750	1.000	0.301	0.812	0.857	0.790	0.545	0.952	0.424
GGG≥2	Lesion (8/33)	0.945	0.173	0.615	1.000	0.301	0.700	0.957	0.469	0.915	0.714	0.789
	Patient (9/30)	0.910	0.219	0.643	1.000	0.262	0.667	0.944	0.548	0.864	0.762	0.718
GGG≥3	Lesion (3/33)	0.856	0.301	0.333	1.000	0.315	0.333	0.963	0.125	0.936	0.238	0.933
	Patient (3/30)	0.840	0.301	0.333	1.000	0.315	0.333	0.958	0.129	0.932	0.238	0.929
(Pros.X) GGG≥1	Lesion (17/69)	0.898	0.028	0.593	0.976	0.053	0.667	0.938	-	-	-	-
	Patient (16/45)	0.866	0.028	0.390	1.000	0.104	0.600	0.950	-	-	-	-
GGG≥2	Lesion (13/69)	0.959	0.028	0.520	1.000	0.108	0.706	0.981	-	-	-	-
	Patient (13/45)	0.865	0.028	0.394	1.000	0.108	0.545	0.957	-	-	-	-
GGG≥3	Lesion (7/69)	0.751	0.195	0.417	0.965	0.195	0.417	0.965	-	-	-	-
	Patient (7/45)	0.767	0.016	0.233	1.000	0.026	0.240	0.950	-	-	-	-

the GGG2 GT lesion but also identified an additional GGG2 lesion. This might be a mistake or might show a real lesion that was missed by the radiologists (we cannot know, as no massive biopsy information is available for the ProstateX dataset). Due to this uncertainty, lesion-level evaluation should not penalize detections for which GT information was not available (such as this one), as discussed in Section 4.5.4.

4.3.1.3 Sequence ablation tests

In Section 4.5.3.3, Random Channel Drop is presented as a training-time data augmentation technique that should help alleviate the problem of missing sequences. For a model trained in such a fashion, we can assess the individual importance of the different sequences by dropping them (i.e.: setting them to 0) at test time and analyzing the performance penalty that the model incurs. The AUCs after dropping different sequences (or combinations of them) are shown in Table 4.4.

As can be seen, removing the low b-valued (b400 for ProstateX/b500 for IVO) DW sequence seems to have minimal impact on both datasets, as is to be expected. Conversely, while removing the high b-valued (b800 for ProstateX/b1000 or b1400 for IVO) DW sequences has little impact on the ProstateX data, it severely affects the performance on the IVO data, likely due to the higher b values employed in this dataset (which may prove more informative). Furthermore, removing all DW sequences

Table 4.3: Quantitative results for IVO (top) and ProstateX (bottom) test data evaluated with different Gleason Grade Group (GGG) significance criteria (e.g.: lesions with GGG≥1, 2, or 3 are considered positive), at lesion- and patient-level ($N_{positives}/N_{total}$), and at two thresholds (t): maximum test sensitivity (which is the lesion-level maximum sensitivity threshold for GGG≥2 classification in the test set, but applied to all GGGs), and the maximum train sensitivity (same as the previous one, but the threshold was obtained from the training data). Results are expressed in terms of Area under the ROC curve (AUC), sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV).

(Dataset) & Significance criterion	*Level*	*AUC*	*Max. train sensitivity*					*Max. test sensitivity*				
			t	*Sens.*	*Spec.*	*PPV*	*NPV*	*t*	*Sens.*	*Spec.*	*PPV*	*NPV*
(IVO) GGG≥1	Lesion (13/33)	0.892	0.164	0.846	0.800	0.733	0.889	0.173	0.846	0.800	0.733	0.889
	Patient (15/30)	0.920	0.164	1.000	0.067	0.517	1.000	0.173	1.000	0.200	0.556	1.000
GGG≥2	Lesion (8/33)	0.945	0.164	1.000	0.600	0.444	1.000	0.173	1.000	0.800	0.615	1.000
	Patient (9/30)	0.910	0.164	1.000	0.571	0.500	1.000	0.173	1.000	0.571	0.500	1.000
GGG≥3	Lesion (3/33)	0.856	0.164	0.714	0.855	0.357	0.964	0.173	1.000	0.667	0.231	1.000
	Patient (3/30)	0.840	0.164	0.714	0.711	0.312	0.931	0.173	1.000	0.630	0.231	1.000
(Pros.X) GGG≥1	Lesion (17/69)	0.898	0.086	0.706	0.923	0.750	0.906	0.028	0.941	0.788	0.593	0.976
	Patient (16/45)	0.866	0.086	0.938	0.586	0.556	0.944	0.028	1.000	0.138	0.390	1.000
GGG≥2	Lesion (13/69)	0.959	0.086	0.923	0.893	0.667	0.980	0.028	1.000	0.786	0.520	1.000
	Patient (13/45)	0.865	0.086	0.923	0.625	0.500	0.952	0.028	1.000	0.375	0.394	1.000
GGG≥3	Lesion (7/69)	0.751	0.086	1.000	0.667	0.231	1.000	0.028	0.714	0.742	0.238	0.958
	Patient (7/45)	0.767	0.086	1.000	0.630	0.231	1.000	0.028	0.714	0.500	0.208	0.905

severely affects the IVO dataset, but has almost no impact on ProstateX. The removal of the ADC map has a similar negative impact on both datasets, although the results vary depending on how they are analyzed (lesion- or patient-wise). Likewise, dropping the K^{trans} sequence on the ProstateX data or the DCE sequences on the IVO data clearly harms the performance. For the final test, all sequences are dropped except for the T2; despite it, the model still has a commendable performance, especially in the ProstateX set, which might indicate that the proposed Random Channel Drop augmentation has served its purpose of making the model more robust to missing sequences.

4.3.2 Prostate zonal segmentation

Regarding the prostate zonal segmentation model, which was developed with the sole purpose of automating the PCa detection system (view Section 4.5.2.2), the results for all datasets can be found in Table 4.5, with mean Sørensen-Dice similarity coefficient (DSC) ranging from 0.894 to 0.941. DSC is a metric between 0 and 1, employed to assess the relative overlap between predicted and ground truth (GT) segmentations. Some qualitative results for this segmentation model can be seen in Figures 4.2, 4.4, and 4.6.

Table 4.4: Area under the ROC curve after dropping one (or several) particular sequences (i.e.: setting the value to 0) in test time for the Gleason Grade Group ≥2 significance criterion.

MRI sequence dropped	*ProstateX*		*IVO*	
	Lesion	Patient	Lesion	Patient
None (Baseline)	0.959	0.865	0.945	0.910
b400/500	0.944	0.861	0.940	0.868
b800/1000/1400	0.946	0.873	0.895	0.783
All b-numbers	0.951	0.844	0.845	0.720
ADC	0.905	0.870	0.940	0.836
K^{trans}	0.894	0.865	-	-
All DCE	-	-	0.895	0.820
All but T2	0.804	0.808	0.782	0.545

Table 4.5: Results for the prostate zonal segmentation model. DSC: Sørensen-Dice similarity coefficient, CG: Central Gland, PZ: Peripheral Zone

Dataset	*N*	*Mean DSC*		
		Prost.	*CG*	*PZ*
Private train	80	0.941	0.935	0.866
Private test	12	0.915	0.915	0.833
NCI-ISBI train	60	0.894	0.860	0.690

4.4 Discussion

Despite mpMRI interpretation being time-consuming and observer-dependent, it is a major clinical decision driver and poses great clinical relevance. In this paper we presented a CAD system developed with two main MRI datasets integrating T2, DW, b-value, and ADC maps in both of them as well as K^{trans} for ProstateX and DCE for the IVO dataset. These were compared against fusion and transperineal template biopsies, which is considered the pre-operative gold standard to evaluate prostate cancer extent (Drost et al., 2019).

Different outcomes can be measured for this system. Regarding lesion detection as exposed in Section 4.3.1.1, the results for lesions GGG≥2 significance criterion can be considered optimal: all csPCa lesions were detected while maintaining a very high specificity, except for the patient-level ProstateX evaluation, and a great AUC ranging from 0.865 to 0.959. Furthermore, the IVO results outperform the PI-RADS scores, especially at the high sensitivity setting (PI-RADS≥ 4) which is of most interest in clinical practice. This can be seen in Figure 4.1, where the ROC is at all instances above and to the left of the PI-RADS scores. For further comparison, several studies have reported radiologist sensitivities/specificities for the detection of csPCa from mpMRI at a patient level of 0.93/0.41 (Ahmed et al., 2017), or 0.58-0.96/0.23-87 as shown in a systematic review (Litjens et al., 2015). The results vary wildly due to their single-center nature, their differing criteria for the definition of csPCa, and the often-inaccurate reference standards employed.

Considering GGG≥3 significance criterion, caution is required when interpreting these results due to the very low number of positive cases (e.g.: only three in the IVO test set). Furthermore, the 0.714 patient-level sensitivity does not mean that the model missed GGG3 lesions, but rather that they were assigned to a lower GGG (such as GGG2) and were therefore ignored for the GGG≥3 classification problem.

In addition to the previous tests, the ongoing ProstateX challenge was used for external lesion-level validation, achieving an AUC of 0.85, which would have been the second-best AUC in the original ProstateX challenge (Armato et al., 2018). Additionally, an identical model trained only on the ProstateX data (which has been made publicly available alongside this paper), achieved an AUC of 0.87, which would have tied with the best contender in the challenge. There are now higher AUCs in the online leaderboard but, unfortunately, we were unable to find any publications regarding them, and hence no further analysis can be performed. In any case, these results must be also interpreted with caution: on one hand, the proposed system solves a much more complex problem (namely detection, segmentation & classification) than the comparatively simpler ROI classification systems which are typically employed for this task, and it is therefore in a disadvantage compared to them. On the other hand, as indicated in Section 4.5.1, the ProstateX challenge mpMRIs were used for training the segmentation and detection components of the model, but not the classification head (as GGG information is kept secret by the challenge, and hence unavailable for training). The inclusion of this data was useful for increasing the number of training samples, but it might have introduced some unknown bias for the evaluation of this dataset.

Outside the ProstateX challenge, one of the very first works on the topic by Litjens et al. (2017) reported a sensitivity of 0.42, 0.75, and 0.89 at 0.1, 1, and 10 false positives per normal case using a classical radiomics-based model. More recently, Xu et al. (2019) used a csPCa segmentation CNN whose output was later matched to GT lesions based on distance (similar to ours). He reported a sensitivity of 0.826 at some unknown specificity; also, despite using the ProstateX data, unfortunately, no ProstateX challenge results were provided. Cao et al. (2019) proposed a segmentation CNN that also included GGG classification as part of its output, reporting a maximum sensitivity of 0.893 at 4.64 false positives per patient and an AUC of 0.79 for GGG≥2 prediction. Interestingly, the authors employed histopathology examinations of whole-mount specimens as GT for the model. Aldoj et al. (2020b) utilized the ProstateX data to perform csPCa classification on mpMRI ROIs around the provided lesion positions, reporting an AUC of 0.91 on their internal 25-patient test set; once again, despite using the ProstateX data exactly as conceived for the challenge, they do not provide any challenge results for comparison.

In an interesting prospective validation study, Schelb et al. (2020) obtained a sensitivity/specificity of 0.99/0.24 using a segmentation CNN, a performance that they found comparable to radiologist-derived PI-RADS scores. Woźnicki et al. (2020) proposed a classical radiomics-based model (no CNNs involved) achieving an AUC of 0.807. As for patient-level csPCa classification results, Yoo et al. (2019) achieved an AUC of 0.84 using slice-wise CNN classifier whose predictions were later combined into a patient-wise total score and Winkel et al. (2020) achieved a sensitivity/specificity of 0.87/0.50 on a prospective validation study using a segmentation-based detection system which is most similar to the one proposed here.

Considering lesion segmentation concordance, as exposed in Section 4.3.1.1, our results are unfortunately not directly comparable to other papers in the literature (as those focus on segmenting exclusively csPCa and benign lesions are ignored) and were mostly added for completeness. For instance, Schelb et al. (2020) reported a DSC of 0.34 for csPCa segmentation, similar to Vente et al. (2021)'s 0.37 DSC. Secondly, the reference segmentations for the ProstateX dataset were generated in an automatic manner; hence, the performance for this dataset is not compared against a proper ground truth. Thirdly, mpMRI lesions tend to be small with ill-defined margins and a very high inter-observer variability (Steenbergen et al., 2015). For all these reasons, these relatively low DSC metrics must be interpreted with caution. Instead, the previously discussed metrics provide a more objective outlook on the actual performance of the model.

With respect to the ablation tests, there is an ongoing debate regarding the need for DCE sequences. Bi-parametric MRI (bpMRI) (without DCE sequences) seems to be a more cost- and time-effective alternative to mpMRI, with little detriment to accuracy (Junker et al., 2019; Zawaideh et al., 2020). Likewise, the role of DCE sequences is currently minor in the final score of the PI-RADS system, being used only in peripheral zone regions with value 3 in the DW sequence (which rises to 4 if an early focal uptake is detected in DCE sequences). Conversely, the results of the present study hint towards a greater importance of DCE sequences, which turned out to be the second most important sequences for the model, only behind b-numbers (T2 does not count as it was always included).

Lastly, regarding prostate zonal segmentation, we observed a great concordance between the model's and expert radiologist's prostate segmentation with a DSC that ranged from 0.894 to 0.941 depending on the MRI dataset. As can be seen, the results in the Private test set are extremely good, better in fact than any other model in the literature when evaluated in its internal test set and when evaluated blindly in the NCI-ISBI dataset. In Qin et al. (2020), for instance, the authors train one CNN on an internal dataset and another identical CNN on the NCI-ISBI train dataset independently, and evaluate them by cross-validation, achieving a DSC of 0.908 and 0.785 at the Central Gland (CG) and Peripheral Zone (PZ) in their internal dataset, and a DSC of 0.901 and 0.806 in the NCI-ISBI dataset. For a fairer comparison with our model, in Rundo et al. (2019), the authors train their model on two internal datasets (achieving a DSC of 0.829/0.859 in CG segmentation, and 0.906/0.829 in PZ segmentation), which then test blindly in the NCI-ISBI dataset, achieving 0.811 and 0.551 in CG and PZ segmentation, respectively. Finally, Aldoj et al. (2020a), training on a larger cohort of 141 patients and evaluating in their internal test set of 47, achieved a DSC of 0.921, 0.895, and 0.781 for whole gland, CG, and PZ segmentation.

The interpretation of mpMRIs based on Artificial Intelligence (AI) represents a very promising line of research that has already been successfully applied to prostate gland segmentation and PCa lesion detection using both transperineal prostate biopsy and radical prostatectomy specimens as GT with varying results (Yoo et al., 2019; Winkel et al., 2020). We went a step further and developed the first algorithm, to the best of our knowledge, that automatically contours the prostate into its zones, performs well at lesion detection and Gleason Grade prediction (identifying lesions of a given grade or higher), and segments such lesions albeit with a moderate overlapping. The model outperformed expert radiologists with extensive MRI experience and achieved top results in the ProstateX challenge.

The code has been made publicly available, including an automatic prostate mpMRI non-rigid registration algorithm and an automatic mpMRI lesion segmentation model. Most importantly, the fact that the code is online might allow future researchers to use this model as a reference upon which to build or to compare their models.

Our work presents some limitations. Firstly, further validation and prospective blinded trial would be required to compare histological results of targeted biopsies to the lesions identified by the model. Secondly, although the model was successfully trained on two datasets, it still behaves differently on each of them (e.g.: the optimal thresholds vary significantly between them), which is not desirable, but probably unavoidable. Obviously, more data from sources as varied as possible would be ideal to overcome such difficulties and further improve the performance and generality of the model. Thirdly, AI systems have proven cumbersome to integrate into clinical practice for a variety of reasons (costs, rejection, etc.); we hope that by making the code freely available some of these obstacles can be more easily overcome.

In any case, this is yet another step in the foreseeable direction of developing a strong collaborative AI net that progressively incorporates as many mpMRIs with the corresponding GT as possible. The clinical applications of this model are countless, amongst which we could consider assisting radiologists by speeding up prostate segmentation, training purposes as well as a safety net to avoid missing PCa lesions. Further, the ability to detect csPCa can easily highlight which MRIs would require prompt reporting and prioritizing biopsy. Moreover, given the recent trend towards conservative PCa approaches such as focal therapy or active surveillance (usually implying a more dedicated prostate biopsy), predicting the Gleason Grade, as well as the number of lesions pre-biopsy, could identify eligible men that could be offered transperineal targeted biopsy in the first place.

4.5 Materials and Methods

4.5.1 Data description

For the development and validation of the model, two main prostate mpMRI datasets were employed: ProstateX (Litjens et al., 2014a), which is part of an ongoing online challenge at https://prostatex.grand-challenge.org and is freely available for download (Litjens et al., 2017); and IVO, from the homonymous Valencian Institute of Oncology. The study was approved by the Ethical Committee of the Valencian

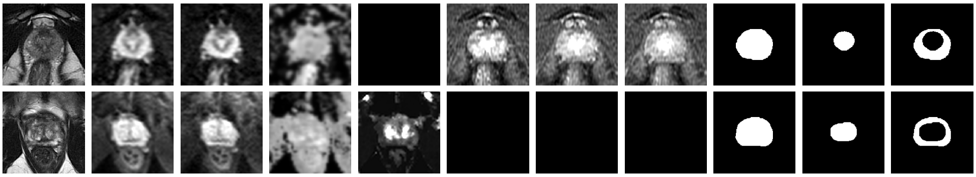


Figure 4.3: Final pre-processed image from a single patient (top: IVO, bottom: ProstateX). Channels (from left to right): T2, b400/b500, b800/b1000/b1400, ADC, K^{trans}, DCE $t = 10$, DCE $t = 20$, DCE $t = 30$, prostate mask, CG mask and PZ mask.

Institute of Oncology (CEIm-FIVO) with protocol code PROSTATEDL (2019-12) and date 17th of July, 2019. All experiments were performed in accordance with relevant guidelines and regulations. Informed consent was obtained from all participants and/or their legal guardians.

For ProstateX, the data consisted of a total of 204 mpMRIs (one per patient) including the following sequences: T2-weighted (T2), diffusion-weighted (DW) with b-values b50, b400, and b800 s/mm^2, apparent diffusion coefficient (ADC) map (calculated from the b-values), and K^{trans} (computed from dynamic contrast-enhanced -DCE- T1-weighted series). For each of these patients, one to four (1.62 per patient on average) lesion locations (i.e.: a point marking their position) and their GGG are provided (GGG is provided as part of the ProstateX2 challenge, which shares the same data with ProstateX). The lesion locations were reported by or under the supervision of an expert radiologist with more than 20 years of experience in prostate MR and confirmed by MR-guided biopsy. Furthermore, 140 additional mpMRIs are provided as part of the challenge set, including all previous information except for the GGG of the lesions. All mpMRIs were acquired by two different Siemens 3-Tesla scanners.

For IVO, there were a total of 221 mpMRIs, including the following sequences: T2, DW with b-values b100, b500, and b1000 s/mm^2 (in 1.36% of the cases, b1400 was available, instead of b1000), ADC (4.52% missing) and a temporal series of 30 DCE T1-weighted images (42.53% missing). For each mpMRI, one to two (1.04 per patient) lesions were segmented by one of several radiologists with two to seven years of experience in PCa imaging, and their PI-RADS were provided. The Gleason Score (GS) (Epstein et al., 2005) was assessed by transperineal fusion-guided with two to three cylinders directed to each of the ROIs. Additionally all patients underwent systematic template biopsy comprising 20-30 cylinders to sample the rest of the prostate.

Four PCa classes were considered: GGG0 or benign (57.32% of all lesions), GGG1 (GS 3+3, 17.28%), GGG2 (GS 3+4, 12.70%), and GGG3+ (GS $\geq$ 4+3, 12.70%); therefore, lesions of GGG$\geq$3 were grouped into a single category to try to balance the classes, and also because the protocol for a suspect GGG 3+ lesion would be similar irrespective of its specific grade (i.e.: the lesion would be biopsied for confirmation).

4.5.2 Pre-processing

After collecting them, mpMRIs had to be pre-processed to accomplish three main objectives, namely: (1) homogenize differences within datasets, (2) homogenize differences between datasets, and (3) enrich the images with extra information that might be useful for the model. Additionally, the preprocessing pipeline was designed to require as little human intervention as possible, in pursuit of developing a system easily implementable in clinical practice.

For the first objective, all images were cropped to an ROI around the prostate of size 160 × 160 × 24 voxels with a spacing of (0.5, 0.5, 3)mm, which corresponds with the median (and mode) spacing of the T2 sequences for both datasets. The rest of the sequences were applied the same processing for the sake of homogeneity. B-Spline interpolation of third order was employed for all image interpolation tasks, while Gaussian label interpolation was used for the segmentation masks. For the IVO dataset, the time series of 30 DCE images per patient was sampled at times 10, 20, and 30, approximately coinciding with the peak, progression, and decay of the contrast agent. Then, all sequences were combined into a single multi-channel image, in which any missing sequences were left blanks (value of 0), such as the three DCE channels in every ProstateX image, or the K^{trans} channel in every IVO image. The intensity was normalized by applying Equation 4.1 to every channel of an image I independently, as introduced in Pellicer-Valero et al. (2021).

$$I_{new} = \frac{I - percentile(I, 1)}{percentile(I, 99) - percentile(I, 1)} \tag{4.1}$$

Regarding objective (2), the procedure for homogenizing lesion representations between datasets is described in Section 4.5.2.1, and a special data augmentation employed to alleviate the problem of missing sequences is presented in Section 4.5.3.3. Additionally, sequences b500 (from IVO) and b400 (from ProstateX) were considered similar enough to conform to the same channel in the final image; likewise, sequences b1000/b1400 (from IVO) and b800 (from ProstateX) were assigned to a single common channel too.

Concerning objective (3), Section 4.5.2.2 argues that prostate zonal segmentation is an important input for PCa assessment and describes the conception of a model for producing such segmentations automatically. Additionally, DW and ADC sequences were found to be misaligned to the rest of the sequences in several patients; hence an automated registration step was added, which is presented in Section 4.5.2.3.

Figure 4.3 shows the channels of one image from each dataset after all the mentioned pre-processing steps.

4.5.2.1 Automated lesion growing

To enable training a single model on both datasets, it was mandatory to homogenize how lesion information was to be provided to the model: while the IVO dataset

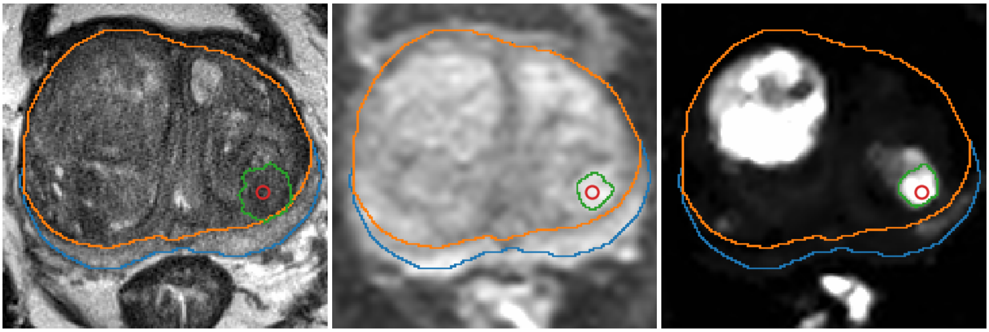


Figure 4.4: Automatic lesion segmentation for a ProstateX patient in sequences (from left to right: T2, b800 and K^{trans}) before combining them. Prostate zonal segmentation and the original lesion position (in red) are shown for reference.

provided the full segmentation mask for each lesion, in ProstateX only the center position of the lesion was available. Although detection systems can be adapted to detect positions, they are typically designed to work with much more semantically rich BBs (He et al., 2017), or segmentations, or both (Ren et al., 2017).

To solve this inconsistency between the datasets, a similar approach to Liu et al. (2019) was employed: for the ProstateX dataset, lesions were automatically segmented by growing them from the provided image position (used as seed), using a threshold level set method from Python library SimpleITK (Yaniv et al., 2018). Concretely, the algorithm was applied independently to sequences T2, b800, and K^{trans}, and all segmented areas present in at least two of these three sequences were kept. Figure 4.4 shows the process of applying this segmentation algorithm to one image. This figure (and several others in this paper) were generated using Python library plot_lib (Pellicer-Valero, 2020).

4.5.2.2 Automated prostate zonal segmentation

Following McNeal's criterion (Selman, 2011), the prostate is typically partitioned into two distinct zones: the Central Gland (CG, including both the transition zone and the central zone, which are difficult to distinguish) and the Peripheral Zone (PZ). PCa lesions vary in frequency and malignancy depending on the zone (Haffner et al., 2009) and, as such, PI-RADS v2 considers them when assessing mpMRIs (Weinreb et al., 2016). Therefore, just like a radiologist, a model for automated PCa detection and classification will likely benefit from having both CG and PZ mask priors provided as inputs, in addition to the mpMRI.

Accordingly, a cascading system of two segmentation CNNs, similar to the one introduced by Zhu et al. (2019b), was developed for automatic CG and PZ segmentation. As it can be seen in Figure 4.5, the first CNN -a published model (Pellicer-Valero et al., 2021) based on the U-Net (Ronneberger et al., 2015) CNN architecture with dense (Huang et al., 2017) and residual (He et al., 2016) blocks-, takes a prostate

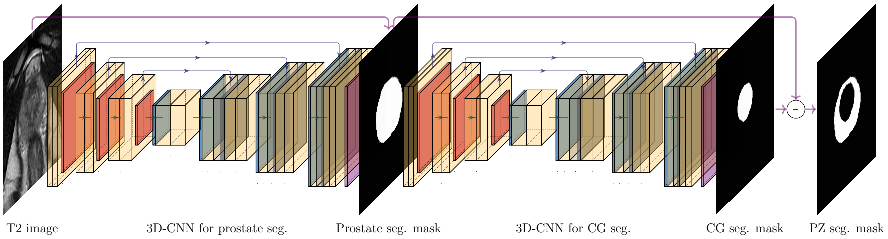


Figure 4.5: Cascading three-dimensional (3D) convolutional neural networks (CNNs) for prostate central gland (CG) and peripheral zone (PZ) segmentation. The first 3D-CNN takes a T2 sequence as input and produces a prostate segmentation mask as output, while the second 3D-CNN takes both the T2 sequence and the prostate segmentation (generated by the previous CNN) as inputs to produce the CG segmentation mask as output. Finally, PZ is computed by subtraction of both output masks.

T2 image as input and produces a prostate segmentation mask as output. Then, the second CNN takes both the T2 image and the prostate segmentation mask obtained in the previous step and generates a CG segmentation mask as output. Finally, the PZ segmentation mask can be computed by subtracting the CG from the prostate segmentation mask.

The second CNN employed an architecture identical to the first one but was retrained on 92 prostate T2 images from a private dataset, in which the CG was manually segmented by a radiologist with two years of experience in PCa imaging. To be more precise, 80 of the 92 images were used for training the CG segmentation model, while the remaining 12 were employed for testing. Additionally, this model was also blindly tested (i.e.: with no retraining or adaptation of any kind) against the NCI-ISBI (N et al., 2015) train dataset, which is freely available at `http://doi.org/10.7937/K9/TCIA.2015.zF0vlOPv`. The results of this prostate zonal segmentation model are very briefly analyzed and compared to others in Section 4.3.2. Once trained and validated, this model was employed to obtain the CG and PZ masks of all the prostates in the current study.

4.5.2.3 Automated sequence registration

In several patients, DW sequences and the ADC map were misaligned to T2 and the other sequences. As a solution, non-rigid registration (based on a BSpline transformation) was applied between the spatial gradient of the T2 and the ADC map using Python library SimpleITK (Yaniv et al., 2018), with Mattes Mutual Information (Mattes et al., 2001) as loss function and gradient descent (Ruder, 2016) as the optimizer for the BSpline parameters. For every mpMRI, the registration algorithm was run 50 times with different parameter initializations, and the correlation coefficient between the spatial gradient of the T2 sequence and the spatial gradient of the registered ADC map was evaluated at the CG and the PZ areas. These custom metrics allowed to place a bigger emphasis to the areas of interest, as compared to image-wide

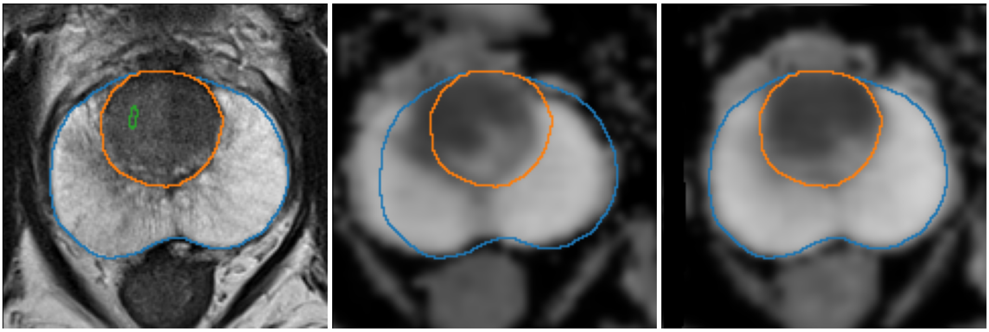


Figure 4.6: Automatic registration between T2 sequence (left) and ADC map (center: before, right: after) for a sample mpMRI.

metrics. Finally, the transformation associated with the run yielding the highest value for the average of all metrics and the loss was chosen as final and applied to both DW and ADC sequences. Figure 4.6 shows the result of applying this procedure to one mpMRI.

4.5.3 Model training and validation

After pre-processing the data, it was used to train a Retina U-Net (Jaeger et al., 2020) CNN architecture, which allows for the simultaneous detection, segmentation, and classification of PCa lesions. Section 4.5.3.1 provides an overview of this architecture, while Sections 4.5.3.2-4.5.3.5 deal with all engineering decisions related to the model training, validation, and testing.

4.5.3.1 Architecture: Retina U-Net

The Retina U-Net (Jaeger et al., 2020) architecture combines the Retina Net (Lin et al., 2017b) detector with the U-Net segmentation CNN and is specifically designed for application to medical images. On one hand, Retina Net is a one-shot detector, meaning that classification and BB refinement (regression) are directly performed using the intermediate activation maps from the output of each decoder block in the Feature Pyramid Network (FPN) that conforms its backbone (Lin et al., 2017a), making it not only more efficient but also better suited for lesion detection in medical images, which have distinct characteristics compared to natural images (e.g.: there is no overlap between detections).

Furthermore, in the Retina U-Net, the FPN has been extended with two more high-resolution pyramid levels leading to a final segmentation layer, hence making the extended FPN architecture extremely akin to that of the U-Net. Therefore, the lesions are segmented independently of the detections (unlike other similar detection+segmentation architectures, such as Mask R-CNN, He et al. (2017)). This simplifies the architecture significantly, while still being a sensible choice for segmenting lesions since they all represent a single entity irrespective of their particular classes. Figure 4.7 shows an overview of the Retina U-net architecture applied to the problem of simultaneous PCa detection, classification, and segmentation.

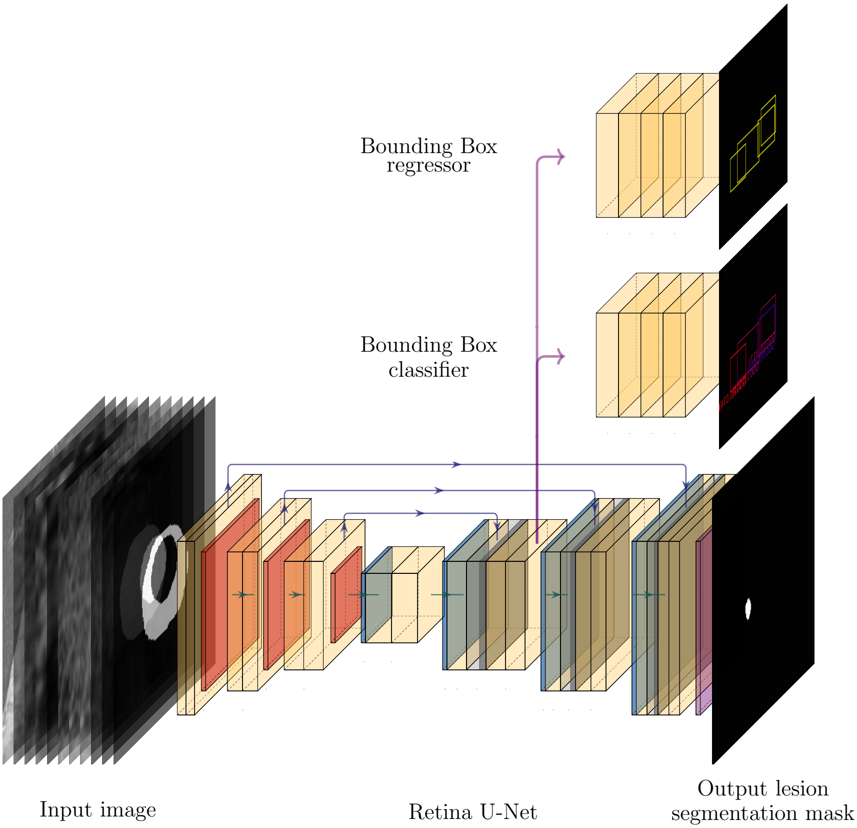


Figure 4.7: Overview of the Retina U-Net architecture. On the bottom, a U-Net-like architecture segments the lesions present in the image irrespective of their class. On the top, a bounding box (BB) regression head takes a feature map from a decoder of the U-Net and refines the coarse detections, while the BB classifier tries to predict their class. These two heads visit all decoder levels, performing detection at different scales transparently.

4.5.3.2 Hyperparameters

An ensemble of five CNNs (see Section 4.5.3.5) was trained with the ResNet101-like backbone (He et al., 2016) with batch normalization (Ioffe and Szegedy, 2015) and a batch size of 6, at 120 batches per epoch, for a total of 115 epochs. Please, refer to Section 4.5.3.4 for more information on how data was split for training and validating the model. A triangular cyclical learning rate (LR) with exponential decay was employed (Smith, 2017), with LRs oscillating between a minimum of $8 \cdot 10^{-5}$ and a maximum of $3.5 \cdot 10^{-4}$. For the BBs, a single aspect ratio of 1 (before BB refinement) was considered sufficient, with scales ranging from $4 \times 4 \times 1$ voxels (i.e.: $2 \times 2 \ \times 3$ mm), all the way to $28 \times 28 \times 9$ voxels (i.e.: $14 \times 14 \times 27$ mm), depending on the pyramid level on which the detection was performed. The rest of the parameters were left at their default values (Jaeger et al., 2020).

In particular, the encoder was a ResNet101-like CNN with the highest-resolution pyramid levels (P_0 and P_1) consisting of a single convolution, and the rest ($P_2, ..., P_5$) consisting of $[3, 7, 21, 3]$ residual blocks, respectively. The stride of the last convolution of each pyramid level $P_0, ..., P_5$ was set to $[1, 2, 2, 2, 2, 2]$, respectively for the x and y dimensions of the feature maps, and to $[1, 1, 1, 2, 2, 2]$ for the z dimension, to account for the non-uniform voxel spacing. The decoder consisted in a single convolution per pyramid level followed by a simple upsampling; feature maps from the skip connections were merged with the upsampled feature maps by addition. Both the BB regressor head and the classifier head consisted of a stack of five convolutions. Convolution kernels were all of size 3×3 and *relu* non-linearity was used as activation function.

4.5.3.3 Online data augmentation

To help with regularization and to expand the limited training data, extensive online 3D data augmentation was employed during training using the Python library Batchgenerators (Fabian et al., 2020). Both rigid and non-rigid transformations, such as scaling, rotations, and elastic deformations were used.

Additionally, a custom augmentation was included to help deal with the issue of missing sequences, either because they never existed (such as K^{trans} images in the IVO dataset), or because they were not available. This augmentation, named Random Channel Drop, consisted in setting any given channel to zero (blanking it) with a certain probability, hence accustoming the model to dealing with missing data. During training, every channel of every image had a 7.5% probability of being dropped, except for the T2 channel and the segmentation masks, which had a probability of 0% (since they are assumed to be always available). The three DCE channels were considered as a whole for the purposes of dropping them (i.e.: they could not be dropped independently of each other).

4.5.3.4 Data partitioning

The mpMRIs were split into two sets: the train/validation set and the test set. The test set only contained "complete" mpMRIs (with no missing sequences), amounting to 30 IVO patients (23.62% of all complete IVO patients) and 45 ProstateX patients (22.17% of all ProstateX patients). This set was kept secret during the development of the model and was only employed eventually to validate it. Instead, for internal validation, five-fold cross-validation (CV) was employed: the train/validation set was split into five disjoint subsets, and five different instances of the same Retina U-Net model were successively trained on four out of the five subsets and validated on the fifth, hence creating a virtual validation dataset that encompassed the totality of the training data (but not the test data, which were kept apart).

As mentioned in Section 4.5.1, there was an additional ProstateX challenge set containing 140 mpMRIs with all the same information as the training set, except for the lesion GGG, which was not available. Hence, this dataset could also be employed for training both the segmentation and the BB regressor components of the Retina U-Net (but not the classifier). As such, this dataset was included as part of the training set (but not in the validation sets, as it contained no GT class information), and the classifier had to be modified to ignore any detection belonging to this dataset (i.e.: the loss was not propagated from such detections).

In summary, the model was trained and five-fold cross-validated with 191 IVO patients (of which only 45.55% were complete) + 159 ProstateX patients (all complete) + 140 ProstateX test patients (those coming from the ProstateX challenge set, for which GGG class information was not available). For testing, a secret subset consisting of 30 IVO patients and 45 ProstateX patients (all complete) was employed. The model was also tested on the ongoing ProstateX challenge.

4.5.3.5 Epoch and CV ensembling during testing

During the final test set prediction, both epoch and CV ensembling were used to boost the capabilities of the model. In general, ensembling consists in training N models for the same task, using them to predict on a given test set, and then combining all N predictions to achieve a better joint performance than that of each model individually. Hence, the five CV models were used for ensembling and, additionally, for every one of these CV models, the weights from the best (i.e.: highest validation mean -over all classes- Average Precision) five epochs were used as further independent models, totaling an equivalent of 25 virtual models.

Then, the predictions from the ensemble on the test set were combined in the following way: for segmentation masks, the average mask (over all 25 proposals) was computed and, for the BBs, the weighted box clustering (WBC) algorithm with an Intersection over Union threshold of $1 \cdot 10^{-5}$ was applied to each class independently. The WBC algorithm is described in the original Retina U-Net paper (Jaeger et al., 2020).

4.5.4 Lesion matching and evaluation

The results were evaluated at three lesion significance thresholds (GGG≥1, GGG≥2, and GGG≥3) and two levels: lesion-level and patient-level. Only predicted BBs with a predicted GGG equal or above the chosen significance threshold (e.g.: GGG≥2) were considered, and the rest were completely ignored.

For lesion-level evaluation, each of the GT lesions was first matched with one (or none) of the detected lesions. First, all predicted BBs whose centroid was less than

15mm away from that of the GT BB were selected as candidates for matching, and assigned a matching score computed as $\widehat{p} + k \cdot (1 - d/15mm)$, where $\widehat{p}$ represents the actual score given by the model to that detection, d is the distance between the GT BB centroid and the candidate BB centroid, and $k = 2$. That way, both the model confidence ($\widehat{p}$) and distance to the GT (d) were considered for matching. The parameters for this matching procedure (e.g.: $k = 2$, 15mm) were adjusted directly on the training set. If no detections existed within a 15mm radius of a GT BB, a score of 0 was assigned to it. This evaluation method measures the performance of the model only on GT lesions for which biopsy confirmation and GGG are available, without assuming anything about the rest of the prostate, which may or may not contain other lesions. Furthermore, it allows the model to compete in the online ProstateX challenge (despite it not being an ROI classification model) since it can assign a score to every GT lesion.

For patient-level evaluation, the patient score was computed as the highest score from any BB predicted for the patient, and the GT GGG of a patient was computed as the highest GGG among all his GT lesions and among all the 20-30 cylinders obtained in the systematic biopsy (which were only available for patients from the IVO dataset). Hence, for the IVO dataset, a patient without any significant GT lesions might still have csPCa; for ProstateX, however, we do not know, and we must assume that this does not happen.

Author contributions

Conceptualization, JDM; Data curation, OJP and VG; Methodology, OJP; Project administration, JDM; Resources, VGP, IG and MB; Software, OJP; Supervision, JLR, JR, MR and JDM; Validation, IG, MB and PG; Visualization, OJP; Writing – original draft, OJP, JLMJ; Writing – review & editing, VG, MJR and JDM.

Funding

This work has been partially supported by a doctoral grant of the Spanish Ministry of Innovation and Science, with reference FPU17/01993.

Data availability

Data from the ProstateX challenge are available at https://doi.org/10.7937/K9TCIA.2017.MURS5CL (Litjens et al., 2017); data from the Valencian Institute of Oncology is not publicly available, since the ethical committee (CEIm-FIVO) only approved its use for the current study. They might be made available for research purposes on reasonable request from the corresponding author. The code of the project is available at https://github.com/OscarPellicer/prostate_lesion_detection.

Competing Interests

The authors declare no competing interests.

Chapter 5

Real-time Biomechanical Modeling of the Liver using Machine Learning Models trained on Finite Element Method Simulations

Reference: Oscar J. Pellicer-Valero, María José Rupérez, Sandra Martínez-Sanchis, and José D. Martín-Guerrero. Real-time biomechanical modeling of the liver using Machine Learning models trained on Finite Element Method simulations. *Expert Syst. Appl.*, 143: 113083, apr 2020b. ISSN 09574174. doi: 10.1016/j.eswa.2019.113083. URL https://linkinghub.elsevier.com/retrieve/pii/S0957417419308000

Oscar J. Pellicer-Valero*[1], María José Rupérez[2], Sandra Martínez-Sanchis[2], José D. Martín-Guerrero[1]

[1] Intelligent Data Analysis Laboratory, Department of Electronic Engineering, ETSE (Engineering School), Universitat de València (UV), Av. Universitat, sn, 46100 Burjassot, València, Spain.
Oscar.Pellicer@uv.es, jose.d.martin@uv.es

[2] Centro de Investigación en Ingeniería Mecánica (CIIM), Universitat Politècnica de València (UPV), Camino de Vera, sn, 46022 València, Spain.
mjrupere@upvnet.upv.es, sanmars1@upv.es

5.1 Abstract

The development of accurate real-time models of the biomechanical behavior of different organs and tissues still poses a challenge in the field of biomechanical engineering. In the case of the liver, specifically, such a model would constitute a great leap forward in the implementation of complex applications such as surgical simulators, computed-assisted surgery or guided tumor irradiation.

In this work, a relatively novel approach for developing such a model is presented. It consists in the use of a machine learning algorithm, which provides real-time inference, trained on tens of thousands of simulations of the biomechanical behavior of the liver carried out by the finite element method on more than 100 different liver geometries.

Considering a target accuracy threshold of $3mm$ for the Euclidean Error, four different scenarios were modeled and assessed: a single liver with an arbitrary force applied (99.96% of samples within the accepted error range), a single liver with two simultaneous forces applied (99.84% samples in range), a single liver with different material properties and an arbitrary force applied (98.46% samples in range), and a much more general model capable of modeling the behavior of any liver with an arbitrary force applied (99.01% samples in range for the median liver).

The results show that the Machine Learning models perform extremely well on all the scenarios, managing to keep the Mean Euclidean Error under $1mm$ in all cases. Furthermore, the proposed model achieves working frequencies above $100Hz$ on modest hardware (with frequencies above $1000Hz$ being easily achievable on more powerful GPUs) thus fulfilling the real-time requirements. These results constitute a remarkable improvement in this field and may involve a prompt implementation in clinical practice.

Keywords: Machine Learning; Finite Element Method; Real Time; Liver; Coherent Point Drift; Biomechanical Modeling

5.2 Introduction

In the last few years, the field of biomechanical engineering has undergone a continued growth as many new technology-driven applications are being developed and introduced in the clinical practice. However, several specific applications, such as Computed Assisted Surgery (CAS), surgical simulators with haptic feedback or directed tumor irradiation (as in gating), all share a requirement for precise and real time biomechanical models of the organ they must interact with, a subject that remains as one of the biggest challenges in biomechanics.

In the case of CAS, the liver is of special interest, since it moves significantly during the respiratory cycle. High precision techniques such as biopsies, tumor ablation, cryotherapy, brachitherapy, tumor embolization, directed irradiation (gating), or vector delivery for genetic therapy (Clifford et al., 2002) could all benefit from a biomechanical model of the liver able to assist clinicians during these procedures.

The first biomechanical models able to work in real time were based on mass-spring simulations (Nedel and Thalmann, 1998; Duysak et al., 2003). Despite their speed, these have been progressively abandoned due to their inability to accurately model the nonlinearities which characterize biological tissue.

Models based on the Finite Elements Method (FEM), on the contrary, have a very well established mechanical and mathematical base, and allow for high accuracy simulations for any kind of geometry or material. However, the increased accuracy comes at the cost of prohibitive computational times, which hinder the application of these methods for real-time systems.

In order to accelerate these biomechanical simulations, there are two main techniques that stand out in the literature, the first trying to exploit the parallelism of the problem with Graphic Processing Units (GPUs) or Central Processing Unit (CPU) clusters (Faure et al., 2012; Peterlík et al., 2012), and the second attempting dimensionality reduction techniques by means of algorithms like Proper Generalized Decomposition (PGD); this algorithm tries to only solve the few most important deformation modes, allowing for very accurate real-time results in many problems where the field evolves smoothly (Chinesta et al., 2013).

In distributed or GPU models, the price to pay in exchange for real-time simulations is the use of very coarse meshes, which lack the required accuracy for most applications (Faure et al., 2012). In PGD-based models, the nonlinearities of hyperelastic materials, which normally describe the mechanical behavior of the soft biological tissue (Fung and Skalak, 1981), may limit the applicability due to the increased number of deformation modes required for an accurate simulation.

Other radically different approaches are the data-based strategies, which consist in training a Machine Learning (ML) model from simulations (e.g.: obtained from FEM) or directly from sensor data. ML algorithms are able to automatically learn nonlinear mappings between several inputs (applied force, application area, node coordinates, etc.), and several outputs (e.g.: displacement, strain or stress fields). Although the training process is relatively slow, once trained, these algorithms provide extremely quick inference times, therefore fulfilling the requirement for real-time simulations. This strategy has been successfully applied in the literature on several organs (Jahya et al., 2013; Deo and De, 2009), being the work presented in (Morooka et al., 2008) the first where this approach was applied to the liver. In all instances, however, the employed meshes were rather coarse, and also, only one liver geometry was considered. Thus, in order to apply this procedure to any other liver, all the FEM simulations should be repeated on any new geometry, and a new ML model should be trained on the results, both processes being very time consuming.

As an exception, the work presented in (Lorente et al., 2017) proposed a ML model that was validated on geometries from different livers. Here, though, the limitation stemmed from the fact that only a very reduced displacement set was considered.

In sight of the current state of the art, the main objective of this work is to develop a data-based model able to simulate the biomechanical behavior of any liver, subjected to any force, with sufficient accuracy, and in real time.

On one hand, for the definition of sufficient accuracy, a $3mm$ threshold was set for the Euclidean error (Lorente et al., 2017), which is considered clinically acceptable for tumor gating. On the other hand, real-time operation implies that the simulation is able to run at least at $25Hz$, while for haptic feedback applications, real time is considered only for frequencies higher than $300Hz$ (Cotin et al., 2000).

In addition to the main ML model (applicable to any liver), three more ML models will be developed in order to show that this procedure could also be used for very high precision scenarios, multiple force interactions, or even scenarios where the material properties of the liver are variable.

A remarkable contribution included in this work is the development of an algorithm able to provide a natural parametrization of the geometry of any liver in only a few variables. Moreover, a simple yet powerful modification to the Coherent Point Drift (CPD) algorithm, which significantly improves its performance when the registered geometries differ substantially, is employed.

The work layout is as follows. First, Section 5.3 introduces the relevant details regarding the research development, focusing on data sources, data processing, FEM simulations and ML model training. Then, in Section 5.4, the results are presented and discussed. Finally, Section 5.5 summarizes the main conclusions drawn from the work, and suggests further lines for research.

5.3 Experimental setup

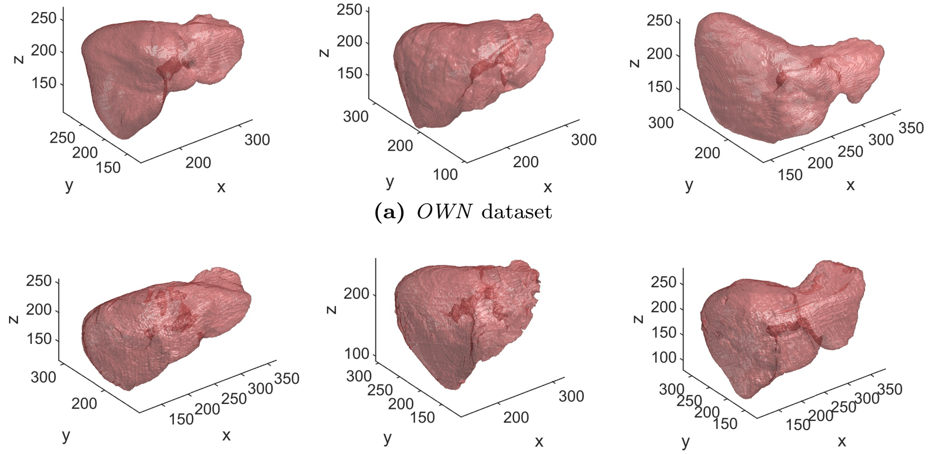
5.3.1 Data acquisition

The Computed Tomography (CT) images used in this project came from two main sources. The first image set (OWN) was provided by Hospital La Fe, in València, and consists of a total of 24 abdominal CT images with their respective segmentation mask (a binary 3D image) for the liver (Figure 5.1a).

The second set ($LITS$) comes from the 2017 Liver Tumor Segmentation Challenge (LiTS)(Christ, 2017). This was a competition organized by the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2017) in conjunction with the IEEE International Symposium on Biomedical Imaging (ISBI 2017) whose objectives were the automatic liver segmentation, tumor segmentation, and tumor load estimation. This dataset is publicly available and consists of 130 scans of abdominal CT images from six different medical centers with their respective segmentation of the liver (Figure 5.1b).

5.3.2 Data processing

Once all the images were acquired, they were resized, put in a same frame of reference, cleaned, and meshed. Matlab 2018a was used for all the image processing steps, while



(a) *OWN* dataset

(b) *LITS* dataset

Figure 5.1: Some liver masks of the *OWN* (a) and the *LITS* (b) datasets.

Simpleware's ScanIP was used for meshing. All these steps were completely automated and can, therefore, be easily applied to new images should they become available.

Firstly, the masks were resized using cubic interpolation, so that the voxels had a size of $1mm$ in all three dimensions. Secondly, some masks were flipped along some axis or some axes were swapped, so that they all shared the same axis configuration. Thirdly, all masks were moved to a $400 \times 400 \times 400$ image and their centroid set to the position $(200, 200, 200)$. All this processing was necessary due to the multiple sources that the images came from. Additionally, images with a voxel size coarser than $2.5mm$ along any dimension were discarded.

At this point, it was decided to discard some outlying liver geometries (15 out of 152) which would probably confound the model. The discarded livers were chosen by visual inspection before any further processing was performed, such as the ones shown in Figure 5.2.

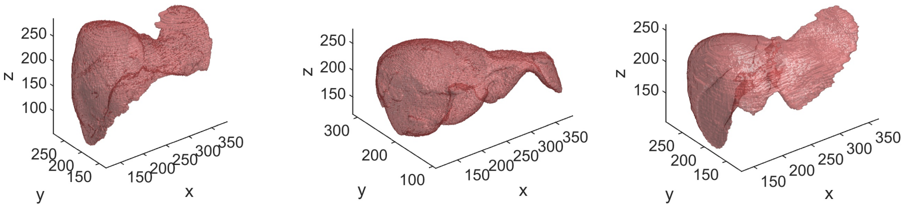
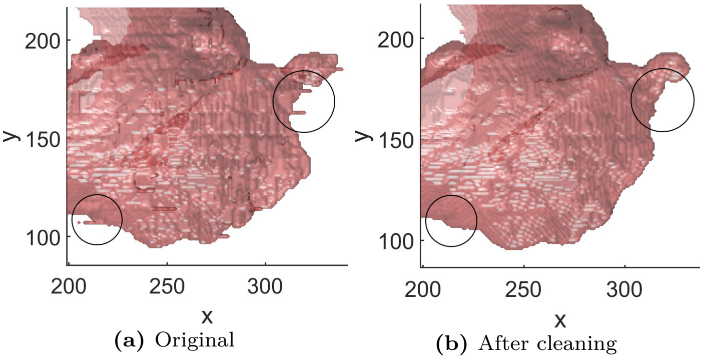


Figure 5.2: Some discarded livers

The next step was to clean the images to avoid meshing issues or later convergence problems. On one hand, some segmentations suffered from artifacts and noise (Figure 5.3 left) which was not a result of the actual geometry of the liver, but rather a result of the employed automatic segmentation method. To solve this issue, an opening followed by a closing morphological operation was applied, using a spherical structuring element with a radius of three voxels.



(a) Original **(b)** After cleaning

Figure 5.3: Cleaning of the binary mask based on an opening-closing morphological operation. Circles highlight areas where some of the artifacts appear.

On the other hand, only a few masks had a segmented hepatic tree. Therefore, in order to homogenize all images, it was decided to fill in the cavities left by the segmented ducts, by means of further morphological operations (Figure 5.4).

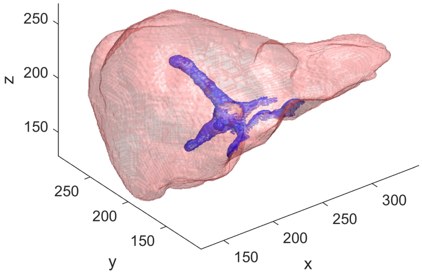


Figure 5.4: The hepatic tree (in blue) is automatically detected and filled in.

The final step was meshing all livers. This could be automated by using ScanIP scripting capabilities. The resulting meshes had $11,736 \pm 3,599$ nodes (Figure 5.5).

5.3.3 FEM simulations

FEM is a numerical method for finding approximate solutions for a particular field ϕ (such as the deformation field) on an arbitrarily shaped geometry (such as that of any liver) given a particular set of boundary conditions (restrictions on how the liver interacts with its surroundings). This is achieved by discretizing this geometry in a

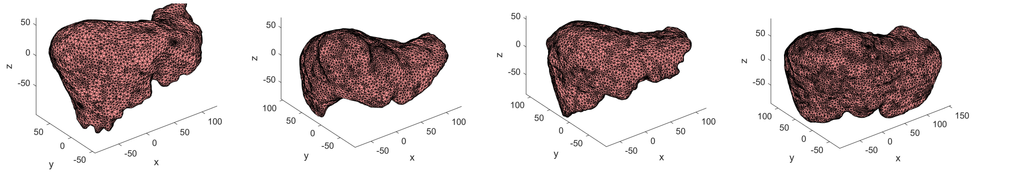


Figure 5.5: Examples of FE liver meshes.

set of finite elements and finding the solutions of the field only for the nodes that comprise it, thus reducing the degrees of freedom of the problem (Strang and Fix, 1973).

The usual formulation for the elastic problem is based on variational methods. If an energy balance is applied on the body of interest, Equation (5.1) is obtained:

$$\Pi_p = W_s - W_p \tag{5.1}$$

where Π_p is the total potential energy of the system, W_s stands for the energy stored in the deformed structure and W_p represents the work exerted by the forces acting upon it.

In virtue of the Minimum Total Potential Energy Theorem, the total potential energy Π_p will be minimum at the equilibrium, namely, for a particular displacement field $\{u\}$ (which will be solution) for which all the differential equations and boundary conditions are simultaneously satisfied. Therefore, if the solution is found when Π_p reaches a minimum, then it must also be true that its derivative with respect to the system parameters (or degrees of freedom) $\{u\}$ must also be null, as Equation (5.2) shows:

$$\frac{\delta \Pi_p(\{u\})}{\delta(\{u\})} = 0 \tag{5.2}$$

Since the field $\{u\}$ has been discretized, the solution must only be found for a finite number of points u_i.

$$\frac{\delta \Pi_p(u_i)}{\delta(u_i)} = 0 \tag{5.3}$$

Developing Equation (5.3), a solution for u_i can be found.

When a solid body is subjected to a large deformation under a comparatively small load, the relationship of positions in deformed and undeformed configurations is described by a deformation gradient tensor $\mathbf{F}$:

$$\mathbf{F} = \sum_{\alpha=1}^{3} \lambda_\alpha \mathbf{n}_\alpha \otimes \mathbf{N}_\alpha := \begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} & \mathbf{F}_{13} \\ \mathbf{F}_{21} & \mathbf{F}_{22} & \mathbf{F}_{23} \\ \mathbf{F}_{31} & \mathbf{F}_{32} & \mathbf{F}_{33} \end{bmatrix} \tag{5.4}$$

where λ_1, λ_2 and λ_3 are the stretches in the three principal directions; and $\lambda = 1 + dL/L$ with L being the undeformed length. $\mathbf{N}_1, \mathbf{N}_2, \mathbf{N}_3$ and $\mathbf{n}_1, \mathbf{n}_2, \mathbf{n}_3$ are material vector triads and spatial vector triads, respectively. The left Cauchy Green deformation tensor, $\mathbf{B}$, describes the strain, while the Cauchy stress tensor T, describes the stress:

$$\mathbf{B} = \mathbf{F}\mathbf{F}^T = \sum_{\alpha=1}^{3} \lambda_\alpha^2 \mathbf{n}_\alpha \otimes \mathbf{n}_\alpha \tag{5.5}$$

$$\mathbf{T} = -p\mathbf{1} + 2\frac{\delta W}{\delta I_1}\mathbf{B} - 2\frac{\delta W}{\delta I_2}\mathbf{B}^{-1} \tag{5.6}$$

where:

$$I_1 = tr\mathbf{B} \tag{5.7}$$

$$I_2 = \frac{1}{2}[(tr\mathbf{B})^2 - tr(\mathbf{B}\mathbf{B})] \tag{5.8}$$

$$I_3 = det\mathbf{B} \tag{5.9}$$

where I_1, I_2 and I_3 are strain invariants. The hydrostatic pressure p is constitutively indeterminate, and hence it is obtained from the underlying equilibrium and boundary conditions of the particular problem. Under the assumption of isotropic behavior, the strain energy density function W_s can be expressed as a function of the strain invariants (Rivlin, 1948a,b):

$$W_s = W_s(I_1, I_2, I_3) \tag{5.10}$$

Alternatively, W_s can also be expressed directly as a function of the three principal stretches (Valanis and Landel, 1967), namely λ_1, λ_2 and λ_3.

5.3.3.1 Biomechanical model

The next step was to choose a suitable constitutive model for the liver. The main tissue found in this organ is the parenchyma, which in its most general form can be considered a visco-poro-hiperelastic material. Moreover, the hepatic tree is comprised of a different material, which should be independently characterized. Finally, some authors also take into account the presence of a collagen capsule wrapping the parenchyma, known as Glisson capsule (Brunon et al., 2010).

Concerning the characterization of the parenchyma's hyperelastic behavior, most recent studies have found a first order Ogden model (Untaroiu and Lu, 2013) specially well suited, requiring only two empirical parameters (Marchesseau et al., 2017) as elastic constants for its construction. Regarding viscoelasticity, the bibliography on the topic shows that traction tests at different deformation velocities give rise to different Young moduli (Untaroiu and Lu, 2013), thus proving its usefulness. Regarding the porous properties of the parenchyma (which should be able to model the capilarization of the liver), its inclusion in a FEM model still remains marginal (Marchesseau et al., 2017). Moreover, some authors model the hepatic tree as a set of truss elements (Faure et al., 2012), or as a different mesh with its own mechanical properties (Plantefève et al., 2016). Finally, with respect to the Glisson capsule, few authors (Lister et al., 2011) consider its inclusion in FEM models due to the difficulty of segmenting it in medical images and/or estimating its material properties.

Following the latest trends, a first order Ogden model was used to model the mechanical behavior of the liver parenchyma. Viscoelasticity was not taken into account based on the hypothesis that the applied forces were slow enough for such effects not to be of importance. Thus, the performed simulations were static. Finally, the model was considered homogeneous and the porosity effects were embedded into the elastic properties of the Ogden model. Regarding the hepatic tree or the Glisson capsule, both were assimilated by the parenchyma due to the lack of segmentation masks for these elements.

The deformation energy density function W_s for an Ogden elastic model is given by Equation (5.11):

$$W_s = \sum_{k=1}^{N} \frac{\mu_k}{a_k} (\lambda_1^{a_k} + \lambda_2^{a_k} + \lambda_3^{a_k}) \tag{5.11}$$

where N is the order of the model, μ_k and a_k are empirical parameters of the material and $\lambda_1, \lambda_2, \lambda_3$ are the principal stretches.

The values for the material parameters μ_k and a_k were obtained from a set of 30 material properties described in (Untaroiu and Lu, 2013). A default material was chosen as the median of all material properties in said paper:

$$a_1 = 10.06, \mu_1 = 4.1kPa$$

For the compressibility modulus (K_0), a value of 100 times the value of μ_1 was selected to ensure a quasi-incompressible behavior, which is a common choice for soft tissue modeling.

5.3.3.2 Boundary conditions

From the perspective of the boundary conditions (BC), the liver is in contact with multiple organs and structures, thus rendering the task of finding anatomically correct BCs extremely challenging.

The most usual solution found in the literature is to resort to simplified BCs, where some nodes are considered fixed and the rest are left free. Following this trend, many authors fix the nodes in contact with the cava vein or with the falciform ligament (Marchesseau et al., 2017; Plantefève et al., 2014). Others do not consider the falciform ligament as a restriction and resort to only fixing the cava vein (Plantefève et al., 2016).

Typically, even more simplified BCs are considered, especially when the objective is to prove the feasibility of a new method rather than to make an anatomically correct simulation of the organ and its interactions within the body, e.g., in (Niroomandi et al., 2012) and (Lister et al., 2011) palpation was simulated with the livers laid against a flat surface, hence sufficing to fix the nodes in contact with this surface.

For this work, the liver was considered attached only to the cava vein. Therefore, a null displacement boundary condition was applied to the liver nodes in contact with it (Figure 5.6).

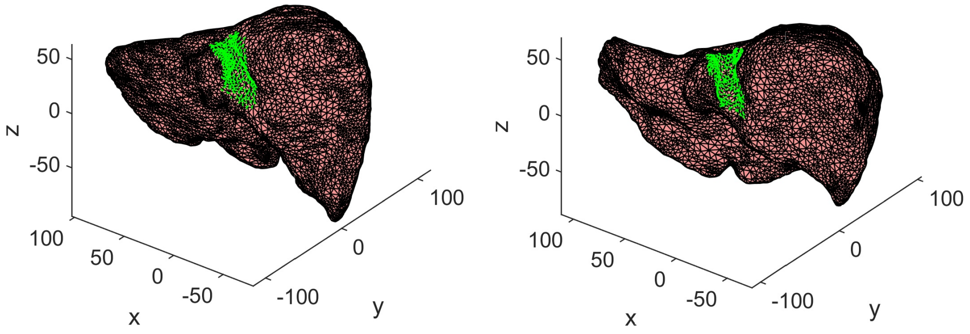


Figure 5.6: Some livers with BCs applied (in green).

5.3.3.3 Applied forces

The last step before launching the simulations consisted in defining the forces to be applied to the liver. In the literature, cylindrical indenters are typically pressed on to the surface of the organ, which also allows for a direct comparison of the simulation results with previous ex vivo identation tests (Lister et al., 2011). Other authors consider indenters with an infinitely small radius, hence the forces becoming nodal forces (Niroomandi et al., 2012).

The forces in this work were also considered to be nodal, and they were applied on a random node of the liver, with a random orientation, and with a magnitude of $0.4N$ (a similar magnitude to the forces applied in Lister et al. (2011) or Niroomandi et al. (2012)). The objective was to provide a simple but challenging set of forces. Nonetheless, for a real-world application, the simulations could be performed with any kind of forces originating from any sort of surgical tool deemed necessary.

5.3.3.4 Performed simulations

Following the proposed objectives, the mechanical behavior of the liver was simulated in four different scenarios.

In Table 5.1, a summary of the setup for each simulated scenario is presented. In the first scenario (the simplest one), only one liver geometry was considered, upon which

400 different forces, each one with a random position and orientation, were applied, hence adding up to a total of 400 deformed livers, stemming from 400 different FEM simulations. Each force was built by sampling its three components (f^x, f^y, f^z) from a uniform distribution $\mathcal{U}[0,1]$, and then normalizing the vector so that its magnitude is $0.4N$. The node upon which each force was applied was also chosen randomly without replacement from the list of all nodes.

Table 5.1: Simulated scenarios

ID	*Random forces*	*Two forces*	*Multiple mater.*	*Multiple livers*	*Simulated forces*
1	Yes	No	No	No	400
2	Yes	Yes	No	No	1,500 (×2)
3	Yes	No	Yes	No	3,000
4	Yes	No	No	Yes	10,200

The second and third scenarios also employed only one liver. In the second scenario two random simultaneous forces were considered in each simulation, instead of just one, while in the third scenario multiple material properties were contemplated. In contrast to the first scenario, more simulations were needed to account for the growing casuistry.

Finally, the fourth scenario posed the biggest challenge, as it was designed to work on any liver geometry.

All the simulations were performed with FEBio (Maas et al., 2012) and automated with Matlab. The analysis type was static (no viscoelasticity was finally considered), and the force was applied in ten steps to help the FEM converge when large deformations are present. Since the magnitude of any force was set to $F = 0.4N$, at each step the magnitude is increased by $\frac{1}{10}F$. After the tenth step, the full force F has been finally applied, and the final deformed geometry can be obtained. Furthermore, the results from each of the ten intermediate steps can be used as extra simulation data, therefore increasing the amount of simulation data tenfold at no additional cost.

To better understand why this four scenarios were specifically selected, an example application to CAS can be considered. If there are at least a couple hours between the acquisition of the image of the liver and the start of the surgery, this time can be used to train a model for that specific liver, which would also be highly accurate (first scenario). Furthermore, the second scenario could be useful if two surgical tools were simultaneously used to interact with the liver during the operation, while the third scenario would be needed if the material properties of the liver are expected to change during the intervention, e.g.: due to inflammation. However, if there is not enough time, it is impossible to have early access to the image, or it is not viable to spend the computing power required to retrain a model from scratch, the model trained from the fourth scenario, will work directly and with sufficient accuracy, as it will be shown.

5.3.4 Training of the ML models

5.3.4.1 Feature selection for the ML models

Before being able to use the data to train a ML model, it should be expressed in terms of an input matrix X and an output matrix Y. Then, a ML algorithm would be trained to learn the mapping $X \rightarrow Y$ as accurately as possible.

The input matrix X was of dimensions: $(S * N * T) \times F$ where S stands for the number of simulations (e.g., 400 for the first scenario), N stands for the number of nodes ($\sim 12,000$), T stands for the number of successfully simulated time steps (usually ten), and F is the number of input features. Thereby, every row (each sample) corresponded to a node of the mesh, and each column contained several features, which are described as follows:

- x, y, z: Coordinates of the considered node, whose displacement is to be obtained.
- f^x, f^y, f^z: Force vector.
- n^x, n^y, n^z: Coordinates of the node where force was applied.
- $d^x_{min}, d^y_{min}, d^z_{min}, d^{euc.}_{min}$: Distance (vector and magnitude) from the considered node to the closest fixed node.
- $d^x_{mean}, d^y_{mean}, d^z_{mean}, d^{euc.}_{mean}$: Distance from the considered node to the centroid of all the fixed nodes.
- $d^x_{load}, d^y_{load}, d^z_{load}, d^{euc.}_{load}$: Distance from the considered node to the node where the force was applied.
- $g_1, g_2, ..., g_{27}$: Additional features that parametrize the liver geometry (to be explained in Section 5.3.4.2).
- a_1, μ_1: Material properties for the Ogden model (only used in the third scenario).

The choice of this particular set of features is not arbitrary, and responds to the requirements of the problem. Indeed, each row of X represents an independent sample and, as such, it must somehow contain all the information that the ML algorithm might need to compute the output. This justifies the extensive use of geometry related features, such as the distance to the fix nodes and to the node where the load is applied, or the final geometry characterizing features.

In addition to the previous features, for the second scenario (where two simultaneous forces were applied) further features related to the second force, with the exact same meaning as for the first force, were included: $(d^x_{min})_2, (d^y_{min})_2, (d^z_{min})_2, (d^{euc.}_{min})_2, (d^x_{mean})_2, (d^y_{mean})_2, (d^z_{mean})_2, (d^{euc.}_{mean})_2, (d^x_{load})_2, (d^y_{load})_2, (d^z_{load})_2, (d^{euc.}_{load})_2$. This also allowed for the number of simulations to be artificially doubled just by swapping the values of the features associated to the first force with those associated to the second one.

Finally, the output matrix Y was of dimensions: $(S * N * T) \times 3$, containing as many rows (samples) as X, but only the following three columns (outputs):

- d^x, d^y, d^z: Node displacement vector.

5.3.4.2 Liver geometry parametrization

For the ML algorithm to be able to generalize to different livers, the whole liver geometry should somehow be introduced into each sample of X. To address this challenge, two consecutive subproblems must be solved:

a) Express each liver geometry as a vector of N features, such that each feature has the same meaning for all livers.

b) Apply dimensionality reduction techniques that reduce the vector from N to n features such that $n \ll N$.

Feature Consistency Regarding the first subproblem, a typical approach consists in directly using the binary mask flattened into a vector. However, this method assumes that two voxels in the same location from two different livers have the same meaning or, in other words, represent the same feature, which is often not true.

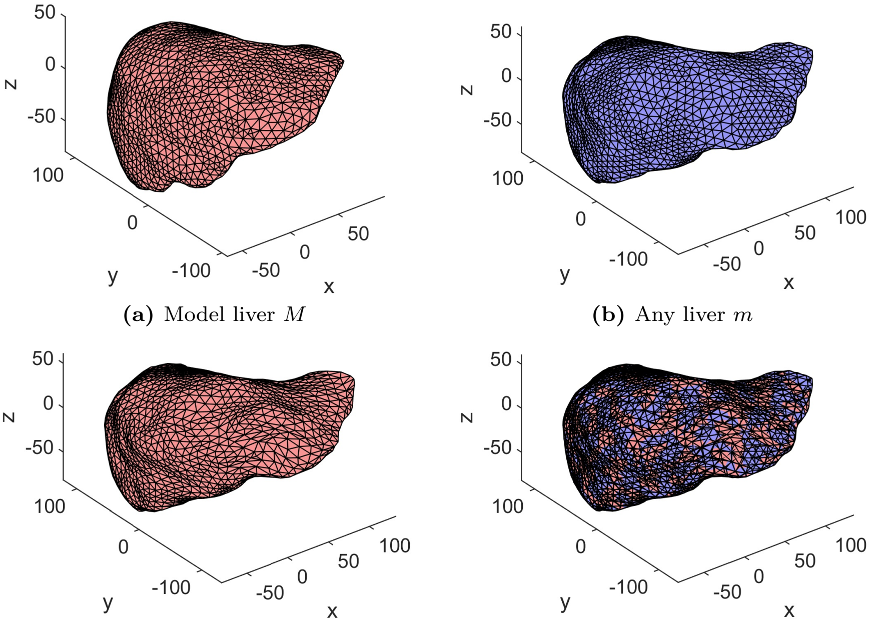
Here, a novel and effective approach is proposed. Firstly, the nodes of the surface of a model liver M (which was chosen beforehand for subjectively being the most regular), were registered by a soft registration algorithm to the nodes of the surface of each liver m to obtain M_r (M registered). Secondly, the displacement field that the nodes of the surface of M underwent to become M_r was obtained. Namely, $field = M_r - M$.

Then, M_r could be flattened to express it as vector, where each element has the same meaning for all livers: how much each node of the surface of M must be displaced in x, y or z in order to become M_r (where M_r is an approximation of the original geometry m).

The bottom right plot of Figure 5.7 shows that this registration process was successful, since M_r approximates m satisfactorily. A slightly coarser mesh was used to speed up the registration process.

For the non-rigid registration, a modified version of CPD was employed. CPD is a point set registration algorithm that finds the spatial transformation that best aligns two point sets and/or finds the correspondence between the points of both sets (Myronenko and Song, 2009). CPD can perform both rigid registration (the transformation is limited to some combination of translation, rotation and scaling, amounting to a total of six parameters), and non-rigid registration (a nonlinear transformation of any number of parameters).

The proposed modification consists in applying scheduled changes to the regularization parameters β, related to the width of a Gaussian smoothing filter, and λ, the trade-off between fit and regularization, which are gradually reduced as the algorithm converges, thus refining the registration by going from low to high spatial frequencies. This change, although relatively trivial, allows the algorithm to significantly outperform its unmodified counterpart, specially for problems where the input and output shapes differ substantially.



(a) Model liver M

(b) Any liver m

(c) M registered to m: M_r

(d) m and M_r superimposed

Figure 5.7: An example of liver registration

The values used for these two regularization parameters can be found in Table 5.2, and a visualization of the convergence process in two different cases can be checked in Figure 5.8.

Table 5.2: Regularization values β and λ for CPD depending on the relative tolerance of the algorithm (Myronenko and Song, 2009); when the minimum relative tolerance is surpassed, the next stage is activated.

Stage	*Minimum relative tolerance to switch to next stage*	β	λ
1	1×10^{-2}	8	6
2	5×10^{-3}	2	1.5
3	1×10^{-3}	1.2	0.9
4	1×10^{-4}	0.6	0.45
5	1×10^{-5}	0.3	0.225

Dimensionality Reduction Concerning the second subproblem, Principal Component Analysis (PCA) was used to reduce the dimensionality of this set of geometry characterizing vectors to only a few parameters. PCA makes a transformation of a possibly correlated data set into a non-correlated orthogonal base, whose variables are

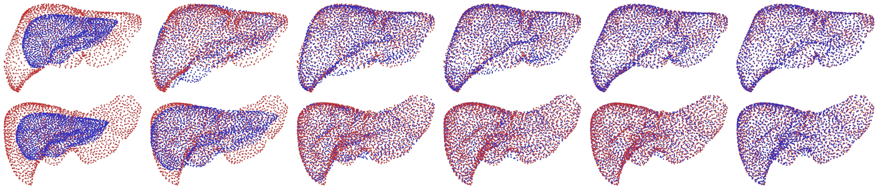


Figure 5.8: Convergence of the modified CPD algorithm for two different livers. First image to the left represents the initial state, while the rest were captured at the end of each stage, right before switching to the next. As it can be seen, the registration is extremely successful even when the geometries differ significantly. The blue point cloud corresponds to the nodes of the model liver M, as they mutate to become M_r, an approximation of the arbitrary-shaped liver m, whose nodes are given by the red point cloud.

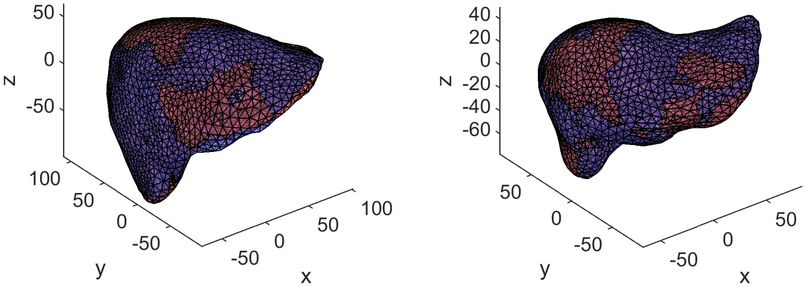


Figure 5.9: Reconstruction (in red) of the original geometry (in blue) given only the first 27 PCs for two different livers.

called principal components (PCs), and are ordered by amount of variance explained. If the N variables in the original space are highly redundant, then it is possible to compress high dimensional data to only a few variables n (such that $n \ll N$) by applying PCA and taking the first n PCs as the new variables.

Figure 5.9 shows the reconstruction of the original liver geometry given the first 27 PCs, which is the number of parameters finally chosen by means of hyper-parameter optimization (see Section 5.3.4.3) to parametrize the liver geometry. Figure 5.10 shows (for all the livers aggregated) the Intersection over Union (IoU) and the DICE scores (which are both intersection scores), as well as the accumulated relative variance given the first n PCs (the rest were set to zero).

Therefore, these 27 PCs will be the 27 features $(g_1, g_2, ..., g_{27})$, that will appear in matrix X to help define the liver geometry. If the ML algorithm is powerful enough, it should be able to learn the meaning of each parameter and use it to internally reconstruct an approximation of the shape of each liver.

Finally, the proposed approach was compared to the few similar approaches found in the literature. In (González et al., 2016), the authors use Locally Linear Embedding (LLE) on meshes generated by applying a rigid transformation (six parameters) to a single initial mesh. LLE successfully compresses this transformation into four

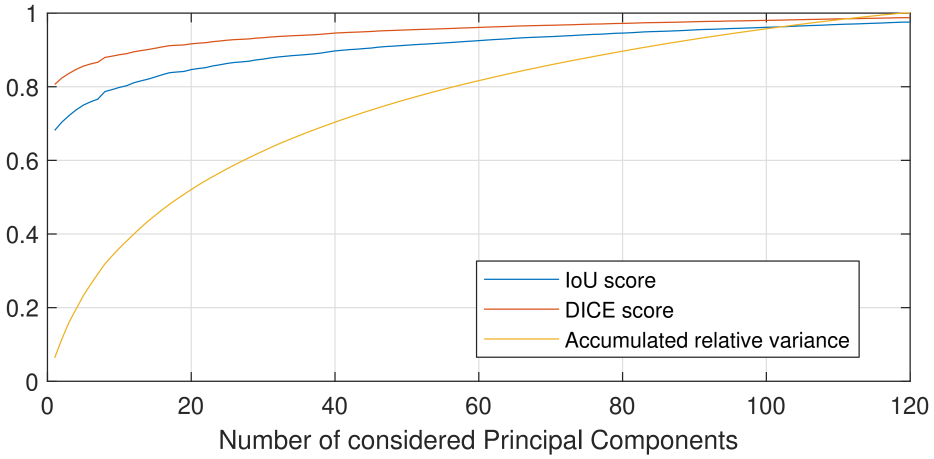


Figure 5.10: IoU score, DICE score and accumulated relative variance depending on the number of PCs used to reconstruct the geometry.

parameters, while keeping most of the variance of the geometry. In (González et al., 2018), a similar experiment is conducted, but this time using Kernel PCA (kPCA) instead of LLE, and applying an affine transformation (twelve parameters) to the initial mesh, instead of a rigid transformation. Although the results were very promising in both papers, the first subproblem is not directly addressed in any of them, thus rendering both approaches ineffective for arbitrary liver meshes.

The method here proposed can be thought as a generalization of the previous methods, in the sense that the proposed transformation is not limited to being rigid or affine; instead, it is a general nonlinear transformation. This approach can therefore be applied to any geometry, not only to simple transformations of a single model mesh.

5.3.4.3 ML model and hyper-parameter optimization

Once the input and output matrices (X and Y, respectively) were built, the next step was to train a ML algorithm able to approximate Y given X with sufficient accuracy, and in real time.

Regarding the kind of ML algorithm to use, two main contenders were tested: a Feedforward NN for regression and a Random Forest (RF) regressor, which is an ensemble of Regression Trees (Breiman, 2017). Even though the second algorithm was successfully employed in (Lorente et al., 2017), for this particular problem the NNs showed a vastly superior performance in the preliminary tests, and were therefore finally chosen to conduct all the experiments.

The feedforward NNs used in this work are supervised learning algorithms able to learn nonlinear mappings (models) between certain inputs (such as the coordinates of the node, the force orientation, the distance to the force, etc.) and certain outputs (such as the displacement in x, y, z of a node).

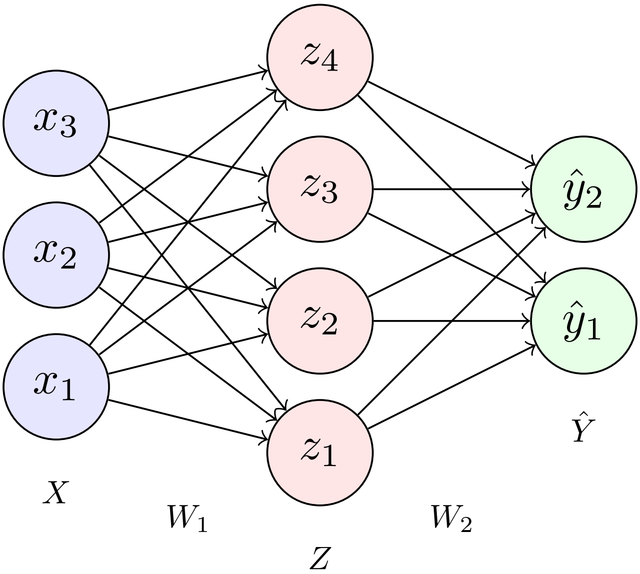


Figure 5.11: A feedforward neural network comprised of an input layer X, a hidden layer Z, an output layer $\hat{Y}$ and the weights (W_1, W_2)

Figure 5.11 offers a visual representation of the algorithm (for a single hidden layer, and ignoring biases) and introduces some notation. To compute the output of the network $\hat{Y}$ given the input X, Equations (5.12) and (5.13) are used:

$$Z = X \times W_1 \tag{5.12}$$

$$\hat{Y} = f(Z) \times W_2 \tag{5.13}$$

where f is a nonlinear activation function, W_1 and W_2 are the weights of the network, or in other words, the parameters that the NN must learn, and X, Y contain the inputs and the actual outputs in matrix form, respectively. Thus, a NN can be seen as a stack of linear transformations with nonlinear activation functions in between. The NN could have as many hidden layers as needed, but only one has been considered in Figure 5.11 for the sake of simplicity.

To train the NN to approximate a certain function, some parameters (or weights) $\theta = \{W_1, W_2\}$ must be found in order to minimize a certain cost function $J(\theta)$ such as the Mean Squared Error (MSE):

$$J(\theta) = grandmean((\hat{Y}(\theta) - Y)^2) \tag{5.14}$$

where *grandmean* represents the mean across all elements of the squared errors matrix $(\hat{Y}(\theta) - Y)^2$, and the exponentiation is applied element-wise.

To minimize J with respect to the parameters θ, Gradient Descent (GD) is typically used (Bottou, 2010):

$$\theta = \theta - \mu \nabla_\theta J(\theta) \tag{5.15}$$

Table 5.3: Optimal values found for the hyper-paramaters of the model in each scenario. Note that for the first three scenarios no PCs are needed: since only a single liver is used, the Neural Network does not need any geometry information.

Scenario	*Architecture*	*Dropout*	*Batch size*	*Noise variance*	*L2 regulariz.*	*Learning rate*	*Number of PCs*	*% sampled*
1 (base case)	$(500, 300, 100, 50)$	$(0.3, 0.5, 0.6, 0.7)$	512	0	0	2.5×10^{-4}	0	50%
2 (two forces)	$(500,) \times 5$	$(0.5,) \times 5$	256	0	3×10^{-5}	1.25×10^{-4}	0	50%
3 (30 materials)	$(500,) \times 5$	$(0.5,) \times 5$	512	0	0	2.5×10^{-4}	0	25%
4 (102 livers)	$(500,) \times 5$	$(0.5,) \times 5$	4096	0.1	2×10^{-4}	1×10^{-3}	27	12%

GD is an iterative algorithm that at each step pushes the parameters θ a small amount μ in the direction opposite to gradient ∇_θ of $J(\theta)$ with respect to θ, thus finally converging to a minimum for J (Ruder, 2016).

NNs are extremely powerful nonlinear approximators that tend to overfit the training set, therefore jeopardizing the generality of the model to new samples that the network has not trained with. To prevent this problem, multiple techniques can be used, such as: adding $L2$ regularization to the parameters, adding Gaussian noise to the inputs, or by using dropout (Srivastava et al., 2014). For all the tasks related to NN training, the Python library Keras (Chollet, 2015) running on top of Tensorflow (Abadi et al., 2016) was employed.

The objective at this point was to optimize the hyper-parameters of the NN model (such as the NN architecture, the learning rate, the batch size, etc., as well as other variables that affect performance such as the number of PCs to use) in order to reduce the error in the validation set.

For the first three scenarios, the validation set contained a randomly chosen $\sim 10\%$ of all simulations, with a different applied force each. Thus, the performance of these models was evaluated on forces that the NN had not been trained with.

For the last scenario, where multiple liver geometries were considered, the validation set contained all the simulations corresponding to a randomly chosen $\sim 12.5\%$ of all livers. Hereby, it is possible to assess the behavior of the model for livers which it has never seen before.

To obtain the final results (view Section 5.4), k-fold cross validation was employed, where $k = 10$ for the first three scenarios and $k = 8$ for the last scenario.

The hyper-parameter optimization was manually conducted, until the optimal values shown in Table 5.3 were reached. Rectified Linear Unit (ReLU) activation functions and MSE cost function were always employed, while noise injection (which consists in adding a random normal noise to the inputs as a way of regularizing the network) was only applied on the last scenario.

Regarding the training, the learning rate was reduced tenfold after the second epoch without improvement on validation loss for the first three scenarios, while for the last one, the learning rate was reduced by a factor of 1.5 after each epoch. For all scenarios, training was stopped if the loss in the validation set stopped improving.

Table 5.4: Results for all scenarios: Mean Absolute Error (MAE); Mean Euclidean Error (MEE); percentage of predictions with a Euclidean Error (EE) below $1mm$ and $3mm$; and correlation coefficient between predicted and actual values.

Scenario	*Algorithm*	*MAE (mm)*			*MEE (mm)*	*% of samples*		*Correlation coefficient*		
		x	y	z		*EE* $< 1mm$	*EE* $< 3mm$	x	y	z
1 (base case)	NN	0.1173	0.1176	0.1224	0.2389	98.2615%	**99.9551**%	0.9973	0.9974	0.9959
	Naive	1.5828	1.5777	1.4020	3.0661	32.5021%	64.2548%	-	-	-
2 (two forces)	NN	0.1862	0.1825	0.1977	0.3816	93.0551%	**99.8382**%	0.9967	0.9967	0.9949
	Naive	2.3247	2.2804	2.0247	4.4828	20.9003%	50.1540%	-	-	-
3 (30 materials)	NN	0.1621	0.1616	0.1548	0.3299	94.7381%	**98.4644**%	0.9911	0.9904	0.9900
	Naive	2.1622	2.1178	1.7933	4.2052	35.0012%	63.7634%	-	-	-
4: *Mean* (102 livers)	NN	0.4130	0.4732	0.4119	0.8643	72.4243%	**95.5784**%	0.9790	0.9814	0.9702
	Naive	1.4444	1.6767	1.2993	2.9534	35.3185%	66.6061%	-	-	-
4: *Median* (102 livers)	NN	0.3377	0.3360	0.3315	0.6619	79.0198%	**99.0115**%	0.9828	0.9874	0.9731
	Naive	1.6341	1.8307	1.2278	3.1332	33.7861%	63.3039%	-	-	-

Finally, it must be noted that all the data coming from the simulations was randomly sampled before being used to train the NN. The percentage of data remaining after the sampling is also shown in Table 5.3. This processing was performed to speed up the training process at practically no cost in performance. In fact, the simulation data is very redundant due to two main reasons. First, for nodes positioned away from the node of application of the force, the field is very similar between neighboring nodes. Second, each force was applied in ten steps (at an increasing magnitude), thus providing very similar results between consecutive steps.

5.4 Results and discussion

In this section, the results for each of the four models will be presented and it will be discussed if both sufficient accuracy and real-time inference were achieved. All final metrics have been computed using all the samples from all k validations sets concatenated.

5.4.1 Scenario 1: Base model

The first scenario is the simplest one, since only one liver, one material and one arbitrary force is considered, although the force may have any orientation, any magnitude, and be applied to any node. Only 360 simulations were employed to train this model, while the remaining 40 made up the validation set.

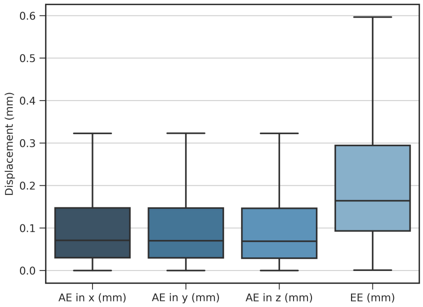


Figure 5.12: Scenario 1: Box plot of the absolute errors (AE) in x, y, z, as well as the Euclidean error (EE). Outliers are excluded.

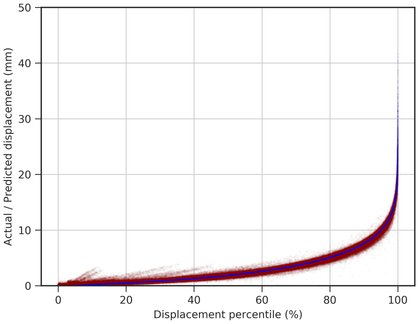


Figure 5.13: Scenario 1: Actual Euclidean displacement (in blue) and predicted Euclidean displacement (in red) for all samples, sorted by ascending actual Euclidean displacement.

The numerical results for all scenarios have been compiled in Table 5.4. A naive model, which always outputs a constant value calculated as the average of all the training outputs, was included for comparison as well.

Analyzing the first row of Table 5.4, which corresponds to this scenario, the first column shows the mean absolute error (MAE) for each output coordinate (x, y, z), followed by the mean euclidean error (MEE), which represents the mean distance between the predicted and the real displacement fields. All these errors are extremely low, staying around $0.12mm$ for all three coordinates, and below $0.25mm$ for the MEE. For reference, both the discretization error of the original mask as well as the maximum ScanIP's mesh error are around $0.5mm$ ($\frac{1}{2}$ the voxel size). Also, the maximum displacements for this particular scenario are above $30mm$, as it can be seen in Figure 5.13, which will later be discussed.

Continuing with Table 5.4, the following two columns show the percentage of samples with an Euclidean error (EE) below $1mm$ and $3mm$ (which was set as the objective threshold). As it can be observed, the results are also very good in this regard, since 99.96% of the samples manage to stay below the $3mm$ error limit, while 98.26% stay below $1mm$. Finally, the last three columns show the correlation coefficients for all three output coordinates, which for this scenario are almost unitary, proving the notable performance of the proposed model.

Figure 5.12 shows the absolute error (AE) distributions in x, y and z, as well as the Euclidean error (EE) distribution. As it can be noted, coordinate errors manage to stay under $0.25mm$, while Euclidean errors remain below $0.6mm$, excluding outliers. These errors are well below the maximum systematic error of $1mm$.

Finally, Figure 5.13 shows how the predicted Euclidean displacements (in red) follow very closely the actual Euclidean displacements (in blue), even when large displacements are present.

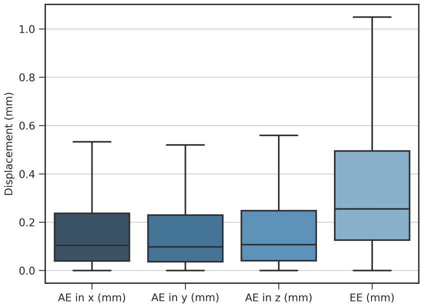


Figure 5.14: Scenario 2: Box plot of the absolute errors (AE) in x, y, z, as well as the Euclidean error (EE). Outliers are excluded.

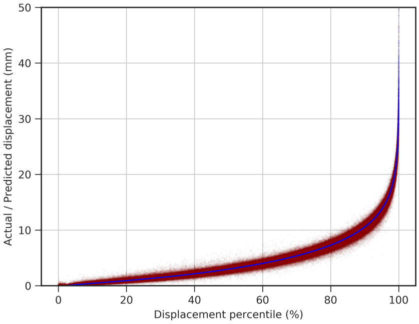


Figure 5.15: Scenario 2: Actual Euclidean displacement (in blue) and predicted Euclidean displacement (in red) for all samples, sorted by ascending actual Euclidean displacement.

5.4.2 Scenario 2: Two simultaneous forces

For the second scenario, two simultaneous random forces were applied in each of the 1,500 simulations. This is a rather complex situation, since the non-linearities prevent the resulting deformation field from being a simple addition of the independent effects of each force.

As it can be seen from the second row of Table 5.4, the results are still accurate, achieving a 99.84% of the samples within the error threshold of $3mm$. It is worth mentioning that the naive model performs much worse in this scenario as it did on the previous one, suggesting that the deformation magnitudes are significantly larger, which makes them more difficult to predict. The box plots in Figure 5.14 show that the Euclidean error distribution mostly stays below $1mm$.

Finally, from the scatter plot (5.15), it can be checked that, in fact, the deformations are larger as compared to the previous case. Nevertheless, the model still achieves a very low dispersion in the predictions.

5.4.3 Scenario 3: 30 materials

The third scenario was designed to test the learning abilities of the model against the change in material properties. For each of the 3,000 simulations, one of 30 different parameter sets was randomly sampled, thus amounting to a total of ~ 100 simulations per material.

Once again, the third row of Table 5.4 shows that very good results were achieved. The percentage of samples below the $3mm$ mark falls slightly to 98.46% due to the presence of many more outliers. Indeed, a few very elastic materials give rise to extreme deformations (above $60mm$) when the force acts upon certain parts of the liver. Figures 5.16 and 5.17 confirm these conclusions; in fact, the box plot of Euclidean errors verifies that the performance is even better than in the first scenario (if outliers are excluded).

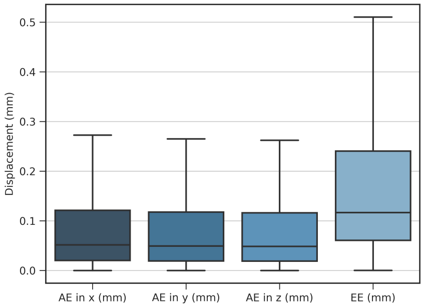


Figure 5.16: Scenario 3: Box plot of the absolute errors (AE) in x, y, z, as well as the Euclidean error (EE). Outliers are excluded.

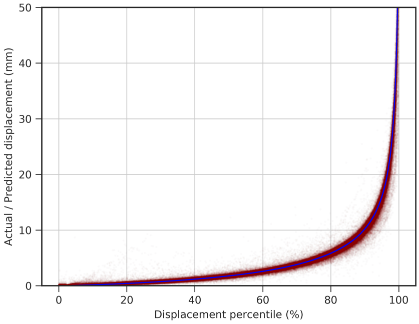


Figure 5.17: Scenario 3: Actual Euclidean displacement (in blue) and predicted Euclidean displacement (in red) for all samples, sorted by ascending actual Euclidean displacement.

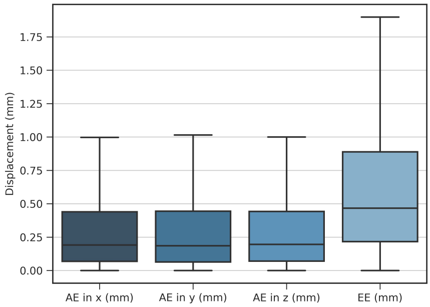


Figure 5.18: Scenario 4: Box plot of the absolute errors (AE) in x, y, z, as well as the Euclidean error (EE). Outliers are excluded.

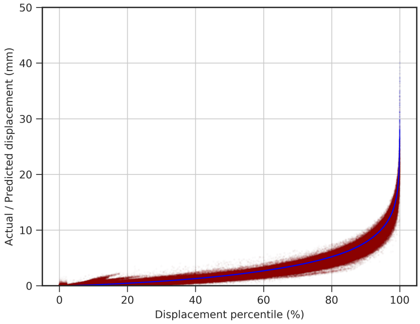


Figure 5.19: Scenario 4: Actual Euclidean displacement (in blue) and predicted Euclidean displacement (in red) for all samples, sorted by ascending actual Euclidean displacement.

5.4.4 Scenario 4: 102 livers

The fourth scenario constitutes the main outcome of this paper. The objective was to train a model able to model the biomechanical behavior of any liver presenting any geometry. To this end, the validation sets were composed of all the simulations

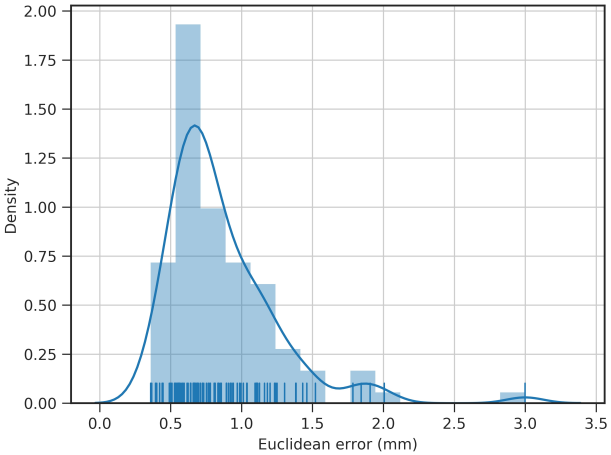


Figure 5.20: Scenario 4: Distribution of MEEs for all the 102 livers in the eight validation sets.

belonging to a $\sim 12.5\%$ of all livers. Hence, the results show how the model performs when tested on new livers.

For this scenario, the results were aggregated differently. On one hand, for each of the 102 livers in the validation set, the same metrics as in all the previous cases were computed. In fact, the last row of Table 5.4 shows these metrics for a median liver (chosen as the liver with a z correlation in the median). Figures 5.18 and 5.19 were obtained for this median liver. Even though the results have worsened, 99.01% of the samples still manage to stay below the allowed error threshold of $3mm$, which is a good finding, considering that the liver under consideration represents the median behavior of a model which was not trained with that particular geometry.

On the other hand, the results were aggregated for all livers by computing the mean, which can also be seen in the fourth row of Table 5.4. Here, the values are not as ideal due to the influence of some outlying liver geometries for which the model did not perform as well. To further prove this point, Figure 5.20 shows the distribution of MEE over all the livers of all validation sets. As it can be observed, the distribution peaks at a MEE of around 0.6 to $0.7mm$, precisely where the median liver lied, but some outlying livers with poorer performance drag the mean of the MEEs distribution towards higher values.

5.4.5 Real-time performance

Next, the ability of the model to work in real time will be assessed. To make an inference (get the predicted output $\hat{Y}$) with a feedforward NN, two steps are required: build the matrix X, and propagate it through the network.

Building X is extremely quick, since most of its features can be computed offline, except for the distance to the node where the force was applied, as well as the force itself. Nonetheless, calculating these features can be considered immediate.

Propagating X through the NN consists of a series of matrix multiplications followed by non-linear activations. Matrix multiplications can be done very efficiently using GPUs, allowing for an improvement in speed of several orders of magnitude with respect to a CPU. The ReLU non-linearities used in this paper are also extremely simple, and have a negligible impact on the final cost.

Oddly enough, the highest computational burden would be the cost of transferring the X matrix from computer memory to the GPU memory, but for such small X matrices, this is is not a problem either.

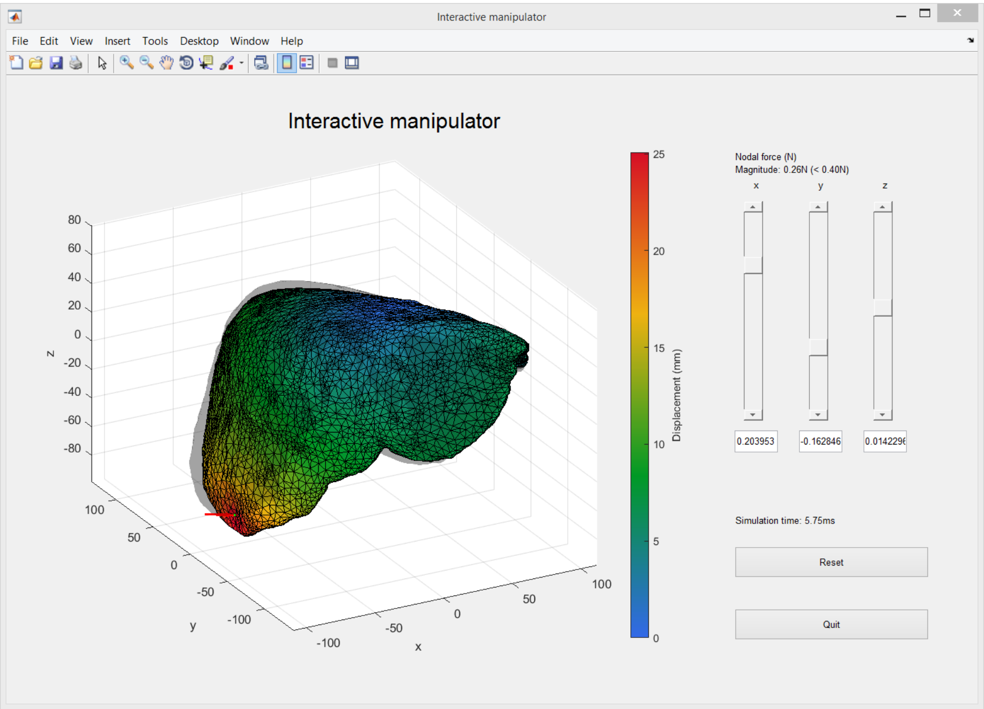
Finally, for a real application, it would not be necessary to compute the displacement field for all nodes, but rather for those which are of interest, such as the visible surface of the liver.

In a 2013 laptop equipped with a two-core i5 processor and a low end GT 840M GPU, the time required for building X and propagating it through the NN to get the displacement of a liver is around $2ms$ for the first scenario, and around $5ms$ for the last scenario (for which the network architecture is more complex). Using slightly better hardware and a more polished implementation, it would be trivial to achieve inference times in the order of hundreds or even tenths of microseconds, thus enabling the model for its use in haptic feedback systems, which require working frequencies above $500Hz$.

5.4.6 Interactive liver manipulator

To conclude the results section, an interactive liver manipulator will be presented. It has been developed to visually assess the correct behavior of the model, as well as to prove its real time capabilities.

Figure 5.21 shows a video of this software being used on the last scenario, on the median liver. It must be emphasized that this is a liver that the model was not trained with. Watching the simulation, it is evident that the displacement field that the liver undergoes given arbitrary forces corresponds with the intuitive expectations, thus proving that the behavior is (at least visually) correct. Furthermore, it can be seen from the video that the time required for inference is usually around the previously stated $5ms$.



Nodal force (N)
Magnitude: 0.26N (< 0.40N)

0.203953

-0.162846

0.014229

Simulation time: 5.75ms

Figure 5.21: Capture of the interactive manipulator simulating the mechanical behavior of the liver of median validation performance given an arbitrary force.

5.5 Conclusion

The main objective of building a general model able to simulate the mechanical behavior of any liver, given an arbitrary force, with sufficient accuracy and in real time has been achieved. However, other options exist too. For instance, the model in the first scenario achieves higher levels of accuracy, as compared to any liver in the last scenario. In practice, for applications requiring such precision, it would be sensible to simulate a particular liver geometry under a few hundred forces and train a NN on top of it, as part of the pre-operative process. For instance, simulating the mechanical behavior of a liver under 200 different forces (which would be almost as accurate as with 360 forces), and training a NN model on these simulations, takes around two hours in a server with an eight-core Intel Xeon E5-2620 CPU and a mid tier Nvidia Maxwell GPU.

Furthermore, the capability of the method to generalize from situations where multiple forces or different material properties are at play has been shown.

Finally, a novel geometry parametrization algorithm has been developed, which allows the NN model to generalize to unknown geometries. Moreover, a simple but effective modification to the CPD has been employed, which enables this algorithm to achieve excellent registration results even when the registered geometries are drastically different.

5.6 Limitations and further work

Although the proposed methods proved successful, some areas of improvement can be detected.

First, the geometry parametrization technique, despite effectively allowing the NN to generalize to any liver geometry, is still approximate. A possible further research line would be to use convolutional layers to input the original segmentation mask directly into the NN.

Regarding convolutional NNs, a topic of active research is the development of automatic segmentation algorithms, which take the CT or Magnetic Resonance (MR) image and directly compute a segmentation mask.

Finally, finding accurate boundary conditions and constitutive models is still a challenge in Biomechanics. Particularly, the improvement of techniques for in-vivo identification of elastic parameters is fundamental for the development of high-accuracy patient-specific models.

Acknowledgments

This work has been funded by the Spanish Ministry of Economy and Competitiveness (MINECO) through research projects TIN2014-52033-R, also supported by European FEDER funds.

Chapter 6

Deep Learning Contributions for Reducing the Complexity of Prostate Biomechanical Models

Reference: Oscar J. Pellicer-valero, Maria José Rupérez, Victor Gonzalez-perez, and José D. Martín-Guerrero. Deep learning contributions for reducing the complexity of prostate biomechanical models. In PhD F. Chinesta, PhD, E. Cueto, PhD, Y. Payan, PhD and J. Ohayon, editor, *Reduc. Order Model. Biomech. Living Organs*. Elsevier, 2022
Accepted chapter, to be published by Elsevier in the last trimester of 2022.

Oscar J. Pellicer-Valero*[1], Maria José Rupérez[2], Victor Gonzalez-Perez[3], José D. Martín-Guerrero[1]

[1] Intelligent Data Analysis Laboratory, Department of Electronic Engineering, ETSE (Engineering School), Universitat de València (UV), Av. Universitat, sn, Bujassot, València 46100, Spain.
`Oscar.Pellicer@uv.es`, `jose.d.martin@uv.es`

[2] Instituto Universitario de Ingeniería Mecánica y Biomecánica (I2MB), Universitat Politècnica de València (UPV), Camino de Vera, sn, València 46022, Spain.
`mjrupere@upvnet.upv.es`

[3] Department of Medical Physics, Fundación Instituto, Valenciano de Oncología (FIVO), Beltrán Báguena, 8, 46009, València, Spain.
`vgonzalezper@hotmail.com`

6.1 Abstract

Prostate cancer (PCa) is the most common malignancy in males in western countries, which is usually detected with prostate-specific antigen (PSA) levels measurement, along with digital rectal examination, and Ultrasound (US) systematic biopsy for confirmation. Recently, multi-parametric Magnetic Resonance -MR- Imaging (mpMRI) has improved PCa diagnosis significantly, allowing for accurate, non-invasive detection of PCa lesions, and opening the door to MR-guided US biopsies that can directly target the lesions, as opposed to classical systematic biopsies which are expensive and error-prone. Targeted biopsies are usually done using very accessible transrectal US (TRUS) probes; it requires finding the correspondence between the pre-acquired prostate mp-MRI and the intraoperative TRUS image, a problem known as MR-TRUS registration or fusion. In this chapter, an automatic system for near-real-time MR-TRUS prostate registration will be developed and validated using prostate MR-TRUS pairs from 204 patients. The dense deformation field (DDF) transforming a patient's MR prostate points (along with marked lesions) to the corresponding TRUS prostate points are calculated using Coherent Point Drift (CPD) to match prostate surfaces, followed by a Finite Element Method (FEM) simulation to obtain mechanically plausible internal deformations. Then, a Convolutional Neural Network (CNN) is trained to directly predict the DDFs from MR-US pairs (along with the corresponding prostate masks, which are automatically segmented), attaining an almost perfect approximation to the CPD+FEM DDF while reaching near-real-time speeds.

Keywords: Registration; Magnetic Resonance; Ultrasound; Prostate; Finite Element Method; Coherent Point Drift; Convolutional Neural Network; Deep Learning

6.2 Introduction

Prostate cancer (PCa) is the most common malignancy in males in western countries, and the second by number of deaths (Bray et al., 2018). The standard diagnostic pathway consists in prostate-specific antigen (PSA) levels measurement, along with digital rectal examination, and Ultrasound (US) systematic biopsy for confirmation and cancer staging (Mottet et al., 2020). In the last decade, multi-parametric Magnetic Resonance Imaging (mpMRI) has become increasingly prevalent due to its ability to accurately detect most PCas in a non-invasive manner (Mehralivand et al., 2018). On one hand, compared to PSA's poor specificity, mpMRI reduces the need for biopsy for many patients (Schröder et al., 2009). On the other hand, it allows for targeted Magnetic Resonance (MR)-guided biopsies (as opposed to standard systematic biopsies) and/or focal ablation therapies, which are extremely attractive due to their low complication profile (Ahdoot et al., 2019).

In particular, systematic biopsies are expensive, since 20-30 samples must be collected and analyzed; they are also error-prone, as it is not unusual to miss significant PCa or to sample the less-aggressive part of the tumor, hence leading to flawed risk stratification for the patient (Campos-Fernandes et al., 2009; Kvåle et al., 2009). By contrast, MR-guided biopsies can directly target the lesions, hence needing much fewer samples, and have shown improved sensitivity and specificity for lesion detection (Marra et al., 2019).

Two main approaches to guided biopsies exist: in-bore mpMRI-guided MR biopsy and mpMRI-guided US biopsy. While the first method has shown significant increase in diagnostic yield as compared to systematic biopsies (Hoeks et al., 2012), it is prohibitively expensive for most medical institutions both in terms of time and cost (Hambrock et al., 2010). mpMRI-guided US biopsies represent a much more accessible alternative, which nonetheless requires solving a complex registration problem, i.e.: finding the full correspondence between the pre-acquired prostate mpMRI and the intraoperative US (TRUS) or MR image. Once this correspondence is found, it is possible to transfer lesion positions, marked by radiologists in mpMRI, to the US image, where unfortunately lesions lack contrast with respect to surrounding tissue, and cannot therefore be located directly (Kaplan et al., 2002).

MR-TRUS prostate registration can either be performed mentally by an expert urologist during the biopsy procedure -which is known as cognitive fusion, and has shown some contradictory results (Puech et al., 2013)-, manually -which is limited to rigid registration and is physician-dependent, time-consuming and irreproducible-, or computationally through some sort of registration method. This last alternative is currently the focus of most research effort, as it has the potential for real-time, fully automatic, reproducible, and highly accurate MR-TRUS fusion which could improve the diagnostic yield of prostate biopsies, while reducing time and costs.

In this chapter, an automatic system for near-real-time MR-TRUS prostate registration is developed and validated using 204 prostate MR-TRUS pairs. In summary, Ground Truth (GT) dense deformation fields (DDF) transforming a patient's prostate MR image to their corresponding US image are calculated using Coherent Point Drift (CPD) (Myronenko and Song, 2009) to match prostate surfaces and then Finite Element Method (FEM) to obtain mechanically-plausible internal deformations. Then, a Convolutional Neural Network (CNN) is trained to directly predict the DDFs from MR-US pairs (along with the corresponding prostate masks), attaining an almost perfect approximation while reaching near-real-time inference speeds.

6.3 State of the art

Over the years, several MR-US prostate registration methods have been proposed. Depending on the output of the method, i.e. the MR-TRUS transformation parameters, there are rigid (Yan et al., 2018) and non-rigid methods (Fu et al., 2021). While true prostate deformations are obviously non-rigid, such transformations require the estimation of many parameters, making it a significantly more complex task.

Depending on the input, there are surface-based (van de Ven et al., 2015), some enforcing mechanically-compatible deformations (Marami et al., 2015), and intensity-based methods (Haskins et al., 2019). While the former mostly exploit the similarity in prostate shape between modalities, the latter use the fact that both MR and TRUS images emerge from the same underlying structures and hence a relationship between their voxel intensities should exist. Although true in theory, TRUS images contain many elements that do not appear in their corresponding MR image, such as US artifacts and shadows, the TRUS probe itself, and possibly the urethral catheter and bladder balloon, the latter of which also induce large deformations in the prostate both from within and from without. Furthermore, TRUS is known to provide very limited intraprostatic anatomical information other than the sparse calcifications and cysts (Fu et al., 2021). On the other hand, surface-based models rely on accurate prostate segmentations, which are not always easily obtainable, and TRUS imaging quality shortcomings have been reported to result in high inter-expert prostate segmentation variability (Smith et al., 2007).

Regarding DL techniques, different algorithms have been employed for varying purposes, such as to exploit the particularities of the MR-TRUS problem to directly predict a non-rigid DDF through a weakly supervised training scheme (i.e. when GT DDFs are not available and only image similarity metrics can be employed) (Hu et al., 2018), to estimate rigid transformation parameters in a supervised manner (i.e., when true transformation parameters are available) (Song et al., 2021), to estimate a differentiable registration error so as to later use for optimizing transformation parameters (Haskins et al., 2019; Czolbe et al., 2021), or to more accurately and efficiently approximate the Mutual Information (MI) metric (Belghazi et al., 2018; Nan et al., 2020), a non-linear generalization of cross-correlation often used in multi-modal medical image registration (Maes et al., 2015). As compared to classical optimization-based registration, DL algorithms trained to predict the transformation parameters have the advantage of being extremely quick in inference, while also being able to exploit dataset-specific information.

As stated, biomechanical constrains have often been used in the literature to regularize surface-based MR-TRUS prostate registration. This approach is supported by several studies showing the feasibility to fully simulate the mechanical behavior of pelvic organs in the context of prostate radiotherapy with low relative errors of 6-8% (Boubaker et al., 2009, 2015). Some authors have proposed to build FEM-based statistical shape models (SSMs) of plausible prostate deformations (Hu et al., 2008, 2012, 2015). For instance, Hu et al. (2012) performed many patient-specific simulations of the mechanical behavior of the prostate gland and surrounding organs, by randomly changing TRUS probe position and orientation, as well as the mechanical properties of involved tissues; then a patient's SSM was built by applying principal component analysis (PCA) to the resulting deformations. Finally, when the TRUS image is acquired, the observed deformation is matched to the simulation SSM, hence allowing to recover the full displacement field.

Similarly, several authors have proposed to extend prostate surface point set registration methods to be FEM-constrained, either by incorporating these constraints directly into the optimization algorithm (Khallaghi et al., 2015a,b), or by using FEM to find within-prostate deformation after performing surface registration (van de Ven et al., 2015; Fu et al., 2021). As an example of the former, Khallaghi et al. (2015b) proposed a modification to the regularization term of the CPD point set registration method (Myronenko and Song, 2009) so that instead of encouraging coherent motion of close points, it encouraged biomechanically compatible prostate deformations; this was achieved by adding the prostate FEM strain energy multiplied by a Tikhonov weight to the CPD objective function. Regarding the latter set of methods, van de Ven et al. (2015) aligned the centroid of both MR and TRUS prostate centroids and used radial projections to find surface points correspondences; those were then employed as constrains to a FEM-based simulation to obtain within prostate displacements. Fu et al. (2021) used a similar approach, by first finding prostate surface correspondences using a variation of the Iterative Closest Point algorithm (Audenaert et al., 2019), and then using those as boundary conditions to a FEM problem to obtain the internal displacement field; finally a point cloud neural network was used to mimic all previous registration steps while lowering inference times significantly.

This chapter proposes to train a CNN on DDFs obtained from mechanically-compatible FEM simulations, hence achieving realistic prostate deformations while attaining very quick inference times. This is very similar to the previously discussed approach by Fu et al. (2021), but employing a CNN instead of a point cloud network, which has the advantage of being able to directly consume imaging data to further improve registration performance and efficiency.

6.4 Materials and methods

Once data was preprocesed (Section 6.4.1), the proposed method could be divided in the following steps: First, the prostate gland was automatically segmented both in MR and TRUS images (Section 6.4.2). Second, prostate segmentations were meshed and MR prostate surface points were non-rigidly registered to US prostate surface points using CPD, hence obtaining the displacement field of the prostate surface (Section 6.4.3). Third, this field was used as boundary condition for a FEM-based simulation of internal prostate displacements (Section 6.4.4). Fourth, a CNN was trained on top of the previously generated data to predict the DDFs directly from MR and US images, along with their corresponding prostate masks, hence foregoing the need for meshing, surface registration, and FEM simulation completely, and attaining near-real-time inference speeds (Section 6.4.5). The complete pipeline is illustrated in Figure 6.1.

6.4.1 Data preprocessing

We employed prostate MR / TRUS pairs from 204 patients scheduled for prostate biopsy (both systematic and guided) from Valencian Institute of Oncology (IVO). Most MR images (96%) came from a 1.5T General Electric scanners, while the rest originated from Philips (3%, 1.5-3T) and Siemens (1%, 1.5T) scanners. The median voxel spacing was 0.47mm, 0.47mm and 4mm along the x, y and z dimensions respectively. TRUS

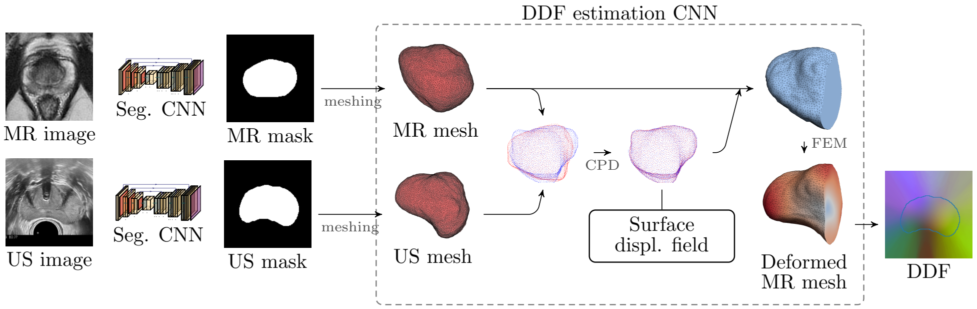


Figure 6.1: Proposed pipeline. Once the DDFs have been obtained for all patients, the elements within the dashed rectangle are replaced by a CNN trained to estimate them. MR: Magnetic resonance, US: Ultrasound, Seg. CNN: Segmentation convolutional neural network, FEM: Finite Element Method, CPD: Coherent Point Drift, DDF: Dense deformation field.

images were acquired with a Hitachi scanner with a median isotropic voxel spacing of 0.33mm. All MR and US images were resampled to a common isotropic voxel spacing of 0.5mm, and a size of 160^3 voxels (i.e. 80mm in every dimension). All images were translated so that the centroid of the prostate (obtained by automatic segmentation, see Section 6.4.2) was located in the centroid of the resulting image, i.e. at coordinates [80, 80, 80] in terms of voxels. This can be seen as translational pre-registration, as the centroids of all MR-US prostate pairs now coincided.

For validation purposes, landmarks (also called fiducials) that were recognizable both in MR and US prostates were manually marked with help of a radiologist with seven years of experience in prostate cancer imaging. For every prostate, two anatomical landmarks were identified: the points where the urethra meets the prostate, e.g. at its base and apex; as well as a variable number of histopathological landmarks (1.57 per prostate on average), e.g. cysts and calcifications that could be matched within modalities. Despite our best efforts, a precise identification of such fiducials was extremely challenging.

6.4.2 Prostate segmentation with Convolutional Neural Networks

The 3-dimensional (3D) CNN model described in Pellicer-Valero et al. (2021) was employed for segmenting the prostate automatically in MR and TRUS.

6.4.2.1 Overview of Convolutional Neural Networks

As a very brief introduction, CNNs are a kind of DL algorithm comprised of a stack of convolutional filters and non-linear activation functions, wherein the filter parameters are learned by stochastic gradient descent. CNNs have been extensively used for all kinds of image processing tasks, achieving state of the art results in virtually all of them. Due to their convolutional architecture, they are very efficient in terms of parameters (as compared to fully connected neural networks) and have inherent translational invariance (Lecun and Bengio, 1995).

For problems where both input and output are images, such as segmentation or DDF estimation, the U-net architecture (Ronneberger et al., 2015) (or one of its many variants) is predominantly employed. It features an encoder-decoder design with skip connections that forward information at several stages, hence allowing high resolution low-level features (from the first stages) and low-resolution high-level features (from the last stages) to be used jointly for predicting the output (Figure 6.2). All DL algorithms are trained using some variant of the stochastic gradient descent optimizer, which works by iteratively pushing the model parameters θ a small amount μ in the direction opposite to the gradient of the loss function J with respect to the parameters (Equation 6.1). Thanks to the openly available auto-grad frameworks, which automatically calculate the gradients of the loss function with respect to any parameter, such as Tensorflow (Abadi et al., 2016) (employed for this segmentation step) and Pytorch (Paszke et al., 2019) (employed for the DDF prediction step, see section 6.4.5), almost any imaginable architecture and loss function can be easily trained as long as everything is kept differentiable.

$$\text{GD}: \ \theta \leftarrow \theta - \mu \frac{\delta J(\theta, x, y)}{\delta \theta} \tag{6.1}$$

6.4.2.2 Transrectal ultrasound image segmentation

During the preprocessing stages, we observed that some TRUS images presented missing slices either near the base, the apex, or both, undermining the usability of such prostates for surface-based registration, as shapes in both modalities would be radically different due to this acquisition oversight. Since this seems to be a common problem in clinical practice, a modification to the standard training procedure of the TRUS prostate segmentation CNN described in Pellicer-Valero et al. (2021) was employed. First, during training, blank regions (where no image had been captured) were masked out from the loss, so that the CNN could predict anything within those regions without penalty. Then, the following augmentation was added: during training, a random amount of up to 20 slices was removed from both the start and the end of the TRUS image (not counting blank slices), hence forcing the CNN to "imagine" what the prostate would most likely look like for those removed slices. With those two simple modifications, the CNN was able to learn to adequately reconstruct missing prostate slices in TRUS by design, thus adding robustness to this kind of acquisition issue. Figure 6.3 shows an example of a prostate segmentation where some slices had to be reconstructed by the trained model.

6.4.2.3 Magnetic resonance image segmentation

Magnetic resonance images were segmented by directly applying the model described by Pellicer-Valero et al. (2021), along with a postprocessing technique known as neural

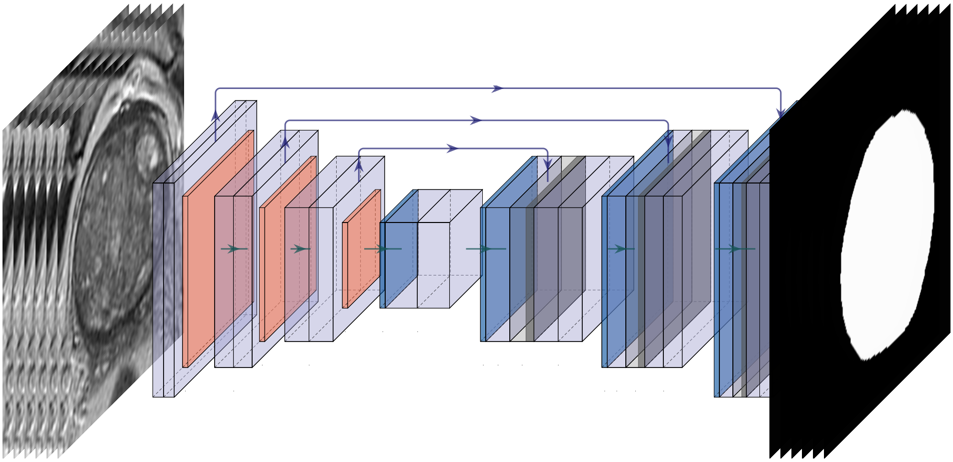


Figure 6.2: 3D U-net CNN for MR prostate segmentation. The U-net takes a 3D MR image as input, and forwards it through a series of convolutional filters and non-linear activation functions. In the encoder, the image resolution is progressively reduced as the number of channels increases. In the decoder, the resolution increases and the number of channels decreases. Skip connections (above arrows) connect same-resolution levels from encoder to decoder, so as to improve the spatial accuracy of the output. 3D: Three-dimensional, CNN: Convolutional neural network, MR: Magnetic resonance

resolution enhancement (Pellicer-Valero et al., 2020a). This method works by shifting the input image by several different sub-voxel amounts, feeding those transformed images to a standard segmentation CNN in order to obtain the corresponding segmentation masks, and then combining them into a single final high resolution mask. For our purposes, this allowed for a six-fold resolution increase along the z-axis, hence correcting the z-anisotropy for the MR prostate segmentation masks, and potentially improving registration accuracy.

6.4.3 Meshing and surface registration with Coherent Point Drift

Prior to further processing steps, high quality meshes were obtained from MR and TRUS prostate segmentation masks by using TetGen (Si, 2010). Final MR meshes contained (mean ± standard deviation) $247,321.7 \pm 122,735.2$ tetrahedral elements and $6,260.9 \pm 2,004.1$ surface vertices; US meshes contained $232,466.7 \pm 105,970.0$ tetrahedral elements and $6,226.1 \pm 1,879.9$ surface vertices.

Surface vertices of the MR prostate mesh were then registered to the US prostate surface vertices using a probabilistic point set registration algorithm known as CPD (Myronenko and Song, 2009). CPD works by solving a probability density estimation problem, wherein a set of moving points (Y, the M MR prostate surface points)

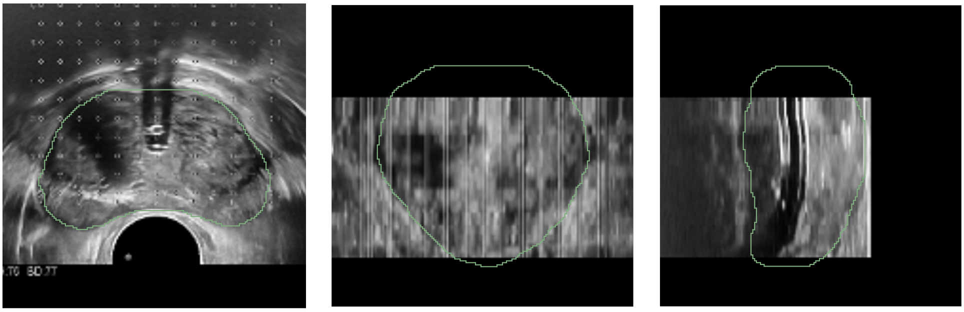


Figure 6.3: Prostate TRUS image with corresponding automatic prostate segmentation overlaid. Notice that, even if the whole prostate was not properly captured at both ends, its shape has been inferred by the segmentation CNN. Views (from left to right): axial, coronal and sagittal. TRUS: transrectal ultrasound, CNN: Convolutional neural network.

make up the centroids of a Gaussian Mixture Model (GMM), and a set of fixed points (X, the N TRUS prostate surface points) represent GMM observations (with some uniform noise of magnitude ω, the same for all of them). The negative log-likelihood E (see Equations 6.2-6.5) of the observations belonging to the GMM is then minimized with respect to the transformation parameters θ and the variance of the Gaussians σ^2 (the same for all of them) by means of the iterative Expectation-Maximization algorithm.

$$p(x) = \sum_{m}^{M+1} p(m)p(x|m) \tag{6.2}$$

$$p(m) = \frac{1-\omega}{M} \text{ if } m \leq M, \text{else } \omega \tag{6.3}$$

$$p(x|m) = \mathcal{N}(x|y_m, \sigma^2) \text{ if } m \leq M, \text{else } \frac{1}{N} \tag{6.4}$$

$$E(\theta, \sigma^2) = -\sum_{n}^{N} \log \sum_{m}^{M+1} p(m)p(x|m) \tag{6.5}$$

During training, Gaussian centroids Y move according to the transformation defined by θ in such a way that the likelihood of the observations X is maximized. Rigid, affine, and non-rigid variants of the algorithm exist. Very informally, the non-rigid version builds a "translation proposal" for each GMM centroid as the sum of all vectors from that centroid to all observations weighted by the likelihood of that centroid given every observation $p(m|x)$; then, the resulting transformation field is regularized so that nearby points (within a Gaussian-weighted neighborhood) are only allowed to move coherently. Non-rigid CPD takes two new parameters: β, which corresponds to the width of the smoothing Gaussian filter that enforces coherence, and γ, which represents the trade-off between the goodness of maximum likelihood fit and regularization.

For our problem, CPD was employed to obtain the prostate surface transformation field. First, rigid CPD was applied, followed by non-rigid CPD with parameters $\omega = 0, \beta = 100, \gamma = 3.3$. Figure 6.4 shows the registration progress of a patients' prostate over several EM iterations.

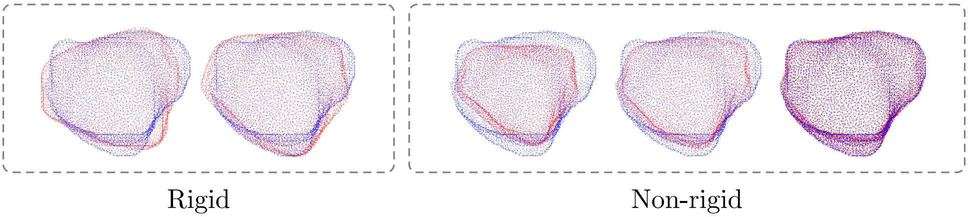


Figure 6.4: Prostate MR to TRUS surface point set registration (first rigid and later non-rigid) employing the CPD algorithm. Every image corresponds with an iteration of the Expectation-Maximization algorithm (only a few iterations were included). MR: Magnetic resonance, TRUS: Transrectal ultrasound, CPD: Coherent Point Drift.

6.4.4 Mechanical simulation with Finite Element Method

After the previous step, CPD had provided a discrete displacement field transforming the MR prostate mesh surface points into US prostate mesh surface points. This nodal displacement field was employed as a boundary condition to a FEM simulation of the mechanical behavior of the whole MR prostate mesh. This approach allowed us to obtain a mechanically-compatible displacement field within the prostate, as opposed to simpler extrapolation.

Shortly summarized, FEM is numerical method for finding approximate solutions to a particular field (in our case, the internal prostate displacement field u) over a geometry that has been discretized in a set of finite elements (a MR prostate mesh), given some boundary conditions (u in the prostate surface). Thanks to the discretization, u must only be found for a finite set of locations $u_0, u_1, ...$ (at element vertices), as the elements act as local interpolators that allow us to compute u at any given point, as well as its spatial derivative and integral.

In the case of a mechanical problem that involves linear elastic behavior, the application of this method can be described as follows: For an element e with nodal displacements u^e_n and shape functions (local interpolators) N^e, the displacement field within e can be approximated as: $u^e \approx \tilde{u}^e = N^e \cdot u^e_n$. Likewise, element strains ϵ^e can be found by applying the differential operator L to $\tilde{u}^e$: $\epsilon^e = L \cdot \tilde{u}^e = B^e \cdot u^e_n$, with $B^e = N^e \cdot L^e$. The element stiffness matrix is defined as $K^e = \int_v B^{eT} C B^e dV$, where C is the material stiffness matrix that relates stresses to strains within the element according to material's properties. The K^e from all elements can be assembled into a global stiffness matrix K that is related to all nodal displacements U and all nodal forces F.

Finally, we can define the total potential energy of the mechanical system in Equation 6.6, where $W_s = \frac{1}{2}U^T KU$ is the strain energy of the system, and $W_p = U^T F$ is the work potential, both expressed with respect to the approximate nodal displacement field U. The equilibrium will be reached when the total potential energy (Π_p) is minimum, hence by taking the derivative of Π_p and equating it to 0, Equation 6.7 appears, which is a simple linear system that, once solved, allows us to obtain U, as well as $u \approx \tilde{u}$ through element interpolation.

$$\Pi_p = W_s - W_p \tag{6.6}$$

$$\frac{\delta \Pi_p(U)}{\delta(U)} = 0 \rightarrow KU = F \tag{6.7}$$

When a body undergoes large deformations, or when the stress-strain relationship is otherwise no longer linear, isotropic and incompressible, different formulations for W_s must be found. In particular, the neo-Hookean material model still assumes perfect elasticity, but incorporates the non-linearities stemming from large deformations. For neo-Hookean materials strain energy W_s is defined as in Equation 6.8, where μ and λ are material properties and I_1 and J are invariants. In particular, $I_1 = \text{trace}(FF^T)$ and $J = \det(F)$, where F is the deformation gradient tensor: $F_{ij} = \frac{\delta x_i}{\delta X_j}$, with X being a material point position and x the transformed material point's position (hence $u = x - X$).

$$W_s = \frac{\mu}{2}(I_1 - 3) - \mu \ln J + \frac{\lambda}{2}(\ln J)^2 \tag{6.8}$$

λ is the first Lamé parameter and μ is the shear modulus or the second Lamé parameter. Both can be computed from the Young modulus (E) and Poisson coefficient (ν) using equivalences 6.9 & 6.10.

$$\lambda = \frac{E \cdot \nu}{(1+\nu) \cdot (1-2\nu)} \tag{6.9}$$

$$\mu = \frac{E}{2 \cdot (1+\nu)} \tag{6.10}$$

For the purposes of simulating the mechanical behavior of the prostate, different authors propose an array of mechanical properties, with Young's moduli ranging from 2-60kPa (Kemper et al., 2004; Boubaker et al., 2009, 2015). Following the constitutive model used by the recent publication by Fu et al. (2021), which pursued a very similar path to ours, the prostate was assumed to behave according to a neo-Hookean model with a Poisson's ratio $\nu = 0.49$ and a Young's modulus $E = 5kPa$ (quite compressible). Khallaghi et al. (2015b) used similar properties as well.

FEBio 3.5.1 (Maas et al., 2012) was employed for finding the FEM solution to the displacement field within the prostate. In particular, the complete problem (geometry, boundary conditions, mechanical properties, etc.) was defined within Python and saved as a .feb file. Then, the FEBio solver was called on this file and the nodal displacements were recovered from the simulation logs. Surface deformations were applied in five steps, with a magnitude that increased by a fifth of the total displacement at each step, so as to facilitate convergence. Figure 6.5 shows the magnitude of the displacement field over the five simulation steps for an arbitrary patient.

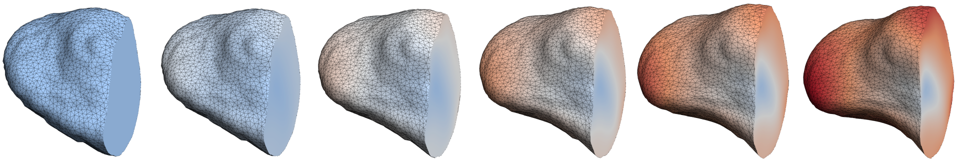


Figure 6.5: Prostate before deformation (leftmost) and after applying the surface displacement field in five FEM simulation steps. Color indicates the displacement magnitude, ranging from blue (0mm) to red (10mm) FEM: Finite element method.

6.4.5 Dense Deformation Field estimation with a Convolutional Neural Network

The final step was to train a U-net CNN to predict the displacement field directly from MR and US images (along with their corresponding prostate masks, see Figure 6.6), hence foregoing the need for meshing, surface registration, and FEM simulation, once trained.

6.4.5.1 Dense Deformation Field definition

A dense deformation field or DDF was built from the obtained displacement fields. A DDF is just an image with the same size (160 voxels per side) and spacing (0.5mm, isotropic) as the preprocessed MR image, whose values at any given position represent the displacement -in x, y and z directions- that the MR image should undergo to match the US image. Since the DDF has three channels -for x, y and z displacements-, a slice of it can be plotted as a color image (see Figures 6.1 or 6.6 for an example). Also, since deformations outside the prostate are of little interest to our purposes, those values in the DDF were just linearly extrapolated from the available prostate displacements. Finally, the resulting DDF was inverted by using SimpleITK (Lowekamp et al., 2013), which is an implementation detail of how DDFs are typically used; i.e. to build the transformed image, sampling is performed form the output image towards the input image (US to MR) so that interpolation is performed in the input space, where information is more complete.

6.4.5.2 Convolutional Neural Network architecture

Once the DDFs were obtained for all patients, a U-Net-like CNN was trained to predict them when given only MR and US images and their corresponding prostate masks as input. In particular, the registration model was implemented as a Pytorch Lightning (Falcon et al., 2019) module, using the V-net (Milletari et al., 2016) architecture, as implemented in the VNetLight module from MedicalZooPytorch (Nikolaos, 2019). Only two modifications were made to this module: batch normalization was replaced by instance normalization (i.e. with instance normalization, standarization is applied instance-wise and channel-wise, instead of being batch-wide), and the PReLU activation function was used (which is just like a ReLU, but the slope c is learned channel-wise: $\text{PReLU}(x) = 0$ if $x \leq 0$, else cx).

Input images were downscaled to a size of 120^3 voxels and a spacing of $0.\widehat{6}$mm, and were normalized in such a way that all values below the $0.1\%^{\text{th}}$ percentile and above the $99.9\%^{\text{th}}$ percentile were cut off, and the rest were rescaled to the range [0,1]. While this downscaling represented a 25% loss in resolution, it resulted in an almost 2.4× decrease in Graphics Computing Unit (GPU) memory requirements, which enabled training with a batch size of three instead of just one. The batch size is the number of images that are simultaneously fed to the CNN every training step, and over which the gradient is averaged when training with gradient descent. Thus, this 25% resolution reduction resulted in a three-fold increase in training speed and in less noisy gradients (although gradient accumulation is also straightforward to implement and would address this second issue as well).

6.4.5.3 Data partitioning

Data was partitioned into three subsets: train (75% of the data, N=153), validation (10%, N=21) and test (15%, N=30). The CNN was trained with different hyperparameters in the training set, and the results were evaluated in the validation set, while the test set was kept secret until the very end, and was only used for reporting the final results. The model was trained with a constant learning rate of 0.001 until validation score did not improve for 19 consecutive epochs (epochs are runs over the whole training dataset); the final model took 509 epochs to converge.

6.4.5.4 Loss functions and training

Three loss functions and a regularization term were combined by weighted sum (with all weights being one, except for the DSC weight, which was 0.5) to train this network (see Figure 6.6). For the following loss equations, image axes $[\text{channel}, x, y, z]$ will be indexed with indices $[c, i, j, k]$ respectively:

mMSE loss: Masked Mean Square Error loss between the predicted DDF (pDDF) and the GT DDF (or just DDF). Both DDFs were masked by multiplying them by a Gaussian-filtered version of the MR prostate mask, m, so that regions far from the prostate (for which GT DDF was just linearly extrapolated) have a rapidly decreasing

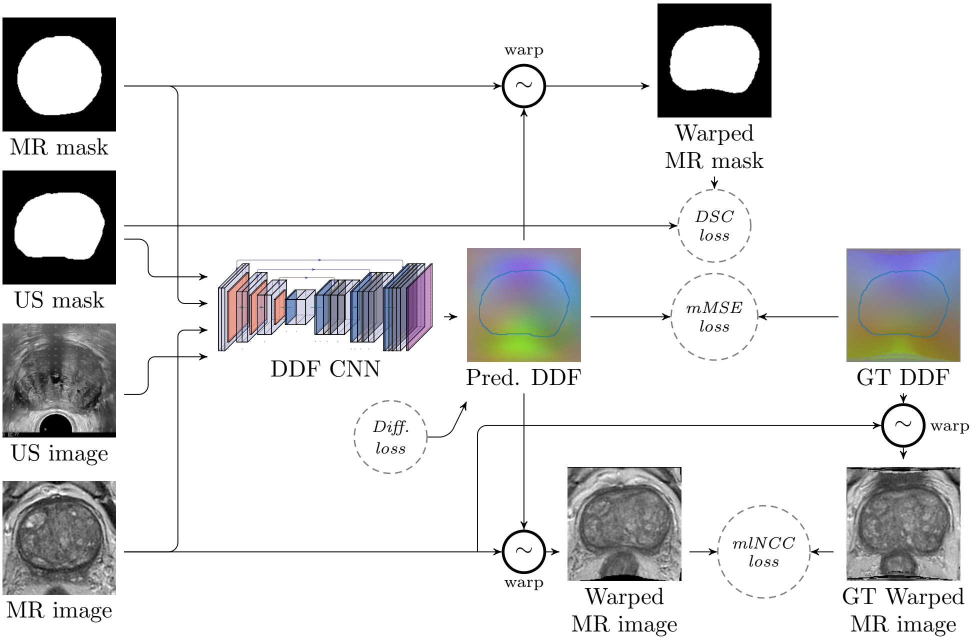


Figure 6.6: Overview of the DDF prediction CNN, along with all its loss functions, which are eventually combined by weighted sum. MR: Magnetic resonance, US: Ultrasound, DDF: Dense deformation field, CNN: Convolutional neural network, GT: Ground truth, Diff.: Diffusion, DSC: Dice similarity coefficient, mMSE: Masked mean square error, mlNCC: masked local normalized cross-correlation.

impact on the loss. In summary, this loss simply encourages the pDDF to be as similar as possible to the DDF:

$$\text{mMSE}(DDF, pDDF, m) = \underset{cijk}{\text{mean}} \left(DDF_{cijk} \cdot m_{1ijk} - pDDF_{ijk} \cdot m_{1ijk} \right)^2 \tag{6.11}$$

Diffusion loss: Diffusion regularization term to discourage high-frequency terms in the pDDF. It is simply the mean square norm of the spatial gradient of the pDDF, implemented using finite differences:

$$\text{Diff. loss}(pDDF) = \underset{cijk}{\text{mean}} \sum_{a=\{i',j',k'\}} \left(\frac{\delta pDDF_{cijk}(a)}{\delta a} \right)^2 \tag{6.12}$$

DSC: Dice Similarity Coefficient loss between the US mask m and the MR mask transformed according to the pDDF pm. I.e., a fully differentiable warping module within the model uses the pDDF to transform the MR mask into the warped MR mask pm. This loss encourages good MR-US surface registration, since it is minimized as the overlap between the warped MR mask and the US mask increases:

$$\text{DSC loss}(pm, m) = 1 - \text{DSC}(pm, m) = 1 - \frac{2 \cdot \sum_{cijk}(pm_{cijk} \cdot m_{cijk})}{\sum_{cijk} pm_{cijk} + \sum_{cijk} m_{cijk}} \tag{6.13}$$

mlNCC loss: Masked Local Normalized Cross-Correlation loss between the MR image warped according to the DDF (I) and the same MR image warped according to the pDDF (pI). It is a masked loss, just like the previous mMSE loss, so as to ignore DDF values that are far from the prostate. It is also local because the normalized cross-correlation is only computed within a $9 \times 9 \times 9$ window and then averaged over the whole image. In summary, it encourages the DDF-transformed MR images to be similar to the pDDF-transformed MR images. Equation 6.14 shows the definition of masked NCC (mNCC), which should be applied over $9 \times 9 \times 9$ image patches and averaged over the whole image. Further discussion on these losses can be found in the VoxelMorph paper (Balakrishnan et al., 2019).

$$\begin{aligned} \text{mNCC loss}(I, pI, m) = 1 - \text{mNCC}(I, pI, m) = \\ 1 - \text{mean}_{cijk} \frac{((I_{cijk} - \text{mean}(I)) \cdot (pI_{cijk} - \text{mean}(pI)))^2 \cdot m_{cijk}}{\text{var}(I \cdot m) \cdot \text{var}(pI \cdot m)} \end{aligned} \tag{6.14}$$

6.4.5.5 Metrics

Several additional metrics will be employed to evaluate the resulting registrations. Shortly summarized:

- **MI**: Mutual Information, a metric derived form information theory that can be understood as a non-linear generalization of cross-correlation. It measures the dependence between A and B (in our case between the transformed MR image and the TRUS) as the distance between the joint distribution $p_{AB}(a, b)$ and the distribution associated to the case of complete independence $p_A(a) \cdot p_A(a)$, by means of the Kullback-Leibler measure. See Maes et al. (2015) for more information.

- **HD95**: 95th percentile Haussdorff distance of two surfaces measures the 95th percentile largest of the minimum distances from all points of one of the surfaces to the other.

- **ABD**: Average Boundary Distance represents the average of the minimum distances from all points of one of the surfaces to the other.

- **TRE**: Target Registration Error is the mean euclidean distance between corresponding MR and US landmarks (in mm). TRE was evaluated separately for all histopathological landmarks, the urethra intersection at prostate base and the urethra intersection at the apex.

6.5 Results

After training the DDF prediction model, it was evaluated in the test set (N=30 patients, 47 histopathological landmarks), which had been kept secret. Table 6.1 shows the resulting test set metrics for three situations: before non-rigid registration or (*Initial*), using the GT DDF (*CPD + FEM*), and using the trained CNN (*CNN*). Paired t-tests were performed against *Initial* reference results, with asterisks indicating significance level: p-value $\leq 0.05^*$, p-value $\leq 0.01^{**}$, p-value $\leq 0.001^{***}$.

Table 6.1: Registration metrics evaluated on the test set (N=30 patients, 47 hist. landmarks). Paired t-tests were performed against *Initial* reference results, with asterisks indicating significance level: p-value $\leq 0.05^*$, p-value $\leq 0.01^{**}$, p-value $\leq 0.001^{***}$. MI: Mutual information, DSC: Dice similarity coefficient, HD95: 95th percentile Haussdorff distance, ABD: Average boundary distance, TRE: Target registration error, Hist.: Histopathological.

		Surface metrics			*TRE (mm)*		
	MI	*DSC*	*HD95*	*ABD*	*Hist.*	*Apex*	*Base*
Initial	-0.0348	0.8657	5.3280	2.0183	4.2704	6.5448	5.4393
CPD + FEM	-0.0429^{**}	0.9837^{***}	1.0637^{***}	0.3225^{***}	3.6754^{*}	3.4671^{*}	5.7651
CNN	-0.0427^{**}	0.9743^{***}	1.1020^{***}	0.3958^{***}	3.5439^{**}	3.9272^{*}	5.6725

As can be seen, all metrics except for base TRE, improved significantly thanks to the non-rigid registration. Furthermore, the *CNN* predictions were able to emulate *CPD + FEM* almost perfectly, both achieving overall very similar metrics (*CNN* being sometimes even better). Regarding individual metrics, all surface-based metrics (DSC, HD95, and ABD) show a very significant improvement, which is expected since the whole registration process was guided exclusively by the surfaces. In fact, these values could be made arbitrarily better by reducing CPD regularization parameters (see Section 6.4.3), but this would hurt registration performance otherwise as long as segmented prostates are not perfectly segmented. The good surface registration performance shown by the CNN demonstrates it was able to learn the problem successfully.

On the other hand, MI is an intensity-based metric, which is interesting to analyze since none of the described methods (CPD, FEM or the CNN) were optimized for any intensity-based metric between transformed MR and US images. Table 6.1 shows that the performance of *CPD + FEM* and *CNN* in terms of MI is similar, and significantly superior to that of *Initial*.

Finally, histopathological and urethra apex TREs seem to improve significantly for either method, but not for urethra base TREs, which remain high or even worsen slightly. This might be related to the fact that many prostates were missing base slices (as seen in Figure 6.3), which had to be reconstructed (see Section 6.4.2.2); yet urethral landmarks were still marked approximately in such cases.

For a visual assessment of the results, Figure 6.7 shows the transformed MRs using the predicted DDF for a sample of four test patients (one within each dashed box). Despite MR and TRUS modalities being radically different, a continuity in shapes between both can be observed in the composite images. In all cases, and despite the different shapes of the prostate, the CNN seems to have learned to solve the registration problem successfully.

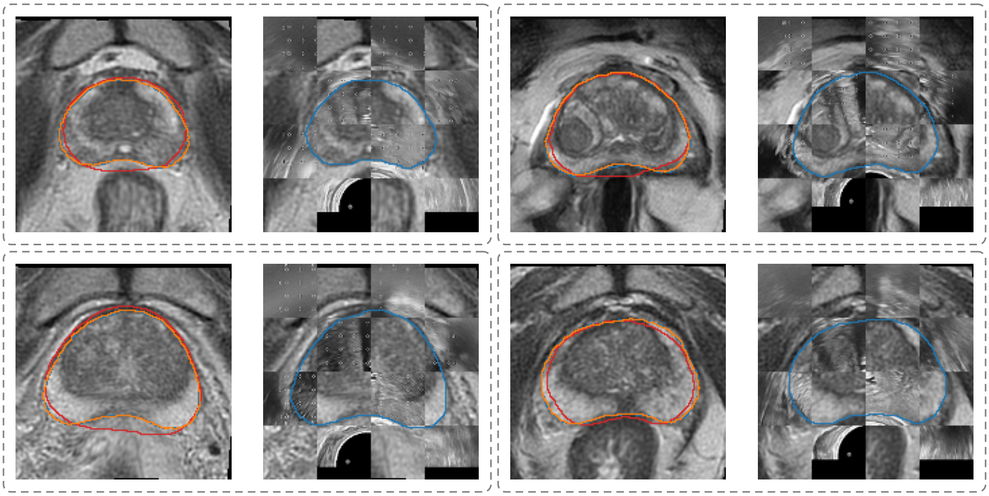


Figure 6.7: Transformed MRs using the predicted DDF for four test patients (one within each dashed box). Left image shows the deformed DDF along with the original MR prostate mask (in red) the US mask (in orange) and the transformed MR mask (in blue). Right image shows a checkerboard composite of the transformed MR image with the US image. MR: Magnetic resonance, US: Ultrasound, DDF: Dense deformation field.

6.6 Discussion

Comparing the results with other authors is not straightforward, as the datasets and validation methods employed are all very different. For instance, Hu et al. (2012) reported an excellent TRE of 2.40mm. Using 8 patients, they performed 500 FEM simulations per patient with different TRUS probe positions and orientations, and then matched those to the TRUS prostate surface. While reasonable, their approach requires segmenting pelvic structures other than prostate (pelvis, bladder, rectal wall, etc.), and performing 500 simulations per patient, which is extremely time-consuming and could only be achieved in the context of a long pre-planning stage, which is uncommon in clinical practice for prostate biopsy. The same authors would go on to solve this problem in a later paper (Hu et al., 2015), by employing a SSM, achieving similar TREs (2.42mm) and needing only around 20 seconds for inference, as opposed to several hours. Still only the same 8 patients were used.

Similarly, van de Ven et al. (2015) achieved a TRE of 2.76mm using 10 patients to develop and test their biomechanical model but did not try to accelerate it by means of DL. Khallaghi et al. (2015b) (similar to Khallaghi et al. (2015a)) used a very interesting FEM-regularized CPD algorithm to directly predict volumetric displacement fields, achieving a TRE of 2.6mm in 29 patients. Although a simpler constitutive model was employed, the method might be extended in a future to consider non-linear models, such as the neo-Hookean.

Fu et al. (2021) followed an approach very similar to ours, but employing a point cloud network instead of a CNN. They achieved DSC, ABD, HD95 and TRE metrics of 0.94, 0.90, 2.96, and 1.57mm, respectively, which are great results. Regarding validation, they built 450 training datasets from 50 available patients, and performed five-fold cross validation to test the network, although it is not specified whether simulations from a single patient could appear in several folds. Finally, Marami et al. (2015) used a state estimation framework to estimate the deformation of the prostate based on the intensity-based modality independent neighborhood descriptors metric (Heinrich et al., 2012), achieving an outstanding TRE of 1.87mm for pre-operative MR to intra-operative MR matching, which a much easier problem than MR-TRUS registration.

In general, the DDF prediction CNN trained on biomechanically-compatible prostate deformations was able to improve TRE significantly (except in the prostate base), while reaching very fast inference times. More precisely, the inference time for a single case was 193.65ms on average: 180.20ms of GPU (Nvidia Titan V) time and 13.45ms of CPU (Intel i7 9700k) time. Assuming similar times from the TRUS segmentation network and leaving some overhead for image interpolation and transformation steps, the total required time for performing segmentation + registration could be of around 500ms, i.e. 2Hz, or near real-time performance for a fully automatic system. No performance optimizations were made prior to measuring these times; hence, with better hardware and more optimization, this 2Hz figure will likely increase. Also, note that MR segmentation can be performed offline, and it therefore adds no computation time in an online scenario. This is the fastest non-rigid MR-TRUS registration model in the literature, the second quickest being Fu et al. (2021)'s, at a forward propagation time of 2-3 seconds, without counting other steps such as point cloud generation or segmentation.

The proposed method could already be easily implemented in clinical practice and opens the door to the possibility of real-time-guided prostate interventions, potentially improving accuracy thanks to its adaptability as prostate shape changes during an intervention, due to patient movement, probe movement, or blood inflow, among others. This work shows how CNNs are able to learn the MR-TRUS prostate registration problem when given GT DDFs, and future research lines should instead focus on trying to find the most accurate GT DDFs possible.

Funding

This work has been partially supported by a doctoral grant of the Spanish Ministry of Innovation and Science, with reference FPU17/01993.

Institutional review

The study was approved by the Ethical Committee of the València Institute of Oncology (CEIm-FIVO) with protocol code PROSTATEDL (2019-12) and date 17th of July, 2019.

Chapter 7

Main results and conclusion

This work set out with the goal of improving the quality of life of PCa suspect patients and of the clinicians caring for them through the use of AI by tackling two concrete objectives. The first one was solved by developing a MR prostate segmentation model (Chapter 2), and then using it for the development of a highly accurate fully automatic lesion detection, segmentation, and classification system (Chapter 4); the second objective was addressed by developing high resolution MR and US prostate segmentation models (Chapter 2 & 3) and using them for obtaining ground truth registration transforms by means of CPD and FEM, which were eventually employed for training a DL model able to speed up the process significantly (Chapters 5 & 6). In all cases, the final systems were fully automatic, were tested to perform on par or better than experts (when comparison was possible), and were extremely quick in inference (real- or near-real-time speeds). Increased inference speeds are of special importance in the MR-US registration model, allowing the registration to adapt as prostate shape changes due to patient movement, probe movement, or blood inflow, among others, hence improving the registration accuracy, which might otherwise degrade as the intervention progresses. The code for both objectives has been made publicly available at `https://github.com/OscarPellicer/Deep-Learning-in-Prostate-PhD`.

Section 7.1 will overview the contributions and novelties introduced by these five publications, both in terms of improvements to patient's well-being and clinician's workflow, and from a scientific-technical point of view, while Section 7.2 will delve into the general limitations of clinical AIs.

7.1 Contributions

In Chapter 2, fully automatic MR and a US prostate segmentation models are proposed, both achieving an excellent performance, with the MR model even outperforming expert radiologists. These models undercut or even eliminate the need for manual segmentation, which is known to require extensive experience and be very time-consuming, and ultimately suffers from high inter- and intra-expert variability. Furthermore, inference (i.e., generating a new segmentation) is extremely quick, which is of special interest for intraoperative US, which currently requires the urologist spending around 10 minutes in the middle of an operation; it can also help alleviate some much-needed time for radiologists. Finally, the models are shown to be robust, and should therefore perform well irrespective of the scanner or medical center where the images are acquired, which is essential should the model be deployed outside the medical center where it was developed.

From a scientific-technical point of view, the chapter proposes a new CNN architecture, along with several design and training choices that, in unison, help create robust, well performing segmentation models. The designed U-Net-like CNN architecture combines a DenseNet encoder, which is extremely efficient parameter-wise, with a ResNet decoder, and incorporates techniques such as checkpoint ensembling, cyclic learning rate, a heavy data augmentation routine, and a very varied training set, to achieve its performance and robustness. Interestingly, many of these techniques are nowadays relatively commonplace, but they were not well established when the model was under development. In general, no other single model was found to perform consistently well in several datasets simultaneously. Finally, neural resolution enhancement, a technique introduced in Chapter 3 is used here for the first time on a real model to successfully improve the resolution of the generated segmentation masks.

Chapter 3 presents a technique for improving the output resolution of segmentation CNNs, even beyond the original image's resolution. This is of special interest for MR or CT image segmentation, since they tend to have reduced resolution along one axis as compared to the other two, due to the slice-wise acquisition procedure, leading to problematic anisotropic voxel sizes. High-resolution segmentations can improve the precision of later tasks, such as registration or simulation of the biomechanical behavior.

Technically speaking, the method is very straightforward and can be applied to any already-trained segmentation CNN. It leverages interpolation in the space of the original input image, where information is still complete, instead of doing so in the much less informative discretized output space. Additionally, it exploits the contextual knowledge of the CNN about the particular segmentation task, arguably improving the results even further. All interpolators used currently in practice disregard this information, which is nonetheless available and could easily lead to more precise results.

Chapter 4 presents a fully automatic model for PCa lesion detection, segmentation, and classification that is shown to perform above expert radiologists for clinically significant (Gleason Grade Group ≥ 2) lesion detection. AI-based prostate mpMRI interpretation has many potential use cases, the most obvious perhaps being a second opinion for assisting radiologists, and reducing the risk of missing clinically significant lesions. It could also be used by radiologists as a criterion for patient prioritization, by sorting the patients according to AI-assessed risk, and hence allowing them to focus on the most urgent cases first. Finally, it could be used to develop viable population-wide screening programs, by employing an AI that automatically refers the patient under the slightest suspicion.

Several novelties are introduced. Firstly, it is the first fully automatic framework to perform this task (to the author's knowledge). It leverages a proper detection network, the Retina U-Net, as compared with the standard approach of using segmentation followed by post-processing steps to obtain the independent lesions. It also makes use of the previously developed MR prostate segmentation model and extends it in a cascading setup to also distinguish between central gland and peripheral zone;

these segmentation information is extremely useful for the detection model, as the appearance and likelihood of lesions differ between zones. An automatic procedure employing Mutual Information and spatial gradient features is also proposed for the non-trivial task of mpMRI sequence registration.

In Chapter 5, a method for simulating the biomechanical behavior of the liver (or any organ) in real-time is proposed. It constitutes a huge leap forward in the implementation of applications such as surgical simulators, computed assisted surgery, or guided tumor ablation. The success of this approach opens the doors to further research in FEM acceleration through DL, and leads directly to the developments in Chapter 6.

When it was published, the employed approach went beyond existing research by allowing the use of any liver as input to the trained DL model, instead of it being limited to a single liver geometry. This was achieved by parameterizing the shape of an arbitrary liver, so that it could be fed to a standard neural network, making it shape-aware. The resulting DL model was proven highly accurate and extremely quick (above 100Hz).

Lastly, Chapter 6 presents an automatic system for non-rigid MR-US prostate registration that improves significantly with respect to baseline rigid registration. Most importantly, the system works in near real-time, which opens the door to real-time guided prostate interventions, potentially raising accuracy by letting the registration adapt to the changing shape of the gland, either due to patient movements, probe motion, or blood accumulation, among other factors.

The main novelty here lies in this being the quickest model ever proposed for MR-US prostate registration. This is achieved by first generating a set of biomechanically compatible registration transformations (using CPD for surface registration and FEM for computing the displacements within the gland), and then training a U-Net-like CNN to predict the final transformation directly from the input images, hence skipping all time-consuming intermediate steps.

7.2 Limitations and further work

Although this work contributes significantly to the existing body of knowledge, some limitations still apply. Since specific limitations are deeply discussed in each of the chapters, the next paragraphs will summarize the most crucial and general ones:

Although CNNs have shown time and again their potential in image processing tasks, they are ultimately black box models, meaning that their inner workings are not understood. As such, there are no guarantees that a CNN will always perform as expected (specially when there is a domain shift from training to real-world data, as is commonly the case), and CNN predictions are generally not explainable. This is specially problematic in a clinical scenario, where wrong decisions taken by an AI might have a huge impact on the life of a patient, and legal accountability might require elucidating the reasons that lead to a specific AI behavior. Furthermore, the domain shift problem is specially notable for medical CNNs, which typically underperform when used on images from significantly different scanners than those on which they were trained.

On one hand, regarding explainability, complex models performing complex tasks (such as cancer-detecting CNNs) will arguably never be fully explainable to humans, because only models that are themselves simple, or models that can be divided into much simpler constitutive elements are. Unfortunately, similar to the human brain, CNNs are complex structures that work by the compound interaction of many parts, and are therefore complex to analyze. Still, progress is being made in this regard, and hopefully soon DL architectures will be explainable to a satisfactory level. On the other hand, the counterpoint to the unpredictable performance of CNNs are the patients who might be already benefiting from AI-based applications, but are not due to regulatory and legal fears. AI systems are not perfect but, in contrast to humans, medical AI systems can and should be thoroughly validated and scrutinized before clinical use. For most applications though, the combination of AI and humans will likely yield the best results, using AI as a second clinical opinion, a prioritization tool or, for very well performing AIs, as a fully autonomous screening referral system. Careful assessment must take place, but without forgetting the costs of delaying medical AI deployment either, which might oftentimes outweigh the risks.

A final limitation, that is perhaps simultaneously the most important future line of work, is the transfer of research results to an actual product that can directly influence patient's lives. Despite the indisputable scientific value, research work with such an obvious and immediate application such as this calls for a practical implementation, and cannot be considered fully realized otherwise. Further work should therefore focus on finding the best way of bringing these projects into practice (such as making the code freely available), so that they can directly materialize the reason for which they were developed in the fist place: improving the life of the patients.

Chapter 8

Resumen amplio en castellano

8.1 Objetivos

En este trabajo se utilizarán algoritmos de inteligencia artificial (IA) para mejorar la calidad de vida de los pacientes con sospecha de cáncer de próstata (CaP), y de los clínicos que los atienden, mejorando la eficiencia y precisión en el diagnóstico, biopsia y tratamiento (como la braquiterapia de alta tasa) de esta patología. Todo ello se conseguirá a través de dos objetivos secundarios:

a) Desarrollo de un sistema de detección, segmentación y clasificación de lesiones de PCa en Resonancia Magnética (RM) multiparamétrica (RMmp). Para ello será necesario crear previamente un modelo de segmentación de próstata sobre RM.

b) Desarrollo de un sistema de registro de próstata entre RM y ultrasonidos (US). Esto requerirá la creación de un modelo de segmentación de la próstata para RM y US que sea capaz de producir máscaras de segmentación de alta resolución.

Todos los sistemas anteriores deberán ser totalmente automáticos (no requerir ninguna intervención), tener un rendimiento igual o mejor que el de los radiólogos expertos y ser muy rápidos de usar (en tiempo real o casi) para poder utilizarlos en la práctica diaria.

8.2 Introducción al problema médico

En 2020, el CaP fue la primera neoplasia por incidencia en la población masculina en Europa (Gandaglia et al., 2021), con un riesgo acumulado del 8,2% de ser diagnosticado antes de los 75 años, y un riesgo acumulado de muerte del 1% (Dyba et al., 2021). A pesar de ser el cáncer más frecuente en los hombres, sólo es el tercero por número de muertes (10% de todas las muertes relacionadas con el cáncer en los hombres), tras el cáncer colorrectal (12,3% del total de hombres) y el cáncer de pulmón (24,2%) (Dyba et al., 2021). De hecho, el 59% de los hombres mayores de 79 años que murieron por causas no relacionadas se encontró que tenían un PCa incidental tras necropsia (Bell et al., 2015). Esta discordancia entre incidencia y mortalidad se debe a la heterogénea agresividad de las lesiones de CaP, siendo generalmente de lenta evolución, así como a la efectividad de los tratamientos actuales. En cualquier caso, el CaP es una importante carga socioeconómica y sanitaria, y cualquier mejora en su diagnóstico, manejo o tratamiento supondrá sin duda un importante impacto positivo en la vida de millones de personas.

La vía clínica estándar para el diagnóstico del CaP suele consistir en mediciones periódicas del Antígeno Prostático Específico (PSA, del inglés *Prostate Specific Antigen*), una proteína producida por las células de la próstata y que se mide en el plasma, junto con los exámenes rectales digitales (ERD). El PSA suele ser producido en mayores cantidades por las células prostáticas malignas, por lo que una elevación de su concentración (por ejemplo, por encima de 4ng/mL) puede ser indicativa de un PCa. Sin embargo, muchos otros factores, como la hiperplasia prostática benigna o el agrandamiento de la próstata, también pueden elevar los niveles de PSA. Por lo tanto, aunque muy sensible, el PSA sigue siendo una prueba muy poco específica para el diagnóstico de CaP, con un valor predictivo positivo de sólo el 24% (es decir, sólo uno de cada cuatro hombres con niveles elevados de PSA tiene realmente un CaP) (Hugosson et al., 2019).

Tradicionalmente, los pacientes con niveles altos de PSA o ERD positivos se someten directamente a una biopsia de confirmación guiada por ultrasonidos (US), en la que se recogen entre 20 y 30 muestras de tejido de la próstata del paciente con el uso de agujas, y la agresividad de la lesión se analiza después mediante una cuidadosa evaluación histopatológica. A cada muestra se le asigna una puntuación de Gleason (GS, de 3 a 5) (Epstein et al., 2005) dependiendo del aspecto de las células, por ejemplo: a las células de aspecto normal se les asigna una GS más baja, es probable que crezcan lentamente y no sean muy agresivas, mientras que las células de aspecto muy anormal reciben una GS más alta y pueden ser extremadamente agresivas. En función de los dos GS más comunes detectados en una muestra, estos se clasifican a su vez en un sistema de graduación de 1 a 5 conocido como grado de la Sociedad Internacional de Patología Urológica (ISUP) (también conocido como Grupo de Grado de Gleason, GGG) (Epstein et al., 2016). La figura 1.1 muestra un ejemplo de varias muestras histológicas de próstata junto con su correspondiente GS y GGG.

En los últimos años, la introducción de la RM y, en particular, la RMmp previa a la biopsia ha cambiado drásticamente este paradigma. La RM es una técnica de imagen médica no invasiva y no ionizante que emplea campos magnéticos muy potentes (de 1,5 a 3T normalmente) para obtener una imagen tridimensional (3D) de las estructuras internas del cuerpo. Dependiendo del protocolo de adquisición (es decir, cómo, cuándo, dónde y durante cuánto tiempo se activan los campos magnéticos), se pueden obtener diferentes secuencias de RM, las cuales ponen de manifiesto distintas propiedades del mismo tejido subyacente; la combinación de varias de estas secuencias produce una RMmp. Las secuencias RMmp de próstata más comunes pueden verse en la Figura 1.2.

Concretamente, un radiólogo entrenado es capaz de identificar las lesiones de PCa mediante la evaluación visual de las RMmps, permitiendo una mejor selección de los pacientes para la biopsia de próstata (Mehralivand et al., 2018), reduciendo significativamente el número de biopsias innecesarias, aumentando el rendimiento diagnóstico del procedimiento (Ahmed et al., 2017), y permitiendo exámenes de biopsia guiados por fusión más precisos, y terapias focales, en comparación con los enfoques de fusión cognitiva (Marra et al., 2019). Una revisión sistemática de 2019, incluyendo 29 publicaciones y 8503 pacientes, descubrió que la RMmp tiene una sensibilidad y especificidad de 0,87 [intervalo de confianza -IC- del 95%, 0,81-0,91] y 0,68 [IC del 95%, 0,56-0,79], respectivamente, y un área bajo la curva ROC (AUC-ROC) de 0,87 [IC del 95%, 0,84-0,90], lo que ayuda a explicar su aceptación generalizada actual como herramienta de diagnóstico estándar del CaP.

A pesar de los aspectos positivos de la RMmp, esta viene con su propio conjunto de problemas. En primer lugar, la interpretación de la RMmp requiere mucho tiempo, depende de la experiencia del experto (Gaziev et al., 2016), y suele ir acompañada de una variabilidad interobservador no despreciable (Sonn et al., 2019). Esto es particularmente relevante fuera de los centros expertos de alto volumen (Kohestani et al., 2019). En segundo lugar, las adquisiciones de RMmp son costosas, al igual que lo es la contratación de los radiólogos necesarios para analizar un número cada vez mayor de RMmp debido a la generalización de las revisiones periódicas de PCa. En tercer lugar, aunque la tecnología de RM es en sí misma no ionizante y segura, los agentes de contraste como el gadolinio, típicamente empleados en las secuencias de RM de contraste dinámico, son cada vez más controvertidos por su acumulación a largo plazo en los tejidos.

La RMmp no sólo ha marcado un punto de inflexión en el diagnóstico del CaP, sino también en las intervenciones de biopsia y tratamiento focal del CaP. Mientras que las biopsias sistemáticas clásicas requerían la recogida y el análisis de 20-30 muestras, las biopsias guiadas por RM pueden dirigirse directamente a las lesiones marcadas por el radiólogo, por lo que se necesitan muchas menos muestras y se mejora la detección de los CaP clínicamente significativos en comparación con las biopsias sistemáticas por sí solas (Marra et al., 2019). Normalmente, las biopsias guiadas por RM se realizan utilizando USTR para guiar las operaciones, ya que el uso de la RM intraoperatoria sería prohibitivo para la mayoría de las instituciones médicas (Hambrock et al., 2010). Por el contrario, las biopsias USTR guiadas por RMmp son mucho más accesibles, pero requieren localizar las posiciones exactas de las lesiones (marcadas por los radiólogos en la RMmp preadquirida) dentro de la imagen US intraoperatoria, donde desafortunadamente las lesiones carecen de contraste con respecto al tejido circundante y por lo tanto no pueden ser identificadas visualmente (Kaplan et al., 2002). El problema de encontrar las correspondencias completas entre las posiciones de la RMmp y la USTR de la próstata se conoce como registro, y puede ser realizado mentalmente por un urólogo experto durante el procedimiento de biopsia (lo que se conoce como fusión cognitiva, y ha mostrado algunos resultados contradictorios, Puech et al. (2013)), o computacionalmente, lo que ofrece una mayor precisión y reproducibilidad, y es un foco de investigación activa. Del mismo modo, se necesitan técnicas precisas de registro de RM-USTR para las terapias focales, que cada vez son más populares.

En resumen, se plantea un importante problema socioeconómico: aunque el cribado del PSA ha demostrado ser muy eficaz para el diagnóstico del PCa, reduciendo la mortalidad por esta causa en más de un 20% (Schröder et al., 2009, 2012), conlleva un alto riesgo de sobrediagnóstico, lo que supone una carga económica que los sistemas sanitarios no pueden asumir. Aunque la RMmp ha mejorado significativamente la situación, reduciendo las biopsias innecesarias y mejorando el rendimiento del procedimiento; la RMmp introduce un nuevo conjunto de problemas.

Este trabajo se compone de cinco publicaciones revisadas por pares que intentan abordar estos problemas con la ayuda de redes neuronales convolucionales (CNNs, del inglés *Convolutional Neural Networks*, Sección 8.3.1) un algoritmo de IA especializado en el procesamiento de imágenes (médicas y de otro tipo), así como otras técnicas útiles, como el método de elementos finitos (MEF, Sección 8.3.2) para simular el comportamiento biomecánico de la próstata, y la deriva de puntos coherente (CPD, del ingles *Coherent Point Drift*, Sección 8.3.3), un método de registro de conjuntos de puntos utilizado para abordar el problema de registro RM-US. Estos se presentarán en las siguientes secciones 8.3.1-8.3.3, mientras que en la Sección 8.4 se hará un repaso de la distribución del resto del documento, incluyendo un resumen de cada una de las publicaciones (Capítulos 2-6).

8.3 Metodología

8.3.1 Inteligencia artificial para el análisis de imágenes médicas

8.3.1.1 Perspectiva histórica

La IA es un término de uso muy amplio, que engloba muchos campos diferentes con el propósito común de desarrollar sistemas capaces de manifestar comportamientos inteligentes. Sin embargo, con frecuencia la IA se utiliza exclusivamente para referirse al Aprendizaje Automático (ML, del inglés *Machine Learning*), que es un subcampo de la IA que estudia los algoritmos capaces de aprender de la experiencia (Benet-Ferrus et al., 2022). Aunque el campo del ML se inició hace más de medio siglo, no fue hasta principios de la década de 2010 cuando se produjo la verdadera revolución, debido a la confluencia de varios factores, a saber: desarrollos teóricos (nuevas arquitecturas, mejor inicialización de los parámetros, *frameworks* de auto-gradiente, etc.) que permitieron entrenar redes neuronales (NN, del inglés *Neural Networks*) más profundas y potentes, de ahí la aparición del campo conocido como aprendizaje profundo (DL, del inglés *Deep Learning*); los datos, que son el combustible necesario para entrenarlas, se han hecho omnipresentes gracias al crecimiento de Internet y la digitalización; y las unidades de computación gráfica (GPU, del inglés *Graphics Processing Units*) comenzaron a emplearse para la computación general (más allá de los juegos de ordenador), acelerando los cálculos requeridos por las NN en varios órdenes de magnitud.

El punto de inflexión fue AlexNet (Krizhevsky et al., 2012), una CNN que ganó el concurso de clasificación de imágenes ImageNet (Russakovsky et al., 2015) de 2012 por un amplio margen, lo que despertó un interés por el DL que ha crecido exponencialmente desde entonces. En el contexto del análisis de imágenes médicas, las CNN, como AlexNet, han sido el motor de la mayoría de los desarrollos actuales.

Esteva et al. (2017) entrenó una CNN de clasificación en $\sim$ 130.000 imágenes de lesiones de la piel, logrando un rendimiento igual al de los expertos; De Fauw et al. (2018) empleó $\sim$ 15.000 imágenes de tomografía de coherencia óptica (TCO) para entrenar un conjunto de dos CNNs capaces de detectar una amplia gama de enfermedades de la retina, con un rendimiento que iguala o supera el de los expertos; Balakrishnan et al. (2019) propuso el *framework* VoxelMorph, un método basado en CNNs para el registro de imágenes médicas, logrando velocidades de registro varios órdenes de magnitud más rápidas que las alternativas clásicas basadas en la optimización imagen a imagen, y permitiendo la inclusión de información de segmentación auxiliar para mejorar aún más la precisión; finalmente, Minaee et al. (2020) entrenó una CNN en 5.000 radiografías de tórax para detectar la presencia de COVID-19, logrando una sensibilidad/especificidad de 0.98/0.90. Como puede verse, el algoritmo único que está detrás de todas estas contribuciones es la CNN; debido a su importancia, en la sección 8.3.1.2 se presentarán los bloques básicos e ideas que hay detrás. Además, en estos trabajos se demuestra que las CNN necesitan muchos datos para aprender, pero, una vez recogidos, los sistemas entrenados logran rendimientos similares (y a veces por encima) de los expertos.

8.3.1.2 Resumen técnico de las redes neuronales convolucionales

En pocas palabras, las CNN pueden verse como una pila de filtros convolucionales aprendibles junto con otras no linealidades, en las que los parámetros del filtro se aprenden por descenso de gradiente. En este contexto, las imágenes deben entenderse como matrices de números; por ejemplo, la imagen bidimensional en escala de grises I a la izquierda de la Figura 1.3 también puede verse como una matriz (de dimensiones 5×5 en este caso), donde cada elemento I_{ij} representa la intensidad del píxel en esa ubicación. Del mismo modo, el filtro convolucional θ es sólo otra matriz de números (dimensiones 3×3 en este caso); aplicando la convolución θ a I, se genera el mapa de activación de salida O (también conocido como mapa de características), es decir, $O = I * \theta$.

Dependiendo de los parámetros, los filtros convolucionales pueden detectar características como líneas, puntos, etc. Además, las convoluciones también pueden aplicarse a los mapas de activación, ya que tanto las imágenes I como las activaciones O son simplemente matrices de números, y apilando convoluciones se pueden detectar patrones cada vez más complejos.

Otra operación muy común en las CNNs es el *downsampling* y *upsampling* de los mapas de activación (es decir, reducir y aumentar su resolución, respectivamente). Una forma muy común de realizar el *downsampling* es cambiando el *stride* de la convolución, es decir, el paso con el que se desliza sobre la imagen. Otra forma de realizar el *downsampling* es utilizando el operador *maxpooling*, que es igual que una convolución, pero en lugar de aplicar un producto escalar, sólo aplica el operador máximo a los elementos de I dentro de la ventana del núcleo. Para el *upsampling*, las convoluciones de transposición operan con un paso > 1 sobre la salida, en lugar de la entrada, consiguiendo así este efecto; aun así, también es habitual utilizar la interpolación lineal o cúbica simple.

Por último, las CNNs emplean no linealidades para mejorar su capacidad de reconocer patrones complejos. La no linealidad más sencilla (o función de activación en este contexto) es la función de activación ReLU (*Rectified Linear Unit*) (Ecuación 8.1), que fue propuesta a finales de los años 60 y tiene motivaciones tanto biológicas como matemáticas. A pesar de su simplicidad, actualmente es la función de activación más utilizada en DL, junto con sus muchas variantes, como la ReLU con fugas (que tiene una pequeña pendiente para $x \leq 0$), la PReLU (que hace de la pendiente un parámetro aprendible), o la GELU (que es una aproximación suave a la ReLU).

$$\mathrm{ReLU}(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{otherwise} \end{cases} \tag{8.1}$$

Combinando todos los elementos anteriores, se pueden obtener diferentes arquitecturas de CNN. Por ejemplo, la figura 1.5 muestra la arquitectura de la CNN AlexNet (Krizhevsky et al., 2012), que no es más que una pila de convoluciones seguidas de activaciones ReLU, con algunas operaciones de max-pooling en medio, reduciendo gradualmente la resolución de la imagen de entrada de 224×224 a 13×13 mientras se aumenta el número de canales de 3 a 256; a continuación, el último mapa de características se aplana y se hace pasar por una NN estándar con una función de activación final *softmax* que predice la probabilidad de cada una de las 1000 clases de ImageNet (Russakovsky et al., 2015) cuando se proporciona una imagen de entrada. La mayoría de las CNNs comparten una estructura similar: a medida que la red se hace más profunda, la resolución espacial se reduce mientras que el número de canales se incrementa, transformando así la información espacial (líneas en una posición determinada, un color, etc.), en información progresivamente más rica semánticamente (una forma, una combinación de colores, etc.), y finalmente en características altamente informativas (una cara, una rueda, una flor, etc.) que pueden ser utilizadas para predecir la salida.

Para entrenar una CNN (y todos los algoritmos de DL, de hecho) se emplea el algoritmo de descenso de gradiente (DG). Después de inicializar los pesos (o parámetros θ) de la CNN a un pequeño valor aleatorio, DG funciona empujándolos iterativamente una pequeña cantidad μ en la dirección opuesta al gradiente de la función de pérdida J (alguna medida de error de predicción) con respecto a ellos (Ecuación 8.2), logrando así minimizar el error paulatinamente. Gracias al DG, las CNNs consiguen aprender filtros convolucionales útiles para un problema determinado, sin necesidad de ser programadas explícitamente para ello; esto contrasta con la visión de ML clásica, que empleaba parámetros de filtro definidos manualmente, y tenía un rendimiento mucho peor en tareas de percepción complejas.

$$\mathrm{DG}: \ \theta \leftarrow \theta - \mu \frac{\delta J(\theta, x, y)}{\delta \theta} \tag{8.2}$$

A lo largo de los años se han introducido varias arquitecturas de CNN, muchas de las cuales siguen en activo. Simonyan and Zisserman (2015) propuso las arquitecturas VGG16 y VGG19, muy similares a AlexNet, pero aumentando tanto la profundidad (número de capas convolucionales, que se aumentaron a 16 y 19 respectivamente) como la anchura (número de canales por capa, que se incrementó hasta 512 canales), consiguiendo así un modelo más potente. Sin embargo, con una mayor profundidad, el flujo de información a través de la red se deterioraba; para solucionar este problema, He et al. (2016) incorporó conexiones residuales a su arquitectura ResNet, conectando las etapas anteriores de la CNN con las posteriores mediante la adición de mapas de características, creando así una vía de baja resistencia para la información, y permitiendo arquitecturas mucho más profundas, como el modelo ResNet-152 de 152 capas. Esta misma idea fue ampliada por la arquitectura DenseNet (Huang et al., 2017). Recientemente Tan and Le (2019) empleó la búsqueda de arquitecturas neuronales (Zoph and Le, 2017), una técnica basada en el aprendizaje por refuerzo para encontrar una arquitectura CNN base óptima, que junto con un escalado simultáneo óptimo de profundidad/anchura/resolución, dio lugar a la familia de arquitecturas conocida como EfficientNet.

Más allá de la clasificación, para problemas como la segmentación, tanto la entrada como la salida son imágenes. En el contexto de las imágenes de próstata, la segmentación consiste en delinear o marcar la próstata dentro de una imagen médica, separándola del resto de órganos o estructuras. Al igual que las imágenes, las máscaras de segmentación pueden verse como una imagen binaria (es decir, que sólo contiene 1s y 0s) con un uno en todas las posiciones dentro de la región de interés (por ejemplo, la próstata) y un cero en el resto. Para este tipo de tareas, se emplea principalmente la arquitectura U-Net (Ronneberger et al., 2015) (o una de sus muchas variantes).

Este trabajo emplea ampliamente las CNNs: El capítulo 2 propone una arquitectura específica tipo U-Net, que combina conexiones densas y residuales, así como muchas otras técnicas, para resolver el problema de la segmentación de la próstata en RM y USTR; el capítulo 3 desarrolla una técnica para mejorar la resolución de salida de cualquier CNN de segmentación; El capítulo 4 hace uso de la Retina-U-Net (Jaeger et al., 2020), una arquitectura que combina una U-Net para la segmentación de lesiones de CaP con la RetinaNet (Lin et al., 2017b) para la detección de lesiones, la clasificación y el refinamiento de los *bounding boxes*; Por último, el capítulo 6 utiliza las mismas ideas del *framework* VoxelMorph (Balakrishnan et al., 2019) para entrenar una CNN que, dado un par correspondiente de RM de próstata y USTR, es capaz de realizar directamente el registro RM-US en tiempo casi real; entendiendo por tiempo casi real que el sistema puede utilizarse sin tener que esperar a que responda respuesta (por ejemplo, menos de medio segundo). De hecho, aunque son lentos en el entrenamiento, los modelos DL suelen tener la ventaja de ser extremadamente rápidos en la inferencia (es decir, en la predicción sobre una nueva muestra).

8.3.2 Simulación del comportamiento biomecánico mediante el método de los elementos finitos

El MEF es un método numérico utilizado para encontrar soluciones aproximadas a problemas de ingeniería y física matemática que no pueden resolverse analíticamente debido a la complejidad de sus ecuaciones constitutivas, la geometría del problema y/o sus condiciones de contorno. Es una herramienta estándar en una gran variedad de industrias, que ayuda a la validación y el diseño de productos, especialmente en los casos en que las pruebas físicas serían muy costosas o incluso inviables (como la simulación de la atmósfera y la gravedad marcianas, o la simulación del comportamiento mecánico de los tejidos vivos).

En el contexto de este trabajo, el MEF se utiliza para obtener el campo de desplazamiento dentro de un hígado (Capítulo 5), o una próstata (Capítulo 6), dadas algunas condiciones de contorno, como las fuerzas externas, o un campo de desplazamiento superficial, respectivamente. A grandes rasgos, la resolución de un problema biomecánico (como es el caso) con el MEF requiere mallar los volúmenes de interés (es decir, discretizarlos en una malla de elementos finitos), proporcionar algunas ecuaciones constitutivas y parámetros para su comportamiento mecánico, y establecer unas condiciones de contorno adecuadas. A continuación, el solucionador del MEF tratará de encontrar el campo de desplazamiento que minimice la energía potencial del sistema, en virtud del teorema de la mínima energía potencial total.

La Figura 1.7 muestra el proceso de obtención de una malla de próstata de forma automática y la simulación de su comportamiento mecánico mediante el MEF: en primer lugar, se emplea una CNN de segmentación para obtener la máscara de segmentación, a continuación, la máscara se malla utilizando TetGen (Si, 2010), y, finalmente, la malla, las propiedades del material y las condiciones de contorno se introducen en el *solver* para obtener el campo de desplazamientos dentro de esta glándula. Nótese que la superficie de la malla se ha dividido en una multitud elementos triangulares superficiales, mientras que en el interior (no mostrado) se utilizan elementos tetraédricos volumétricos. El MEF hace un buen uso de esta discretización, encontrando los desplazamientos sólo en los vértices de la malla (también llamados nodos), y luego interpolando al resto del volumen, reduciendo así los grados de libertad de un número infinito a tantos como vértices. Evidentemente, cuanto más fina sea la malla, más precisa será la solución, pero también más lenta será de calcular.

Para un problema mecánico (como en la Figura 1.7), la Ecuación 8.3 establece que la energía potencial total Π_p es igual a la energía de deformación del sistema W_s menos el potencial de trabajo W_p. En particular, para materiales elásticos lineales (es decir, con una relación tensión-deformación lineal), $W_s = \frac{1}{2}U^T KU$, donde U es una matriz con el desplazamiento de los nodos y K es la matriz de rigidez global que se ha construido ensamblando las matrices de rigidez de cada elemento; y $W_p = U^T F$, donde F representa las fuerzas nodales. La minimización de la energía potencial total puede lograrse tomando la derivada de Π_p con respecto al campo de desplazamientos nodales U e igualándola a cero, siendo U simplemente la solución de un sistema lineal (Ecuación 8.4). Se puede encontrar una explicación más detallada sobre el MEF en la sección 6.4.4.

$$\Pi_p = W_s - W_p \tag{8.3}$$

$$\frac{\delta \Pi_p(U)}{\delta(U)} = 0 \rightarrow KU = F \tag{8.4}$$

En la práctica, los tejidos blandos (como los de la próstata o el hígado) suelen presentar un comportamiento mecánico no lineal, por lo que deben emplearse formulaciones más complejas para W_s. Además, las ecuaciones constitutivas para estos materiales son muy difíciles de parametrizar, ya que no suele ser factible realizar mediciones mecánicas directas dentro del cuerpo, y las propiedades cambian significativamente cuando estas se determinan ex vivo. Además, las condiciones de contorno adecuadas (es decir, cómo interactúa un órgano con el tejido circundante) suelen ser aún más difíciles de obtener. Por lo general, hay que hacer fuertes suposiciones y simplificaciones, como se discute en las secciones 5.3.3.1 y 5.3.3.2 para el caso del hígado.

En este trabajo, el MEF se emplea en el Capítulo 5 para simular el comportamiento biomecánico del hígado, y luego utilizar esas simulaciones para entrenar un modelo de DL que logre velocidades de inferencia en tiempo real sin sacrificar una alta precisión; tanto la alta velocidad como la precisión son necesarias en el contexto de los simuladores quirúrgicos, la cirugía asistida por ordenador y la irradiación tumoral guiada, entre otras aplicaciones. En el capítulo 6, se utiliza el MEF para obtener el campo de desplazamiento que experimenta la próstata durante una biopsia (o cualquier procedimiento dirigido, como la braquiterapia) con respecto a la próstata en reposo (a partir de la RM), resolviendo así el problema de registro RM-US; de forma similar al capítulo anterior, se entrena finalmente una CNN para imitar esas simulaciones y así poder realizar el registro RM-US en tiempo casi real, lo que abre la puerta a un registro continuo y a una mayor precisión gracias a esta adaptabilidad, de la que carecen los métodos actuales.

8.3.3 Registro de conjuntos de puntos con deriva de puntos coherente

El registro de conjuntos de puntos consiste en encontrar la correspondencia entre dos conjuntos de puntos y/o recuperar la transformación que mapea el conjunto de puntos móviles $X_{N\times d}$ al conjunto de puntos fijos $Y_{M\times d}$, donde N y M es el número de muestras de X y Y, respectivamente, y d son las dimensionalidades de los conjuntos de puntos (típicamente 2 ó 3) (Figura 1.8).

La transformación buscada puede ser rígida (consistente sólo en rotación y/o traslación), o no rígida. Para el caso rígido simple, un objetivo más formal sería encontrar los parámetros de transformación $\theta = \{R, t\}$ (donde R es una matriz de rotación, y t es un vector de traslación) que más se aproximen a X a Y. Esto se puede lograr mediante la optimización de la función de coste J con respecto a θ

(Ecuación 8.6), sujeto a que R sea una matriz de rotación, donde $\|\cdot\|_F$ denota la norma de Frobenius y $\hat{X}$ es X transformado (Ecuación 8.5). Se debe tener en cuenta que este problema de optimización puede dar lugar a soluciones donde X está reflejada, lo que ocurre cuando $\det(X) = -1$.

$$\hat{X} = R \cdot X + t \tag{8.5}$$

$$J = \|\hat{X}(\theta) - Y\|_F \ \text{ s.t. } R^T \cdot R = \mathbb{I} \tag{8.6}$$

Si el número de puntos tanto en X como en Y es el mismo (es decir los puntos están emparejados), esto se conoce como el problema de Procrustes, que tiene una solución de forma cerrada: el vector de traslación óptimo t^* es simplemente el vector que une los centroides de ambos conjuntos de puntos $t^* = \bar{Y} - \bar{X}$, mientras que la matriz de rotación óptima R^* puede obtenerse como $R^* = UV^T$, con U, V procedentes de la descomposición en valores singulares de $(Y - \bar{Y})(X - \bar{X}) = U\Sigma V^T$, donde a Y, X se les han restando sus respectivos centroides.

Si el número de puntos en X e Y es diferente, el método de registro de conjuntos de puntos más sencillo es el algoritmo del punto más cercano iterativo (ICP, del inglés *Iterative Closest Point*), desarrollado por (Besl and McKay, 1992), que funciona de la siguiente manera: primero, se inicializan los parámetros (por ejemplo: $R = \mathbb{I}, t = [0, ..., 0]^T$); segundo, se empareja cada punto de X con su punto más cercano en Y, obteniendo así un subconjunto de Y que llamamos Y_s que ahora se empareja con X; tercero, se resuelve el problema de Procrustes entre X y Y_s. Finalmente, X se transforma según los parámetros encontrados al resolver Procrustes, y los pasos dos y tres se repiten hasta que X no cambie entre iteraciones.

Desgraciadamente, el algoritmo ICP básico presenta varios defectos importantes, a saber: X e Y deben estar lo suficientemente cerca para que el ICP converja a la solución óptima, no es robusto contra el ruido o la presencia de valores atípicos, y no realiza registro no rígido. Se han propuesto muchos métodos para superar estas limitaciones, como los métodos probabilísticos, que asignan correspondencias de puntos que no son binarias, sino probabilísticas, es decir, ya no hay un punto "más cercano", sino que todos los puntos de Y están "algo" cerca de cada punto de X según alguna ponderación probabilística.

CPD (Myronenko and Song, 2009) es un método probabilístico de registro de conjuntos de puntos formulado como un problema de estimación de la densidad de probabilidad, en el que el conjunto de puntos móviles Y constituye los centroides de un modelo de mezcla gaussiana (GMM, del inglés *Gaussian Mixture Model*, una distribución probabilística definida como una suma de gaussianas), y el conjunto de puntos fijos X representan las observaciones del GMM, que tienen algo de ruido uniforme (nótese que los conjuntos de puntos fijos y móviles han cambiado de nombre para ser coherentes con la notación de Myronenko and Song (2009)). En CPD, el objetivo es maximizar la probabilidad de que las observaciones X pertenezcan al GMM definido por los puntos en Y. Para obtener más información, consulte la sección 6.4.3. En virtud de su naturaleza probabilística, el CPD se comporta mejor que el ICP, especialmente en presencia de ruido y valores atípicos, además incluye variantes rígidas y no rígidas del algoritmo. Un ejemplo de CPD aplicado al registro de conjuntos de puntos de superficie no rígidos en el hígado puede verse en la Figura 5.8 y, para la próstata, en la Figura 6.4.

En el presente trabajo, el CPD se emplea en dos ocasiones. En el capítulo 5, se emplea para hacer coincidir una malla de referencia común (cuyos vértices son el conjunto de puntos móviles) con un conjunto de mallas de hígado (cuyos vértices son el conjunto de puntos fijos); las primeras componentes principales de la transformación que debe sufrir la malla del hígado de referencia pueden utilizarse entonces para parametrizar de forma eficiente la forma de cualquier malla de un hígado. En el capítulo 6, se utiliza el CPD para realizar el registro entre las mallas de próstata obtenidas de la segmentación de una RM y un USTR del mismo paciente.

8.4 Resumen del documento

El documento está organizado como sigue:

Capítulo 2: *Robust Resolution-Enhanced Prostate Segmentation in Magnetic Resonance and Ultrasound Images through Convolutional Neural Networks*, publicado en *Applied Sciences*, 2021 (2021 *Journal Impact Factor* -JIF- 2.84, Q2, percentil 58.15 en *Engineering, Multidisciplinary*). Se propone un modelo rápido, robusto, preciso y generalizable para la segmentación de próstata en RM y USTR empleando CNNs. El modelo logra un rendimiento consistente e incluso supera la variabilidad entre expertos en la segmentación por RM. La segmentación de la próstata se realiza de forma rutinaria en la RM y el USTR, ya que es necesaria para analizar la RMmp y realizar el registro RM-USTR. Segmentaciones más precisas pueden dar lugar a un mejor registro y pronóstico, mientras que unos resultados casi instantáneos son de especial interés para los urólogos, que actualmente tienen que dedicar en torno a diez minutos a realizar manualmente la segmentación en medio de la operación de biopsia o ablación del tumor. Véase en la Figura 2.6 algunos ejemplos de segmentación automática.

Capítulo 3: *Cost-free Resolution Enhancement in Convolutional Neural Networks for Medical Image Segmentation*, publicado en las actas de la ESANN, 2021 (2021 *Computer Research and Education* -CORE- Rango B). Esta publicación propone un método sencillo pero eficaz para mejorar la resolución de salida de cualquier CNN de segmentación ya entrenada (como las desarrolladas en el capítulo 2), incluso más allá de la de la imagen original. Las segmentaciones de próstata de alta resolución pueden conducir a un mejor registro y/o a simulaciones más precisas de su comportamiento biomecánico. Véase en la Figura 3.3 un ejemplo de esta técnica.

Capítulo 4: *Deep Learning for Fully Automatic Detection, Segmentation, and Gleason Grade Estimation of Prostate Cancer in Multiparametric Magnetic Resonance Images* publicado en *Scientific Reports*, 2022 (2021 JIF 5.00, Q2, percentil 74.66 en *Multidisciplinary Sciences*). Este artículo presenta un modelo basado en CNN para el análisis automático de RMmp, logrando un excelente AUC-ROC/sensibilidad/especificidad a nivel de lesión de 0,95/1,00/0,80 para el criterio de significancia de PCa GGG $\geq$ 2, superando a los radiólogos expertos. Utiliza el modelo del capítulo 2 para realizar la segmentación automática de la próstata. Las aplicaciones clínicas de este modelo son innumerables: podría utilizarse como segunda opinión clínica -una red de seguridad para reducir la probabilidad de omitir lesiones de PCa-, para la priorización del análisis de RMmp y/o para la priorización de la biopsia, o incluso como una herramienta de sugerencia de derivación totalmente automática en el contexto de futuros programas de cribado de PCa a nivel poblacional. Véase la Figura 4.2 para ver algunos ejemplos de este modelo en acción.

Capítulo 5: *Real-time Biomechanical Modeling of the Liver using Machine Learning Models trained on Finite Element Method Simulations* publicado en *Expert Systems with Applications*, 2020 (JIF 2020 6,95, Q1, percentil 91,39 en *Electrical and Electronic Engineering*). Los tejidos y órganos vivos presentan un comportamiento biomecánico complejo, cuya simulación es, sin embargo, de gran interés en el contexto de la planificación quirúrgica, la cirugía asistida por ordenador o el registro con restricciones mecánicas. Aunque el MEF se emplea habitualmente con este fin, suele ser demasiado lento para su uso en tiempo real. Aquí se propone utilizar el aprendizaje automático (ML) para acelerar las simulaciones del MEF, conservando una alta precisión y mejorando la velocidad de simulación en varios órdenes de magnitud. Véase en la Figura 5.21 una prueba de concepto de la simulación del comportamiento mecánico del hígado ejecutada en tiempo real.

Capítulo 6: *Deep Learning Contributions for Reducing the Complexity of Prostate Biomechanical Models* aceptado para su publicación en *Reduced Order Models for the Biomechanics of Living Organs*, Elsevier, 2022 (*Recognized Publisher, Book Citation Index, Thomson Reuters*). Esta última publicación aborda el complejo problema del registro MR-US utilizando una CNN para aprender la tarea de registro, de forma similar a como se hizo en el capítulo 5. Como referencia para el entrenamiento, se utiliza el CPD para hacer coincidir primero las superficies de la próstata de la RMmp-US (generadas automáticamente por los modelos de segmentación de los capítulos 2 y 3), seguido de una simulación con el MEF para obtener deformaciones internas mecánicamente plausibles. La CNN entrenada logró una aproximación casi perfecta, reduciendo así significativamente el error de registro, a la vez que alcanzaban velocidades cercanas al tiempo real.

Por último, en la Sección 8.5 se analiza si los objetivos originales se cumplieron finalmente, se resumen las aportaciones de cada uno de los trabajos tanto desde el punto de vista del paciente como desde el punto de vista técnico (Sección 8.5.1), y se concluye discutiendo algunas limitaciones generales de las IAs médicas, así como trabajos futuros (Sección 8.5.2).

Las referencias de todos los capítulos se han recogido al final del documento para evitar duplicidades. El código de este trabajo se ha puesto a disposición del público en https://github.com/OscarPellicer/Deep-Learning-in-Prostate-PhD.

8.5 Principales resultados y conclusiones

Este trabajo se planteó con el objetivo de mejorar la calidad de vida de los pacientes con sospecha de CaP y de los clínicos que los atienden mediante el uso de la IA, abordando dos objetivos concretos. El primero se resolvió mediante el desarrollo de un modelo de segmentación de próstata por RM (Capítulo 2), y su posterior utilización para el desarrollo de un sistema de detección, segmentación y clasificación de lesiones totalmente automático y de alta precisión (Capítulo 4); el segundo objetivo se abordó mediante el desarrollo de modelos de segmentación de próstata de alta resolución en RM y US (Capítulo 2 & 3) y su utilización para la obtención de transformaciones de referencia para registro RM-US mediante CPD y MEF, que finalmente se emplearon para el entrenamiento de un modelo de DL capaz de acelerar este proceso de forma significativa (Capítulos 5 & 6). En todos los casos, los sistemas finales han sigo totalmente automáticos, funcionaban a la par o mejor que los expertos (cuando la comparación era posible), y eran extremadamente rápidos en la inferencia (velocidades de tiempo real o casi real). El aumento de la velocidad de inferencia es de especial importancia en el modelo de registro de RM-US, ya que permite que el registro se adapte a medida que la forma de la próstata cambia debido al movimiento del paciente, el movimiento de la sonda o la afluencia de sangre, entre otros, mejorando así la precisión del registro, que de otro modo podría degradarse a medida que avanza la intervención. El código para ambos objetivos se ha hecho público en https://github.com/OscarPellicer/Deep-Learning-in-Prostate-PhD.

En la sección 8.5.1 se repasarán las aportaciones y novedades introducidas por estas cinco publicaciones, tanto en términos de mejoras para el bienestar del paciente y el flujo de trabajo del clínico, como desde el punto de vista científico-técnico, mientras que en la sección 8.5.2 se profundizará en las limitaciones generales de las IAs clínicas.

8.5.1 Contribuciones

En el capítulo 2 se proponen modelos de segmentación de próstata totalmente automáticos de RM y de US, logrando ambos un rendimiento excelente, con el modelo de RM incluso superando a los radiólogos expertos. Estos modelos pueden reducir o incluso eliminar la necesidad de la segmentación manual, que se sabe que requiere una amplia experiencia y consume mucho tiempo, y en última instancia, sufre una alta variabilidad inter e intra-experto. Además, la inferencia (es decir, la generación de una nueva segmentación) es extremadamente rápida, lo cual es de especial interés para la ecografía intraoperatoria, que actualmente requiere que el urólogo dedique unos 10 minutos en medio de una operación a generar la segmentación; también puede ayudar a aliviar algo de tiempo muy necesario para los radiólogos. Por último, los modelos han demostrado ser robustos y, por tanto, deberían funcionar bien independientemente del escáner o del centro médico donde se adquieran las imágenes, lo cual es esencial en caso de que el modelo se despliegue fuera del centro médico donde se ha desarrollado.

Desde un punto de vista científico-técnico, el capítulo propone una nueva arquitectura de CNN, junto con varias decisiones de diseño y entrenamiento que, en conjunto, ayudan a crear modelos de segmentación robustos y de buen rendimiento. La arquitectura CNN diseñada, similar a la U-Net, combina un codificador DenseNet, que es extremadamente eficiente en cuanto a parámetros, con un decodificador ResNet, e incorpora técnicas como el *ensembling* de puntos de control del modelo, el uso de una tasa de aprendizaje cíclica, la rutina de aumentación de datos, y un conjunto de entrenamiento muy variado, para lograr su rendimiento y robustez. Curiosamente, muchas de estas técnicas son hoy en día relativamente comunes, pero no estaban bien establecidas cuando el modelo estaba en desarrollo. En general, no se encontró ningún otro modelo que tuviera un rendimiento consistente en varios conjuntos de datos simultáneamente. Por último, la mejora de la resolución neuronal, una técnica introducida en el Capítulo 3 se utiliza aquí por primera vez en un modelo real para mejorar con éxito la resolución de las máscaras de segmentación generadas.

El capítulo 3 presenta una técnica para mejorar la resolución de salida de las CNNs de segmentación, incluso más allá de la resolución de la imagen original. Esto es de especial interés para la segmentación de imágenes de RM o Tomografía Axial Computerizada (TAC), ya que tienden a tener una resolución menor a lo largo de un eje en comparación con los otros dos, debido al procedimiento de adquisición por cortes, lo que lleva a tamaños de voxel anisotrópicos problemáticos. Las segmentaciones de alta resolución pueden mejorar la precisión en tareas posteriores, como el registro o la simulación del comportamiento biomecánico.

Técnicamente, el método es muy sencillo y puede aplicarse a cualquier CNN de segmentación ya entrenada. Este se basa en aprovechar la interpolación en el espacio de la imagen de entrada original, donde la información sigue siendo completa, en lugar de hacerlo en el espacio de salida discretizado, y mucho menos informativo. Además, aprovecha el conocimiento contextual de la CNN sobre la tarea de segmentación concreta, lo que podría mejorar aún más los resultados. Todos los interpoladores utilizados actualmente en la práctica ignoran esta información, que sin embargo está disponible y podría conducir fácilmente a resultados más precisos.

El capítulo 4 presenta un modelo totalmente automático para la detección, segmentación y clasificación de lesiones de CaP que ha demostrado tener un rendimiento superior al de los radiólogos expertos en la detección de lesiones clínicamente significativas (Grupo de Grado Gleason ≥ 2). La interpretación de la RMmp de próstata basada en la IA tiene muchos casos de uso potenciales, siendo quizás elmás obvio una segunda opinión para ayudar a los radiólogos y reducir el riesgo de omitir lesiones clínicamente significativas. También podría ser utilizada por los radiólogos como criterio para priorizar a los pacientes, clasificándolos según el riesgo evaluado por la IA, y permitiéndoles así centrarse primero en los casos más urgentes. Por último, podría utilizarse para desarrollar programas de cribado para toda la población, empleando una IA que derive automáticamente al paciente bajo la más mínima sospecha.

Se introducen varias novedades. En primer lugar, se trata del primer *framework* totalmente automático para realizar esta tarea (a conocimiento del autor). Utiliza una red de detección, la Retina U-Net, en comparación con el enfoque estándar de utilizar la segmentación seguida de pasos de posprocesamiento para obtener las lesiones independientes. También hace uso del modelo de segmentación de próstata en RM desarrollado anteriormente y lo amplía en una configuración en cascada para distinguir también entre la glándula central y la zona periférica; esta información de segmentación es extremadamente útil para el modelo de detección, ya que el aspecto y la probabilidad de las lesiones difieren entre zonas. También se propone un procedimiento automático que emplea información mutua y características de gradiente espacial para la tarea no trivial del registro de secuencias de RMmp.

En el capítulo 5, se propone un método para simular el comportamiento biomecánico del hígado (o de cualquier órgano) en tiempo real. Constituye un gran avance en la implementación de aplicaciones como los simuladores quirúrgicos, la cirugía asistida por ordenador o la ablación tumoral guiada. El éxito de esta aproximación abre las puertas a nuevas investigaciones en la aceleración del MEF mediante DL, y conduce directamente a los desarrollos del capítulo 6.

Cuando se publicó, el enfoque empleado fue más allá de la investigación existente al permitir el uso de cualquier hígado como entrada al modelo DL entrenado, en lugar de limitarse a una única geometría de hígado. Esto se consiguió parametrizando la forma de un hígado arbitrario, de modo que pudiera alimentarse a una red neuronal estándar, haciéndola así conocedora de su forma. El modelo DL resultante demostró ser muy preciso y extremadamente rápido (por encima de 100 Hz).

Por último, el capítulo 6 presenta un sistema automático para el registro no rígido de la próstata en RM-US que mejora significativamente con respecto al registro rígido de referencia. Lo más importante es que el sistema funciona casi en tiempo real, lo que abre la puerta a las intervenciones guiadas de próstata en tiempo real, aumentando potencialmente la precisión al permitir que el registro se adapte a la forma cambiante de la glándula, ya sea debido a los movimientos del paciente, el movimiento de la sonda o la acumulación de sangre, entre otros factores.

La principal novedad radica en que se trata del modelo más rápido que se ha propuesto para el registro de la próstata mediante RM-US. Esto se consigue generando primero un conjunto de transformaciones de registro compatibles con la biomecánica (utilizando CPD para el registro de la superficie y MEF para calcular los desplazamientos dentro de la glándula), y luego entrenando una CNN tipo U-Net para predecir la transformación final directamente a partir de las imágenes de entrada, omitiendo así todos los pasos intermedios que requieren mucho tiempo.

8.5.2 Limitaciones y trabajo adicional

Aunque este trabajo contribuye de forma significativa al cuerpo de conocimientos existente, todavía existen algunas limitaciones. Dado que las limitaciones específicas se discuten en profundidad en cada uno de los capítulos, los siguientes párrafos resumirán las más cruciales y generales:

Aunque las CNNs han demostrado una y otra vez su potencial en tareas de procesamiento de imágenes, en última instancia son modelos de caja negra, lo que significa que no se entiende su funcionamiento interno. Por lo tanto, no hay garantías de que una CNN funcione siempre como se espera (especialmente cuando hay un cambio de dominio de los datos de entrenamiento a los del mundo real, como suele ser el caso), y las predicciones de la CNN generalmente no son explicables. Esto es especialmente problemático en un escenario clínico, en el que las decisiones erróneas tomadas por una IA pueden tener un gran impacto en la vida de un paciente, y la responsabilidad legal puede requerir la elucidación de las razones que condujeron a un comportamiento específico de la IA. Además, el problema del cambio de dominio es especialmente notable en el caso de las CNN médicas, que suelen tener un rendimiento inferior cuando se utilizan con imágenes procedentes de escáneres significativamente diferentes de aquellos con los que fueron entrenadas.

Por un lado, en lo que respecta a la explicabilidad, los modelos complejos que realizan tareas complejas (como las CNNs que detectan cánceres) posiblemente nunca serán totalmente explicables para los humanos, porque sólo pueden serlo modelos que o bien son en sí mismos simples, o que pueden dividirse en elementos constitutivos mucho más simples. Desgraciadamente, al igual que el cerebro humano, las CNNs son estructuras complejas que funcionan mediante la interacción compuesta de muchas partes y, por tanto, son complejas de analizar. Aun así, se está avanzando en este sentido, y es de esperar que pronto las arquitecturas de DL puedan explicarse a un nivel satisfactorio. Por otro lado, la contrapartida al imprevisible rendimiento de las CNNs son los pacientes que podrían estar ya beneficiándose de las aplicaciones basadas en la IA, pero que no lo hacen debido a miedos normativos y legales. Los sistemas de IA no son perfectos pero, a diferencia de los humanos, los sistemas de IA médica pueden y deben ser validados y examinados a fondo antes de su uso clínico. Sin embargo, para la mayoría de las aplicaciones, la combinación de IA y seres humanos probablemente dará los mejores resultados, utilizando la IA como una segunda opinión clínica, una herramienta de priorización o, en el caso de IAs con muy buen rendimiento, como un sistema de cribado totalmente autónomo. En definitiva, debe realizarse una evaluación cuidadosa, pero sin olvidar tampoco los costes de retrasar el despliegue de la IA médica, que a menudo podrían superar a los riesgos.

Una última limitación, que quizá sea al mismo tiempo la línea de trabajo futura más importante, es la transferencia de los resultados de la investigación a un producto real que pueda influir directamente en la vida de los pacientes. A pesar del indiscutible valor científico, un trabajo de investigación con una aplicación tan obvia e inmediata como éste exige una aplicación práctica, y este no puede considerarse plenamente realizado de otro modo. Por ello, los trabajos posteriores deberían centrarse en encontrar la mejor manera de llevar estos proyectos a la práctica (como poner el código a disposición del público), para que puedan materializar directamente la razón por la que se desarrollaron en primer lugar: mejorar la vida de los pacientes.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: A system for large-scale machine learning. *arXiv*, may 2016. URL http://arxiv.org/abs/1605.08695.

Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. *Proc. - Int. Symp. Biomed. Imaging*, 2019-April: 683–687, 2019. ISSN 19458452. doi: 10.1109/ISBI.2019.8759329.

Michael Ahdoot, Amir H. Lebastchi, Baris Turkbey, Bradford Wood, and Peter A. Pinto. Contemporary treatments in prostate cancer focal therapy, may 2019. ISSN 1531703X.

Hashim U. Ahmed, Ahmed El-Shater Bosaily, Louise C. Brown, Rhian Gabe, Richard Kaplan, Mahesh K. Parmar, Yolanda Collaco-Moraes, Katie Ward, Richard G. Hindley, Alex Freeman, Alex P. Kirkham, Robert Oldroyd, Chris Parker, and Mark Emberton. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. *Lancet*, 389(10071):815–822, feb 2017. ISSN 1474547X. doi: 10.1016/S0140-6736(16)32401-1. URL http://dx.doi.org/10.1016/.

Hyun-Mo Ahn, Yeon-Ho Oh, Joong-Kyoung Kim, Jae-Sung Song, and Sung-Chin Hahn. Experimental Verification and Finite Element Analysis of Short-Circuit Electromagnetic Force for Dry-Type Transformer. *IEEE Trans. Magn.*, 48(2):819–822, feb 2012. ISSN 0018-9464. doi: 10.1109/TMAG.2011.2174212. URL http://ieeexplore.ieee.org/document/6136629/.

Alexandra Branzan Albu, Trevor Beugeling, and Denis Laurendeau. A morphology-based approach for interslice interpolation of anatomical slices from volumetric images. *IEEE Trans. Biomed. Eng.*, 55(8):2022–2038, aug 2008. ISSN 00189294. doi: 10.1109/TBME.2008.921158.

Nader Aldoj, Federico Biavati, Florian Michallek, Sebastian Stober, and Marc Dewey. Automatic prostate and prostate zones segmentation of magnetic resonance images using DenseNet-like U-net. *Sci. Rep.*, 10(1):1–17, 2020a. ISSN 20452322. doi: 10.1038/s41598-020-71080-0. URL https://doi.org/10.1038/s41598-020-71080-0.

Nader Aldoj, Steffen Lukas, Marc Dewey, and Tobias Penzkofer. Semi-automatic classification of prostate cancer on multi-parametric MR imaging using a multi-channel 3D convolutional neural network. *Eur. Radiol.*, 30(2):1243–1253, feb 2020b. ISSN 14321084. doi: 10.1007/s00330-019-06417-z. URL https://doi.org/10.1007/s00330-019-06417-z.

P. D. Allen, J. Graham, D. C. Williamson, and C. E. Hutchinson. Differential segmentation of the prostate in MR images using combined 3D shape modelling and voxel classification. In *2006 3rd IEEE Int. Symp. Biomed. Imaging From Nano to Macro - Proc.*, volume

2006, pages 410–413. IEEE, 2006. ISBN 0780395778. doi: 10.1109/isbi.2006.1624940. URL http://ieeexplore.ieee.org/document/1624940/.

Muhammad Arif, Ivo G. Schoots, Jose Castillo Tovar, Chris H. Bangma, Gabriel P. Krestin, Monique J. Roobol, Wiro Niessen, and Jifke F. Veenland. Clinically significant prostate cancer detection and segmentation in low-risk patients using a convolutional neural network on multi-parametric MRI. *Eur. Radiol.*, 30(12):6582–6592, dec 2020. ISSN 14321084. doi: 10.1007/s00330-020-07008-z. URL https://doi.org/10.1007/s00330-020-07008-z.

Samuel G. Armato, Henkjan Huisman, Karen Drukker, Lubomir Hadjiiski, Justin S. Kirby, Nicholas Petrick, George Redmond, Maryellen L. Giger, Kenny Cha, Artem Mamonov, Jayashree Kalpathy-Cramer, and Keyvan Farahani. PROSTATEx Challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. *J. Med. Imaging*, 5(04):1, nov 2018. ISSN 2329-4302. doi: 10.1117/1.jmi.5.4.044501. URL https://www.spiedigitallibrary.org/journals/journal-of-medical-imaging/volume-5/issue-4/044501/PROSTATEx-Challenges-for-computerized-classification-of-prostate-lesions-from-multiparametric/10.1117/1.JMI.5.4.044501.full.

Emmanuel A. Audenaert, Jan Van Houcke, Diogo F. Almeida, Lena Paelinck, M. Peiffer, Gunther Steenackers, and Dirk Vandermeulen. Cascaded statistical shape model based segmentation of the full lower limb in CT. *https://doi.org/10.1080/10255842.2019.1577828*, 22(6):644–657, apr 2019. ISSN 14768259. doi: 10.1080/10255842.2019.1577828. URL https://www.tandfonline.com/doi/abs/10.1080/10255842.2019.1577828.

Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *IEEE Trans. Med. Imaging*, 38(8):1788–1800, 2019. ISSN 1558254X. doi: 10.1109/TMI.2019.2897538.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mutual information neural estimation. In *35th Int. Conf. Mach. Learn. ICML 2018*, volume 2, pages 864–873, jan 2018. ISBN 9781510867963. URL https://arxiv.org/abs/1801.04062v5.

Katy J.L. Bell, Chris Del Mar, Gordon Wright, James Dickinson, and Paul Glasziou. Prevalence of incidental prostate cancer: A systematic review of autopsy studies. *Int. J. cancer*, 137(7):1749–1757, oct 2015. ISSN 1097-0215. doi: 10.1002/IJC.29538. URL https://pubmed.ncbi.nlm.nih.gov/25821151/.

David Benet-Ferrus, Oscar J. Pellicer-Valero, David Benet, and Oscar J. Pellicer-Valero. Artificial intelligence: the unstoppable revolution in ophthalmology. *Surv. Ophthalmol.*, 67(1):252–270, jan 2022. ISSN 18793304. doi: 10.1016/j.survophthal.2021.03.003. URL https://pubmed.ncbi.nlm.nih.gov/33741420/.

Paul J. Besl and Neil D. McKay. A Method for Registration of 3-D Shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(2):239–256, 1992. ISSN 01628828. doi: 10.1109/34.121791.

Pradeep Bhandari and Kevin Anderson. CFD analysis for assessing the effect of wind on the thermal control of the mars science laboratory curiosity rover. In *43rd Int. Conf. Environ. Syst.*, Pasadena, California, 2013. Jet Propulsion Laboratory, California Institute of Technology. ISBN 9781624102158. doi: 10.2514/6.2013-3325.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proc. COMPSTAT 2010 - 19th Int. Conf. Comput. Stat. Keynote, Invit. Contrib. Pap.*, pages 177–186. Physica-Verlag HD, Heidelberg, 2010. URL http://www.springerlink.com/index/10.1007/978-3-7908-2604-3_16.

Mohamed Bader Boubaker, Mohamed Haboussi, Jean-François Ganghoffer, and Pierre Aletti. Finite element simulation of interactions between pelvic organs: Predictive model of the prostate motion in the context of radiotherapy. *J. Biomech.*, 42(12):1862–1868, aug 2009. ISSN 00219290. doi: 10.1016/j.jbiomech.2009.05.022. URL https://linkinghub.elsevier.com/retrieve/pii/S0021929009002851.

Mohamed Bader Boubaker, Mohamed Haboussi, Jean-François Ganghoffer, and Pierre Aletti. Predictive model of the prostate motion in the context of radiotherapy: A biomechanical approach relying on urodynamic data and mechanical testing. *J. Mech. Behav. Biomed. Mater.*, 49:30–42, sep 2015. ISSN 17516161. doi: 10.1016/j.jmbbm.2015.04.016. URL https://linkinghub.elsevier.com/retrieve/pii/S1751616115001368.

Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.*, 68(6): 394–424, 2018. ISSN 1542-4863. doi: 10.3322/caac.21492.

Leo Breiman. *Classification And Regression Trees*. Routledge, oct 2017. doi: 10.1201/9781315139470. URL https://www.taylorfrancis.com/books/9781315139470.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. *arXiv*, 1(May):1–7, may 2020. URL https://arxiv.org/abs/2005.14165.

A. Brunon, K. Bruyère-Garnier, and M. Coret. Mechanical characterization of liver capsule through uniaxial quasi-static tensile tests until failure. *J. Biomech.*, 43(11):2221–2227, aug 2010. ISSN 00219290. doi: 10.1016/j.jbiomech.2010.03.038. URL https://www.sciencedirect.com/science/article/pii/S0021929010001831.

Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, Jose Neira, Ian Reid, and John J Leonard. Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age. *IEEE Trans. Robot.*, 32(6):1309–1332, dec 2016. ISSN 1552-3098. doi: 10.1109/TRO.2016.2624754. URL http://ieeexplore.ieee.org/document/7747236/.

Ewen Callaway. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures, dec 2020. ISSN 14764687.

Jean Louis Campos-Fernandes, Laurence Bastien, Nathalie Nicolaiew, Grégoire Robert, Stéphane Terry, Francis Vacherot, Laurent Salomon, Yves Allory, Dimitri Vordos, Andras Hoznek, René Yiou, Jean Jacques Patard, Claude Clément Abbou, and Alexandre de la Taille. Prostate Cancer Detection Rate in Patients with Repeated Extended 21-Sample

Needle Biopsy. *Eur. Urol.*, 55(3):600–609, mar 2009. ISSN 0302-2838. doi: 10.1016/J.EURURO.2008.06.043.

Ruiming Cao, Amirhossein Mohammadian Bajgiran, Sohrab Afshari Mirak, Sepideh Shakeri, Xinran Zhong, Dieter Enzmann, Steven Raman, and Kyunghyun Sung. Joint Prostate Cancer Detection and Gleason Score Prediction in mp-MRI via FocalNet. *IEEE Trans. Med. Imaging*, 38(11):2496–2506, nov 2019. ISSN 1558254X. doi: 10.1109/TMI.2019.2901928.

Numa Cellini, Alessio G. Morganti, Gian C. Mattiucci, Vincenzo Valentini, Mariavittoria Leone, Stefano Luzi, Riccardo Manfredi, Nicola Dinapoli, Cinzia Digesu', and Daniela Smaniotto. Analysis of intraprostatic failures in patients treated with hormonal therapy and radiotherapy: Implications for conformal therapy planning. *Int. J. Radiat. Oncol. Biol. Phys.*, 53(3):595–599, jul 2002. ISSN 03603016. doi: 10.1016/S0360-3016(02)02795-5.

Ian Chan, William Wells, Robert V. Mulkern, Steven Haker, Jianqing Zhang, Kelly H. Zou, Stephan E. Maier, and Clare M. C. Tempany. Detection of prostate cancer by integration of line-scan diffusion, T2-mapping and T2-weighted magnetic resonance imaging; a multichannel statistical classifier. *Med. Phys.*, 30(9):2390–2398, aug 2003. ISSN 00942405. doi: 10.1118/1.1593633. URL http://doi.wiley.com/10.1118/1.1593633.

Hugh Chen, Scott Lundberg, and Su-In Lee. Checkpoint Ensembles: Ensemble Methods from a Single Training Process. *arXiv*, oct 2017. URL http://arxiv.org/abs/1710.03282.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training Deep Nets with Sublinear Memory Cost. *arXiv*, apr 2016. URL http://arxiv.org/abs/1604.06174.

F. Chinesta, A. Leygue, F. Bordeu, J. V. Aguado, E. Cueto, D. Gonzalez, I. Alfaro, A. Ammar, and A. Huerta. PGD-Based Computational Vademecum for Efficient Design, Optimization and Control. *Arch. Comput. Methods Eng.*, 20(1):31–59, mar 2013. ISSN 1134-3060. doi: 10.1007/s11831-013-9080-x. URL http://link.springer.com/10.1007/s11831-013-9080-x.

Francois Chollet. Keras: Deep Learning library for Theano and TensorFlow. *GitHub Repos.*, pages 1–21, 2015.

Patrick Christ. LiTS - Liver Tumor Segmentation Challenge. *LITS-Challenge*, 2017. URL https://competitions.codalab.org/competitions/17094.

Dan C. Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Adv. Neural Inf. Process. Syst.*, volume 4, pages 2843–2851, 2012. ISBN 9781627480031. URL http://www.idsia.ch/.

Mark A. Clifford, Filip Banovac, Elliot Levy, and Kevin Cleary. Assessment of hepatic motion secondary to respiration for computer assisted interventions. *Comput. Aided Surg.*, 7(5):291–299, 2002. ISSN 1092-9088. doi: 10.1002/igs.10049. URL http://doi.wiley.com/10.1002/igs.10049.

Stéphane Cotin, Hervé Delingette, and Nicholas Ayache. A hybrid elastic model for real-time cutting, deformations, and force feedback for surgery training and simulation. *Vis.*

Comput., 16(8):437–452, dec 2000. ISSN 0178-2789. doi: 10.1007/PL00007215. URL http://link.springer.com/10.1007/PL00007215.

Steffen Czolbe, Oswin Krause, and Aasa Feragen. Semantic similarity metrics for learned image registration. *Proc. Mach. Learn. Res.*, 2021. URL http://arxiv.org/abs/2104.10051.

Jeffrey De Fauw, Joseph R. Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, George van den Driessche, Balaji Lakshminarayanan, Clemens Meyer, Faith Mackinder, Simon Bouton, Kareem Ayoub, Reena Chopra, Dominic King, Alan Karthikesalingam, Cían O. Hughes, Rosalind Raine, Julian Hughes, Dawn A. Sim, Catherine Egan, Adnan Tufail, Hugh Montgomery, Demis Hassabis, Geraint Rees, Trevor Back, Peng T. Khaw, Mustafa Suleyman, Julien Cornebise, Pearse A. Keane, and Olaf Ronneberger. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.*, 24(9):1342–1350, 2018. ISSN 1546170X. doi: 10.1038/s41591-018-0107-6. URL http://dx.doi.org/10.1038/s41591-018-0107-6.

Dhanannjay Deo and Suvranu De. PhyNeSS: A Physics-driven Neural Networks-based Surgery Simulation system with force feedback. In *World Haptics 2009 - Third Jt. EuroHaptics Conf. Symp. Haptic Interfaces Virtual Environ. Teleoperator Syst.*, pages 30–34. IEEE, 2009. doi: 10.1109/WHC.2009.4810896. URL http://ieeexplore.ieee.org/document/4810896/.

Frank-Jan H Drost, Daniël F Osses, Daan Nieboer, Ewout W Steyerberg, Chris H Bangma, Monique J Roobol, and Ivo G Schoots. Prostate MRI, with or without MRI-targeted biopsy, and systematic biopsy for detecting prostate cancer. *Cochrane Database Syst. Rev.*, 4(4):CD012663, apr 2019. ISSN 1469-493X. doi: 10.1002/14651858.cd012663.pub2. URL http://doi.wiley.com/10.1002/14651858.CD012663.pub2.

Alpaslan Duysak, Jian J. Zhang, and V. Ilankovan. Efficient modelling and simulation of soft tissue deformation using mass-spring systems. *Int. Congr. Ser.*, 1256(C):337–342, jun 2003. ISSN 05315131. doi: 10.1016/S0531-5131(03)00423-0. URL https://www.sciencedirect.com/science/article/pii/S0531513103004230.

Tadeusz Dyba, Giorgia Randi, Freddie Bray, Carmen Martos, Francesco Giusti, Nicholas Nicholson, Anna Gavin, Manuela Flego, Luciana Neamtiu, Nadya Dimitrova, Raquel Negrão Carvalho, Jacques Ferlay, and Manola Bettio. The European cancer burden in 2020: Incidence and mortality estimates for 40 countries and 25 major cancers. *Eur. J. Cancer*, 157:308, nov 2021. ISSN 18790852. doi: 10.1016/J.EJCA.2021.07.039. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8568058/.

Jonathan I Epstein, William C Allsbrook, Mahul B Amin, and Lars L Egevad. The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. *Am. J. Surg. Pathol.*, 29(9):1228–1242, sep 2005. ISSN 0147-5185. doi: 10.1097/01.pas.0000173646.99337.b1. URL http://journals.lww.com/00000478-200509000-00015.

Jonathan I. Epstein, Lars Egevad, Mahul B. Amin, Brett Delahunt, John R. Srigley, and Peter A. Humphrey. The 2014 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma definition of grading patterns and proposal for a new grading system. *Am. J. Surg. Pathol.*, 40(2):244–252, 2016. ISSN 15320979. doi: 10.1097/PAS.0000000000000530. URL https://pubmed.ncbi.nlm.nih.gov/26492179/.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nat. 2017 5427639*, 542(7639):115–118, jan 2017. ISSN 1476-4687. doi: 10.1038/nature21056. URL https://www.nature.com/articles/nature21056.

Isensee Fabian, Jäger Paul, Zimmerer David Wasserthal Jakob, Petersen Jens, Kohl Simon, Schock Justus, Klein Andre, Roß Tobias, Wirkert Sebastian, Neher Peter, Dinkelacker Stefan, Köhler Gregor, and Maier-Hein Klaus. batchgenerators - a python framework for data augmentation, jan 2020. URL https://zenodo.org/record/3632567.

William Falcon et al. PyTorch Lightning, 2019. URL https://github.com/PyTorchLightning/pytorch-lightning.

François Faure, Christian Duriez, Hervé Delingette, Jérémie Allard, Benjamin Gilles, Stéphanie Marchesseau, Hugo Talbot, Hadrien Courtecuisse, Guillaume Bousquet, Igor Peterlik, and Stéphane Cotin. SOFA: A Multi-Model Framework for Interactive Physical Simulation. In Yohan Payan, editor, *Soft Tissue Biomech. Model. Comput. Assist. Surg.*, volume 11, pages 283–321. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-29013-8. doi: 10.1007/8415_2012_125. URL http://link.springer.com/10.1007/8415_2012_125.

Daniel Freedman, Richard J. Radke, Tao Zhang, Yongwon Jeong, D. Michael Lovelock, and George T.Y. Chen. Model-based segmentation of medical imagery by matching distributions. *IEEE Trans. Med. Imaging*, 24(2):281–292, mar 2005. ISSN 02780062. doi: 10.1109/TMI.2004.841228.

Yabo Fu, Yang Lei, Tonghe Wang, Pretesh Patel, Ashesh B. Jani, Hui Mao, Walter J. Curran, Tian Liu, and Xiaofeng Yang. Biomechanically constrained non-rigid MR-TRUS prostate registration using deep learning based 3D point cloud matching. *Med. Image Anal.*, 67:101845, jan 2021. ISSN 1361-8415. doi: 10.1016/J.MEDIA.2020.101845.

Y. C. Fung and Richard Skalak. Biomechanics: Mechanical Properties of Living Tissues. *J. Biomech. Eng.*, 103(4):231, 1981. ISSN 01480731. doi: 10.1115/1.3138285. URL http://biomechanical.asmedigitalcollection.asme.org/article.aspx?articleid=1395445.

Giorgio Gandaglia, Riccardo Leni, Freddie Bray, Neil Fleshner, Stephen J. Freedland, Adam Kibel, Pär Stattin, Hendrick Van Poppel, and Carlo La Vecchia. Epidemiology and Prevention of Prostate Cancer. *Eur. Urol. Oncol.*, 4(6):877–892, dec 2021. ISSN 2588-9311. doi: 10.1016/J.EUO.2021.09.006. URL https://pubmed.ncbi.nlm.nih.gov/34716119/.

Gabriele Gaziev, Karan Wadhwa, Tristan Barrett, Brendan C. Koo, Ferdia A. Gallagher, Eva Serrao, Julia Frey, Jonas Seidenader, Lina Carmona, Anne Warren, Vincent Gnanapragasam, Andrew Doble, and Christof Kastner. Defining the learning curve for multiparametric magnetic resonance imaging (MRI) of the prostate using MRI-transrectal ultrasonography (TRUS) fusion-guided transperineal prostate biopsies as a validation tool. *BJU Int.*, 117(1):80–86, jan 2016. ISSN 14644096. doi: 10.1111/bju.12892.

Maryellen L. Giger and Kenji Suzuki. Computer-aided diagnosis. In *Biomed. Inf. Technol.*, pages 359–374. Elsevier Inc., jan 2008. ISBN 9780123735836. doi: 10.1016/B978-012373583-6.50020-7.

D. González, J. V. Aguado, E. Cueto, E. Abisset-Chavanne, and F. Chinesta. kPCA-Based Parametric Solutions Within the PGD Framework. *Arch. Comput. Methods Eng.*, 25(1): 69–86, jan 2018. doi: 10.1007/s11831-016-9173-4. URL http://link.springer.com/10.1007/s11831-016-9173-4.

David González, Elías Cueto, and Francisco Chinesta. Computational Patient Avatars for Surgery Planning. *Ann. Biomed. Eng.*, 44(1):35–45, jan 2016. ISSN 1573-9686. doi: 10.1007/s10439-015-1362-z. URL http://link.springer.com/10.1007/s10439-015-1362-z.

Jérémie Haffner, Eric Potiron, Sébastien Bouyé, Philippe Puech, Xavier Leroy, Laurent Lemaitre, and Arnauld Villers. Peripheral zone prostate cancers: Location and intraprostatic patterns of spread at histopathology. *Prostate*, 69(3):276–282, feb 2009. ISSN 02704137. doi: 10.1002/pros.20881. URL http://doi.wiley.com/10.1002/pros.20881.

Thomas Hambrock, Diederik M. Somford, Caroline Hoeks, Stefan A.W. Bouwense, Henkjan Huisman, Derya Yakar, Inge M. van Oort, J. Alfred Witjes, Jurgen J. Fütterer, and Jelle O. Barentsz. Magnetic Resonance Imaging Guided Prostate Biopsy in Men With Repeat Negative Biopsies and Increased Prostate Specific Antigen. *J. Urol.*, 183(2):520–528, feb 2010. ISSN 00225347. doi: 10.1016/J.JURO.2009.10.022. URL www.jurology.com.

Grant Haskins, Jochen Kruecker, Uwe Kruger, Sheng Xu, Peter A. Pinto, Brad J. Wood, and Pingkun Yan. Learning deep similarity metric for 3D MR–TRUS image registration. *Int. J. Comput. Assist. Radiol. Surg.*, 14(3):417–425, mar 2019. ISSN 18616429. doi: 10.1007/S11548-018-1875-7/TABLES/1. URL https://link.springer.com/article/10.1007/s11548-018-1875-7.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *2015 IEEE Int. Conf. Comput. Vis.*, volume 2015 Inter, pages 1026–1034. IEEE, dec 2015. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.123. URL http://ieeexplore.ieee.org/document/7410480/.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016-Decem:770–778, dec 2016. ISSN 10636919. doi: 10.1109/CVPR.2016.90. URL http://arxiv.org/abs/1512.03385.

Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick, Piotr Dollar, and Ross Girshick. Mask R-CNN, mar 2017. ISSN 19393539. URL https://ieeexplore.ieee.org/document/8372616/.

Mattias P. Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V. Gleeson, Sir Michael Brady, and Julia A. Schnabel. MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Med. Image Anal.*, 16(7):1423–1435, oct 2012. ISSN 1361-8415. doi: 10.1016/J.MEDIA.2012.05.008.

Caroline M.A. Hoeks, Martijn G. Schouten, Joyce G.R. Bomers, Stefan P. Hoogendoorn, Christina A. Hulsbergen-Van De Kaa, Thomas Hambrock, Henk Vergunst, J. P.Michiel Sedelaar, Jurgen J. Fütterer, and Jelle O. Barentsz. Three-Tesla Magnetic Resonance–Guided Prostate Biopsy in Men With Increased Prostate-Specific Antigen and Repeated, Negative, Random, Systematic, Transrectal Ultrasound Biopsies: Detection of Clinically Significant Prostate Cancers. *Eur. Urol.*, 62(5):902–909, nov 2012. ISSN 0302-2838. doi: 10.1016/J.EURURO.2012.01.047.

Yipeng Hu, Dominic Morgan, Hashim Uddin Ahmed, Doug Pendsé, Mahua Sahu, Clare Allen, Mark Emberton, David Hawkes, and Dean Barratt. A Statistical Motion Model Based on Biomechanical Simulations for Data Fusion during Image-Guided Prostate Interventions. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 5241 LNCS(PART 1):737–744, 2008. ISSN 03029743. doi: 10.1007/978-3-540-85988-8_88. URL https://link.springer.com/chapter/10.1007/978-3-540-85988-8_88.

Yipeng Hu, Hashim Uddin Ahmed, Zeike Taylor, Clare Allen, Mark Emberton, David Hawkes, and Dean Barratt. MR to ultrasound registration for image-guided prostate interventions. *Med. Image Anal.*, 16(3):687–703, apr 2012. ISSN 1361-8415. doi: 10.1016/J.MEDIA.2010.11.003.

Yipeng Hu, Eli Gibson, Hashim Uddin Ahmed, Caroline M. Moore, Mark Emberton, and Dean C. Barratt. Population-based prediction of subject-specific prostate deformation for MR-to-ultrasound image registration. *Med. Image Anal.*, 26(1):332–344, dec 2015. ISSN 1361-8415. doi: 10.1016/J.MEDIA.2015.10.006.

Yipeng Hu, Marc Modat, Eli Gibson, Wenqi Li, Nooshin Ghavami, Ester Bonmati, Guotai Wang, Steven Bandula, Caroline M. Moore, Mark Emberton, Sébastien Ourselin, J. Alison Noble, Dean C. Barratt, and Tom Vercauteren. Weakly-supervised convolutional neural networks for multimodal image registration. *Med. Image Anal.*, 49:1–13, oct 2018. ISSN 1361-8415. doi: 10.1016/J.MEDIA.2018.07.002.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, 2017-Janua:2261–2269, aug 2017. doi: 10.1109/CVPR.2017.243. URL http://arxiv.org/abs/1608.06993.

Jonas Hugosson, Monique J. Roobol, Marianne Månsson, Teuvo L.J. Tammela, Marco Zappa, Vera Nelen, Maciej Kwiatkowski, Marcos Lujan, Sigrid V. Carlsson, Kirsi M. Talala, Hans Lilja, Louis J. Denis, Franz Recker, Alvaro Paez, Donella Puliti, Arnauld Villers, Xavier Rebillard, Tuomas P. Kilpeläinen, Ulf H. Stenman, Rebecka Arnsrud Godtman, Karin Stinesen Kollberg, Sue M. Moss, Paula Kujala, Kimmo Taari, Andreas Huber, Theodorus van der Kwast, Eveline A. Heijnsdijk, Chris Bangma, Harry J. De Koning, Fritz H. Schröder, and Anssi Auvinen. A 16-yr Follow-up of the European Randomized study of Screening for Prostate Cancer. *Eur. Urol.*, 76(1):43–51, jul 2019. ISSN 1873-7560. doi: 10.1016/J.EURURO.2019.02.009. URL https://pubmed.ncbi.nlm.nih.gov/30824296/.

Bulat Ibragimov and Lei Xing. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Med. Phys.*, 44(2):547–557, feb 2017. ISSN 00942405. doi: 10.1002/mp.12045. URL http://doi.wiley.com/10.1002/mp.12045.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *32nd Int. Conf. Mach. Learn. ICML 2015*, volume 1, pages 448–456. International Machine Learning Society (IMLS), feb 2015. ISBN 9781510810587.

Fabian Isensee, Paul F. Jaeger, Simon A.A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods*, 2020. ISSN 15487105. doi: 10.1038/s41592-020-01008-z. URL http://dx.doi.org/10.1038/s41592-020-01008-z.

Paul F Jaeger, Simon A A Kohl, Sebastian Bickelhaupt, Fabian Isensee, Tristan Anselm Kuder, Heinz-Peter Schlemmer, and Klaus H Maier-Hein. Retina U-Net: Embarrassingly Simple Exploitation of Segmentation Supervision for Medical Object Detection. In *Proc. Mach. Learn. Res.*, volume 116, pages 171–183, 2020. URL http://proceedings.mlr.press/v116/jaeger20a/jaeger20a.pdf.

Alex Jahya, Mark Herink, and Sarthak Misra. A framework for predicting three-dimensional prostate deformation in real time. *Int. J. Med. Robot. Comput. Assist. Surg.*, 9(4): e52–e60, dec 2013. ISSN 14785951. doi: 10.1002/rcs.1493. URL http://doi.wiley.com/10.1002/rcs.1493.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Kathryn Tunyasuvunakool, Olaf Ronneberge, Russ Bates, Augustin Zídek, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Anna Potapenko, Andrew J Ballard, Andrew Cowie, Bernardino Ro, and Demis Hassabis. High Accuracy Protein Structure Prediction Using Deep Learning, 2020. URL https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology.

Daniel Junker, Fabian Steinkohl, Veronika Fritz, Jasmin Bektic, Theodoros Tokas, Friedrich Aigner, Thomas R.W. Herrmann, Michael Rieger, and Udo Nagele. Comparison of multiparametric and biparametric MRI of the prostate: are gadolinium-based contrast agents needed for routine examinations? *World J. Urol.*, 37(4):691–699, aug 2019. ISSN 14338726. doi: 10.1007/s00345-018-2428-y. URL https://link.springer.com/article/10.1007/s00345-018-2428-y.

Irving Kaplan, Nicklas E. Oldenburg, Paul Meskell, Michael Blake, Paul Church, and Edward J. Holupka. Real time MRI-ultrasound image guided stereotactic prostate biopsy. *Magn. Reson. Imaging*, 20(3):295–299, apr 2002. ISSN 0730-725X. doi: 10.1016/S0730-725X(02)00490-3.

Davood Karimi, Qi Zeng, Prateek Mathur, Apeksha Avinash, Sara Mahdavi, Ingrid Spadinger, Purang Abolmaesumi, and Septimiu E. Salcudean. Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images. *Med. Image Anal.*, 57:186–196, 2019. ISSN 13618423. doi: 10.1016/j.media.2019.07.005. URL https://doi.org/10.1016/j.media.2019.07.005.

Xiang ke Niu, Xue hui Chen, Zhi fan Chen, Lin Chen, Jun Li, and Tao Peng. Diagnostic performance of biparametric MRI for detection of prostate cancer: A systematic review and meta-analysis. *Am. J. Roentgenol.*, 211(2):369–378, jun 2018. ISSN 15463141. doi: 10.2214/AJR.17.18946. URL www.ajronline.org.

Jörn Kemper, R. Sinkus, J. Lorenzen, C. Nolte-Ernsting, A. Stork, and G. Adam. MR elastography of the prostate: initial in-vivo application. *Rofo*, 176(8):1094–1099, aug 2004. ISSN 1438-9029. doi: 10.1055/S-2004-813279. URL https://pubmed.ncbi.nlm.nih.gov/15346284/.

Siavash Khallaghi, C. Antonio Sanchez, Abtin Rasoulian, Saman Nouranian, Cesare Romagnoli, Hamidreza Abdi, Silvia D. Chang, Peter C. Black, Larry Goldenberg, William J. Morris, Ingrid Spadinger, Aaron Fenster, Aaron Ward, Sidney Fels, and Purang Abolmaesumi. Statistical Biomechanical Surface Registration: Application to MR-TRUS Fusion for Prostate Interventions. *IEEE Trans. Med. Imaging*, 34(12):2535–2549, dec 2015a. ISSN 1558254X. doi: 10.1109/TMI.2015.2443978.

Siavash Khallaghi, C. Antonio Sánchez, Abtin Rasoulian, Yue Sun, Farhad Imani, Amir Khojaste, Orcun Goksel, Cesare Romagnoli, Hamidreza Abdi, Silvia Chang, Parvin Mousavi, Aaron Fenster, Aaron Ward, Sidney Fels, and Purang Abolmaesumi. Biomechanically Constrained Surface Registration: Application to MR-TRUS Fusion for Prostate Interventions. *IEEE Trans. Med. Imaging*, 34(11):2404–2414, nov 2015b. ISSN 1558254X. doi: 10.1109/TMI.2015.2440253.

Christopher C. Khoo, David Eldred-Evans, Max Peters, Mariana Bertoncelli Tanaka, Mohamed Noureldin, Saiful Miah, Taimur Shah, Martin J. Connor, Deepika Reddy, Martin Clark, Amish Lakhani, Andrea Rockall, Feargus Hosking-Jervis, Emma Cullen, Manit Arya, David Hrouda, Hasan Qazi, Mathias Winkler, Henry Tam, and Hashim U. Ahmed. Likert vs PI-RADS v2: a comparison of two radiological scoring systems for detection of clinically significant prostate cancer. *BJU Int.*, 125(1):49–55, jan 2020. ISSN 1464410X. doi: 10.1111/bju.14916. URL https://pubmed.ncbi.nlm.nih.gov/31599113/.

Andy Kitchen and Jarrel Seah. Support vector machines for prostate lesion classification. In Samuel G. Armato and Nicholas A. Petrick, editors, *Med. Imaging 2017 Comput. Diagnosis*, volume 10134, page 1013427. SPIE, mar 2017. doi: 10.1117/12.2277120. URL http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2277120.

Stefan Klein, Uulke A. van der Heide, Irene M. Lips, Marco van Vulpen, Marius Staring, and Josien P. W. Pluim. Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information. *Med. Phys.*, 35(4):1407–1417, mar 2008. ISSN 00942405. doi: 10.1118/1.2842076. URL http://doi.wiley.com/10.1118/1.2842076.

Kimia Kohestani, Jonas Wallström, Niclas Dehlfors, Ole Martin Sponga, Marianne Månsson, Andreas Josefsson, Sigrid Carlsson, Mikael Hellström, and Jonas Hugosson. Performance and inter-observer variability of prostate MRI (PI-RADS version 2) outside high-volume centres. *Scand. J. Urol.*, 53(5):304–311, sep 2019. ISSN 21681813. doi: 10.1080/21681805.2019.1675757.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, may 2012. ISSN 15577317. doi: 10.1145/3065386. URL http://dl.acm.org/citation.cfm?doid=3098997.3065386.

Rune Kvåle, Bjørn Møller, Rolf Wahlqvist, Sophie D. Fosså, Aasmund Berner, Christer Busch, Anne E. Kyrdalen, Aud Svindland, Trond Viset, and Ole J. Halvorsen. Concordance between Gleason scores of needle biopsies and radical prostatectomy specimens:

a population-based study. *BJU Int.*, 103(12):1647–1654, jun 2009. ISSN 1464-410X. doi: 10.1111/J.1464-410X.2008.08255.X. URL https://onlinelibrary.wiley.com/doi/full/10.1111/j.1464-410X.2008.08255.x.

Mariane Le Fur and Peter Caravan. The biological fate of gadolinium-based MRI contrast agents: a call to action for bioinorganic chemists. *Metallomics*, 11(2):240–254, feb 2019. ISSN 1756591X. doi: 10.1039/c8mt00302e. URL https://academic.oup.com/metallomics/article/11/2/240/5957487.

Yann Lecun and Yoshua Bengio. Convolutional Networks for Images, Speech, and Time-Series. *Handb. brain theory neural networks*, 3361(10), 1995.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, volume 2017-Janua, pages 105–114, 2017. ISBN 9781538604571. doi: 10.1109/CVPR.2017.19.

Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-Supervised Nets. *J. Mach. Learn. Res.*, 38:562–570, sep 2014. ISSN 15337928. URL http://arxiv.org/abs/1409.5185.

Yang Lei, Sibo Tian, Xiuxiu He, Tonghe Wang, Bo Wang, Pretesh Patel, Ashesh B. Jani, Hui Mao, Walter J. Curran, Tian Liu, and Xiaofeng Yang. Ultrasound prostate segmentation based on multidirectional deeply supervised V-Net. *Med. Phys.*, 46(7):3194–3206, jul 2019. ISSN 0094-2405. doi: 10.1002/mp.13577. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.13577.

Guillaume Lemaître, Robert Martí, Jordi Freixenet, Joan C. Vilanova, Paul M. Walker, and Fabrice Meriaudeau. Computer-Aided Detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: A review. *Comput. Biol. Med.*, 60:8–31, 2015. ISSN 18790534. doi: 10.1016/j.compbiomed.2015.02.009. URL http://dx.doi.org/10.1016/j.compbiomed.2015.02.009.

R. W. Lewis and K. Ravindran. Finite element simulation of metal casting. *Int. J. Numer. Methods Eng.*, 47(1-3):29–59, 2000. ISSN 00295981. doi: 10.1002/(SICI)1097-0207(20000110/30)47:1/3<29::AID-NME760<3.0.CO;2-X. URL https://onlinelibrary.wiley.com/doi/pdf/10.1002/(SICI)1097-0207(20000110/30)47:1/3{%}3C29::AID-NME760{%}3E3.0.CO;2-X.

Xiaochun Liao, David Reutens, and Zhengyi Yang. Morphology-based interslice interpolation using local intensity information for segmentation. In *Proc. - 2011 4th Int. Conf. Biomed. Eng. Informatics, BMEI 2011*, volume 1, pages 384–389, 2011. ISBN 9781424493524. doi: 10.1109/BMEI.2011.6098315. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp={&}arnumber=6098315.

Tsung Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, volume 2017-Janua, pages 936–944, 2017a. ISBN 9781538604571. doi: 10.1109/CVPR.2017.106.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):318–327, aug 2017b. URL http://arxiv.org/abs/1708.02002.

Kevin Lister, Gao Zhan, and P. Jaydev, Desai. Development of in vivo Constitutive Models for Liver: Application to Surgical Simulation. *Ann Biomed Eng*, 39(March 2011):1060–1073, 2011.

Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman. Computer-aided detection of prostate cancer in MRI. *IEEE Trans. Med. Imaging*, 33(5): 1083–1092, 2014a. ISSN 1558254X. doi: 10.1109/TMI.2014.2303821.

Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, Robin Strand, Filip Malmberg, Yangming Ou, Christos Davatzikos, Matthias Kirschner, Florian Jung, Jing Yuan, Wu Qiu, Qinquan Gao, Philip Eddie Edwards, Bianca Maan, Ferdinand van der Heijden, Soumya Ghose, Jhimli Mitra, Jason Dowling, Dean Barratt, Henkjan Huisman, and Anant Madabhushi. Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge. *Med. Image Anal.*, 18(2):359–373, feb 2014b. ISSN 13618415. doi: 10.1016/j.media.2013.12.002.

Geert Litjens, Futterer Jurgen, Henkjan Huisman, and Jurgen; Huisman Henkjan; Litjens Geert; Futterer. Data From Prostate-3T, 2015. URL http://dx.doi.org/10.7937/K9/TCIA.2015.QJTV5IL5.

Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman. ProstateX Challenge data, 2017.

Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, Robin Strand, Filip Malmberg, Yangming Ou, Christos Davatzikos, Matthias Kirschner, Florian Jung, Jing Yuan, Wu Qiu, Qinquan Gao, Philip Eddie Edwards, Bianca Maan, Ferdinand van der Heijden, Soumya Ghose, Jhimli Mitra, Jason Dowling, Dean Barratt, Henkjan Huisman, and Anant Madabhushi. PROMISE12 Results, 2020. URL https://promise12.grand-challenge.org/evaluation/results/.

Saifeng Liu, Huaixiu Zheng, Yesu Feng, and Wei Li. Prostate cancer diagnosis using deep learning with 3d multiparametric MRI, 2017. ISSN 0277-786X.

Zhiyu Liu, Wenhao Jiang, Kit-Hang Lee, Yat-Long Lo, Yui-Lun Ng, Qi Dou, Varut Vardhanabhuti, and Ka-Wai Kwok. A Two-Stage Approach for Automated Prostate Lesion Detection and Classification with Mask R-CNN and Weakly Supervised Deep Neural Network. *LNCS*, 11850:43–51, 2019. doi: 10.1007/978-3-030-32486-5_6. URL https://doi.org/10.1007/978-3-030-32486-5_6.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651, nov 2014. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2572683. URL http://ieeexplore.ieee.org/document/7478072/.

D. Lorente, F. Martínez-Martínez, M.J. Rupérez, M.A. Lago, M. Martínez-Sober, P. Escandell-Montero, J.M. Martínez-Martínez, S. Martínez-Sanchis, A.J. Serrano-López,

C. Monserrat, and J.D. Martín-Guerrero. A framework for modelling the biomechanical behaviour of the human liver during breathing in real time using machine learning. *Expert Syst. Appl.*, 71:342–357, apr 2017. ISSN 09574174. doi: 10.1016/j.eswa.2016.11.037. URL https://linkinghub.elsevier.com/retrieve/pii/S0957417416306728.

Bradley C. Lowekamp, David T. Chen, Luis Ibáñez, and Daniel Blezek. The design of simpleITK. *Front. Neuroinform.*, 7(DEC):45, dec 2013. ISSN 16625196. doi: 10.3389/FNINF.2013.00045/BIBTEX.

Steve A. Maas, Benjamin J. Ellis, Gerard A. Ateshian, and Jeffrey A. Weiss. FEBio: finite elements for biomechanics. *J. Biomech. Eng.*, 134(1):011005, jan 2012. ISSN 1528-8951. doi: 10.1115/1.4005694. URL http://www.ncbi.nlm.nih.gov/pubmed/22482660.

F. Maes, D. Loeckx, D. Vandermeulen, and P. Suetens. Image registration using mutual information. In *Handb. Biomed. Imaging Methodol. Clin. Res.*, pages 295–308. Springer, Boston, MA, jan 2015. ISBN 9780387097497. doi: 10.1007/978-0-387-09749-7_16. URL https://link.springer.com/chapter/10.1007/978-0-387-09749-7_16.

Bahram Marami, Shahin Sirouspour, Suha Ghoul, Jeremy Cepek, Sean R.H. Davidson, David W. Capson, John Trachtenberg, and Aaron Fenster. Elastic registration of prostate MR images based on estimation of deformation states. *Med. Image Anal.*, 21(1):87–103, apr 2015. ISSN 1361-8415. doi: 10.1016/J.MEDIA.2014.12.007.

Stéphanie Marchesseau, Simon Chatelin, Hervé Delingette, Stéphanie Marchesseau, Simon Chatelin, Hervé Delingette, Biomechanical Model, Yohan Payan, and Jacques Ohayon. Non linear Biomechanical Model of the Liver. *HAL*, 2017. URL https://hal.inria.fr/hal-01536406.

Giancarlo Marra, Guillaume Ploussard, Jurgen Futterer, and Massimo Valerio. Controversies in MR targeted biopsy: alone or combined, cognitive versus software-based fusion, transrectal versus transperineal approach? *World J. Urol.*, 37(2):277–287, feb 2019. ISSN 0724-4983. doi: 10.1007/s00345-018-02622-5. URL http://link.springer.com/10.1007/s00345-018-02622-5.

David Mattes, David R. Haynor, Hubert Vesselle, Thomas K. Lewellyn, and William Eubank. Nonrigid multimodality image registration. In Milan Sonka and Kenneth M. Hanson, editors, *Med. Imaging 2001 Image Process.*, volume 4322, pages 1609–1620. SPIE, jul 2001. doi: 10.1117/12.431046. URL http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=906829.

Sherif Mehralivand, Joanna H. Shih, Soroush Rais-Bahrami, Aytekin Oto, Sandra Bednarova, Jeffrey W. Nix, John V. Thomas, Jennifer B. Gordetsky, Sonia Gaur, Stephanie A. Harmon, Mohummad Minhaj Siddiqui, Maria J. Merino, Howard L. Parnes, Bradford J. Wood, Peter A. Pinto, Peter L. Choyke, and Baris Turkbey. A magnetic resonance imaging–based prediction model for prostate biopsy risk stratification. *JAMA Oncol.*, 4(5):678–685, may 2018. ISSN 23742445. doi: 10.1001/jamaoncol.2017.5667. URL https://jamanetwork.com/journals/jamaoncology/fullarticle/2673079.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmad Ahmadi. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In *2016 Fourth Int. Conf. 3D Vis.*, pages 565–571. IEEE, oct 2016. ISBN 978-1-5090-5407-7. doi: 10.1109/3DV.2016.79. URL http://ieeexplore.ieee.org/document/7785132/.

Shervin Minaee, Rahele Kafieh, Milan Sonka, Shakib Yazdani, and Ghazaleh Jamalipour Soufi. Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning. *Med. Image Anal.*, 65:101794, oct 2020. ISSN 1361-8415. doi: 10.1016/J.MEDIA.2020.101794.

Ken'Ichi Morooka, Xian Chen, Ryo Kurazume, Seiichi Uchida, Kenji Hara, Yumi Iwashita, and Makoto Hashizume. Real-time nonlinear FEM with neural network for simulating soft organ model deformation. *Med. Image Comput. Comput. Assist. Interv.*, 11 (Pt 2):742–9, 2008. ISSN 03029743. doi: 10.1007/978-3-540-85990-1-89. URL http://www.ncbi.nlm.nih.gov/pubmed/18982671.

Marilyn J. Morton, Dana H. Whaley, Kathleen R. Brandt, and Kimberly K. Amrami. Screening mammograms: Interpretation with computer-aided detection-prospective evaluation, may 2006. ISSN 00338419.

N Mottet, J Bellmunt, E Briers, M Bolla, L Bourke, P Cornford, M Santis, A Henry, S Joniau, T Lam, M D Mason, H G Van Der Poel, T H Van Der Kwast, O Rouvière, T Wiegel, N Arfi, R C N Van Den Bergh, T Van Den Broeck, M Cumberbatch, N Fossati, T Gross, M Lardas, M Liew, P Moldovan, I G Schoots, and P M Willemse. EAU - ESTRO - ESUR - SIOG Guidelines on Prostate Cancer. *Eur. Assoc. Urol.*, pages 12–18, 2017.

N Mottet, PJ Bastian, J Bellmunt, RCN van den Bergh, M Bolla, NJ van Casteren, P Cornford, S Joniau, V Matveev, TH van der Kwast, H van der Poel, O Rouvière, and T Wiegel. EAU - EANM - ESTRO - ESUR - SIOG: Guidelines on Prostate Cancer. *Eur. Assoc. Urol.*, pages 1–182, 2020.

Andriy Myronenko and Xubo Song. Point-Set Registration: Coherent Point Drift. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(12):2262–75, may 2009. ISSN 1939-3539. doi: 10.1109/TPAMI.2010.46. URL http://ieeexplore.ieee.org/document/5432191/.

Bloch N, Madabhushi A, Huisman H, Freymann J, Kirby J, Grauer M, Enquobahrie A, Jaffe C, Clarke L, and Farahani K. NCI-ISBI 2013 Challenge: Automated Segmentation of Prostate Structures. *Cancer Imaging Arch.*, 2015. doi: http://doi.org/10.7937/K9/TCIA.2015.zF0vlOPv.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve Restricted Boltzmann machines. In *ICML 2010 - Proceedings, 27th Int. Conf. Mach. Learn.*, pages 807–814, 2010. ISBN 9781605589077.

Abhishek Nan, Matthew Tennant, Uriel Rubin, and Nilanjan Ray. DRMIME: Differentiable Mutual Information and Matrix Exponential for Multi-Resolution Image Registration. *arXiv*, 2020. URL http://arxiv.org/abs/2001.09865.

L.P. Nedel and Daniel Thalmann. Real time muscle deformations using mass-spring systems. In *Proceedings. Comput. Graph. Int. (Cat. No.98EX149)*, volume 1998-Janua, pages 156–165. IEEE Comput. Soc, 1998. ISBN 0-8186-8445-3. doi: 10.1109/CGI.1998.694263. URL http://ieeexplore.ieee.org/document/694263/.

Adaloglou Nikolaos. *Deep learning in medical image analysis: a comparative analysis of multi-modal brain-MRI segmentation with 3D deep neural networks.* PhD thesis, University of Patras, 2019.

S. Niroomandi, I. Alfaro, E. Cueto, and F. Chinesta. Accounting for large deformations in real-time simulations of soft tissues based on reduced-order models. *Comput. Methods Programs Biomed.*, 105(1):1–12, jan 2012. ISSN 01692607. doi: 10.1016/j.cmpb.2010.06.012. URL http://dx.doi.org/10.1016/j.cmpb.2010.06.012.

Ozan Oktay, Wenjia Bai, Matthew Lee, Ricardo Guerrero, Konstantinos Kamnitsas, Jose Caballero, Antonio De Marvao, Stuart Cook, Declan O'Regan, and Daniel Rueckert. Multi-input cardiac image super-resolution using convolutional neural networks. In *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, volume 9902 LNCS, pages 246–254, 2016. ISBN 9783319467252. doi: 10.1007/978-3-319-46726-9_29.

OpenAI. DALL·E 2, 2022. URL https://openai.com/dall-e-2/.

Nathan Orlando, Derek J. Gillies, Igor Gyacskov, Cesare Romagnoli, David D'Souza, and Aaron Fenster. Automatic prostate segmentation using deep learning on clinically diverse 3D transrectal ultrasound images. *Med. Phys.*, 47(6):2413–2426, jun 2020. ISSN 00942405. doi: 10.1002/mp.14134. URL https://pubmed.ncbi.nlm.nih.gov/32166768/.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning, 2010. ISSN 10414347.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H Wallach, H Larochelle, A Beygelzimer, F d\textquotesingle Alché-Buc, E Fox, and R Garnett, editors, *Adv. Neural Inf. Process. Syst. 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL https://dl.acm.org/doi/10.5555/3454287.3455008.

Oscar J. Pellicer-Valero. plot_lib, 2020. URL https://doi.org/10.5281/zenodo.4395271.

Oscar J. Pellicer-Valero, José D. Martín-Guerrero, and M.J. Rupérez. Cost-free resolution enhancement in Convolutional Neural Networks for medical image segmentation. In *Proc. ESANN*, pages 145–150, 2020a. URL https://www.esann.org/sites/default/files/proceedings/2020/ES2020-51.pdf.

Oscar J. Pellicer-Valero, María José Rupérez, Sandra Martínez-Sanchis, and José D. Martín-Guerrero. Real-time biomechanical modeling of the liver using Machine Learning models trained on Finite Element Method simulations. *Expert Syst. Appl.*, 143:113083, apr 2020b. ISSN 09574174. doi: 10.1016/j.eswa.2019.113083. URL https://linkinghub.elsevier.com/retrieve/pii/S0957417419308000.

Oscar J. Pellicer-Valero, Victor Gonzalez-Perez, Juan Luis Casanova Ramón-Borja, Isabel Martín García, María Barrios Benito, Paula Pelechano Gómez, José Rubio-Briones, María José Rupérez, and José D. Martín-Guerrero. Robust Resolution-Enhanced Prostate Segmentation in Magnetic Resonance and Ultrasound Images through Convolutional Neural Networks. *Appl. Sci.*, 11(2):844, jan 2021. ISSN 2076-3417. doi: 10.3390/app11020844. URL https://www.mdpi.com/2076-3417/11/2/844.

Oscar J. Pellicer-Valero, José L. Marenco Jiménez, Victor Gonzalez-Perez, Juan Luis Casanova Ramón-Borja, Isabel Martín García, María Barrios Benito, Paula Pelechano

Gómez, José Rubio-Briones, María José Rupérez, and José D. Martín-Guerrero. Deep learning for fully automatic detection, segmentation, and Gleason grade estimation of prostate cancer in multiparametric magnetic resonance images. *Sci. Rep.*, 12(1):2975, dec 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-06730-6. URL https://www.nature.com/articles/s41598-022-06730-6.

Oscar J. Pellicer-valero, Maria José Rupérez, Victor Gonzalez-perez, and José D. Martín-Guerrero. Deep learning contributions for reducing the complexity of prostate biomechanical models. In PhD F. Chinesta, PhD, E. Cueto, PhD, Y. Payan, PhD and J. Ohayon, editor, *Reduc. Order Model. Biomech. Living Organs*. Elsevier, 2022.

Igor Peterlík, Christian Duriez, and Stéphane Cotin. Modeling and real-time simulation of a vascularized liver tissue. *Med. Image Comput. Comput. Assist. Interv.*, 15(Pt 1):50–7, 2012. ISSN 16113349. doi: 10.1007/978-3-642-33415-3_7. URL http://www.ncbi.nlm.nih.gov/pubmed/23285534.

Rosalie Plantefève, Igor Peterlik, Hadrien Courtecuisse, Raffaella Trivisonne, Jean-Pierre Radoux, and Stephane Cotin. Atlas-based transfer of boundary conditions for biomechanical simulation. *Med. Image Comput. Comput. Assist. Interv.*, 17(Pt 2): 33–40, 2014. ISSN 16113349. doi: 10.1007/978-3-319-10470-6_5. URL http://www.ncbi.nlm.nih.gov/pubmed/25485360.

Rosalie Plantefève, Igor Peterlik, Nazim Haouchine, and Stéphane Cotin. Patient-Specific Biomechanical Modeling for Guidance During Minimally-Invasive Hepatic Surgery. *Ann. Biomed. Eng.*, 44(1):139–53, jan 2016. ISSN 1573-9686. doi: 10.1007/s10439-015-1419-z. URL http://link.springer.com/10.1007/s10439-015-1419-z.

Philippe Puech, Olivier Rouvière, Raphaele Renard-Penna, Arnauld Villers, Patrick Devos, Marc Colombel, Marc-Olivier Bitker, Xavier Leroy, Florence Mège-Lechevallier, Eva Comperat, Adil Ouzzane, and Laurent Lemaitre. Prostate cancer diagnosis: multiparametric MR-targeted biopsy with cognitive and transrectal US-MR fusion guidance versus systematic biopsy–prospective multicenter study. *Radiology*, 268(2):461–9, aug 2013. ISSN 1527-1315. doi: 10.1148/radiol.13121501. URL http://www.ncbi.nlm.nih.gov/pubmed/23579051.

Xiangxiang Qin, Yu Zhu, Wei Wang, Shaojun Gui, Bingbing Zheng, and Peijun Wang. 3D multi-scale discriminative network with multi-directional edge loss for prostate zonal segmentation in bi-parametric MR images. *Neurocomputing*, 418:148–161, dec 2020. ISSN 18728286. doi: 10.1016/j.neucom.2020.07.116. URL https://linkinghub.elsevier.com/retrieve/pii/S0925231220313229.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, jun 2017. ISSN 01628828. doi: 10.1109/TPAMI.2016.2577031. URL http://image-net.org/challenges/LSVRC/2015/results.

R. S. Rivlin. Large Elastic Deformations of Isotropic Materials. IV. Further Developments of the General Theory. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, 241 (835):379–397, oct 1948a. ISSN 1364-503X. doi: 10.1098/rsta.1948.0024. URL http://rsta.royalsocietypublishing.org/cgi/doi/10.1098/rsta.1948.0024.

R. S. Rivlin. Large Elastic Deformations of Isotropic Materials. I. Fundamental Concepts. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, 240(822):459–490, jan 1948b. ISSN 1364-503X. doi: 10.1098/rsta.1948.0002. URL http://rsta.royalsocietypublishing.org/cgi/doi/10.1098/rsta.1948.0002.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 9351:234–241, may 2015. ISSN 16113349. doi: 10.1007/978-3-319-24574-4_28. URL http://arxiv.org/abs/1505.04597.

Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv*, pages 1–14, sep 2016. URL http://arxiv.org/abs/1609.04747.

Leonardo Rundo, Changhee Han, Yudai Nagano, Jin Zhang, Ryuichiro Hataya, Carmelo Militello, Andrea Tangherloni, Marco S. Nobile, Claudio Ferretti, Daniela Besozzi, Maria Carla Gilardi, Salvatore Vitabile, Giancarlo Mauri, Hideki Nakayama, and Paolo Cazzaniga. USE-Net: Incorporating Squeeze-and-Excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets. *Neurocomputing*, 365:31–43, nov 2019. ISSN 09252312. doi: 10.1016/j.neucom.2019.07.006. URL https://linkinghub.elsevier.com/retrieve/pii/S0925231219309245.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.*, 115(3):211–252, dec 2015. ISSN 15731405. doi: 10.1007/s11263-015-0816-y.

Patrick Schelb, Xianfeng Wang, Jan Philipp Radtke, Manuel Wiesenfarth, Philipp Kickingereder, Albrecht Stenzinger, Markus Hohenfellner, Heinz Peter Schlemmer, Klaus H. Maier-Hein, and David Bonekamp. Simulated clinical deployment of fully automatic deep learning for clinical prostate MRI assessment. *Eur. Radiol.*, pages 1–12, aug 2020. ISSN 14321084. doi: 10.1007/s00330-020-07086-z. URL https://doi.org/10.1007/s00330-020-07086-z.

Fritz H. Schröder, Jonas Hugosson, Monique J. Roobol, Teuvo L.J. Tammela, Stefano Ciatto, Vera Nelen, Maciej Kwiatkowski, Marcos Lujan, Hans Lilja, Marco Zappa, Louis J. Denis, Franz Recker, Antonio Berenguer, Liisa Määttänen, Chris H. Bangma, Gunnar Aus, Arnauld Villers, Xavier Rebillard, Theodorus van der Kwast, Bert G. Blijenberg, Sue M. Moss, Harry J. de Koning, and Anssi Auvinen. Screening and Prostate-Cancer Mortality in a Randomized European Study. *N. Engl. J. Med.*, 360(13):1320–1328, mar 2009. ISSN 0028-4793. URL https://www.nejm.org/doi/full/10.1056/nejmoa0810084.

Fritz H. Schröder, Jonas Hugosson, Monique J. Roobol, Teuvo L.J. Tammela, Stefano Ciatto, Vera Nelen, Maciej Kwiatkowski, Marcos Lujan, Hans Lilja, Marco Zappa, Louis J. Denis, Franz Recker, Alvaro Páez, Liisa Määttänen, Chris H. Bangma, Gunnar Aus, Sigrid Carlsson, Arnauld Villers, Xavier Rebillard, Theodorus van der Kwast, Paula M. Kujala, Bert G. Blijenberg, Ulf-Hakan Stenman, Andreas Huber, Kimmo Taari, Matti Hakama, Sue M. Moss, Harry J. de Koning, and Anssi Auvinen. Prostate-Cancer Mortality at 11 Years of Follow-up. *N. Engl. J. Med.*, 366(11):981–990, mar 2012. ISSN 0028-4793. doi: 10.1056/NEJMOA1113135/SUPPL_FILE/NEJMOA1113135_DISCLOSURES.PDF. URL https://www.nejm.org/doi/full/10.1056/nejmoa1113135.

Steven H. Selman. The McNeal prostate: A review, dec 2011. ISSN 00904295.

Maysam Shahedi, Martin Halicek, Qinmei Li, Lizhi Liu, Zhenfeng Zhang, Sadhna Verma, David M. Schuster, and Baowei Fei. A semiautomatic approach for prostate segmentation in MR images using local texture classification and statistical shape modeling. In Baowei Fei and Cristian A. Linte, editors, *Med. Imaging 2019 Image-Guided Proced. Robot. Interv. Model.*, volume 10951, page 91. SPIE, mar 2019. ISBN 9781510625495. doi: 10.1117/12.2512282. URL https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10951/2512282/A-semiautomatic-approach-for-prostate-segmentation-in-MR-images-using/10.1117/12.2512282.full.

Hang Si. A quality tetrahedral mesh generator and a 3d delaunay triangulator. Technical Report 13, Weierstraß-Institut für Angewandte Analysis und Stochastik, 2010. URL http://wias-berlin.de/software/tetgen/1.5/doc/manual/manual.pdf.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, jan 2016. ISSN 14764687. doi: 10.1038/nature16961. URL https://www.nature.com/articles/nature16961.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.* International Conference on Learning Representations, ICLR, sep 2015.

Leslie N. Smith. Cyclical Learning Rates for Training Neural Networks. In *2017 IEEE Winter Conf. Appl. Comput. Vis.*, pages 464–472. IEEE, mar 2017. ISBN 978-1-5090-4822-9. doi: 10.1109/WACV.2017.58. URL http://ieeexplore.ieee.org/document/7926641/.

Wendy L. Smith, Craig Lewis, Glenn Bauman, George Rodrigues, David D'Souza, Robert Ash, Derek Ho, Varagur Venkatesan, Dónal Downey, and Aaron Fenster. Prostate volume contouring: A 3D analysis of segmentation using 3DTRUS, CT, and MR. *Int. J. Radiat. Oncol. Biol. Phys.*, 67(4):1238–1247, mar 2007. ISSN 03603016. doi: 10.1016/j.ijrobp.2006.11.027.

Xinrui Song, Hengtao Guo, Xuanang Xu, Hanqing Chao, Sheng Xu, Baris Turkbey, Bradford J. Wood, Ge Wang, and Pingkun Yan. Cross-Modal Attention for MRI and Ultrasound Volume Registration. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 12904 LNCS:66–75, sep 2021. ISSN 16113349. doi: 10.1007/978-3-030-87202-1_7. URL https://link.springer.com/10.1007/978-3-030-87202-1_7.

Geoffrey A. Sonn, Richard E. Fan, Pejman Ghanouni, Nancy N. Wang, James D. Brooks, Andreas M. Loening, Bruce L. Daniel, Katherine J. To'o, Alan E. Thong, John T. Leppert, Katherine J. To'o, Alan E. Thong, John T. Leppert, Katherine J To, Alan E. Thong, and John T. Leppert. Prostate Magnetic Resonance Imaging Interpretation Varies Substantially Across Radiologists. *Eur. Urol. Focus*, 5(4):592–599, jul 2019. ISSN 24054569. doi: 10.1016/j.euf.2017.11.010. URL https://pubmed.ncbi.nlm.nih.gov/29226826/.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15:1929–1958, 2014. ISSN 15337928. URL http://jmlr.org/papers/v15/srivastava14a.html.

Peter Steenbergen, Karin Haustermans, Evelyne Lerut, Raymond Oyen, Liesbeth De Wever, Laura Van Den Bergh, Linda G.W. Kerkmeijer, Frank A. Pameijer, Wouter B. Veldhuis, Jochem R.N. Van Der Voort Van Zyp, Floris J. Pos, Stijn W. Heijmink, Robin Kalisvaart, Hendrik J. Teertstra, Cuong V. Dinh, Ghazaleh Ghobadi, and Uulke A. Van Der Heide. Prostate tumor delineation using multiparametric magnetic resonance imaging: Inter-observer variability and pathology validation. *Radiother. Oncol.*, 115(2):186–190, may 2015. ISSN 18790887. doi: 10.1016/j.radonc.2015.04.012. URL http://dx.doi.org/10.1016/j.radonc.2015.04.012.

Gilbert Strang and George J Fix. *An analysis of the finite element method.* Prentice-hall Englewood Cliffs, NJ, 1973.

Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv*, may 2019. URL http://arxiv.org/abs/1905.11946.

Minh Nguyen Nhat To, Dang Quoc Vu, Baris Turkbey, Peter L. Choyke, and Jin Tae Kwak. Deep dense multi-path neural network for prostate segmentation in magnetic resonance imaging. *Int. J. Comput. Assist. Radiol. Surg.*, 13(11):1687–1696, 2018. ISSN 18616429. doi: 10.1007/s11548-018-1841-4. URL https://doi.org/10.1007/s11548-018-1841-4.

Baris Turkbey and Peter L. Choyke. Prostate Magnetic Resonance Imaging: Lesion Detection and Local Staging. *Annu. Rev. Med.*, 70(1):451–459, jan 2019. ISSN 0066-4219. doi: 10.1146/annurev-med-053117-123215. URL https://www.annualreviews.org/doi/10.1146/annurev-med-053117-123215.

Costin D. Untaroiu and Yuan-Chiao Lu. Material characterization of liver parenchyma using specimen-specific finite element models. *J. Mech. Behav. Biomed. Mater.*, 26:11–22, oct 2013. ISSN 1878-0180. doi: 10.1016/j.jmbbm.2013.05.013. URL http://www.ncbi.nlm.nih.gov/pubmed/23800843.

K. C. Valanis and R. F. Landel. The Strain-Energy Function of a Hyperelastic Material in Terms of the Extension Ratios. *J. Appl. Phys.*, 38(7):2997–3002, jun 1967. ISSN 0021-8979. doi: 10.1063/1.1710039. URL http://aip.scitation.org/doi/10.1063/1.1710039.

Wendy J. M. van de Ven, Yipeng Hu, Jelle O. Barentsz, Nico Karssemeijer, Dean Barratt, and Henkjan J. Huisman. Biomechanical modeling constrained surface-based image registration for prostate MR guided TRUS biopsy. *Med. Phys.*, 42(5):2470–2481, may 2015. ISSN 2473-4209. doi: 10.1118/1.4917481. URL https://onlinelibrary.wiley.com/doi/full/10.1118/1.4917481.

Bram Van Ginneken, Cornelia M. Schaefer-Prokop, and Mathias Prokop. Computer-aided diagnosis: How to move from the laboratory to the clinic, dec 2011. ISSN 00338419.

Coen De Vente, Pieter Vos, Matin Hosseinzadeh, Josien Pluim, and Mitko Veta. Deep Learning Regression for Prostate Cancer Detection and Grading in Bi-Parametric MRI. *IEEE Trans. Biomed. Eng.*, 68(2):374–383, feb 2021. ISSN 15582531. doi: 10.1109/TBME.2020.2993528.

Yi Wang, Haoran Dou, Xiaowei Hu, Lei Zhu, Lei Zhu, Xin Yang, Ming Xu, Jing Qin, Pheng-Ann Heng, Tianfu Wang, and Dong Ni. Deep Attentive Features for Prostate Segmentation in 3D Transrectal Ultrasound. *IEEE Trans. Med. Imaging*, pages 1–1, apr 2019. ISSN 0278-0062. doi: 10.1109/tmi.2019.2913184.

Jeffrey C. Weinreb, Jelle O. Barentsz, Peter L. Choyke, Francois Cornud, Masoom A. Haider, Katarzyna J. Macura, Daniel Margolis, Mitchell D. Schnall, Faina Shtern, Clare M. Tempany, Harriet C. Thoeny, and Sadna Verma. PI-RADS Prostate Imaging - Reporting and Data System: 2015, Version 2. *Eur. Urol.*, 69(1):16–40, jan 2016. ISSN 18737560. doi: 10.1016/j.eururo.2015.08.052.

David J. Winkel, Christian Wetterauer, Marc Oliver Matthias, Bin Lou, Bibo Shi, Ali Kamen, Dorin Comaniciu, Hans-Helge Seifert, Cyrill A. Rentsch, and Daniel T. Boll. Autonomous Detection and Classification of PI-RADS Lesions in an MRI Screening Population Incorporating Multicenter-Labeled Deep Learning and Biparametric Imaging: Proof of Concept. *Diagnostics*, 10(11):951, nov 2020. ISSN 2075-4418. doi: 10.3390/diagnostics10110951. URL https://www.mdpi.com/2075-4418/10/11/951.

Piotr Woźnicki, Niklas Westhoff, Thomas Huber, Philipp Riffel, Matthias F. Froelich, Eva Gresser, Jost von Hardenberg, Alexander Mühlberg, Maurice Stephan Michel, Stefan O. Schoenberg, and Dominik Nörenberg. Multiparametric MRI for Prostate Cancer Characterization: Combined Use of Radiomics Model with PI-RADS and Clinical Parameters. *Cancers (Basel).*, 12(7):1767, jul 2020. ISSN 2072-6694. doi: 10.3390/cancers12071767. URL https://www.mdpi.com/2072-6694/12/7/1767.

Helen Xu, John S.H. Baxter, Oguz Akin, and Diego Cantor-Rivera. Prostate cancer detection using residual networks. *Int. J. Comput. Assist. Radiol. Surg.*, 14(10):1647–1650, oct 2019. ISSN 18616429. doi: 10.1007/s11548-019-01967-5.

Pingkun Yan, Sheng Xu, Ardeshir R Rastinehad, and Brad J Wood. Adversarial image registration with application for MR and TRUS image fusion BT. *9th Int. Work. Mach. Learn. Med. Imaging, MLMI 2018 held conjunction with 21st Int. Conf. Med. Image Comput. Comput.*, 11046 LNCS:197–204, 2018. URL http://dx.doi.org/10.1007/978-3-030-00919-9_23.

Ziv Yaniv, Bradley C. Lowekamp, Hans J. Johnson, and Richard Beare. SimpleITK Image-Analysis Notebooks: a Collaborative Environment for Education and Reproducible Research. *J. Digit. Imaging*, 31(3):290–303, jun 2018. ISSN 1618727X. doi: 10.1007/s10278-017-0037-8. URL https://doi.org/10.1007/s10278-017-0037-8.

Sunghwan Yoo, Isha Gujrathi, Masoom A. Haider, and Farzad Khalvati. Prostate Cancer Detection using Deep Convolutional Neural Networks. *Sci. Rep.*, 9(1), dec 2019. ISSN 20452322. doi: 10.1038/s41598-019-55972-4.

Jeries P. Zawaideh, Evis Sala, Nadeem Shaida, Brendan Koo, Anne Y. Warren, Luca Carmisciano, Kasra Saeb-Parsy, Vincent J. Gnanapragasam, Christof Kastner, and Tristan Barrett. Diagnostic accuracy of biparametric versus multiparametric prostate MRI: assessment of contrast benefit in clinical practice. *Eur. Radiol.*, 30(7):4039–4049, mar 2020. ISSN 14321084. doi: 10.1007/s00330-020-06782-0. URL https://link.springer.com/article/10.1007/s00330-020-06782-0.

Liang Zhen, Xiaoqiang Liu, Chen Yegang, Yang Yongjiao, Xu Yawei, Kang Jiaqi, Wang Xianhao, Song Yuxuan, Hu Rui, Zhang Wei, and Ou Ningjing. Accuracy of multiparametric magnetic resonance imaging for diagnosing prostate Cancer: A systematic review and meta-analysis. *BMC Cancer*, 19(1):1–15, dec 2019. ISSN 14712407. doi: 10.1186/S12885-019-6434-2/FIGURES/5. URL https://link.springer.com/article/10.1186/s12885-019-6434-2.

Qikui Zhu, Bo Du, Baris Turkbey, Peter L. Choyke, and Pingkun Yan. Deeply-supervised CNN for prostate segmentation. *Proc. Int. Jt. Conf. Neural Networks*, 2017-May: 178–184, 2017. doi: 10.1109/IJCNN.2017.7965852.

Qikui Zhu, Bo Du, and Pingkun Yan. Boundary-weighted Domain Adaptive Neural Network for Prostate MR Image Segmentation. *IEEE Trans. Med. Imaging*, pages 1–1, feb 2019a. ISSN 0278-0062. doi: 10.1109/TMI.2019.2935018. URL http://dx.doi.org/10.1109/TMI.2019.2935018.

Yi Zhu, Rong Wei, Ge Gao, Lian Ding, Xiaodong Zhang, Xiaoying Wang, and Jue Zhang. Fully automatic segmentation on prostate MR images based on cascaded fully convolution network. *J. Magn. Reson. Imaging*, 49(4):1149–1156, apr 2019b. ISSN 1053-1807. doi: 10.1002/jmri.26337. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.26337.

Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. In *5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track Proc.*, 2017. ISBN 1611.01578v2.

Contributions of biomechanical modeling and machine learning to the automatic registration of Multiparametric Magnetic Resonance and Transrectal Echography for prostate brachytherapy

Oscar José Pellicer Valero

València, July 2022