Universitatea
Transilvania
din Brașov

ȘCOALA DOCTORALĂ INTERDISCIPLINARĂ

Facultatea: Inginerie Electrică și Știința Calculatoarelor

Ing. Andrei PUIU

Valorificarea învățării auto-supervizate, a datelor sintetice și a soluțiilor de inteligență artificială sigură pentru gestionarea inovativă a cancerului

Leveraging Self-Supervised Learning, Synthetic Data, and Trustworthy Artificial Intelligence for Innovative Cancer Care

ABSTRACT

Conducător științific

Prof. univ. dr. ing. Constantin SUCIU

BRAȘOV, 2024

Universitatea
Transilvania
din Brașov

D-lui (D-nei) .....

Componența
Comisiei de doctorat

Numită prin ordinul Rectorului Universității Transilvania din Brașov

Nr. ..... din .....

PREȘEDINTE:	**Prof. univ. dr. ing. MORARU Sorin-Aurel** **Director de departament** **Universitatea Transilvania din Brașov**
CONDUCĂTOR ȘTIINȚIFIC:	**Prof. univ. dr. ing. SUCIU Constantin** **Universitatea Transilvania din Brașov**
REFERENȚI:	**Prof. univ. dr. ing. MICLEA Liviu** **Universitatea Tehnică din Cluj Napoca**
	Prof. univ. dr. ing. NECOARĂ Ion **Universitatea Națională de Știință și Tehnologie Politehnica București**
	Prof. univ. dr. fiz. URSUȚIU Doru **Universitatea Transilvania din Brașov**

Data, ora şi locul susținerii publice a tezei de doctorat: 13.09.2024, ora 9:00, sala VIII9

Eventualele aprecieri sau observații asupra conținutului lucrării vor fi transmise electronic, în timp util, pe adresa andrei.puiu@unitbv.ro

Totodată, vă invităm să luați parte la şedința publică de susținere a tezei de doctorat.

Vă mulțumim.

Contents

1. Introduction

Deep learning on restricted healthcare data: potential and challenges
Scope of the thesis
Objectives
Thesis structure and content

1.1 Deep learning on restricted healthcare data: potential and challenges

Cancer: The modern health crisis

Cancer is one of the most prevalent and devastating diseases worldwide, affecting millions of people each year. According to the World Health Organization (WHO), cancer is the second leading cause of death globally, accounting for nearly one in six deaths [1]. From a terminology standpoint, cancer refers to a cluster of diseases characterized by an unregulated growth and spread of abnormal cells. Due to its complexity, managing cancer involves a multidisciplinary approach that includes prevention, early detection, effective diagnosis, treatment and palliative care, each aspect playing a significant role in obtaining favorable patient outcomes and quality of life.

Amidst many unknowns that we nowadays face in the fight with cancer, prevention has been shown to play a crucial role in reducing its incidence. Daily habits such as regular physical activity, a healthy diet based on fruits and vegetables, avoidance of smoking and excessive alcohol consumption were all associated with a reduced likelihood of developing cancer. However, other risk factors such as inheritance of genetic variations that create predispositions to developing certain types of cancer remains uncontrollable[2].

Nevertheless, timely detection of cancer is essential for maximizing the chances of a favorable outcome. This is typically done through regular screening, allowing for an early stage diagnosis that boosts the treatment effectiveness[3]. Diagnostic procedures often include a combination of imaging, laboratory tests, biopsies, and histopathological examinations. Advanced imaging techniques like MRI, CT, and PET scans provide detailed visualizations of tumors, which are essential for staging and treatment planning. However, biopsies remain the goal standard in confirming the presence and type of cancer. Nonetheless, the advent of molecular diagnostics has revolutionized cancer diagnosis, allowing for the identification of specific genetic mutations and molecular markers. This precision medicine approach enables the customization of treatment plans based on the unique characteristics of an individual's cancer, enhancing the effectiveness of therapies and minimizing adverse effects.

Cancer treatment encompasses various modalities, each tailored to the type, stage, and location of the cancer, as well as the patient's overall health and preferences. The primary treatment options include surgery, radiation therapy, chemotherapy, immunotherapy, and targeted therapy.

Palliative care is an integral component of cancer management, focusing on improving the quality of life for patients with advanced cancer. It addresses physical symptoms like pain, nausea, and fatigue, as well as emotional, social, and spiritual needs. Palliative care can be provided alongside curative treatments, ensuring comprehensive support for patients and their families throughout the cancer journey.

Effective cancer management requires a multidisciplinary team of healthcare professionals, including oncologists, surgeons, radiologists, pathologists, nurses, and supportive care specialists. This collaborative approach ensures that all aspects of patient care are addressed, from diagnosis through treatment and beyond, providing a holistic and personalized treatment plan.

Deep learning in medicine

Over the past decades, outstanding technological advancements have remodeled the landscape of medical practices, placing the healthcare industry in a transformative phase. The exponential increase in computational power and data storage capacity has led to many innovative ideas that currently represent building blocks for delivering superior healthcare services. This is well reflected in patient outcomes, therefore playing, along with other factors, a significant role in maintaining a positive trend in the life expectancy's global average [1].

For instance, medical imaging, that became routinely employed in clinical practices, plays a major role in every aspect of health, significantly contributing to accurate diagnosis, treatment planning, delivery and follow-up [4]. Bruls and Kwee [5] investigated the number of imaging studies performed during on-call hours between 2006 and 2020, reporting a dramatic increase: the absolute number of X-rays (XR), ultrasounds (US) and computed tomographies (CT) performed within a month increased from 1105 to 1805 (63%), 36 to 118 (227%) and from 112 to 817 (629%) respectively. However, while incontestable benefits come from the intensive usage of imaging in clinical practices, radiologists nowadays experience an enormous workload, being required to interpret the scans and deliver a report to the referring clinician in a timely matter [5, 6].

In the same time, artificial intelligence (AI) has met an outstanding growth by enjoying a huge amount of attention from the scientific community. Due to its potential in modeling non-linear complex concepts, its feasibility spans across various domains, enabling a large scale of applications that were previously deemed as unapproachable. Therefore, integration of AI technologies in the healthcare industry is currently broadly explored, holding many promises in optimizing clinical workflows, and thus, improving patient care. Consequently, many deep learning (DL) based solutions have been proposed to solving various types of clinical problems, including regression, classification, segmentation and image generation [7, 8].

Deep learning based computer-aided diagnosis (DL-CAD) systems are designed to support caregivers in efficiently diagnosing patients based on a certain data type or modality, while upholding reasonable workload levels. Therefore, integration of AI in healthcare could bridge the gaps existent in current clinical practices, supporting caregivers in delivering the best possible services, improving diagnosis accuracy and therefore, patient outcomes. While a difference in opinions was observed at clinicians across various levels of experience [9, 10, 11], adoption of DL-CAD systems in clinical routines could potentially lead to a streamlined approach, advantageous for clinicians across various seniority levels and their patients.

Nevertheless, from a development standpoint, all machine learning algorithms heavily rely on sufficient, qualitative and complete data to produce reliable outputs. The training paradigm of such algorithms could be roughly classified into supervised and unsupervised learning. In the unsupervised setting, the modeling mainly refers to grouping training samples based on their similarities into a predefined number of clusters. Further, when the model is presented with a new case, it will assign in to a certain training group, thus indicating the appartenance to a specific class. However, the supervised learning paradigm have been established as the preferred methodology in training neural networks. In this setting, input-output pairs are leveraged to steer the optimization, following a trial and error framework: the model is iteratively trained by propagating the partial derivatives of an error signal through the network, thus optimizing it to perform a certain prediction task. By generally exhibiting better performance, the latter learning paradigm is widely employed in practice, especially when large scale labeled datasets are accessible.

However, a reliable supervised learning setting depends upon a series of requirements:

- Data sufficiency. The same modeling capacity that established deep learning algorithms as state of the art solutions to solving complex problems are making them susceptible to poor knowledge generalization on unseen samples when provided with insufficient training data (a phenomenon broadly referred to as "overfitting").

- Label accuracy. DL models are designed to capture patterns in the data and further leverage them to predict the outcome in accordance to the corresponding labels: Therefore, accurate annotations are essential to a reliable training process, and thus, to obtaining a robust and accurate model.

- Data completeness. The set of input features provided to the model should possess sufficient predictive capacity w.r.t. the label.

- Data quality. Samples affected by noise or artifacts should be excluded from the training database since they could prevent the model from reaching an optimal state.

Challenges in developing AI based solutions for healthcare

Despite the incontestable potential held by AI in providing solutions to overcoming emerging problems in healthcare industry, product development in this area is often hindered by a set of obstacles.

To begin with, while biomedical data is abundant its circulation is restricted due to ethical principles stemming from patient privacy infringement concerns [12, 4]. Constraints imposed by General Data Protection and Regulations (GDPR) in Europe and Health Insurance Portability and Accountability Act (HIPAA) in the US drastically restrict the access to biomedical data in order to protect patient's confidentiality. Although being fundamentally correct, these constraints are significantly obstructing the path of AI based innovations that could markedly improve clinical practices, eventually leading to an improved health care system. In order to use patient data for research and product development purposes, an informed consent must be obtained. However, an eventual permission only concerns the initial purposes, which prevents any further usage or exploration in subsequent developments. Consequently, the data sufficiency requirement for robustly training DL models is seriously affected due to difficulties in setting up large scale datasets, which should nowadays be rather prospectively collected to meet the privacy related constraints.

Secondly, the supervised learning paradigm heavily relies on accurate and qualitative annotations to reliably steer the optimization process to an optimal point. However, from one use-case to another, data annotation might require a high expertise level, thus being exclusively attainable by health care practitioners [13]. Given the already mentioned increased workload experienced by domain experts [5, 6], the prospective creation of labels for a sufficiently large dataset is rather unfeasible, requiring an extra allocated time that might not be possessed nowadays by caregivers. Nevertheless, in relatively simpler use-cases, non-proffesionel annotators could be trained in labeling data, thus allowing technological stakeholders to set up input-output training pairs based on the raw data available. Noteworthy, annotations made by non-expert personnel hold a relatively increased risk of error, potentially hindering the model training process.

Thirdly, in spite of recent improvements in data storage capacities, cloud services, and adoption of Electronic Health Records (EHRs) systems in many hospitals, sharing patient data across various healthcare providers is currently obstructed by the lack of a clear set of exchange standards and interoperability solutions. As a consequence, since patients might have encounters at different institutions for the same underlying disease, longitudinal data acquisition often suffer from incompleteness, preventing the usage of a complete patient pathway landscape in developing machine learning based solutions for certain use-cases. Therefore, the sparsity in EHRs storage currently represents a blocker in gathering coherent longitudinal data, obstructing developments that could have an outstanding potential impact on patient care.

Lastly, DL algorithms tend to be overconfident in their predictions, potentially leading to providing unreliable outputs when possessing insufficient expertise. Due to their relative increased complexity, deep learning algorithms are often deemed as black-boxes, raising skepticism around their adoption in clinical routines. Indisputable, many clinicians along with their patients are being reluctant in following suggestions coming from an abstract source without a proper understanding of the underlying reasoning processes. Consequently, the inherent lack of transparency manifested by deep learning solutions prevents them from being widely embraced in clinical practices.

State of the art solutions to overcoming healthcare specific challenges

As a response to the aforementioned challenges, technological stakeholders have proposed several solutions to address various aspects of the health care reform, including (1) anonymization, decentralized learning [14], training on encrypted data [15, 16, 17, 12] and synthetic data generation to address data scarcity issues, (2) unsupervised, semi-supervised and self-supervised learning paradigms to address the labeling related challenges, and (3), feature importance estimation and uncertainty quantification to increase trustworthiness.

Synthetic data generation (SGD) has been widely explored as an alternative or adjacent solution to gathering large scale medical datasets. The idea of fabricating virtual data to support technological developments in healthcare has an outstanding potential due to its privacy preserving properties. Therefore, many strategies have been proposed to address the trade-off between privacy and usability, roughly categorized into partially, fully and hybrid synthetic methods [18, 19, 20]. Partially synthetic data generation concept (highly related with non-perturbative anonymization methods) refers to only fabricating patient identifiable features while preserving the non-sensitive information, thus maximizing the usability of synthetically generated samples. On the contrary, fully synthetic data generation emphasizes more on the privacy-preserving properties while trading off a certain amount of practicability. Consequently, hybrid data generation aims at finding an optimal compromise between data usability and privacy preservation properties, combining real information with purely fabricated entities.

Self-supervision is currently one of the most employed techniques in handling partially labeled datasets. Initially proposed for natural language processing (NLP) problems where text data is abundant but typically lacks annotations [21], self-supervised learning paradigm has been widely embraced in healthcare technological developments due to its outstanding potential in substantially increasing the number of usable training samples required by AI algorithms to reach a reliable generalization power. The overall idea of this approach is to use unlabeled data in learning useful representations that can be repurposed afterwards to solve specific biomedical tasks based on the relatively reduced portion of annotated samples. Typically, self-supervision is jointly used with knowledge transfer techniques [13]: firstly, rich latent representations are achieved by employing surrogate problems, often referred to as pretext tasks, where ground truth information could be either inferred or synthesized from the unlabeled inputs; further, the main learning episode employs the self-supervised based pre-trained models as initial checkpoints, thus starting the optimization from an advantageous point. However, in some scenarios self-supervision is often sufficient to reaching the final solution, thus not requiring any additional fine-tuning steps. For instance, image completion (also known as inpainting) could be trained by randomly removing patches from the input and employing convolutional neural network in restoring the information in a semantically consistent way [22, 23].

However, while the aforementioned approaches could act as enablers to embedding AI in clinical practices, a reliable and transparent setting must be ensured to increase the confidence of caregivers and their patients in such solutions. Widely referred to as black boxes, deep learning models needed mechanisms that, besides providing ways of interpreting their reasoning processes, identify cases where their predictions might be error-prone. Therefore, various methods have been proposed to address these requirements.

In contrast to classical machine learning approaches, DL algorithms manifest an inherent lack of

transparency stemming from their complexity. To elucidate how neural networks capture and use patterns in the data to infer certain outcomes several approaches have been proposed [24, 25, 26, 27, 28], mainly aiming at assigning relative importance scores to each input feature with respect to the final prediction. For instance, Lundberg et. al. [27] proposed a method to rank model inputs based on their contribution to the output by estimating Shapely values. Originating from the game theory, Shap analysis was initially designed to split a reward across players based in their relative contribution to the outcome. From a deep learning perspective, the set of input features are deemed as players while the model prediction represents the reward, thus enabling the computation of a relative importance score for each predictor. Moreover, the Shap analysis can be employed in providing either cohort-level or instance-level explanations. The overall analysis is highly suitable for model evaluation and characterization, allowing developers and domain experts to early identify and fix potential malfunctions in a pre-deployment phase, significantly increasing the system trustworthiness. Equally important, instance level explanations could accompany predictions in real time, allowing caregivers to efficiently validate or disregard model's reasoning.

Ultimately, several approaches have been proposed as safe-guards, preventing neural networks from making predictions when possessing insufficient expertise. Since deep learning algorithms heavily rely on the datasets employed in their training, predictions made on new cases underrepresented in their underlying training distribution are rather uncertain. Therefore, mechanisms for allowing neural networks to provide an "I don't know" response rather than attempting to infer are tremendously important in creating a trustworthy setting, preventing potential disastrous consequences to patients well being. Many efficient approaches have been proposed to fulfill this requirements [29, 30, 31, 32], roughly categorized into probabilistic and ensemble based methods. For instance, in addition to providing state of the art solutions in many clinical applications [33], model ensembles [34], by design, can also provide uncertainty estimations by quantifying the level of agreement across various model instances, at the cost of increasing the computational overhead. Overall, incorporating such mechanisms in all DL based solutions significantly ameliorate safety related concerns, leading to an overall reliable and trustworthy system that holds many promises in improving patient care.

1.2 Scope of the thesis

This PhD thesis is aimed at studying how challenges currently faced in developing DL based solutions for the Healthcare industry can be addressed by employing a series of techniques proposed in literature. Despite the outstanding potential of AI in solving various problems to enhance current clinical practices, its immediate adoption is hindered by a set of obstacles stemming from data quality and completeness, annotations availability and the inherent lack of transparency. Therefore, we herein investigate how techniques such self-supervision, synthetic data generation, feature importance assignation and uncertainty quantification could bridge these gaps and enable developments on a set of clinically relevant use-cases. Due to their ability in modeling non-linear complex concepts, trustworthy deep neural networks could significantly reduce the clinical work burden while improving diagnosis accuracy, ultimately leading to superior patient outcomes. Therefore, enabling explorations in this area is essential to support the transformation of Healthcare to a new era, where recent technological are wisely put on patients service.

1.3 Objectives

Specifically, the following objectives were pursued throughout the thesis:

- Improve registration accuracy in a guidance system by employing a self-supervised learning approach to train DL models in medical image extrapolation.

  - Develop a framework to simulate thin slabs acquisitions from CT volumes, thus to create input-output pairs for supervised training.
  - Produce appropriate datasets to resolve (1) an asymmetric and (2) a symmetric extrapolation problems.
  - Employ a generative adversarial framework in training a deep learning model to extrapolate simulated intra-operative CTs.

- Develop a prostate cancer (PCa) patient stratification method based on synthetic longitudinal electronic health records (EHR).
  - Define workups for PCa diagnosis, treatment and follow-up.
  - Collect various statistical properties from literature or derive them from the limited available real data to ensure coherence.
  - Build a synthetic data generator that can produce reliable and realistic longitudinal electronic health records.
  - Study the feasibility of training a TNM staging prediction network on the synthetically generated patient data.

- Employ feature importance techniques and uncertainty quantification the enhance transparency and reliability of neural networks.
  - Develop a model to enhance prediction accuracy of a state of the art malignancy detection DL-CAD system by associating clinical and demographics information.
  - Develop a model to classify non small cell lung cancer (NSCLC) into its sub-types, namely lung adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC), based on the mutational status of a sparse set of genes.
  - Perform a post hoc feature importance analysis to boost explainability and interpretability of the two aforementioned DL models.
  - Investigate how uncertainty estimations could identify samples where the NSCLC subtype classification model is error-prone, increasing its reliability.

1.4 Thesis structure and content

This thesis is structured as follows: Chapter 2 stresses the importance of self-supervision by presenting an approach to train generative adversarial neural networks in extrapolating thin CT volumes, symmetrically and asymmetrically. Exclusively enabled by the self-supervised learning paradigm, our method significantly improved the accuracy of a subsequent registration phase, leading to a superior robustness of the entire pipeline. Chapter 3 presents a systematic methodology to generate realistic synthetic electronic health records. Being free of any privacy concerns, our synthetically generated data do not carry any circulatory restrictions, hence being suitable for developing AI based solutions for various needs in prostate cancer management. We exemplified this by training a neural network in assigning a TNM stage for the patients based on the longitudinal information randomized in their EHRs. Steps towards creating trustworthy AI are presented in chapter 4, where we boost the transparency of two deep learning models specialized in (1) clinically significant prostate cancer identification based on biparametric MRI images and additional clinical and/or demographics parameters, and (2) classifying NSCLC into its sub-types based on patients genomics profiles. Explainability, interpretability and uncertainty quantification techniques are employed in increasing models reliability and trustworthiness. Lastly, final conclusions and future directions of this work are drawn in chapter 5, summarizing all the findings presented throughout the thesis and their potential impact on paving the path of AI in Healthcare.

2. Self supervised learning for medical image extrapolation and registration

---

2.1 Introduction

The recent developments in Machine Learning (ML) algorithms have released an outstanding potential in crafting useful solutions to support processes in the heath care industry, aimed at improving patient outcomes, reducing the workload of clinicians and optimizing costs [4]. For instance, radiologists are nowadays highly susceptible to making errors [35] since they experience an enormous workload, the mean image reading time requirements currently being 3.5 seconds [6]. Therefore, artificial intelligence based solutions could reliably be employed in automating this process, hence reducing time and financial costs while making the entire process less error-prone.

Among all learning paradigms, the most established one remains the supervised approach where input-output annotations pairs are iteratively used in guiding algorithms to model features and correlations within the data. Training is performed by minimizing the error produced by the model based on a set of inputs at certain iteration w.r.t. the corresponding ground-truth annotation.

However, oftentimes in practice gathering large scale datasets with qualitative annotations is extremely difficult due to complexity, costs and time related reasons. From one use-case to another, labeling data may require certain levels of expertise and domain know-how, thus urging for trained annotators or clinical experts [13]. The inherent complexity in annotating large datasets often prevents the scientific community from disposing of sufficient data samples to reliably develop machine learning based solution to solve various clinical challenges, significantly restricting potential benefits for clinicians and their patients.

Due to the great potential held by machine learning algorithms in health care, scientific community proposed a series of methods to overcoming the aforementioned challenges. Among all, the self-supervised learning paradigm was suggested to produce meaningful representations of unlabeled data[13] that can be repurposed afterwards to train neural networks in performing relevant tasks based on a limited amount of annotated data, through fine-tuning. It was initially designed to handle natural language processing (NLP) problems [21], where the text data is abundant but usually lacks of labeling and structure . Due to the outstanding impact of self-supervision in the NLP area, a lot of interest emerged for this technique to be adapted in computer vision field.

Image inpainting (also known as image interpolation), and image outpainting (also referred to as image extrapolation) represent examples of the self-prediction strategy. In this chapter we demonstrate the benefits of this approach by presenting a self-supervised learning method to extrapolate

simulated thin volumes that mimic intra-operative CT (iCT) acquisitions for an improved registration to the high resolution pre-operative volumes. For this particular use-case, a classical supervised approach is not an option since, by its nature, the ground-truth information could not be established.

This chapter is organized as follows: Section 2.2 presents an introduction of the use-case followed throughout this chapter addressed by means of self-supervision, which is the essential component to which this use-case became addressable. Sections 2.3 and 2.4 present and compare two different extrapolation approaches, where volumes are asymmetrically, or symmetrically outpainted, respectively. Conclusions on the current work are drawn in section 2.5 while limitations and future outlook are discussed in section 2.6.

2.2 Thin CT extrapolation for liver needle therapies[1]

Over the past years, the use of medical imaging in computer aided interventions has become more and more popular, supporting clinicians in their workflow and thus reducing the procedural associated risks [37].

This chapter is focused on increasing the trustworthiness of liver needle therapies such as Radiofrequency Ablation (RFA) or biopsy, where real time imaging plays a main role in guiding the intervention confidently. Although it is well known that there is a trade-off between radiation dose, acquisition time and image quality, during such surgical interventions all procedures must be carried out as quickly and accurately as possible. A possible solution to this problem is to intraoperatively acquire thin images-that provide low - resolution visualizations of a small liver region - and register them with complete high resolution preoperative images [38].

Registration is a technique used to align two images with respect to the patient's internal structures. Formally, having a reference and a template image $R, T : \mathbb{R}^d \rightarrow \mathbb{R}$, registration objective is to find a transformation $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ such that $R \approx T \circ \varphi$ [39]. Therefore, registration techniques are employed to retrieve high resolution preoperative information such as lesion location and appearance and aggregate it with the thin intraoperative images revealing the real-time needle localization, thus increasing navigation confidence. Based on the operands, there are multiple types of registration including slice-to-volume, projection-to-volume, volume-to-volume, etc. [40]. Herein we focus on the latter, aiming to boost the performance of two Computer Tomograph (CT) volumes rigid registration. Two volumes can be registered using a feature-based approach, an intensity-based approach or a combination of the two techniques. In feature-based registration, a set of corresponding features (e.g., landmarks, center of mass, etc.) are used to compute the transformation φ to register a volume (called the moving or template volume, T) to the space of the other volume (fixed or reference volume, R) [41]. The intensity-based approach can be formulated as an optimization problem, seeking the best set of parameters for the transformation φ to minimize a predefined distance measure: $argmin_{\varphi}[D(R, T \circ \varphi)]$ [39, 42]. However, this approach is not robust due to the potential presence of local minimums caused by image artifacts and sub-optimal distance metrics. Combinations of the two approaches might be used to improve registration accuracy and robustness (e.g., using intensity-based registration as a refinement step for the feature-based registration).

To the best of our knowledge, registration of thin images has been overlooked so far. Since all the registration techniques are highly dependent on the amount of mutual information (common data presented by both images from different perspectives), analysis of thin images is very challenging due to their reduced field of view (FOV). However, during surgeries low-resolution thin CT slabs are acquired to mitigate the patient's exposure risk. In this context, despite performing an initial alignment based on center of mass or geometric center, intensity-based registration is prone to failure given to the distinct fields of view of the operands. To reliably retrieve the corresponding high resolution preoperative data, a feature-based approach must be considered. However landmark detection algorithms might also be affected by the thin volume quality thus yielding a poor registration

[1] This section describes experiments done in [36], which represents previously published work of the author, under the PhD research program.

performance.

We therefore propose a method to extrapolate thin CT slabs, generating additional slices from the few existing ones, hence providing enhanced context information required by registration algorithms to work robustly.

Generative adversarial networks (GANs) [43] are a state of the art method for solving tasks such as synthetic image generation [44, 45, 46], segmentation [47], super-resolution [48], denoising [49, 50], style-transfer [51, 52] and inpainting [22, 23].

Image interpolation, also known as image completion or inpainting [53], aims at filling missing regions within an image with coherent and realistic content based on the surrounding information. Thus, in image interpolation, the field of view is well defined. In contrast, image extrapolation [54, 55, 56] is a more challenging task since the field of view has to be extended by hallucinating coherent and realistic content outside the boundaries of the existing information.

In this chapter, we introduce an extrapolation methodology based on a generator network which increases the field of view of thin intraoperative CT volumes, and improves the accuracy and robustness of a subsequent registration process. To prove the efficiency of the proposed method we focus on the liver area and assume that a thin acquisition would have a thickness of approximately 5 cm. However, this can be easily adjusted for other thicknesses or use-cases.

2.3 Asymmetric extrapolation[2]

The first option we explored was an asymmetric extrapolation approach aiming at reconstructing the entire liver field of view based on the thin slab, regardless to the structures it is displaying (e.g. the gallbladder area which is located just beneath the liver). Therefore, in the context of guiding invasive interventions extrapolated images always display complete visualizations of the surgical field, thus making the subsequent registration step mostly rely on that specific area of interest. Depending on the surgical site, this property might be beneficial in terms of minimizing possible artifacts (e.g. motion, which is more prominent in lungs or bowel as compared to liver). However, extrapolating volumes asymmetrically yields a set of challenges that must be addressed for usability reasons while increasing the algorithm complexity.

First of all, the problem difficulty can be formulated as a function of the distance between areas to be extrapolated and the actual information area.Figure 2.1 shows three candidate examples of synthesized thin slabs, each depicting different areas of the liver, which is fully visible in the last column (ground-truth for extrapolation). In the second image, the thin slice is centrally localized within the liver bounding box, in which case the extrapolation problem is rather symmetrical - same amount of information must be synthetically generated in each direction. On the contrary, images 1 and 3 show extreme cases, where the location of thin are exactly at the liver's top, or the bottom. In this case the problem becomes more difficult, since it should produce synthetic information relatively away from the existing one.

Therefore, extrapolation problem difficulty increases with the distance between pixels or voxels being regressed at some location and the actual information area. In figure 2.1 those distances are qualitatively explained by the lengths of the red arrows. For instance, in the first scenario (left) regressing intensities of pixels representing the liver's top is relatively more difficult than it is in the second scenario (mid-left).

Moreover, when employing convolutional neural networks (CNNs) in performing extrapolation their feasibility is conditioned by choosing the right architecture of each specific problem. Particularly, encoder's receptive field of view at the bottleneck must be large enough to capture sufficient real information for all areas that need to be synthetically filled. This issue becomes even more pronounced in the inference phase, where one might prefer to employ the model in extrapolating higher resolution volumes.

[2]This section describes experiments done in [36], which represents previously published work of the author, under the PhD research program.

Figure 2.1: Asymmetric extrapolation scenarios. The first three columns show possible candidates for the thin slab while the last column display the full liver's field of view.

Secondly, in real world applications thin slab's relative position in the liver's grid is not available (e.g. black areas in the first three columns of figure 2.1). Therefore, although this information can be calculated and/or randomized at training time in a self-supervised setting, at inference time it is impossible to establish the extent of extrapolation before feeding the thin slab to the network. Possible approaches to overcome this limitations exist, but at the cost of increasing the pipeline complexity through addition of extra processing steps.

In this section we propose a self-supervised generative-adversarial approach to increase the thin slab's field of view by means of extrapolation, of which extent is only specified at training time through a conditional discriminator. Moreover, to infer the metadata of extrapolated volumes we employ an extra-registration step allowing us compute their spatial information.

2.3.1 Methods

In this section we introduce a self-supervised approach for extrapolating axial slices, thus enhancing the context information required by the registration algorithms to obtain a good alignment. Due to the lack of real intraoperative data, we synthesize thin images by extracting approximately 5 cm thick sub-regions (see Section 2.3.1.1) from full CT field of views.

As depicted in Figure 2.2, given a CT volume $f : \mathbb{R}^d \to \mathbb{R}$ we first use an uniform distribution to build a binary mask $m : \mathbb{R}^d \to \{0, 1\}$ to randomly remove 75% of the information through a voxel-wise multiplication, thus yielding an image $g : \mathbb{R}^d \to \mathbb{R}$. We further refer to this image as the grid image, which is defining the extrapolation extent. Next, we simulate a thin acquisition $t : \mathbb{R}^d \to \mathbb{R}$ by extracting a region of interest (ROI) out of the grid image and then employ a deep neural network to restore the missing information, thus extrapolating the thin slab across z direction.

2.3.1.1 Dataset

The dataset consisted of 1400 high resolution CT images, each of which provided a complete visualization of the liver. Furthermore, from each of these images we only considered an ROI determined by the liver bounding box with respect to the z axis (further, we refer to this as the full image). To generalize the model, we stochastically set the thickness of the full image to the height of the liver bounding box, adding $\pm$25 mm in each direction. All these images have a constant resolution of 512 $\times$ 512 in the x-y plane, with a voxel spacing of 0.8 mm, while the mean resolution for the z-axis is of 179.2 voxels (ranging from 24 to 796, with a mean voxel spacing of 1.49 mm). All the images were resampled to a spacing of [3,3,1.5] mm. Further, to create an isotropic grid of size 128 $\times$ 128 $\times$ 128 voxels, either padding or cropping was performed. To avoid numerical instability and arithmetic overflow when computing the variance, we normalized our data using the Welford's online algorithm [57].

The data were employed to develop a self-supervised learning framework, automatically creating input-output pairs from the ground-truth images: at training time, a quarter of the full volume FOV

was randomly extracted simulating an intraoperative volume of varying thickness. Further, a deep neural network was employed when reconstructing the original volume, thus extrapolating the thin slab across the z-axis.

We randomly split the data into a training set representing 80% of the data and a testing set representing the remaining 20% of the data. Additionally, we used 100 CT pairs for quantifying the registration performance.

2.3.1.2 Proposed Method

We trained our extrapolation network (also referred to as generator) within an adversarial framework, optimizing it to "fool" another neural network (called critic or discriminator) regarding the authenticity of generated samples.

The generator network first performed a repetition of the thin slab across the z-axis, increasing the thickness of the input with a factor of four, thus defining the target FOV of the extrapolated image. This repetition adapts the encoder's feature maps to the decoder's dimensions such that we can take advantage of the long term skip connections propagating the information through the network. Moreover, this strategy is beneficial in terms of expanding the receptive field of view at the bottleneck, thus using the limited amount of real information efficiently. The rest of the generator is a variation of U-net, where each block consists of a sequence of convolution, activation function and instance normalization layers [58]. In the encoder part, downsampling was performed using 2-strided convolutions, until a receptive field of view of $255 \times 255 \times 255$ voxels was obtained at the bottleneck. Nonlinearities are provided by LeakyReLU activations, while the decoder employs ReLUs. Upsampling was performed through interpolation layers followed by 1-strided convolutions.

We used similar blocks as in the generator to create a patch-discriminator [59] conditioned on the grid image (Figure 2.2—g), which, besides the image to be discriminated, was provided as an input. This image helped the critic to penalize the generator in regards to finding the right extrapolation extent. Instead of outputting a single value, the critic outputs a $8 \times 8 \times 8$ feature-map on which each element discriminates $31 \times 31 \times 31$ voxels patches in the input.

Optimization Strategy

We trained the critic to distinguish between fake ($\tilde{e}$) and real samples (f), thus maximizing the Wasserstein distance between the real (P_r) and fake (P_g) data distribution [60]:

$$L_{critic} = E_{\tilde{e}\sim P_g}[D(\tilde{e}, g)] - E_{f\sim P_r}[D(f, g)] + \lambda E_{\hat{e}\sim P_{\hat{e}}}[(|| \bigtriangledown_{\hat{e}} (D(\hat{e}, g))||_2 - 1)^2] \tag{2.1}$$

Equation (2.1) displays the objective function used to train the critic, where the third term is a gradient penalty term used to improve the training stability [61].

Secondly, we trained the generator to produce images which are indistinguishable from the real ones, thus minimizing L_{critic} by optimizing:

$$L_{adv} = -E_{\tilde{e}\sim P_g}[D(\tilde{e}, g)] \tag{2.2}$$

To further stimulate the generation of image details and consistent internal structures, in addition to the adversarial component, we also used a feature loss [62] penalty. This component aims at minimizing the L_1 distance between features F extracted from real and fake samples, respectively. The feature maps are provided by the third convolution layer of a 3D network trained in brain tumor segmentation [63].

$$L_{feat} = E_{\tilde{e},f}[||F(\tilde{e}) - F(f)||_1] \tag{2.3}$$

As depicted in Figure 2.2, the grid information (volume g) was only used at training time by the critic to constrain the generator to find the right position of the thin slab within the target field of view.

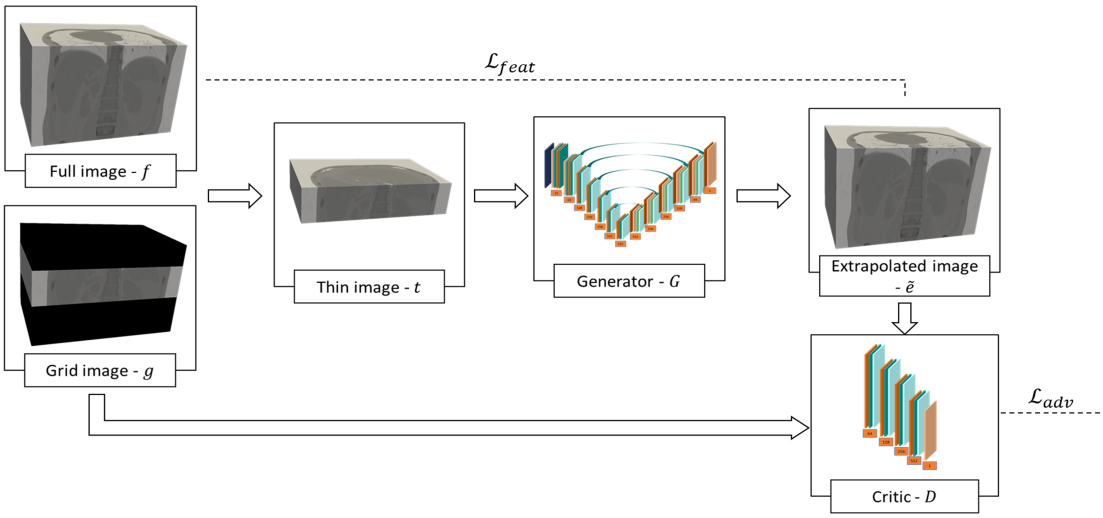


Figure 2.2: Generator optimization workflow. A conditional GAN was employed in extrapolating thin input volumes, expanding their FOV with a factor of 4.

The objective of the generator represents a weighted combination of the two terms of Equations (2.2) and (2.3). The weights have been empirically chosen such that the components take values in the same range: $\lambda_{adv} = 1$ and $\lambda_{feat} = 1$, which has been shown to lead to a better performance of the model. When using a larger weight for the supervision signal, as suggested in [59], the adversarial loss became unstable in the early stages of the training, hindering an improvement of the generated images over time.

$$L_{gen} = \lambda_{adv} L_{adv} + \lambda_{feat} L_{feat} \tag{2.4}$$

Since the cost function used to train GANs stems from another neural network trained jointly, the loss alone can be misleading when trying to identify the best performing model. Therefore, for the current experiment, model selection was performed through a visual inspection of the samples produced by the generator over time.

Image Metadata Retrieval

Since our convolutional neural network (CNN) generator operates on voxel intensity information only, we needed to perform an extra-step to retrieve the metadata of the extrapolated images.

Intuitively, the extrapolated image will have the same spacing and orientation as the thin one. However, the origin and dimension of the image changes due to the addition of synthetic information. Determining the grid dimension of the expanded volume is straight forward since we always quadruple the input field of view on the z-axis:

$$(d\tilde{e}_x, d\tilde{e}_y, d\tilde{e}_z) = (dt_x, dt_y, dt_z \times 4) \tag{2.5}$$

To compute the origin of the extrapolated volume, we first needed to determine thin slab's location within the extrapolation grid. In the current work, we addressed this issue in the post-processing phase, performing an extra-registration step to determine the extent of extrapolation as further described:

We overlapped the thin slab (sliding it across z direction) at each possible location of the extrapolated volume, calculating the voxel-wise mean squared error (Figure 2.3 - $d_{1..k}$). Next, we determined the extent extrapolation by picking the index which minimized this penalty.

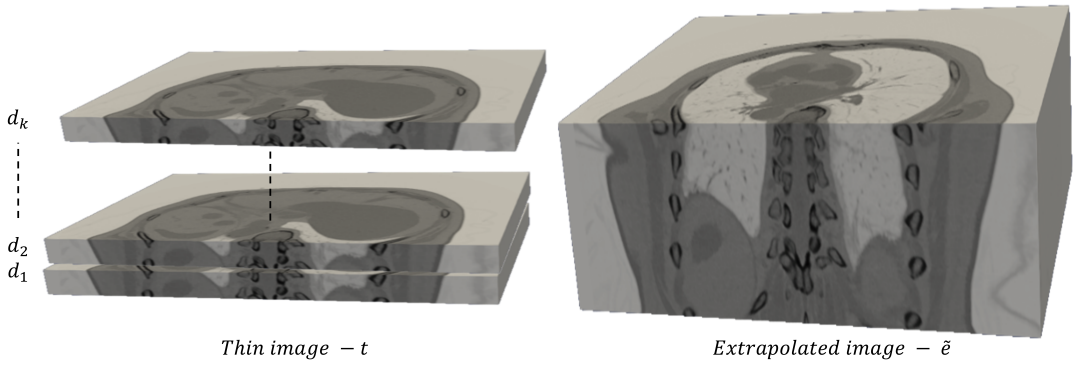


Figure 2.3: Position regression; left - thin slab t; right - extrapolated volume $\tilde{e}$.

Further, the origin of the extrapolated image was calculated using the following expression:

$$(o\tilde{e}_x, o\tilde{e}_y, o\tilde{e}_z) = (ot_x, ot_y, ot_z - argmin_{i=1..k}(d_i) \times st_z) \tag{2.6}$$

where st_z is the spacing of thin volume across z direction.

In our tests, this simple registration step was always accurate because the extrapolation network only had to copy the thin slab's intensities into the output volume without modifying them at all, hence generating relatively few errors.

2.3.1.3 Performance Quantification

One of the major challenges in image generation tasks is the lack of a goal standard method to quantify the performance of the generative models. Hence, we herein propose a goal oriented quantification method consisting in two tests: landmark detection [64, 65] and registration errors [66].

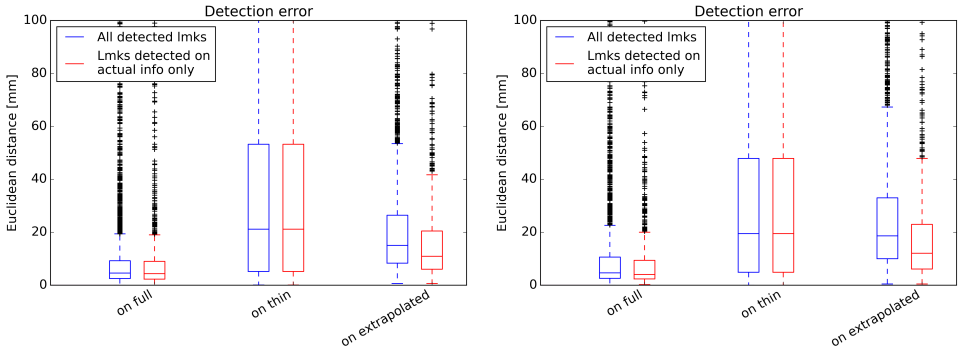
As we want to perform a feature-based registration of two volumes based on a set of corresponding landmarks, we must encourage accurate detection on the synthetic images. Hence, we first evaluate our extrapolation models based on the euclidean distance between the manual annotations and the landmarks detected on the thin, extrapolated and ground-truth volumes, respectively.

For the registration test, the 100 additional CT pairs mentioned in Section 2.3.1.1 were used as follows: we randomly extracted thin slabs from the fixed images and then employed our models for extrapolation. Further, we compared the performance between the registration of ground-truth fixed and full moving images, thin-fixed and full moving images and extrapolated-fixed and full-moving images. We used two metrics for this evaluation: surface distance and DICE, both computed on the liver masks, obtained by using the same segmentation model employed for data preprocessing.

2.3.2 Results

2.3.2.1 Landmark Detection Test

We ran a pretrained landmark detection model [65] on three variants of each test image: full, thin and extrapolated. Next, we calculated the Euclidean distance between each detected landmark and the corresponding manual annotation. The results are depicted in Figure 2.4: the proposed method reduces the median detection error by approximately 40% (from 19.51 mm to 12.08 mm, p-value = $7.38e^{-37}$) while the interquartile range (IQR) is reduced by more than a half, which means that our method increases landmark detection robustness significantly (Table 2.1).



(a) Train set. (b) Test set.

Figure 2.4: Detection errors. For the blue boxplots all detected landmarks were considered, while the red boxplots only take into account landmarks detected on the region containing the actual information.

Table 2.1: Landmark detection results on the test set.

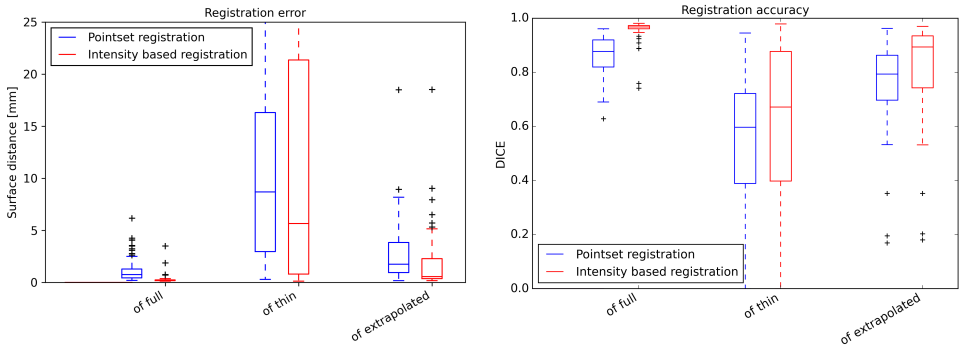
Image	Median (±IQR) [mm]	
	All Detected Landmarks	**Landmarks Detected on Actual Info**
Full volume	$4.64(\pm 8.02)$	$4.04(\pm 7.05)$
Thin volume	$19.51(\pm 43.0)$	$19.51(\pm 43.0)$
Extrapolated volume	$18.62(\pm 22.96)$	$12.08(\pm 16.86)$

Since a quarter of the full volume thickness is always used as an input, each extrapolated image should contain (1) that quarter of the FOV (we will refer to it as actual information region) and (2) three quarters of extrapolated (hallucinated) information. All detected landmarks were considered for the blue boxplots, including the ones detected in the extrapolated region. On the other hand, the red boxplots display the detection error on the actual information only, which is more relevant, since we only employed extrapolation to provide more context for detection algorithms, rather than generating synthetic points to be used for registration.

2.3.2.2 Registration Test

Figure 2.5 displays the registration results of the full moving images with all three variants of the fixed images-full, thin and extrapolated. The blue boxplots display the results of landmark-based registration which is then used as an initialization for the intensity-based registration, depicted in red.

As expected, the best performance was obtained when the full-moving images are registered with full-fixed images (having a median SD of 0.20(±0.08) mm after intensity-based registration), and the worst results were obtained when the full-moving images were registered with thin-fixed images (5.66(±20.56) mm). However, we obtained a registration performance comparable to the one corresponding to full-fixed images (0.57(±2.05) mm) by using the proposed extrapolation method as a prior step, thus reducing the thin slab registration error with a factor of 10 (p-value = $4.18e^{-6}$). The same holds true when considering the DICE score (Figure 2.5b), which increased due to the extrapolation from 0.67 to 0.88 (median).



(a) Surface distance.

(b) DICE score.

Figure 2.5: Registration results: landmark-based registration in blue, intensity-based registration in red. Each figure has three groups: left - registration of full-fixed with full-moving images; middle - registration of thin-fixed with full-moving images; right - registration of extrapolated-fixed with full-moving images.

2.4 Symmetric extrapolation

The second strategy we explored was to formulate the task as a symmetrical extrapolation problem. The main idea of this approach is to generate the same amount of synthetic information in each extrapolating direction (see second image in figure 2.1), thus minimizing the distance between hallucinated pixels/voxels and the actual information. Considering extrapolation quality as a function of the gap between regressed intensities and existing ones, we expect a better performance of this approach as compared to the one described in section 2.3. Therefore, thin slabs field of views depicting regions of the liver are enhanced regardless of exceeding its boundaries. If a thin slab describes the top of the liver extrapolation extent will expand into thoracic area, while extrapolations of volumes describing the bottom part will exhibit the lower abdominal region.

In all experiments described herein extrapolation is designed to always quadruple the thin volumes field of view. Therefore, the largest possible gap between hallucinated areas and actual ones is $1.5 \times t_{thin}$, where t_{thin} represents the thin slab's thickness. The same rationale could be applied for the asymmetric extrapolation problem, where the largest possible gap could go up to $3 \times t_{thin}$ in extreme cases. Generalizing the above example to any extrapolation extent, the maximum distance between synthetic and real voxels for the symmetric and asymmetric approaches would be $\frac{(e_r-1)}{2} \times t_{thin}$ and $(e_r - 1) \times t_{thin}$ respectively, where e_r stands for extrapolation extent ratio ($e_r = 4$ for quadrupling the thin volume FOV). Besides minimizing extrapolation extent, another benefit of this approach stems from providing all the information required to compute the spatial features of volumes, such as origin, thus not requiring extra registration steps in the pipeline. Avoiding additional error-prone steps increases robustness of the entire process, which can lead to an enhanced trustworthiness supporting adoption of such solutions in clinical routine. Therefore, this section presents a symmetric extrapolation self-supervised method that aims at improving the registration performance through expansion of thin volumes field of view as a preprocessing phase.

2.4.1 Methods

2.4.1.1 Dataset

Since symmetric extrapolation require volumes not limited to the liver field of view, for this approach we dispose of a relatively smaller dataset of 983 volumes depicting the thoracic and abdominal regions. However, the self-supervised learning framework allows us to create a large number of training samples by simultaneously randomizing the thickness of simulated thin acquisitions and

their exact location w.r.t. the entire liver bounding box. The top row of figure 2.6 shows how thin slabs of different thicknesses can be sampled from the same full volume: while the asymmetric approach was constrained at producing a synthetic version of the entire liver based on a quarter of it (varying slab thicknesses was solely based on the anatomy, e.g. liver's height divided by 4), the symmetric counterpart allowed us to sample any thickness from an uniform distribution, particularly $t_{thin} \approx U[30mm, 50mm]$. The second randomization point is the actual location of the thin slab within the liver grid, as depicted in the bottom part of figure 2.6: a thin slab of certain thickness could display different regions of the liver. Therefore, a large number of input-output pairs can be created from the same original volume, allowing us to synthetically create sufficient training samples.

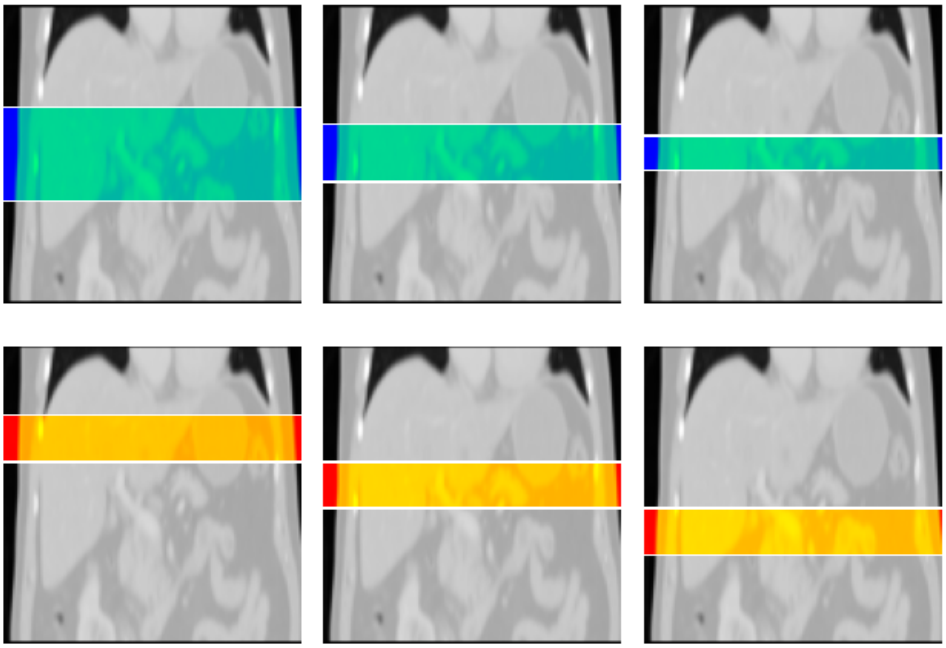


Figure 2.6: Synthetic input-output pair generation for training the symmetric extrapolation model.

Experiments described in this chapter used 5 different synthetic input-output pairs from each image, thus creating a dataset of 4915 examples. All volumes have a constant grid size of 512 in x-y plane with a mean spacing between voxels of $0.8mm$ (ranging from $0.45mm$ to $0.97mm$). The mean voxel count across z direction is 473 (from 59 to 2241) with a mean spacing of $1.44mm$ ($0.3mm$ to $8mm$). All images were downsampled to a constant spacing of $3mm \times 3mm \times 1.5mm$ and a constant grid-size of $128 \times 128 \times 128$ was obtain for the full images by padding or cropping.

Data was further split at cohort level into training and test sets, comprising 80% and 20% of the patients respectively. Similarly to the experiments carried for asymmetric extrapolation, we used the same additional set of 100 CT pairs to evaluate registration performance, thus obtaining a one-to-one comparison of the two approaches.

2.4.1.2 Proposed method

In terms of optimization framework and model architecture we used the same approach as in section 2.3.1.2. However, the image metadata retrieval presented there, particularly the registration of thin volume to the extrapolated one, is not longer required in the symmetric extrapolation setting, where all spatial information can be computed directly. While the voxel spacing and volume dimension are still retrieved consistently, computing the origin is straightforward, as depicted in equation 2.7.

$$(o\tilde{e}_x, o\tilde{e}_y, o\tilde{e}_z) = (ot_x, ot_y, ot_z - \frac{gsf_z - gst_z}{2} \times st_z) \tag{2.7}$$

where st_z is the thin image spacing, gsf_z is the grid size of the full image and gst_z is the grid size of the thin image, all across z axis.

2.4.2 Results

2.4.2.1 Landmark detection test

Table 2.2 shows de median errors and the corresponding IQRs for body markers detected on 1) entire images and 2) thin simulation areas. In comparison to asymmetric extrapolation setting, errors of body markers detected on thin images are larger, mainly due the uniformly distributed thicknesses, ranging from 3 to 5 cm in the current approach. However, extrapolation reduces these errors by a factor of 2, decreasing the median Euclidean distance from 21.73 to 10.86 mm, when only taking into account landmarks detected on the actual region. Moreover, when considering all body markers regardless to area they have been detected in, the median error is still improved by 30% compared to the thin image detection per se, while the asymmetric extrapolation only produced an 5% improvement.

Table 2.2: Landmark detection results on the test set for symmetric extrapolation.

Image	Median (±IQR) [mm]	
	All Detected Landmarks	Landmarks Detected on Actual Info
Full volume	$4.68(\pm 8.49)$	$4.60(\pm 9.21)$
Thin volume	$21.73(\pm 41.46)$	$21.73(\pm 41.46)$
Extrapolated volume	$15.52(\pm 21.48)$	$10.86(\pm 16.45)$

2.4.2.2 Registration test

As depicted in table 2.3, registration is superior when simulated thin acquisitions are extrapolated in the preprocessing phase, regardless to the registration type (e.g. landmark based or intensity based). To quantify alignment accuracy, we computed the surface distances and DICE scores on binary liver masks of fix and moving CTs, aligned with transforms obtained by registering the full-moving volumes to the full, thin and extrapolated fixed counterparts. For the intensity based registration we used landmarks detected on extrapolated volumes to create an initial alignment and then performed the intensity based refinement step using the simulated thin CT. This approach is more reliable as compared to intensity based registering the full-moving to extrapolated-fixed since it does not take into account hallucinated voxels that only had the role to enhance the contextual information of simulated intra-operative volumes.

From the perspective of both, surface distance and dice metrics, synthetically extrapolated volumes yield a superior quality of the registration. Alignment of intensity based registered thin is even inferior to the one of pointset based registered extrapolated volumes, which after the intensity based refinement also outperforms the landmark based registration of full, ground truth volumes.

Table 2.3: Registration results for the symmetric extrapolation approach.

Volume	Landmark based		Intensity based	
	Median SD (±IQR)[mm]	Median DICE (±IQR)	Median SD (±IQR)[mm]	Median DICE (±IQR)
Full volume	$0.49(\pm 1.43)$	$0.85(\pm 0.14)$	$0.04(\pm 0.24)$	$0.93(\pm 0.02)$
Thin volume	$11.55(\pm 19.12)$	$0.52(\pm 0.32)$	$10.37(\pm 24.90)$	$0.54(\pm 0.52)$
Extrapolated volume	$\mathbf{1.67(\pm 3.73)}$	$\mathbf{0.76(\pm 0.17)}$	$\mathbf{0.14(\pm 1.77)}$	$\mathbf{0.9(\pm 0.18)}$

The fact that the testset used to run the analysis presented herein is the same used in section 2.3 allows us to directly compare the two approaches through the registration performance metrics. In terms of thickness distribution, the dataset used by symmetric extrapolation experiments is more challenging since thin images field of view along z axis was uniformly sampled to exhibit regions between 30 and 50 mm thick, while synthesized thin volumes in the asymmetric counterpart were constrained on showing a quarter of the liver's height, thus heavily depending on the anatomy (thicknesses follow a Gaussian distribution with a mode of 5 cm). This observation is confirmed by the current results, thin volumes being registered with an inferior accuracy regardless to the registration type. For instance, the intensity based approach produced a median DICE score of 0.54, as compared to 0.67 obtained when extrapolating asymmetrically. Nevertheless, despite operating on relatively more difficult inputs, symmetric extrapolation outperformed the asymmetric approach, decreasing the median SD from 0.57 to 0.14 mm while increasing the median DICE score from 0.88 to 0.9.

2.4.3 Usecase conclusions

This chapter presented a self supervised symmetric extrapolation approach to enhance the field of view of thin-intraoperative images as a prior step to registration. As compared the asymmetric counterpart, the problem is simplified when extrapolating the same amount of information in each direction by 1) minimizing the distance between voxels to be estimated and the actual ones and 2) not requiring workarounds to estimate the spatial information of synthetic thick images. While the overall approach in terms of architecture and training strategy remained unchanged as presented in section 2.3.1.2, the benefits of symmetrically extrapolating information are confirmed by the current results, yielding improvements in both common image generation (SSIM 0.794 from 0.719) and task specific metrics assessed.

First, consistently to asymmetric results, our proposed extrapolation methodology improved the overall pipeline performance and stability. Landmark detection accuracy is crucial in providing a good initial alignment of volumes since the intensity based approach is an ill-posed problem (multiple overlaps could produce the same error). We showed that extrapolation reduced the median detection error from 21.73 obtained on the thin, to 10.86 mm, thus encouraging a superior initial alignment. Indubitably, these improvements are reflected in the subsequent pointset registration performance, significantly increasing the median DICE score from 0.52 to 0.76. Nonetheless, improvements in the quality of initial transforms are propagated through the intensity based registration that yields a final median DICE of 0.9, which is relatively close to the one produced by ground-truth full volumes (0.93).

Secondly, when compared to the asymmetric approach, despite of using a more challenging input that has a relatively higher density of 30 mm thick synthesized thin slabs, the proposed symmetric extrapolation method reduced the median detection error from 12.08 to 10.86 mm when only considering landmarks detected on the actual region of the image, and from 18.62 to 15.52 mm when all body markers were used, regardless to the region of origin. This indicates a better quality of extrapolated images in the symmetric setting, thus yielding a better initial alignment. The assumption that any improvement in quality of landmark based registration has substantial benefits in the intensity based outcomes is confirmed by the current results that show improvements in median SD and DICE, from 0.57 to 0.14 mm and from 0.88 to 0.9 respectively.

Overall, results presented herein could represent a building block for increasing the robustness of image-guided therapies, and therefore reflect in improved outcomes for the end-used, namely the patient.

2.4.4 Usecase discussion

We found that the current proposed symmetric extrapolation methodology is superior to its asymmetric counterpart, leading to a better detection of landmarks, and thus, an improved initial alignment of images, that supports the final intensity based registration step to be more accurate. Sym-

metric extrapolation problem is relatively more accessible as compared to asymmetric version due to the following properties:

1. Extrapolation extent is established a priori, only depending on the FOV enhancement factor (e.g. 4 across experiments presented herein) setting. Therefore, it does not require any additional registration steps to enable the computation of spatial meta-information for extrapolated volumes.

2. By design, it is minimizing the distance between areas to be estimated and the actual information region, thus improving the structural consistency across hallucinated axial frames.

In the context of guiding invasive procedures symmetric extrapolation seems a reasonable choice since acquired intraoperative images will always be regarded as the center of extrapolation, possibly improving the near real-time registration to preoperative acquisitions. However, the aforementioned property also implies extrapolating outside boundaries of the body structure being under intervention that may lead additional challenges. For instance, from one use-case to another the pipeline requires models trained to accurately detect landmarks on the entire extrapolation domain. The availability of such models could represent a challenge, since it requires large scale datasets of annotated body markers which are non-trivial to build due to complexity, time requirements and costs.

From model perspective, Unets were recently outperformed by other architectures that employ, for instance, the attention mechanism in the encoder side to capture long term relationships in the volumes [67]. In spite of not being explored in the current work, limitations related to the receptive field of view at the bottleneck could be addressed with such methods, potentially leading to an improved realness of generated volumes. Therefore, employing attention based architectures represent a possible promising lead for future developments.

However, the methods proposed in this chapter represent a proof of concept that generative adversarial networks can be successfully employed in enhancing the robustness of CT based image guided therapies, supporting the registration phase in producing more reliable alignments.

2.5 Conclusions

This chapter described a self-supervised approach to enable development of machine learning algorithms that aim at increasing the robustness of image guidance in liver needle interventions. Particularly, since registration of thin intra-operative images is a very challenging task, simulated thin acquisitions were sampled to create input-output pairs, that can be used in training neural networks to extrapolate thin slabs as a prior step to registration.

We quantified the benefits of this extrapolation step through two task specific metrics, namely landmark detection error and registration performance. Accurate landmark detection is critical in computing initial transforms that can align the field of views of the two volumes to some decent extent, thus facilitating the subsequent intensity based rigid registration. The results presented herein prove that our proposed method increase registration robustness significantly, leading to an improved image guidance for liver needle therapies.

However, feasibility of this approach is highly conditioned by the availability of thin input - thick output pairs, which could not be produced in practice during interventions. Therefore, self-supervision is the main component to enable these developments, allowing us to create a large number of synthetic input - output pairs from CT data, and thus train volume extrapolation models. The strategy employed herein was a self-prediction scheme, where full CT volumes were altered by removing 75% of axial slices to mimic a thin acquisition, with a random thickness of 30 to 50 mm. Next, a convolutional neural network with an encoder-decoder architecture was employed in restoring initial volumes, thus extrapolating the relatively few remained axial slices across z direction.

In conclusion, self-supervision is not only a powerful paradigm when limited amount of data is available, but also in extreme scenarios, as the one presented in this chapter, where datasets for supervised learning are impossible to obtain.

2.6 Discussion

Self supervised learning was the key component that enabled developments presented herein. Following a self-prediction strategy we were able to synthetically create input-output pairs of CT volumes to train a deep neural network in extrapolating thin acquisitions. The main outcome of this approach is an improved registration, and thus a superior guidance of surgical interventions on the liver.

We quantified improvements through downstream task specific metrics, reporting the median Euclidean distance between body-markers detected on the volumes and corresponding annotations, and the median surface distance and DICE score of subsequent registration. The evident improvements generated by the proposed approach can be attributed to the extrapolation network, of which training was exclusively enabled by self-supervision. However, while providing a clear evidence that our proposed method improved registration significantly, the current analysis could not quantify the impact of self-supervision per se, since real testing data which would serve as a baseline is impossible to obtain: the gain in performance could be attributed to the architecture choice, training framework, and lastly, employed landmark detection models and rigid registration engines. However, none of this analysis would have been possible without self-supervision, that enabled developments presented in this chapter.

Usually, self-supervision is employed in pretraining deep neural networks on large scale unlabeled datasets, thus creating so called foundational models capable of deriving high level representations of the data in form of a latent space. Next, from one use-case to another, a knowledge transfer can be employed by fine-tuning these foundational models on the limited amount of available data. Generally, this approach is superior to the classic paradigm where models are trained from scratch only based on a relatively small number of training examples [13]. However, improvements in performance generated by self-supervision are also dependent on the pretext task used in the pretraining phase. From one application to another, a pretext task could be preferred over others.

To reliably quantify the impact of self-supervision per se, first, the classical approach of training from scratch only based on the limited available labeled data should be exploit and considered as a baseline. Next, models pretrained using self-supervision can be either directly employed or fine-tuned to perform the same task and then evaluated on the same testing data as the baseline. In context of constant architecture and testing data, all improvements in performance can be exclusively attributed to self-supervision, hence providing a better quantification of its impact on the overall performance. However, specifics of the use-case followed throughout this chapter does not allow us to conduct such analyses since a baseline is unachievable. On the other hand, as previously mentioned, developments presented herein would have been unreachable without self-supervision, thus steering a large amount of credit for the current results towards this paradigm.

Other approaches to increase trustworthiness of image guidance during interventions might also be employed, where classical training paradigm is still a viable option. For instance, landmark detection models could be specialized in reliably detecting body markers on thin volumes, acquired intra-operatively. However, annotating large datasets accordingly would be extremely exhaustive, hence making this approach less scalable as compared to our proposed extrapolation method.

To conclude, the results presented in this chapter represent a proof of concept that self-supervision could be successfully employed in areas where gathering large scale labeled datasets is problematic, or even unfeasible. The main benefits stem from facilitating the scientific community to be prepared (to some extent) to fastly response to requirements coming from the clinical side and run feasibility studies of diverse use-cases, that can be held a great potential for clinicians and their patients.

3. Synthetic data generation for prostate cancer patient stratification

Introduction
Prostate cancer patient stratification
Methods
Results
Conclusions
Discussion

3.1 Introduction

When provided with sufficient and high quality data, artificial intelligence is currently the most promising approach in solving any emerging problem, thus holding an extremely high potential in increasing the quality of life, improving processes, reducing costs, etc. However, collecting and exploring large scale qualitative datasets could be challenging in some industries due to a series of specific concerns. For instance, healthcare information is widely protected by GDPR in Europe and HIPAA in the US to maintain patient confidentiality. As a consequence, healthcare data is often suffering from incompleteness, poor quality or insufficient data points due to privacy constrains [12, 68], preventing effective development of machine learning based solutions to be adopted in clinical routines [4].

Scientific community has proposed a series of techniques and workarounds to overcome the aforementioned challenges, including anonymization , self-supervision [13], encryption [12], synthetic data generation [69], federated learning [14], etc.

Synthetic data generation (SGD) is one of the most promising approaches to overcome challenges posed by the regulatory constrains in accessing clinical restricted data. A qualitative synthetic dataset does not contain any indication of a real person while maintaining the distribution of parameters realistic as well as natural correlations between features. A synthetic dataset that mimics and preserves statistical properties of real cohorts could be used for modeling, educational purposes, simulation and prediction research, hypothesis and algorithm testing, information technology (IT) developments, etc. [20, 70].

Therefore, since this type of data could be free of privacy threats it can be widely shared with third parties or scientific community to enable timely developments in healthcare, potentially leading to a better care for the patients and also a reduced workload for clinicians. However, although efficient development of trustworthy models to be adopted in clinical practice is highly conditioned on the quality and realness of synthetic data, usually there is a trade-off between privacy preserving properties and usability. Purely synthetic datasets that are completely free of privacy related threats usually have less usability in contrast to synthetic datasets that still preserve parts of real information while also holding patient re-identification risks.

Although some scientists consider that the term synthetic should be exclusively referring to entirely fabricated data [20], SGD could be mainly categorized in three categories [18, 19]:

- Fully synthetic refers to purely fabricated data that has no indication to individuals, thus not carrying privacy risks. Despite being the strongest in terms of confidentiality preserving, its usability is usually relatively low due to an inferior representation of real world distributions and correlations between features.

- Partially synthetic data uses non-sensitive information from real world examples while synthetically samples values for parameters that could support identification of the subjects. However, since it still contain certain amount of real parameters it carries a larger risk of re-identification, while exhibiting a superior usability.

- Hybrid synthetic data refers to matching a real record to a purely fabricated one through statistical tools, and then combine features from the two records to create a synthetic sample. While disposing of relatively better privacy preservation properties in contrast to partially synthetic data, it is computationally more expensive and introduces noise.

In healthcare, synthetic data generation has been successfully employed for a variety of data types and modalities including images [71, 72], text [73, 74], electronic health records (EHRs) [69, 75], Electrocardiograms (ECG)[76, 77] and even genomics [78, 79] (with certain limitations[80] w.r.t. privacy preservation and usability trade-off). Other studies aimed at creating synthetic medical images of certain modality based on the anatomy revealed by a different imaging technique [81] through style transfer [82] rather than fabricating images from noise, which is of high importance in the context of multimodal modeling or registration.

Synthea[69] is the current state of the art technology in generating realistic fully synthetic EHRs, providing high quality fabricated clinical data completely free of privacy concerns. The framework uses general clinical care maps or guidelines and public health information such as disease incidence and statistics to create clinical disease modules that generate synthetic populations. Subjects belonging to this synthetic cohort could follow different pathways based on randomized variables sampled from data distributions revealed in literature, thus ensuring data completeness and quality if the clinical disease modules are properly designed.

In this chapter we propose a Synthea based methodology to create clinical disease modules. Our aim is to generate realistic electronic health records for patients suffering from prostate cancer, ranging from low to high risk disease consistently handled with various treatment options that fluctuate from active monitoring to definitive therapy or even palliative care. A properly generated synthetic dataset that mimic real data distributions and correlations between various parameters have an outstanding potential to enable timely developments of applications to be adopted in clinical routines, increasing the quality of care. Therefore, feasibility of our fully synthetic generated data is assessed through employing deep neural networks in stratifying prostate cancer patients in risk categories, assigning a disease stage based on the clinical information presented in the synthetic EHRs.

This chapter is organized as follows: Section 3.2 presents an overall introduction of the use-case followed throughout this chapter, stating the importance of reliably automating stage assignation for prostate cancer. Next, section 3.3.1 provides the details of our proposed synthetic data generator together with a description of data concepts and elements available in the resulted cohort and a fidelity analysis for the synthetic data. Section 3.3.2 describes a natural language processing (NLP) approach to assign TNM stages for patients based on medical codes descriptions as well as observations yielded in synthetic EHRs. The results of this approach are presented in section 3.4.2, while conclusions and overall discussion are drawn in sections 3.5 and 3.6 respectively.

3.2 Prostate cancer patient stratification

Prostate cancer is very common in men worldwide. For example, 1 in 8 men in the US will be diagnosed as having prostate cancer during his lifetime. It is estimated that there will be almost

250,000 new prostate cancer cases in the United States this year[83], similar to the incidence of lung and breast cancer, the other most common malignancies.

Aggregate European data is very similar to that of the United States[84] and the incidence and mortality of prostate cancer are rising in Asia. Worldwide at least 360,000 men die of prostate cancer every year[85].

The National Comprehensive Cancer Network (NCCN) is an alliance of 31 cancer centers that publishes and periodically updates cancer treatment guidelines that are now accepted as state-of-the-art recommendations at most cancer treatment centers. The NCCN prostate cancer treatment recommendations depend on which of five prostate cancer risk groups best describes a given patient. These risk groups are defined in terms of clinical stage (including radiological findings), blood tumor markers (e.g., prostate-specific antigen — PSA), histologic tumor grade as determined from a biopsy, and most recently, genetic subtype(s).

Clinical staging captures the amount and spread of cancer in a patient's anatomy. Staging usually consists of three components, T, N, and M, called the TNM system. T describes the size of the tumor and any spread of cancer into nearby tissue; N describes the spread of cancer to nearby lymph nodes; and M describes metastases (spread of cancer to other parts of the body). This system was created and is updated by the American Joint Committee on Cancer (AJCC)[86] and the International Union (UI)[87].

Clinical staging is often determined from multiple clinical diagnostic tests which are administrated longitudinally. Assigning a clinical stage is generally difficult and time-consuming as pertinent findings from the patient history and physical examination as well as radiographic interpretation are usually recorded in unrestricted clinical text, which could include a suggestion for the patient clinical stage solely on basis of the specific test diagnostic results. Due to this, staging ambiguities are not uncommon[88] and are usually resolved by an institutional tumor board (TB).

The effect of errors in determining the correct clinical tumor stage can range from a nuisance to the assigning of a patient to a wrong risk category and have him receive a less than optimal treatment. Most commonly, the difficulties in retrieving prostate cancer staging information from the electronic health record (EHR) pose significant challenges and increased costs for tumor registrars whose aim is to create structured databases for research and outcomes review.

Usually systems that extract clinical staging from medical records do so by combining the conclusions of the various clinicians that are taking care of the patient, which are often affected by ambiguities and conflicting statements[89] and incomplete work-up at that point. Our system differs from others in that it does not extract the staging conclusions of the various clinicians attending the patient and base the output on that. Rather, it deduces the proper stage from the original clinical and radiologic notes. To do this we employ a neural net, well-trained for this task.

3.3 Methods

Privacy constraints limit access to longitudinal clinical data, required for training artificial intelligence (AI) systems. In addition, institutions are becoming more restrictive in allowing clinical data to be used for research in general. Even if anonymized retrospective patient data access is granted, it is often limited due to the lack of patient consent which is required by most institutional review boards (IRBs). These necessary measures to protect patient privacy make it hard to training AI systems which inherently require large amounts of labeled clinical data.

To overcome this problem, we have devised an approach to generate and use a synthetic dataset of 10,000 records for training and 4000 records for testing, as described in the next sections.

3.3.1 Synthetic data generation

We have utilized the Synthea framework and added prostate cancer specific modules to create synthetic prostate cancer data. There are two inputs required for module creation, clinical care-maps

and disease statistics. Carefully guided by an experienced radiation oncologist, we drew clinical care-maps for diagnosis, localized and advanced therapy, and for follow-up.

Prostate cancer diagnosis workup

Figure 3.1 shows an overview of the diagnostic workup which is a combination of stochastic and deterministic states. Stochastic states capture possible variations which are expected in real clinical data, whereas the deterministic states consider rules based on common practice of guiding patient through, for example, diagnostic work-up for clinical staging. At the end, the clinical cancer stage and the risk category are established based on diagnostic reports generated stochastically throughout the work-up. We used statistical distributions for sampling diagnostic reports which were either collected from literature [90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106] or derived from the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial[107] that reveals, longitudinal screening information such as prostate specific antigen (PSA) levels, digital rectal exams (DRE) results, histologic grade (Gleason score) and staging information[108].

Table 3.1: Internal dataset properties.

Parameter	Mean (± std)	Minimum	Maximum
Age [years]	66.3 (±7.75)	43	89
PSA [ng/mL]	20.3 (±69.3)	0.29	1545
Volume [cc]	42.2 (±22.82)	0.57	164
PI-RADS	4.2 (± 0.88)	2	5
Gleason Group	2.3 (±1.42)	1	5

In addition, we have utilized an internal dataset to derive statistical properties that could not be derived from PLCO (e.g. PI-RADS scoring), containing 768 per-case annotations of PI-RADS, Gleason geade, PSA, gland volume and age. Table 3.1 shows a description of this collection, further referred as the internal dataset.

The simulated patient diagnostic work-up always starts with a digital rectal examination and a PSA measurement. First, a value for the PSA is sampled from a distribution derived from the PLCO screening data using a Kernel Density Estimation (KDE). For the subsequent preoperative PSA measurements, the sampling is done to ensure realistic dynamics[90]. Similarly, some of the DRE results, such as enlargements and volume estimations are sampled consistently with patient age and PSA levels, following the subgroup analysis presented in [109]. Any abnormality identified in one of these initial screening tests triggers subsequent diagnostic procedures, such as imaging or the more invasive biopsy. In contrast, if no suspicious findings emerge the screening will continue, repeating PSA measurements and/or DRE procedure in approximately 1 year, which is the recommended screening interval for prostate cancer[96].

To gain consistency, all subsequent diagnostic reports are sampled based on the correlations exposed in literature between the current diagnostic procedure (or test) and the synthesized prior ones. For example, in the screening phase all subsequent PSA measurements will be constrained on the prior ones based on the PSA velocities[101, 90] identified in various risk categories, or PSA doubling time [102]. Similarly, for patients who need additional diagnostic procedures findings are drawn either based on literature descriptions or statics derived from real datasets, as further described.

Depending on the already synthesized PSA measurements, a patient might be recommended for additional diagnostic procedures - such as mpMRI and/or biopsy - in case of abnormal outcomes (e.g. PSA > 4 ng/mL or an abnormal DRE) or a continuation of screening if no abnormalities occur (e.g., negative DRE and a low PSA value). In case of prostate cancer suspicion through screening, the patient will randomly undergo a systematic biopsy (sextant or double-sextant) or a mpMRI. In the first case, biopsy results will be sampled consistently to the only known information from the EHR, namely the PSA values. Figure 3.2 shows the Gleason grade distribution with respect to the

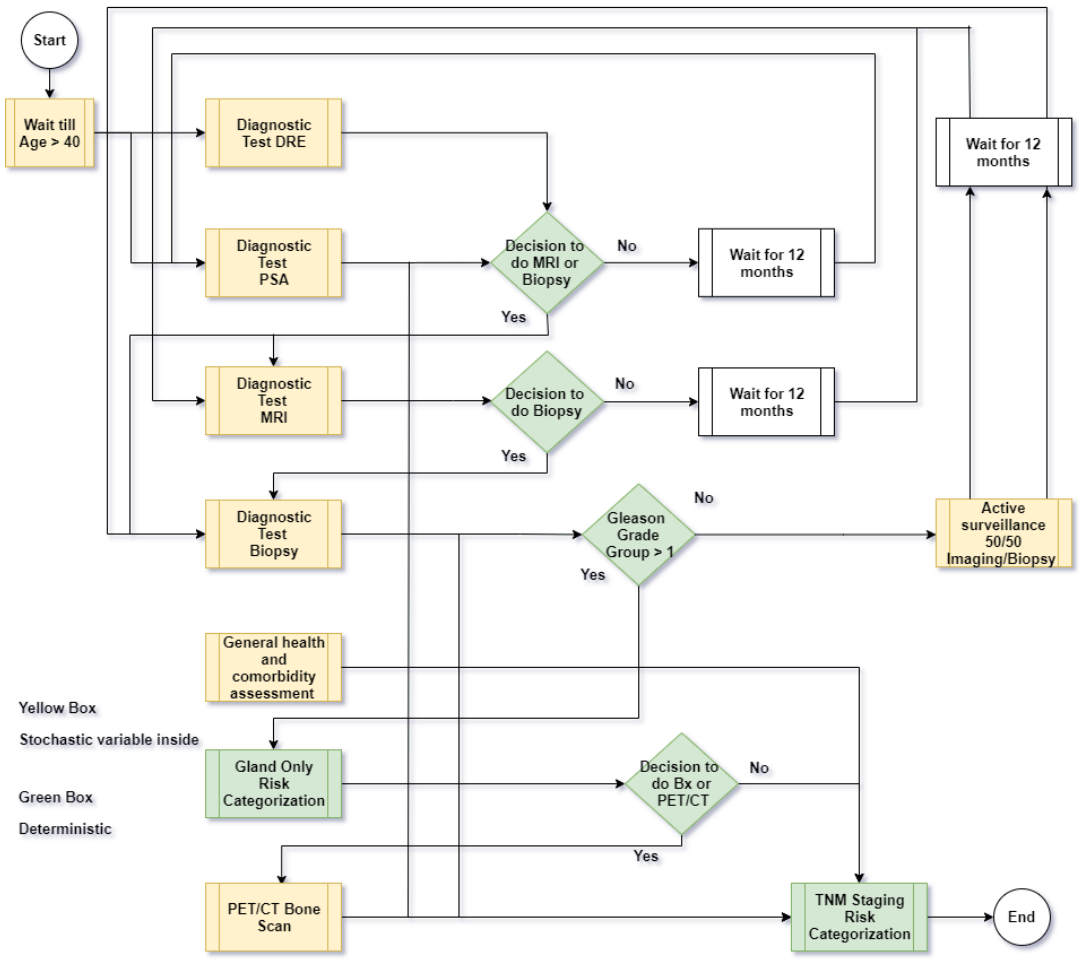


Figure 3.1: Prostate cancer diagnostic work-up: Yellow states have stochastic variables inside while the green ones are deterministic. Abbreviations: DRE, digital rectal examination; PSA, prostate specific antigen; MRI, magnetic resonance imaging; BX: biopsy; PET, positron emission tomography; CT, computed tomography;

discretized PSA level as derived from the PLCO dataset. This distributional landscape is incorporated in our proposed modules allowing them to generate cohorts that mimic statistical properties of real clinical data.

Similarly, we used the same methodology to sample a PI-RADS score w.r.t. discretized PSA levels when a patient is randomized to be investigated through imaging rather than biopsy. Nonetheless, malignant lesions revealed by mpMRI (PI-RADS >=3) must be confirmed through targeted biopsies, consistently assigning a Gleason grade to each suspicious area exposed by imaging. Apart from initial randomization, an mpMRI could be employed in other scenarios, such as following-up a negative systematic biopsy, in which case a PI-RADS score is selected based on the odds derived for each Gleason group separately. Since the PI-RADS score has an overdiagnosis tendency (high sensitivity with relatively low specificity), it will possibly reveal highly suspicious lesions even in patients with a Gleason group of 1, systematically assigned. To that extent, a targeted biopsy could be employed in establishing a final diagnostic, revealing the histopathology of suspicious areas.

At a more granular level, the number of positive biopsied cores is sampled based on the cohort distribution presented in Vallette et. al.[91], while the corresponding locations are randomly chosen. For each positive location, a cancer spread indicator, namely percentage of cancer in core, is sampled using statistical properties from [92]. When a Gleason group is sampled as described in figure 3.2, at least one of the lesions will be assigned with an appropriate Gleason score while others could be

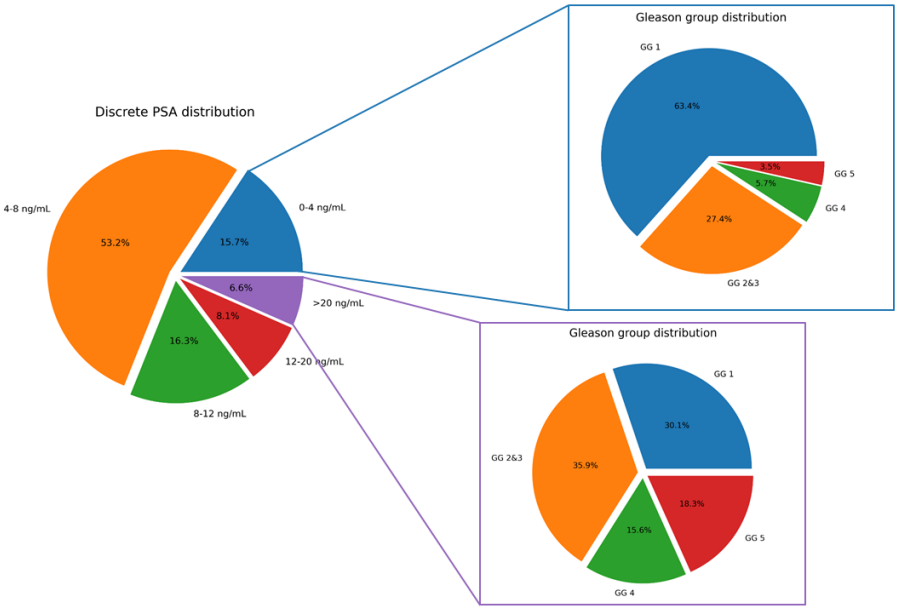


Figure 3.2: Gleason grade distribution with respect to discrete PSA ranges derived from the PLCO dataset.

the same, or less aggressive.

All imaging findings (e.g. mpMRI or PET/CT) are correlated with the prior diagnostic reports through the Partin tables[93], providing risk scores for extra-capsular extensions (ECE), seminal vesicle invasions (SVI) and lymph nodes involvement (LN+) stratified based on PSA values, Gleason scores and the clinical T stage. Therefore, we randomly sample findings based on the probability of organ-confined disease consistently to real world datasets. If imaging reveals a non-localized disease, a confirmation biopsy is performed to identify possible false-positives of PET/CT[94, 95].

Therefore, our proposed synthetic data mimic real world patterns and correlations by covering a broad range of variability, from patients with low PSA levels and high Gleason grades to patients with high PSA levels and low Gleason grades.

When all required diagnostic procedures are completed, we apply the NCCN guidelines[96] to establish the clinical cancer stage and risk category based on randomly sampled diagnostic reports exported throughout the staging workup, thus creating a ground-truth for the training stratification models.

It is worth emphasizing that the patient pathways are very diverse due to the stochastic processes imposed in most of the stages (yellow boxes in figure 3.1). For instance, diagnosis phase could yield patients who only have been assessed with repeated PSA measurements and DREs (if no suspicious findings or very low PSA levels), patients who have been assessed through imaging with or without subsequent procedures depending on the results, patients who went through systematic or targeted biopsies, patients who followed the first line of treatment directly after a suite of diagnostic procedures, or patients who have been actively monitored before definitive therapy, etc.

3.3.2 Clinical TNM stage prediction

While the synthetic data described in the previous section has a good variability and consistency at the cohort level, at the patient level it still has some drawbacks. EHRs generated by Synthea are very well structured and they use the same coding all the time. However, real world data is not always that organized, different sites might be using different coding systems and thus certain data elements might be denoted by multiple codes, names or even different spellings. Therefore, a simple attempt to extract all the ingredients required for the staging could suffer from lack of robustness. On the other hand, despite this variability in the coding systems, code descriptions of the same data

element always have a lot of communalities. We try to exploit these commonalities by utilizing a natural language processing model that works directly on the plain text from the descriptions rather than the structured codes from an EHR.

We first need to translate the patient resources (e.g., represented in FHIR or Health Level Seven - HL7 - queries) into plaintext. Therefore, we extract diagnostic procedures code descriptions along with the time they have been performed. We then extract all the observations recorded within these diagnostic procedures through a rule-based key-value pair extractor, where the key is the observation code description and the value is determined from the findings. Finally, we concatenate the text to obtain a paragraph comprising all the ingredients generated throughout the staging workup in form of plaintext, also preserving the temporal dimension of the longitudinal data. To further enhance the variability in the text data, we augment the dataset by randomly replacing certain words or phrases with synonyms.

3.3.3 Experiment setup

In the training phase, we use the FHIR bundles exported by Synthea as an annotated dataset for supervised learning. As depicted in Figure 3.3, each patient data is preprocessed by a rule-based key-value pair extractor script to create a text block describing all diagnostic procedures along with their results. We train the stage classification system as a fine-tuning task for a Bidirectional Encoder Representations from Transformers (BERT) model[21], starting from the clinical BERT checkpoint[110]. BERT is a bidirectional language model pretrained to provide 768 dimensional contextual embeddings on a couple of unsupervised training tasks: masked language modelling (MLM) and next sentence prediction (NSP).

We fine-tune the model for 5 epochs - empirically determined from the learning curves - to minimize a categorical cross-entropy loss function with a learning rate of 10^{-5}. Although not necessary in the majority of cases, BERT's input sequence length is set to 512 words to maximize the chances of capturing all the information required for clinical stage assignation. Since the GPU memory required by the training process is disproportionally larger with the sequence length, the maximum batch size we can set is 16 on a NVIDIA Volta V100 GPU with 32 GB of memory. To avoid severe class imbalance (e.g. T1cM1 patients -those with metastatic disease- being relatively very rare as compared to T3M1 patients) we trained different classifiers for each staging component, but we refer to these as a single predictor for simplicity. However, since our synthetic data is consistent with real cohort statistics extracted from literature, we still deal with unbalanced class distributions within each stage component. Therefore, we use sample weights within the loss function to provide stronger signals for samples from under-represented classes.

Once trained, the models can be used for inference as depicted in the bottom part of Figure 3.3. Patient's EHR stored in the hospital's FHIR database will be converted to plaintext using the preprocessing module and then pushed through the model to obtain an estimation of the clinical stage, which can be then properly inserted back in the EHR.

3.4 Results

3.4.1 High fidelity of the synthetic data

To assess the realness of our synthetically generated data we compared various purely synthetic features to real world counterparts available in the PLCO as well as the internal dataset described in table 3.1.

First, we compared the distributions of PSA measurements and gland volumes generated by our Synthea modules against to the ones recorded in the internal dataset. As illustrated in figure 3.4, despite the fact that PSA values were sampled based on PLCO and the volume was randomized consistently to PSA and age [109], there is a very strong distributional similarity between synthetic and real PSA measurements (figure 3.4(a)) as well as prostate volumes (figure 3.4(b)).

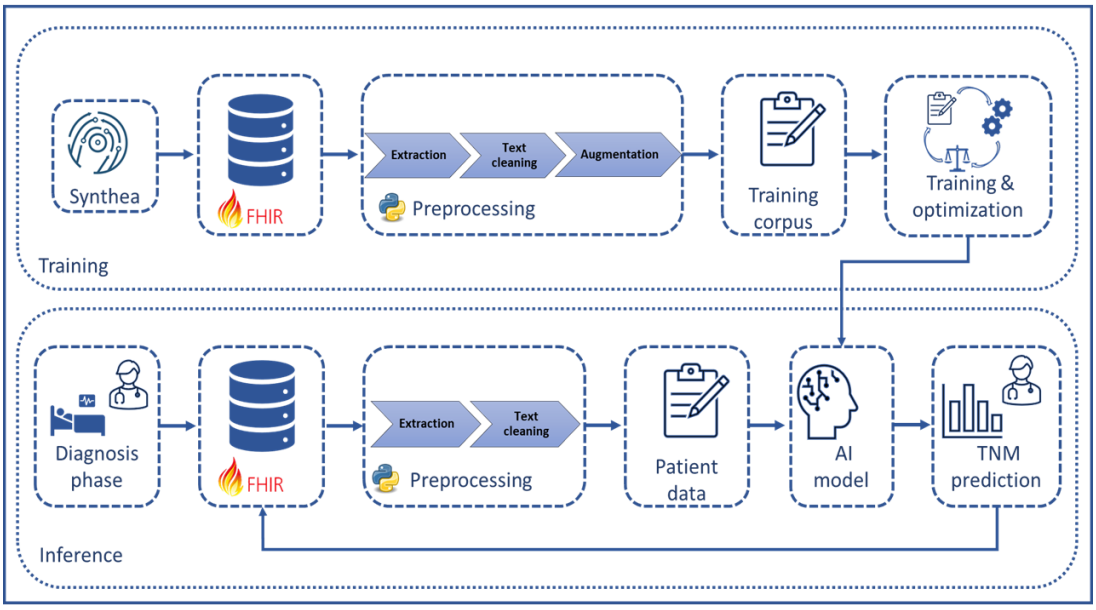
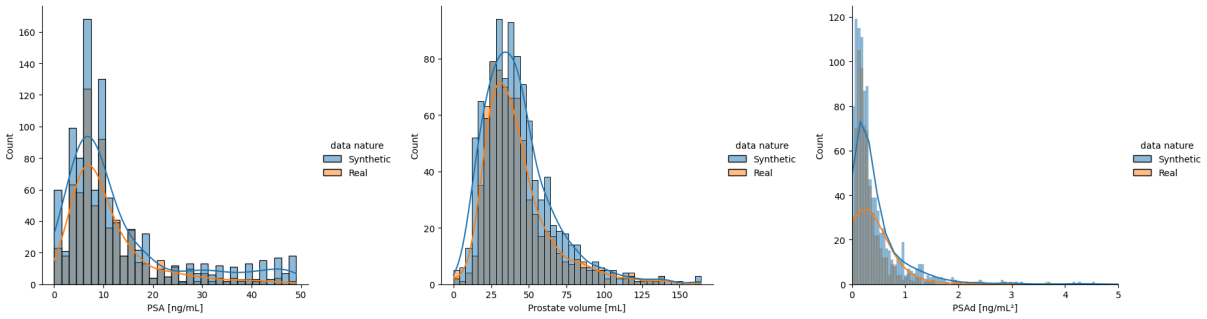


Figure 3.3: Method overview: Training (up) and inference (bottom) pipelines.

However, assessing similarity from an univariate perspective only might be misleading since data should be consistent across different diagnostic tests, exhibiting plausible correlations between various parameters. Particularly, it is well known that the PSA levels correlate with prostate volumes, larger glands tending to produce more antigens. Therefore, figure 3.4(c) demonstrates consistency across the two aforementioned features by also showing similar distributions of PSA density, which is computed by dividing the PSA level to the volume at patient level.



(a) (b) (c)

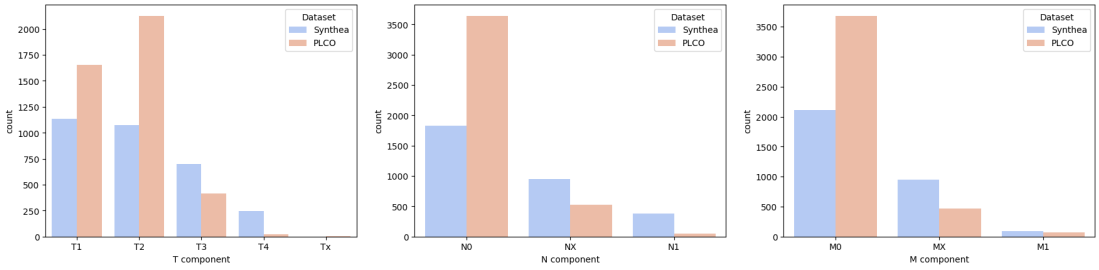
Figure 3.4: Realness assessment of PSA, volume and PSA density.

Similarly, we studied the realness and consistency of synthetically generated Gleason Grades (GG) and PI-RADS distributions from an univariate and a bivariate perspective respectively. From a prevalence ranking perspective we could conclude that our synthetic data has a perfect fidelity.

The PI-RADS system had become the main tool used in reducing the number of unncecessary biopsies, but it still has a relatively low specificty rate as compared to its sensitivity. In other words, it tends to overcall for biopsies in clinically insignificant prostate cancer patients. This inherent predictive capacity of PI-RADS is well reflected in our synthetically generated data, once again demonstrating realness and consistency across longitudinal diagnostic results.

All stochastic processes employed in the entire staging road-map are based on either statistics derived from real datasets or extracted from literature descriptions of certain behaviors and correlations between parameters. Therefore, a perfect fidelity to a certain source cannot be expected.

Nevertheless, since our data generator is designed in a strong causal setting (reports at certain timestamp are always sampled consistently to findings randomized in previous timestamps), the overall landscape of our synthetic cohort is expected to match the real clinical data distribution to some extent. The TNM stage resulted from the staging work-up provides a very good intuition of how realistically various reports have been randomized during staging since it relies on multiple data points, each sampled based on a different statistical source (e.g. based on PLCO, internal dataset or literature). Figure 3.5 illustrates the reliability of our proposed data generator by jointly plotting the TNM stage distribution of synthetic and PLCO cohorts. Although Synthea data has a better granularity of staging (e.g. T2 is further categorized into T2a, T2b or T2c) we aggregated similar stages to match the format of PLCO for a meaningful analysis. In comparison to PLCO, our fabricated data seems to over represent high stages for the T and N components: T3 and T4 are relatively more frequent in Synthea data as compared to PLCO, which is also the case of N1 patients. Moreover, assessments of non-localized components (N and M) is done more frequently in PLCO as compared to Synthea. Considering the methodology used to collect statistical properties embedded in the generator, some level of discrepancy between real and synthetic cohorts is expected due to inductive biases. However, as shown in figure 3.5, the resulted synthetic cohort is reliably representing all stages while maintaining the relative prevalence ranking w.r.t. PLCO, except for T1 and T2 stages. Moreover, it is noteworthy that PLCO study was designed to assess effectiveness of screening protocols, therefore not specifically presenting a population at risk of prostate cancer. On the opposite, our synthetic cohort was intended to better reflect various disease aggressiveness levels, ranging from very low to very high risk, and from localized to metastatic PCa.



(a) (b) (c)

Figure 3.5: Realness assessment of TNM stage distributions based on the PLCO data.

3.4.2 Clinical TNM prediction

We herein report the precision, recall and F1 scores for each class independently, as well as their micro/macro averages and the overall accuracy computed on the synthetic test set. Predicting the T component of the staging system is the most challenging due to the necessity to interpret and correlate a larger number of longitudinal diagnostic reports, and also due to a higher level of granularity in the classification (8 class classification problem). Therefore, the T stage predictor had an overall accuracy of only 98.7% compared to N and M predictors (3 class classification problem) which reached over 99.5%, as depicted in Table 3.2.

3.5 Conclusions

This chapter presents a method to estimate the clinical cancer stage from diagnostic reports in EHRs using realistic synthetic data generated by Synthea. To the best of our knowledge, prediction of clinical cancer stage directly from EHRs has not yet been undertaken by others, possibly because of the major challenges in data annotation pipelines and privacy constraints.

Table 3.2: Performance evaluation of the TNM stage classifiers.

Stage component		Precision	Recall	F1 score	Occurrences
T	Tx	1	0.997	0.998	590
	T1c	0.997	0.984	0.991	1062
	T2a	0.971	0.962	0.966	208
	T2b	0.972	0.991	0.981	428
	T2c	0.978	0.994	0.986	311
	T3a	0.990	0.986	0.988	417
	T3b	0.994	0.991	0.992	333
	T4	0.952	0.983	0.967	240
	Accuracy	0.987	0.987	0.987	3589
	Macro average	0.982	0.986	0.984	3589
	Micro average	0.987	0.987	0.987	3589
N	Nx	0.992	0.996	0.994	1643
	N0	0.996	0.992	0.994	1849
	N1	1	1	1	399
	Accuracy	0.995	0.995	0.995	3891
	Macro average	0.996	0.996	0.996	3891
	Micro average	0.995	0.995	0.995	3891
N	Mx	0.993	0.998	0.995	1643
	M0	0.998	0.995	0.997	2176
	M1	1	1	1	72
	Accuracy	0.996	0.996	0.996	3891
	Macro average	0.997	0.998	0.997	3891
	Micro average	0.996	0.996	0.996	3891

We therefore developed Synthea modules to generate realistic health records for patients with prostate cancer, covering a broad range of disease extent and aggressiveness, ranging from low risk to very high risk and from localized to regional or even metastatic cancer. With data scarcity being a very common blocker in deploying robust AI based solutions targeting different areas of healthcare, a reliable synthetic data generator holds the potential to enable early developments of prototypes and proof of concepts that may be leading to an improved patient experience and outcome, while also reducing the workload of clinical personnel. Besides being able to generate an unlimited number of samples, our fabricated data holds a set of tremendously important properties:

- **Privacy preservation.** Due to its purely synthetic nature, our proposed dataset is free of any privacy preservation concerns or re-identification risks, thus being suitable for a broad range of applications such as hypothesis formulation and testing, software development and testing, educational purposes and early developments of smart features to support clinicians in handling PCa patients more efficiently.

- **Data completeness**. By carefully defining care-maps for diagnosis, treatment and follow-up, our synthetic EHRs contain all the established tests and procedures routinely employed in identifying and managing patients suffering from prostate cancer, at the moment of writing this thesis.

- **Data variability**. Our synthetic cohorts exhibit a great variability in terms of disease aggressiveness as well as patient pathways. We employed various statistical properties either de-

rived from real available datasets, or published in literature, to ensure a realistic distributional landscape of our cohort, including stage incidence, treatment options distribution, follow-up behaviors and recurrence risks.

- **Data consistency**. Despite distributional coherence, the longitudinal nature of EHR data urges for adoption of techniques to maintain consistency across all various observations. Therefore, correlations between different features were extracted from scientific publications and employed in all stochastic steps of our patient generator.

A qualitative evaluation of the synthetic cohort generated by the proposed Synthea modules was carried in two different ways. Firstly, we have inspected distributional properties of various features and compared those with observed incidences and statistics collected from literature or derived from real datasets. As demonstrated in this chapter, our synthetic data exhibit a realistic variability across both disease specifics and patient pathways. Secondly, we have iteratively improved the synthetic data generator based on the feedback provided by Dr. J.R., an experienced radiation oncologist. This played a major role in increasing the reliability and realness of generated electronic health records, revealing inconsistencies in observations, time distribution, treatment options, follow-up dynamics of PSA and side-effects, as well as unrealistic frequencies in encounters, etc.

Finally, we demonstrated that the proposed methodology can be successfully used to develop high performance predictors by providing evaluations on large scale synthetic datasets: TNM staging achieved 98.7% accuracy for clinical T stage and over 99.5% accuracy for the non-localized components on a synthetic test set.

3.6 Discussion

Synthetic data generation

We have herein proposed a systematic method to generate large-scale datasets using Synthea. Particularly, we have developed modules to fabricate electronic health records of patients with prostate cancer that can be then used to , for example, develop smart features to support clinicians in their workflows. Our generated data does not hold any privacy concerns since it is fully synthetic by nature, while at the same time it provides (1) completeness in terms of investigations and procedures, (2) variability across observations, treatment options and patient pathway, and (3) consistency across various diagnostic procedures, treatments and follow-up.

However, from a completeness standpoint the work presented in this chapter has a series of limitations. Firstly, despite exhibiting completeness in terms of diagnosis, treatment and follow-up workups, some applications might require reliable data beyond the first line of treatment. For instance, prediction of the best next treatment option based on the follow-up data in case of a relapse might carry a huge potential in providing the best care to patients across different countries, especially the underdeveloped ones. However, our proposed generator only cover the treatment of first occurring cancer and the subsequent follow-up visits, so extending the modules represents a future direction of the presented work. Secondly, although Synthea generates reports for biopsies or imaging studies such as mpMRI or PET/CT, images coupled with these findings do not exist. This prevents our proposed synthetic dataset to be immediately adopted in developing applications that imply the fusion of different modalities and data types, such as mpMRI and PET scans, pathology images and EHRs. Nevertheless, one straight-forward option to overcome this limitation is to register real anonymized images to the synthetic diagnostic reports based on the findings (e.g. coupling mpMRI images to the synthetic report by matching the number of lesions, their PI-RADS scores, their location, etc.). Becoming a hybrid synthetic dataset, this workaround would increase the usability of our fabricated cohorts even further, but at the cost of weakening the privacy preservation properties owned by the current setting.

In terms of variability and consistency, our proposed Synthea modules might have incorporated inductive biases from the sources we used to collect disease incidence and statistical properties. For instance, we have used the PLCO data, which was collected in a screening trial, to derive joint distributions of PSA measurements and Gleason groups. Therefore, we inherited the PSA distribution of a cohort not specifically suspicious at the screening time, which might be different from the data available in a clinical center specialized in treating prostate cancer.

Moreover, we have designed Synthea modules to generate cohorts for all intents and purposes, thus mimicking as much as possible distributional landscapes of real clinical data. Therefore, the current implementation does not allow a selective generation of patients matching certain properties (e.g. only high risk disease patients treated with radiation and hormone therapy).

Lastly, while meeting our realness expectations from a cohort level perspective, at patient level Synthea data is unrealistically well structured, always storing the EHRs using a specific format and coding system. In contrast, real clinical data could be stored differently from one site to another, using various coding systems to store information in various formats, or even depicting the findings in free-text form. However, given the recent advancements in generative language models (such as GPT), AI models could be employed in augmenting structured synthetic EHRs, translating findings in free-text clinical reports, and thus boosting the usability of our synthetic data.

Overall, in our opinion the method proposed in this chapter is able to generate high quality synthetic electronic health records that have the potential to overcome the inherent data scarcity based limitations in health care. The main strengths of our fabricated data are the lack of any privacy concerns, data completeness w.r.t. currently established procedures and tests routinely used in clinical practice, data variability and longitudinal consistency of electronic health records.

Prostate cancer patient stratification

Despite the very promising results herein reported, we have only given a prove of concept here, not a developed and established workflow. Although the generalization test was successful on a synthetic test set, evaluation on real clinical data is mandatory to make a definitive statement regarding the feasibility of model adoption in real world applications. Besides the model bias, the pre-processing technique which translates FHIR bundles to plaintext represents a possible source of errors, where extra heuristics are used to filter out non prostate cancer related diagnostic reports ensuring a maximum text sequence length of 512 words. Therefore, all these elements should be extensively assessed on a real cohort gathered from multiple sites to determine the model's robustness.

In spite of only providing a proof of concept through TNM stage prediction, our Synthea modules could supply the underlying data for exploring various use-cases where prediction models could bridge gaps in clinical routines and significantly reduce the workload of clinicians. For example, one could use all available information in the EHR to predict the best treatment option for a certain patient. Moreover, since our synthetically generated longitudinal records are not limited to the diagnosis work-up, one could use the follow-up data to early predict the chance of encountering a biochemical recurrence.

In conclusion, we showed that synthetic data generated with Synthea can be used to develop a variety of smart features usable in clinical practice, while avoiding the privacy related constraints. Neural networks can be fully trained on synthetic data (or pretrained) and evaluated (or fine-tuned) only on a relatively small fraction of real patients. An example of use-case is provided by this chapter, where we demonstrated that models could yield 98.7% (T stage prediction) and over 99.5% (N and M stages) accuracy in stratifying prostate cancer patients, when qualitative and complete data is available. However, this work can be extended to a variety of clinical use-cases where processing longitudinal EHR data can be automated to reduce clinician's workload, make the entire process more efficient and less error prone. Therefore, we aim at extending the work to predicting the patient's risk category, which also plays a major role in patient stratification process enabling the selection of the best treatment option.

4. Advancements in Trustworthy AI for Clinical Cancer Applications

Introduction
Clinical association network for clinically significant prostate cancer prediction
Non-small cell lung cancer sub-type classification
Conclusions
Discussion

4.1 Introduction

Integration of AI technologies in the healthcare industry is currently one of the most active areas for researchers across the world, mainly due to its outstanding potential in increasing diagnosis accuracy, providing personalized care plans and ultimately improving patient outcomes.

Deep learning based computer-aided diagnosis (DL-CAD) systems are intended to automate time-consuming and error-prone processes, significantly reducing clinicians workload while improving the overall diagnosis accuracy [111]. However, the immediate adoption of such DL-CAD systems in clinical routines is hindered by a set of concerns related to reliability and trustworthiness. Due to their complexity, AI models are often regarded as black-boxes which may contravene with ethical principles of health care delivery. Among all industries, decisions made in the health care sectors carry tremendous risks for patients well being through unintended consequences. The inherent lack of transparency specific to deep learning models is raising skepticism across clinicians and their patients in following suggestions coming from AI, without a solid understanding of the rationale behind its decision making process. The increased complexity of DL algorithms as compared to classical machine learning approaches make them less transparent, preventing them from being widely adopted despite their potential. Therefore, the current dilemma stems from a chain of trade-offs: (1) between architectural complexity and the potential of solving complex problems and (2) between algorithm complexity and decision making process transparency.

Moreover, AI algorithms rely extensively on learning patterns from the underlying training data [4], making them highly susceptible to inductive biases. Therefore, a rigorous model evaluation and characterization should be performed to assess its robustness, including, for instance, testing on data collected from multiple institutions, and also reflecting all possible scenarios that could occur in clinical practice. However, DL prediction models by nature tend to be overconfident in their reasoning when provided with data samples under-represented in their underlying training database, which might lead to incorrect decisions that could harm patients well being.

Understanding the reasoning of a machine learning model to reach at a certain prediction is crucially important for increasing its trustworthiness, and thus moving closer towards embedding such solutions in various stages of clinical pipelines. In the past decade, multiple studies attempted to attribute importance scores to input features with respect to model predictions as an indicative of the relative impact they have on the reasoning process [24, 25, 26, 27, 28]. Lundberg et. al. [27] proposed a stochastic method to assign Shapely values to each input of a machine learning model by

randomly sampling coalitions between features and assessing their effect on the prediction. Coming from the collaborative game theory, Shapely values split the reward (model prediction) across players (input features) based on their contribution to the outcome. Therefore, in the context of ML they represent quantitative measures of the relative impact that each considered feature had on the inference.

By incorporating such techniques in the overall pipeline, predictions made by AI algorithms could be accompanied by explanations that would allow clinicians to assess relatively difficult or uncommon cases where models could be prone to failure. Nevertheless, transparency gains could build trust in such DL-CAD systems unlocking potential benefits in terms of diagnosis accuracy and patient outcome. However, besides the interpretability and explanability properties, for a complete recipe to trustworthy AI, a deep learning solution should also be conferred with capabilities to identify difficult cases where predictions are rather uncertain. Data samples could have high uncertainty levels due to multiple factors, including noisy or corrupted acquisitions, miss-labeling or inter-user variability effects on the model training [4], new cases being out of the underlying training data, etc.

Model ensembles [34] have been proven to achieve state of the art performance on a variety of biomedical segmentation challenges [33] by aggregating predictions of multiple model instances trained on different data fractions. Besides a generally improved performance as compared to standard training, ensembles are advantageously able to quantify an uncertainty score by computing the prediction variability across all model instances, or the fraction of all predictions being in agreement [31, 32]. Therefore, an AI based solution could identify difficult cases based on these uncertainty estimations and request radiologists assistance, thus increasing the overall robustness of the entire pipeline.

Overall, DL-CAD systems that exhibit transparency by providing explainable and interpretable predictions as well as uncertainty estimates could bridge the trust related gaps that are currently preventing these methods from general adoption in clinical routine, unlocking the potential promised by AI in improving patient care while significantly reducing the workload experienced nowadays by health care practitioners. In this chapter we explore how standard deep learning models could benefit from being enriched with explainability and uncertainty quantification capabilities, that could potentially bring them closer to the concept of trustworthy AI. Specifically, our contribution could be outlined as follows:

- We explore the feasibility of employing a Shap analysis [27] in explaining the reasoning of DL models in solving two clinically relevant tasks, namely clinical significant prostate cancer (csPCa) prediction and non-small cell lung cancer (NSCLC) subtype classification.
- We propose ensemble based uncertainty estimations to identify uncommon or contradicting mutational patterns in a publicly available genomics dataset and thus elucidate limitations of current predictive models in representing heterogeneous data.

This chapter is organized as follows: Section 4.2 describes how a state of the art lesion detection algorithm can benefit from additional clinical information in identifying clinically significant cases of prostate cancer. Besides an overall improved performance, Shapely values are providing excellent insights into how the additional inputs influenced the model reasoning process. Section 4.3 presents a deep learning solution to classify NSCLC into subtypes based on genomics data. With an accurate classification being extremely important for treatment planning, uncertainty estimations and Shapely based explanations were employed in increasing the model's robustness. Finally, overall conclusions are drawn in section 4.4 while all approaches presented herein are discussed in 4.5.

4.2 Clinical association network for clinically significant prostate cancer prediction

4.2.1 Use-case introduction

Prostate cancer (PCa) is the second leading cause of cancer death among men in the US and the first in terms of estimated new cases (1 in 8 men will be so diagnosed during his lifetime) [83]. Prostate specific antigen (PSA) is the established way to initially identify a patient as being suspicious of having prostate cancer and recommend a confirmatory prostate biopsy, an invasive and risky procedure. However, since PSA can also fluctuate due to non-malignant factors its specificity is low [112], leading to a significant number of overdiagnosis and overtreatment. Multi-parametric Magnetic Resonance Imaging (mpMRI) has been widely adopted in clinical routine to better triage the patients with abnormal PSA levels helping 27% of them to avoid an unnecessary primary biopsy while detecting up to 18% more clinically significant prostate cancers (cs-PCa) [113]. Prostate Imaging Reporting and Data System (PI-RADS v2) [114] is a commonly used system to localize and report suspicious lesions within the prostate ranking them with a score from 1 to 5 (any lesion scored above 2 is regarded as malignant). However, reading an mpMRI series is a time-consuming task, it requires a high level of expertise [115] and the inter-user agreement in assigning a PIRADS score is often rather small [116]. Moreover, according to the guidelines, radiologists are trained to assign a PI-RADS score while being blinded to the clinical and demographics parameters, which in addition to the mpMRI study might carry valuable clues in diagnosing cs-PCa.

Computer aided diagnosis (CAD) systems could be employed to bridge the current gaps within the PI-RADS scoring system by automating the process while preserving (or even outperforming) the radiologists performance (in a consensus framework). While many CAD systems candidates are reaching comparable sensitivity to the clinicians, they are still not sufficiently specific [117].

Multiple studies have demonstrated the effectiveness of jointly assessing clinical parameters and imaging information for a better patient stratification [118, 119, 120]. However, these models are still dependent on the reading accuracy and might not work consistently across various sites or various radiologist experience levels. Nonetheless, the idea of using clinical information to guide imaging interpretation could be successfully adopted in CAD systems to enhance their robustness [112].

In this section we're aiming at improving the performance of CAD systems by adding an extra clinical correction phase in the pipeline. Specifically, we employ a deep neural network to adjust the prediction of a proposed state of the art system [117] based on a series of clinical (e.g. PSA, PSA density, gland volume) and demographics (e.g. age) information. Moreover, we conduct a Shap analysis to attribute a relative importance to each input feature with respect to the final prediction to increase transparency of the DL model by providing explainable outputs.

4.2.2 Methods

We herein propose a clinical correction step to an established pipeline [11] that automates the diagnosis of PCa and cs-PCa through a series of machine learning algorithms that sequentially perform gland segmentation, lesion detection, false positive reduction and lesion qualification. First, a preprocessing phase is extracting the T2-Weighted (T2W) and Dynamic Weighted Images (DWI) series from the DICOMs and computes an ADC map and a synthetic high-b DWI (b=2000). Next, as depicted in Figure 4.1 a full-gland segmentation model [121] is used to detect the prostate in T2W and DWI series and create a 3D mask which will serve as an input to the subsequent lesion detection model. Following the methods presented in [117, 111], a 3D fully convolutional neural network is used to detect lesion candidates within the prostate and then a multiscale false positive reduction (FPR) network is employed to adjust the prediction while preserving the overall sensitivity.

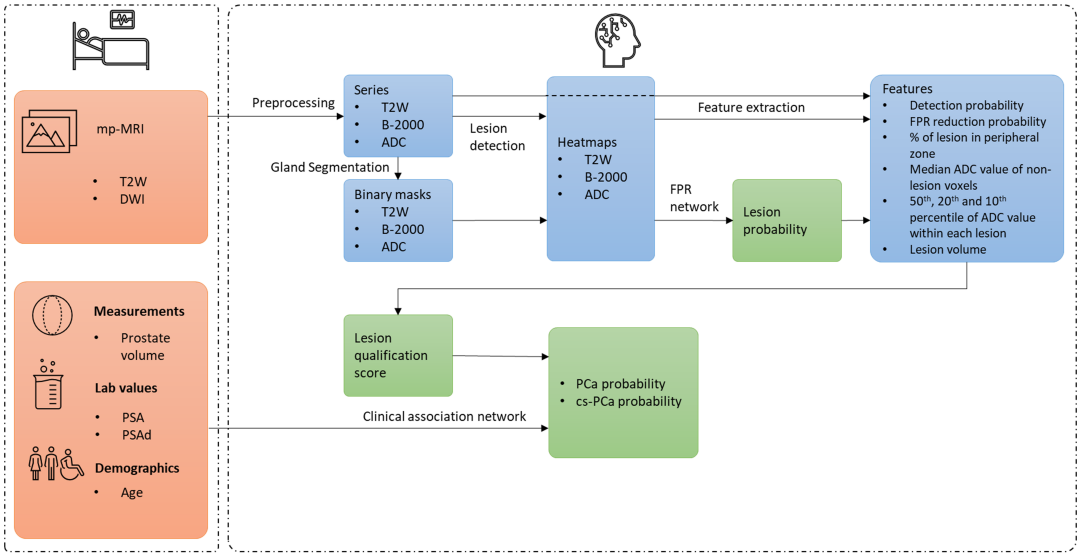


Figure 4.1: Proposed workflow.

The fully automated pipeline described above outputs a qualification score (QS) that describes the malignancy of a lesion solely based on bi-parametric MRI images. Given that clinical information might hold additional predictive capacity, we propose an extra correction phase where QS score is combined with additional features to refine the overall malignancy prediction. To that extent, we combine outputs of the initial pipeline with additional clinical parameters and employ a neural network, further referred to as clinical association net, in predicting whether a certain lesion is clinically significant.

4.2.2.1 Dataset

We gathered anonymized data from 11 sites from different countries ensuring heterogeneity, leading to an increased generalizability potential of the final model. The dataset consists in 2261 patients with clinical (e.g. PSA, PSAD), demographics (e.g. age) and imaging data (multiparametric MRI series along with a PIRADS score assigned by experienced radiologists prospectively with PSA) recorded. 48.7% of these patients have been diagnosed with prostate cancer through a targeted biopsy as a consequence of an abnormal mpMRI (PIRADS >= 3) out of which 59% have been found with clinically significant disease (GG > 2). Since most of the sites do not perform systematic biopsies as a part of their clinical routine, we considered negative cases as the ones with a PIRADS score less than 2. All the cases where the PIRADS was above 2 and no Gleason score available have been excluded. Table 4.1 shows the overall statistics of the dataset.

Table 4.1: Dataset information.

	Gleason group 0 N = 1160	Gleason group 1 N=451	Gleason group >= 2 N = 650
PSA [ng/dL]	9.08 (±9.25)	8.56 (±7.48)	16.55 (±30.72)
PSAD [ng/dL/cc]	0.16 (±0.21)	0.22 (±0.28)	0.45 (±0.74)
Volume [cc]	66.14 (±39.3)	46.88 (±24.99)	41.65 (±23.03)
Age [years]	63.83 (±7.72)	64.65 (±7.23)	67.06 (±7.52)
PI-RADS	2.12 (±0.83)	3.54 (±1.15)	4.31 (±0.82)
Gleason group	0	1	3.03(±1.13)

4.2.2.2 Lesion Qualification network

As proposed in [117] all lesions identified and confirmed by the detection and FPR models are then inspected by a qualification model, which assigns a qualification score (QS) based on a series of 8 input features: detection probability, FPR reduction probability, proportion of lesion extent in the peripheral zone, median ADC value of all non-lesion voxels, 50th, 20th and 10th percentile of ADC value within each lesion, and lesion volume computed from the 3D heatmap produced by the detection network. The qualification model has a simple fully-connected architecture with 2 hidden layers of 32 neurons each. A dropout layer with a droppage rate of 0.5 has been used between the second hidden layer and the output for regularization. The qualification score produced by this model can be used per-se to predict PCa and/or cs-PCa solely based on the mpMRI series. Therefore, the system proposed in this section can still infer when additional clinical data is not available, but with limited performance as shown in section 4.2.3.

4.2.2.3 Clinical association network

Both the FPR reduction probability and qualification score can be thresholded per-se to predict PCa or cs-PCa. However, we herein propose an extra phase, where we aim to add available clinical information in the analysis to further enhance the system's performance. We further refer to this step as a clinical correction phase, where we employ another neural network in adjusting malignancy prediction based on the extra data the network is provided with. This is yet another simple fully-connected neural network with 3 hidden layers activated by hyperbolic tangent functions, while the output layer consists of one neuron activated by a sigmoid function. Similar to the qualification network, a dropout layer has been used to regularize the model, stochastically cutting-off 20% of the connections between the last hidden layer and the output. To assess the impact of each clinical/demographic feature independently we perform a multivariate analysis consisting in 10 different experiments: QS + Age, QS + PSA, QS + PSA density (PSAD), QS + Volume, QS + PSAD + Volume, QS + Age + PSA, QS + Age + PSAD, QS + Age + Volume, QS + Age + PSAD + Volume, QS + Age + PSA + PSAD + Volume.

The clinical association network has been trained to minimize a binary cross entropy (BCE) loss function for 300 epochs. However, a model checkpoint callback was used to perform a better model selection, saving the best set of parameters w.r.t. the loss on the validation set. Optimization has been done using the Adam optimizer with an initial learning rate of 0.0001 and a batch size of 64. To account for class-imbalance (especially in cs-PCa prediction use-case) we defined class weights to emphasize on the examples from the minority class.

A stratified 5-fold cross-validation scheme has been used to mitigate train-validation split biases, hence distinct batches of 20% of the data are used iteratively as a validation set. To increase model robustness, the final score is obtained by averaging the 5 per-fold scores yielded by each individual model. To avoid any biases stemming from institutional differences, the final generalization capability of the ensemble was performed on a held-out set provided by a data site not seen during the training. Additionally, an 100 times bootstrapping with repetition procedure was applied in the performance computation phase to assess the AUC stability across various testing sets.

4.2.3 Results

Multivariate analysis

As depicted in Table 4.2, the clinical PIRADS as assigned by the radiologists can predict PCa with an AUC of 0.904 while the QS only achieves 0.852. However, the clinical correction network improves over the QS significantly in most of the cases, especially when the total gland volume is taken into consideration: when only adding the volume to the QS, the correction network achieves in AUC of 0.902, thus reaching the radiologist performance (p-value = 7.87e-01, indicating that there is no statistically significant difference between our model output and the clinical PIRADS).

Table 4.2: Multivariate analysis results.

	PCa			cs-PCa		
	PIRADS AUC	AIQua Score AUC	DL AUC	PIRADS AUC	AIQua Score AUC	DL AUC
QS + Age	0.904	0.852	0.854	0.899	0.841	0.865
QS + Volume	0.904	0.852	0.902	0.899	0.841	0.899
QS + PSA	0.904	0.852	0.866	0.899	0.841	0.857
QS + PSAD	0.904	0.852	0.869	0.899	0.841	0.867
QS + PSAD + Volume	0.904	0.852	0.897	0.899	0.841	0.890
QS + Age + Volume	0.904	0.852	0.895	0.899	0.841	0.904
QS + Age + PSA	0.904	0.852	0.855	0.899	0.841	0.867
QS + Age + PSAD	0.904	0.852	0.852	0.899	0.841	0.873
QS + Age + PSAD + Volume	0.904	0.852	0.894	0.899	0.841	0.908
QS + Age + PSA + PSAD + Volume	0.904	0.852	0.897	0.899	0.841	0.906

In terms of cs-PCa the clinical PI-RADS have an AUC of 0.899 while the QS have an AUC of 0.841. Consistent with the results on PCa prediction, volume information seems to offer the largest increment in cs-PCa prediction accuracy, improving over the QS with 5.8 points in the AUC, comparable to the radiologist performance. However, in contrast to PCa prediction, age seem to carry additive predictive power further improving the AUC to 0.904, outperforming the clinical PIRADS. Nevertheless, PSAD used in conjunction with the QS, age and gland volume further improves the AUC to 0.908, outperforming clinical PIRADS by 0.9 points in the AUC.

4.2.3.1 Explainability and Interpretability

While the benefits of such an autonomous solution are incontestable there is a high level of suspicion in regards to the way a DL model is reaching at a certain conclusion. Although their effectiveness have been heavily demonstrated by the scientific community within the last decade, deep learning algorithms are mainly seen as "black boxes" often preventing them from adoption in the clinical routine [4]. Therefore, this section presents the results of a Shap analysis [27] employed in providing explanations on how our proposed clinical association network make use of additional parameters in its reasoning process.

Figure 4.2 shows an overall Shap analysis that describes the relative impact that each considered input parameter has on the model output. Crisp values of various parameters are reflected by the color-map, low values being depicted in blue, while red dots represent high values. Each of the considered parameters employed herein can influence model predictions up to certain extent, which is denoted on the x axis. Therefore, concerning the relative impact they have on cs-PCa prediction, the following ranking of our proposed parameters is obtained: QS, Volume, Age, PSAD and PSA.

Apart from the qualification score, the most predictive feature is the total gland volume which can shift the model output by approximately ± 20% as follows: the large glands seem to reduce the risk of csPCa while the small ones seem to do the opposite. In contrast, the patient age shows a positive correlation with the QS, positively shifting the prediction with age. The same applies to PSA and PSAD which are positively correlated with the outcome, but with relatively lower impact. Overall, figure 4.2 demonstrates that each additional parameter we considered has a certain impact on predictions, and thus, the performance gain produced by the clinical correction phase has a steady foundation.

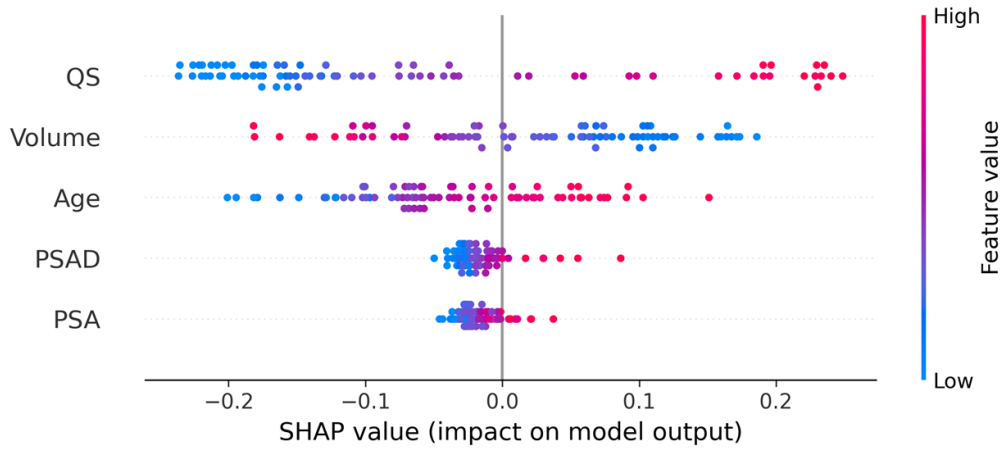


Figure 4.2: Global Shap analysis for increased explainability.

Such global level analysis increases the transparency through providing explanations of how each input plays its role in the inference, across the entire testing cohort. By assessing these results, clinicians could better understand and link the patterns learned by neural networks during the training with clinical situations they have experienced, and therefore increase their trust in such solutions. However, Shap analysis can also be performed at sample level and delivered as a tool to make real time model predictions interpretable.

4.2.4 Conclusions

As noted in section 4.2.3, the clinical correction neural network improves over the qualification score significantly making the entire CAD system outcome comparable to the clinical PIRADS assigned in consensus by experienced radiologists. Moreover, when gland volume is taken into account, the proposed CAD system outperform the clinical PIRADS by a small margin while being completely autonomous.

Following a similar approach, we have demonstrated that not only QS can benefit from additional clinical information, but also the PI-RADS score: from a median AUC of 0.904, our multivariate analysis revealed an improvement of 0.012 points in AUC when age and PSA were taken into account, while the addition of gland volume further improved the median AUC to over 0.93.

Nonetheless, we increased model transparency by running a Shap analysis that assigns a relative importance score to each input feature w.r.t. the predicted risk of clinically significant prostate cancer. Moreover, we have provided instance level explanations for all classifications scenarios, namely TP, TN, FP and FN predictions, showing the inter-dependencies across features. For instance, age, volume and QS individual impact on the output probability differ based on the state of all other variables considered, which in conjunction with the improved performance we obtained proves that our model learned relevant patterns and correlations in the training data.

4.2.5 Discussion

This section provided an example of how additional information can be incorporated, when available, in a fully autonomous solution to increase its overall performance and stability. Specifically, we have demonstrated how clinical and demographics features could be used to improve the accuracy of a DL-CAD system that analyzes bi-parametric MRI images and assigns a malignancy score, and also, the discriminative power of the PI-RADS score as assigned by radiologists. Furthermore, a feature importance inferring algorithm was employed in elucidating the reasoning of our proposed clinical association network, increasing transparency, and thus, the trustworthiness of this solution.

Noteworthy, this work represents a proof of concept rather than an established solution, and therefore must be carefully polished to achieve its highest potential. Firstly, in the current work we have investigated how Shap analysis can provide explainable predictions of deep neural networks by only considering the final module of the pipeline, namely the clinical association network proposed in this section. Therefore, the exact way that the initial DL-CAD system produced the qualification score still remains unclear, hindering the overall interpretability and explainability of the pipeline. Extending this analysis to the entire end-to-end solution represents a future direction of this work.

Secondly, the entire workflow is currently covered by multiple independent modules. The advantage of this approach is that intermediate outcomes can be used to make a prediction if data required by subsequent models is not available for a specific patient. On the other hand, this approach can suffer from error propagation if one of the initial modules fails. An alternative solution would be an end-to-end framework to create a single model covering all the tasks described above, while taking full advantage of a richer feedback signal provided through a multi-task training scheme. However, such approach could hinder the overall pipeline transparency due to an increased complexity.

Lastly, for the clinical association network we only took into consideration a limited set of features that were available in our data. However, in clinical practice additional parameters might be valuable in discriminating clinically significant prostate cancer, including other clinical parameters (e.g. testosterone levels), other imaging studies, genomics profiles, etc. Extending the current analysis with additional patient information might yield significant improvements in accuracy by providing the ground for a better disease stratification. Nevertheless, some types of data might not be widely assessed in clinical routine, and therefore, from a feasibility perspective, should be integrated as optional parameters.

Beyond the DL-CAD system being employed herein, our findings suggest that even the PI-RADS scoring system may benefit from additional information available in EHRs. While the current guidelines suggest that the readers should be blindfolded to any clinical or demographics features, thus assigning a PI-RADS score solely based on multi-parametric MRI images, our results prove that meaningful patterns could be learned to enhance the PI-RADS performance in separating clinically significant lesions from the others.

Overall, we have demonstrated that an increased performance can be obtained when imaging is combined with other types of data and further jointly modeled by machine learning algorithms that can identify meaningful patterns and use them to predict certain outcomes. Besides a rigorous evaluation, DL-CAD systems should be enriched with explainability and interpretability properties to become trustworthy, potentially unlocking tremendous benefits for clinicians and their patients.

4.3 Non-small cell lung cancer subtype classification[1]

4.3.1 Use-case introduction

Next-generation sequencing (NGS)-based genotyping technologies are increasingly being employed to support clinical decision-making. The standard approach to developing diagnostic, prognostic, or predictive models based on NGS-generated high-dimensional data is to preselect a small number of biomarkers, e.g., gene mutations that are identified to be independently associated with the phenotype of interest, and restrict training of a basic statistical model to this restricted set of biomarkers, e.g., to predict survival benefit from immunotherapies[123, 124].

In fact, disease etiology and treatment response are often complex, affected by a multitude of genomic alterations, each exerting a small effect on the phenotype. Therefore, it is expected that the consideration of all identified alterations broadens the range of phenotype risk. However, due to the high dimensionality of genomic data, inherent genetic heterogeneity, as well as small patient cohorts, the robust training of predictive models is challenging[125].

[1] This section describes experiments done in [122], which represents previously published work of the author, under the PhD research program.

Functional readouts of cellular activity and physiological status as provided by 'omics technologies such as RNA sequencing, measuring genome-wide changes in mRNA expression, are expected to have greater capacity to inform clinical management decisions, including diagnosis, prognosis, treatment selection, and monitoring. Ideally, 'omics technologies that capture the complex molecular interplay within and across different biological levels should be combined[126]. Still, due to practical and financial reasons, genomics is the most clinically adopted 'omics modality to date[127, 128, 129, 126].

In cancer, histology-based classification of tumors reflects different clinical presentation and course of the disease. For instance, lung cancers are classified as small-cell (SCLC) or non-small cell (NSCLC). In addition, NSCLCs are further subdivided into subtypes such as lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). While histology-based determination may be inconclusive, LUADs and LUSCs show distinct genetic drivers and cellular signaling activities[130], influenced by the cell type of origin[131].

Importantly, different prognostic determinants have been identified in LUAD vs. LUSC, which also partly show opposite impact on clinical outcome[132]. In addition, NSCLC histology was found to be predictive of treatment response. For instance, platinum-based adjuvant chemotherapy for patients with completely resected early-stage NSCLC conferred a survival benefit in LUSCs but not LUAD[133]. In contrast, pemetrexed chemotherapy only showed improved efficacy compared to other standard treatment options in patients with advanced non-squamous NSCLC[134]. NSCLC histology is also considered predictive of response to targeted therapies and immunotherapies in the latest ASCO guidelines for late-stage NSCLC patients with[135] and without[136] actionable driver alterations. The combined treatment of NSCLC with chemotherapy and immunotherapy is the subject of ongoing investigations as part of clinical trials[137].

Analogous to the differential response to anticancer drugs, the dose effect of stereotactic body radiation therapy (SBRT) has also been described to differ between NSCLC subtypes[138, 139], with an increased local failure rate in patients with LUSC, indicating the need for histology-specific treatment adjustment. Therefore, knowledge of NSCLC histology is essential in the optimized selection from available therapy options.

In this work, we aimed at improving NSCLC subtype classification from mutational data and developed a deep genomic profiling model that, in addition to LUAD and LUSC samples, simultaneously learns from adenocarcinoma (AD) and squamous cell carcinoma (SCC) samples of other tissue types and that is regularized using a neural network model trained from gene expression data. Notably, classification performance can be improved on samples with confident predictions, identified with an ensemble approach capturing prediction uncertainties. Moreover, uncertainty estimates of misclassified samples indicate limitations of the current NSCLC classification scheme in representing mutational heterogeneity within subtypes, potentially impeding the prediction of treatment outcome.

4.3.2 Results

To establish a baseline, we trained a genomic profiling model consisting of a multilayer perceptron (MLP) to classify NSCLC samples into LUAD and LUSC subtypes using mutational data. This baseline model achieved an area under the receiver operating characteristic (ROC) curve (AUC) of 0.82 as a measure of classification performance on test samples using a 10-fold cross-validation scheme.

Expression-based regularization improves the manifold, without improving classification performance

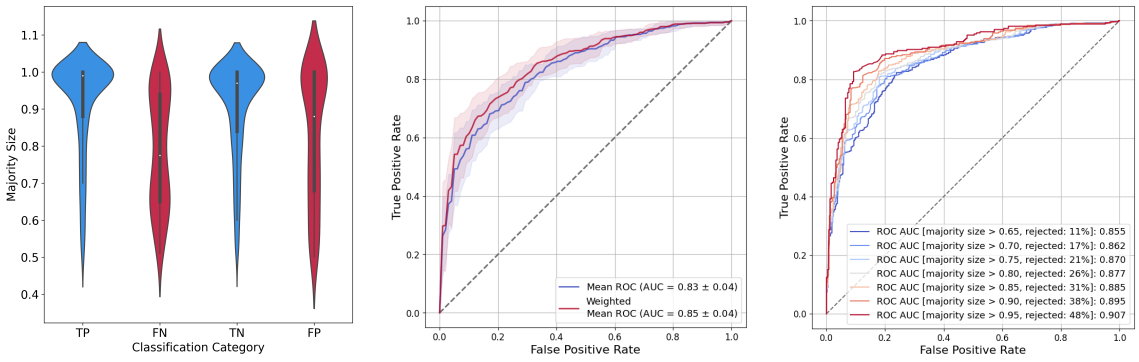
The classification performance of the baseline genomic profiling model could not be exceeded, irrespective of different training and model parameter configurations tested, demonstrating the challenge this sparse, high-dimensional genomic dataset presents. To overcome this challenge, we created an extended dataset by augmenting the NSCLC dataset with additional AD and SCC samples of non-NSCLC histology. This extended dataset served as a regularizer by training a genomic profiling

model to simultaneously classify AD and SCC samples of the lung as well as other tissue types using two prediction heads. However, the classification performance on NSCLC samples did not significantly improve over the baseline genomic profiling model.

Since NSCLC subtypes were found to be distinct at the transcriptomic level[130], we further aimed at improving NSCLC classification accuracy by regularizing training of the genomic profiling model with the latent representation learned by a gene expression-based profiling model, thereby obtaining an expression-aware genomic profiling model. Like previous reports, performance of the expression-based profiling model in classifying NSCLC subtypes, achieves an AUC of 0.98 (± 0.01). However, regularized training did not succeed in leveraging the prediction capacity of the expression-based profiling model to improve the classification performance of the expression-aware over the baseline genomic profiling model.

Prediction uncertainty estimates enable increased performance on confident samples

The results indicate an inherent complexity and ambiguity in terms of how the genomic profile translates into cellular activity and physiological status. Capturing this uncertainty within the model might allow ambiguous samples to be identified and rejected, while improving the performance on remaining samples. To this end, we employed a bootstrap aggregating (also called bagging) approach to train an ensemble of one hundred expression-aware genomic profiling models and calculated an aggregated prediction score by averaging prediction values of all models predicting the majority predicted NSCLC subtype.



(a) (b) (c)

Figure 4.3: Performance of an ensemble of expression-aware genomic profiling models trained to classify NSCLC subtypes using mutational data of the extended dataset. (a) Distribution of the majority size as uncertainty estimate for different classification categories, with TP, FN, TN, and FP corresponding to true positive, false negative, true negative, and false positive predictions, respectively. (b) Performance of the model calculated with (red) and without (blue) uncertainty-based weighting of samples. (c) Performance of the model when applying different uncertainty estimate thresholds and rejecting respective test samples. In addition to the applied threshold, the legend lists the fraction of rejected in all samples.

Although the classification performance of the ensemble remained unchanged (Fig. 4.3(b), blue ROC curve), the ensemble enabled inconclusive samples to be identified by calculating the fraction of models predicting the majority predicted NSCLC subtype (termed majority size) as a measure of prediction uncertainty.

Figure 4.3(a) shows the agreement between the models of the ensemble separately for each classification category. Since the agreement is greater for correctly classified (TP and TN) as opposed to misclassified (FP and FN) samples, the majority size turned out to be a reasonable measure of prediction uncertainty.

Indeed, the AUC could be increased by 0.02 when utilizing the uncertainty estimates as weights in the calculation of the ROC curve (Fig. 4.3(b), red ROC curve). Besides, a threshold value can be applied to reject samples with high prediction uncertainty and, as a result, increase classification performance on remaining samples, as demonstrated in Fig. 4.3(c). Crucially, the AUC monotonically increased the more restrictive the threshold applied. However, there is a tradeoff between classification performance and the fraction of rejected samples. For instance, imposing a minimum majority size of 0.75 increased the AUC to 0.87 but also implicated the rejection of 21 % of the test samples. Likewise, the AUC exceeded 0.9 when applying a more restrictive threshold of 0.95, with the downside that nearly half of the test samples were rejected.

Conflicting mutational patterns limit classification performance

To estimate the contribution of each feature, i.e., the mutational status of each gene, to model predictions, we conducted a SHAP (SHapley Additive exPlanations) analysis using KernelSHAP[27]. This local explainability model trains a surrogate model to learn the Shapley values of different feature combinations (termed coalitions) as its weights and calculates the average contribution of each feature to the predictions of different coalitions in comparison to the average prediction across all samples.

We also estimated the contribution of the mutational status of individual genes to prediction uncertainty. Overall, 19 of the 20 most important genes, corresponding to genes that are recurrently mutated in NSCLC, are identical. However, gene mutations can show opposite effects on prediction uncertainty, depending on co-occurring mutations.

This effect cannot be detected when contrasting SHAP analyses of correctly to incorrectly classified LUAD samples with respect to model predictions (Figs. 4.4(a) and 4.4(b)). However, a SHAP analysis with respect to prediction uncertainties (Figs. 4.4(e) and 4.4(f)) reveals that mutations in genes indicative of LUAD, e.g., KRAS, can increase prediction uncertainty due to conflicting mutations in other genes (Fig. 4.4(f)). Similar effects can be observed in correctly vs. incorrectly classified LUSC samples (Figs. 4.4(c), 4.4(d), 4.4(g), and 4.4(h)).

To explore such mutational patterns on the instance level, the waterfall plots in Figs. 4.5(a) and 4.5(b) illustrate the contribution of the mutational status of individual genes to a correct and an incorrect prediction of two selected LUAD samples, respectively. While the mutation in KRAS has a large impact on correctly classifying TCGA-05-4390-01 as a LUAD sample, the inconclusive mutational pattern in LUAD sample TCGA-93-A4JN-01, comprising mutated ATM but also non-mutated KRAS, increases prediction uncertainty (Fig. 4.5(f)) due to opposing effects on model prediction (Fig. 4.5(b)). In contrast, in LUAD sample TCGA-05-4390-01, mutated ATM decreases prediction uncertainty due to a co-occurring mutation in KRAS (Fig. 4.5(e)).

Similarly, correct vs. incorrect classifications of LUSC samples can also be attributed to more vs. less consistent mutational patterns, as exemplified with LUSC samples TCGA-85-7698-01 (Figs. 4.5(c) and 4.5(g)) and TCGA-66-2727-01 (Figs. 4.5(d) and 4.5(h)), respectively.

From these observations, we deduce samples with higher prediction uncertainty to exhibit mutational patterns that are ambiguous with respect to NSCLC histology, which could be indicative of mixed-type histologies. To investigate this further, and due to the lack of a corresponding label (samples are annotated as either LUAD or LUSC), we selected the most confident driver genes for LUAD (BRAF, EGFR, KRAS, and STK11) and LUSC (CDKN2A, NFE2L2, PIK3CA, and PTEN) and show a comparative analysis between samples carrying a mutated LUAD driver, a mutated LUSC driver gene, or both (mixed driver mutations) as a proxy of mixed-type histology. Remarkably, the classification performance of our genomic profiling model is comparable across all subgroups (Fig. 4.6(a)). Moreover, predictions of mixed driver samples show a trend towards intermediate prediction uncertainty estimates (Fig. 4.6(b)). Finally, the separation of progression-free survival (PFS) curves underlines the clinical relevance of the mixed driver subgroup (Fig. 4.6(c)).

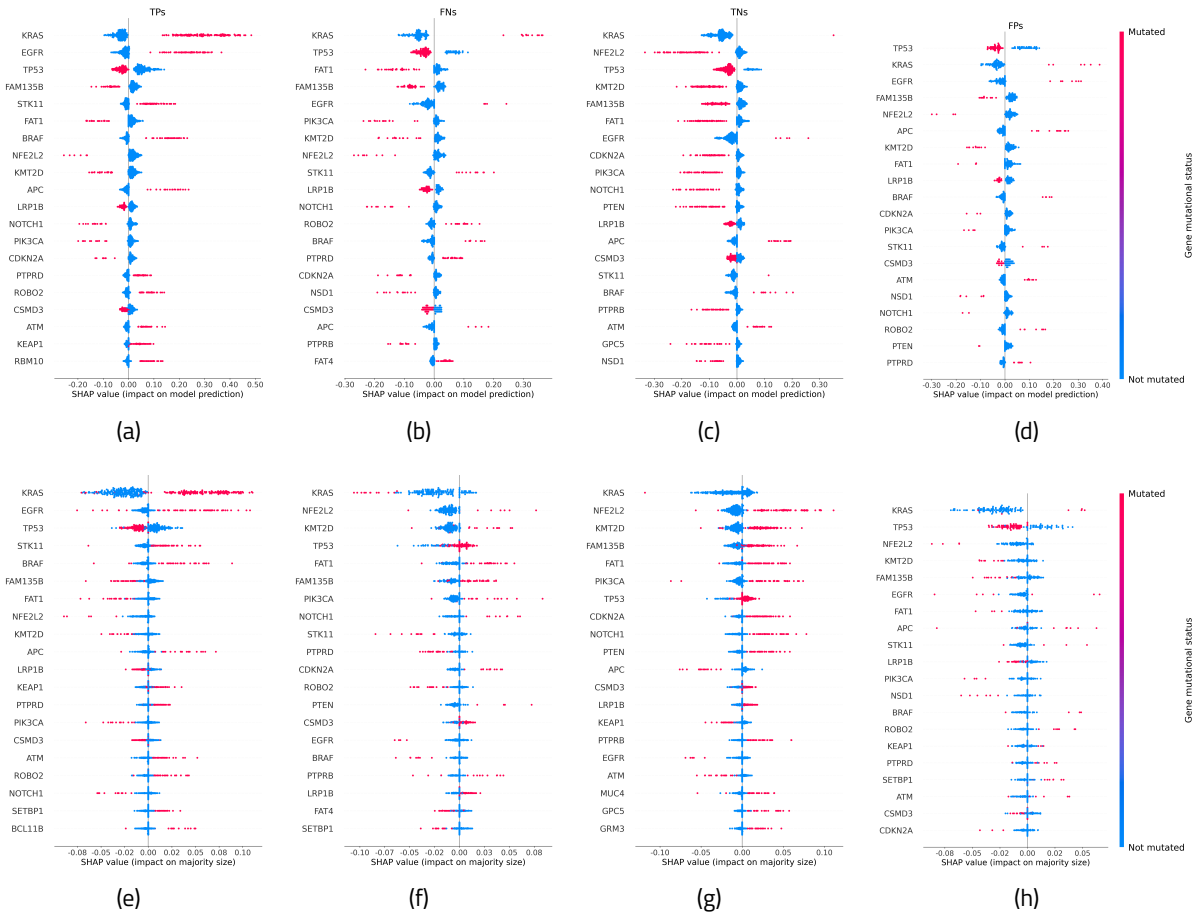


Figure 4.4: Cohort-level SHAP summaries showing the contribution of the mutational status of the 20 genes with greatest impact on (a-d) model prediction and (e-h) prediction uncertainty for different prediction categories: (a, e) true positive (TP), (b, f) false negative (FN), (c, g) true negative (TN), and (d, h) false positive (FP) predictions, respectively. The analysis is based on predictions of the ensemble of expression-aware genomic profiling models trained to classify NSCLC subtypes using mutational data of the extended dataset.

4.3.3 Discussion

The capability of our deep genomic profiling model to assess the confidence in a prediction permits the identification of indeterminate or out-of-distribution samples, which is indispensable for deep learning approaches to become accepted and implemented in clinical practice. In this work, we trained an ensemble of one hundred genomic profiling models. While the size of the ensemble can still be optimized to reduce computing resources, we can also imagine a teacher-student learning setup, providing a distilled model that is more practical for clinical application[140].

Overall, the estimation of prediction uncertainties facilitates the investigation of model predictions, as the contributions of features cannot only be assessed with respect to model predictions but also prediction uncertainties. In fact, such an analysis identified co-occurring mutations indicative of both NSCLC subtypes in misclassified samples, which explains the limited performance observed in classifying NSCLC subtypes using mutational data and questions the current NSCLC subtype classification to adequately represent inherent mutational heterogeneity. This observation is of particular importance as specific mutational patterns in NSCLC have also been linked to clinical heterogeneity[141]. For instance, in non-squamous NSCLC, KRAS mutation has been shown to interact with co-occurring mutations in TP53, STK11, PTPRD, RBM10, and ATM with respect to immune checkpoint inhibitor efficacy[142]. Most of such interactions originate from tumor-initiating mutations in

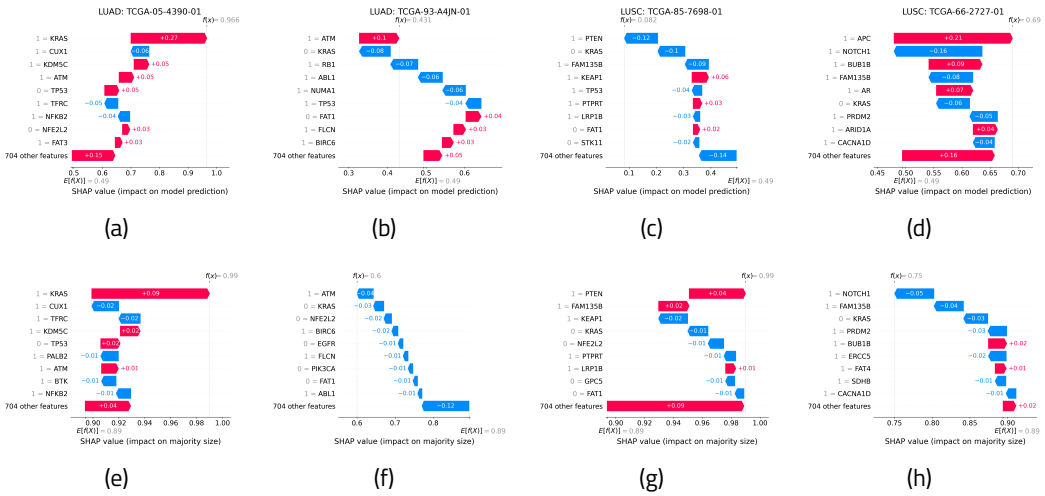


Figure 4.5: Waterfall plots of SHAP feature attributions showing the contribution of the mutational status of the nine genes with greatest impact on (a-d) model prediction and (e-h) prediction uncertainty for selected samples, representing (a, e) a correctly and (b, f) an incorrectly classified LUAD sample with low and high prediction uncertainty, as well as (c, g) a correctly and (d, f) an incorrectly classified LUSC sample with low and high prediction uncertainty, respectively. f(x) corresponds to (a-d) the prediction value (1 = LUAD, 0 = LUSC) and (b) its associated prediction uncertainty (1 = lowest, 0.5 = highest uncertainty), respectively. Plots are aligned at the average prediction across all test samples, E[(f(x)]. Feature values of 0 and 1 correspond to non-mutated and mutated genes, respectively. Predictions are based on the ensemble of expression-aware genomic profiling models trained to classify NSCLC subtypes using mutational data of the extended dataset.

KRAS, TP53, and EGFR[143], resulting in exclusivity patterns that have been found to be associated with response to both targeted[144] and immunotherapy[145].

As input to our model, we used a binary encoding of the mutational status of each gene. However, more details on the functional impact could be used, e.g., by separately considering low, medium, and high impact mutations, as obtained from Variant Effect Predictor (VEP)[146]. Furthermore, summary statistics could be constrained to mutational hotspots or gene regions encoding structural domains. However, both approaches may increase sparsity.

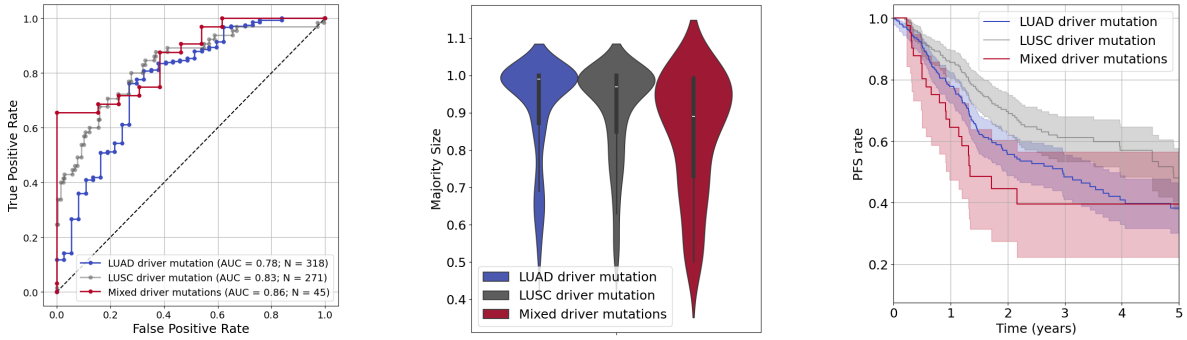
To deal with the curse of dimensionality, where the variance between samples becomes large and sparse[126], the integration of prior knowledge about direct[147] or indirect [148] protein-protein interactions as relational inductive bias may help to effectively reduce the parameter space relative to the naive approach of modeling all interactions terms and, thus, may allow robust training of complex models on small cohorts.

Since our expression-aware genomic profiling model captures mutational patterns linked to histology and treatment efficacy, we anticipate our model to be suited as foundational model of genomic data for clinically relevant prognostic or predictive downstream tasks. To enhance its generalizability, regularization based on additional modalities and tasks may need to be integrated to enable learning a more holistic representation of the phenotype, which emanates from variation across all 'omics levels[149].

4.3.4 Materials and Methods

4.3.4.1 Data

The results of our work are based upon data generated by the TCGA Research Network: `https://www.cancer.gov/tcga`. The derived Pan-Cancer Atlas datasets[131] were downloaded from cBioPortal[150]. We only included primary tumor samples of AD and SCC histology for which both muta-



(a) (b) (c)

Figure 4.6: Comparative analysis of NSCLC subgroups based on the presence of a mutated LUAD driver gene (BRAF, EGFR, KRAS, or STK11), a mutated LUSC driver gene (CDKN2A, NFE2L2, PIK3CA, or PTEN), or both (mixed driver mutations). (a) Performance of the ensemble of expression-aware genomic profiling models across subgroups. The dashed line represents the ROC curve of a random classifier. (b) Distribution of the prediction uncertainty estimate (majority size) across subgroups. (c) Kaplan-Meier curves of progression-free survival (PFS) across subgroups.

tion and expression data were available. Furthermore, we restricted our analysis to genes annotated in the Cancer Gene Census (CGC) of the Catalogue of Somatic Mutations In Cancer (COSMIC, version 95)[151], excluding genes for which either mutation or expression data was not available, yielding 713 genes. Finally, we removed samples with incomplete data for the selected genes.

Progression-free survival analyses were conducted on a subset of 497 LUAD and 476 LUSC samples for which progression-free status annotations were available.

4.3.4.2 Methodology

Model training comprised two phases. First, a gene expression-based model (termed expression-based profiling model) was trained to classify samples into NSCLC subtypes by minimizing the sum of two binary cross entropy (BCE) loss terms (one for each prediction head). Subsequently, a gene mutation-based model (termed expression-aware genomic profiling model) was trained on the same classification task by adding to the loss an L2 regularization term that is calculated between the latent representation of the mutation-based model and the latent representation of the expression-based model.

Due to the small number of available samples, all experiments were conducted using a stratified 10-fold cross-validation scheme. While, in each iteration, one fold was held back as testing set, the samples of the remaining nine folds were further partitioned into a training (80%) and a validation set (20%), again stratified with respect to NSCLC subtype. To train an ensemble of one hundred models, the training set was bootstrapped a hundred times, resulting in one hundred bootstrap samples (each serving as training set for one model). The bootstrap samples were the same size as the training set. The prediction score of the ensemble of models was obtained by averaging prediction values of all models predicting the majority predicted NSCLC subtype. To aggregate the results across testing folds, we averaged the true and false positive rates at each possible operating point in the ROC curve and calculated the AUC afterwards. Prediction uncertainties were calculated as the fraction of models predicting the majority predicted NSCLC subtype in all models of the ensemble (termed majority size).

The models were trained using the Adam optimizer with an initial learning rate of 10^{-4} and a batch size of 64. To counteract potential class imbalance effects, we sampled each training batch using a weighted random sampler[152]. Within each cross-validation iteration, the best model was selected as the one with minimum loss on the validation set, using a patience of 20 epochs. To enable optimal knowledge transfer from the expression-based profiling model, the expression-aware

genomic profiling model was trained using only the L2 regularization term for 30 epochs (empirically determined from learning curves), before switching on the task-specific loss terms.

To create NSCLC subgroups based on the presence of a mutated LUAD driver gene, a mutated LUSC driver gene, or both (mixed driver mutations), we selected the most confident driver genes in either LUAD (BRAF, EGFR, KRAS, STK11) or LUSC (CDKN2A, NFE2L2, PIK3CA, PTEN) based on a Pan-Cancer Atlas analysis[153] using a consensus score greater than four. TP53, which is a recurrently mutated driver gene in both subtypes, was excluded.

4.4 Conclusions

In this chapter we approached the current trustworthiness limitations of deep learning based solutions, exploring the efficiency of two established techniques that address (1) the inherent lack of transparency and (2) the overconfidence in predicting uncertain samples. While these concerns have represented a blocker in general adoption of DL-CAD systems in clinical routines, our findings suggest that trustworthy AI can be created by accounting for these limitations when designing specific solutions.

From an explainability and interpretability standpoint, we herein employed a Shap analysis to assign relative importance scores to input features, quantifying their impact on the final prediction. Specifically, we investigated how this approach elucidates the reasoning of two deep neural networks in (1) adjusting the output of an existing DL-CAD system to better identify clinically significant prostate cancer by adding additional clinical and demographics information, and (2) classifying non-small cell lung cancer into sub-types - namely lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) - based on the patients mutation profiles. For each scenario, we have conducted two types of analysis, one providing an overview of the feature importance ranking at a cohort level (overall analysis) and one describing the relative influence of each input on the prediction at a case level.

The overall analysis provides a clear picture of how selected predictors correlate with the outcome, being suitable for an initial model evaluation and characterization. By delivering such results, domain experts could get better insights into what patterns and correlations are getting leveraged by the model in reaching at certain conclusions, thus significantly increasing its transparency. Therefore, the reasoning could be either validated or disregarded from an early stage significantly improving safety related aspects in embedding such DL-CAD solutions in the existing pipelines. On the other hand, instance level explanations could be employed in real time to further ensure a safe functionality of the deployed model. If the reporting is designed to also describe how the system reached at a certain conclusion clinicians could quickly confirm or infirm its output, thus significantly improving reliability while maintaining their workload at reasonable levels.

Moreover, uncertainty estimations could be employed to further reduce clinical work burden while maintaining a trustworthy setting. This chapter presents how model ensembles can be leveraged to produce such uncertainty scores, allowing the model to identify corner cases that would require a special attention from clinicians. Examples of such situations could represent uncommon cases that were not well represented in the training data, ambiguous cases where conflicting patterns occur, etc. Therefore, uncertainty quantification techniques empower the model to say "I don't know" rather than attempting to provide error-prone predictions, which represents a mandatory property of a trustworthy autonomous system. Moreover, certainty estimates accompanying the prediction significantly reduce the work volume of health care practitioners who could shift their attention only to the fraction of cases where the DL-CAD system is prone to failure.

Overall, we conclude that each AI based solution should possess the properties described above, namely explainability, interpretability and uncertainty estimation. We herein demonstrated the outstanding benefits of delivering a trustworthy system that ensures a safe functionality in practice, while not drastically increasing the overall complexity of the development process. Therefore, it is worth to deliver trustworthy AI!

4.5 Discussion

The findings reported in this chapter reflect that deep learning models, previously regarded as black-boxes, can be enriched to better reflect the underlying reasoning process that stands behind their predictions. We showed that explainable AI can be obtained by employing a post-hoc analysis that provides estimates of how much each input feature weights in building a certain decision, regardless to its architectural or optimization complexity. Moreover, besides the initial benefits stemming from a better model characterization, intepretability techniques can support a trustworthy functionality of the system when employed in clinical practice.

Since machine learning algorithms heavily rely on the underlying training data quantity, quality and completeness, uncertainty estimations represent safeguards that ensure a proper utilization in practice. This chapter provided evidence on how such estimations identify ambiguous cases that should be rather handled by domain experts to ensure an adequate health care delivery. Considering the potential patient harm, each autonomous system should possess such safeguards to minimize the risk of making errors while maximizing the outstanding benefits promised by AI on patient care.

Noteworthy, the results provided herein only represent a proof of concept, hence not an established set of best practices. Other techniques or approaches to make DL-CAD systems explainable, interpretable and aware of uncertainty might be better suitable from one use-case to another. For instance, ensemble based uncertainty estimations may not be suitable in real-time applications, where a maximum inference time limit must be fulfilled. While a trade-off between the number of models in the ensemble and inference time exists, uncertainty estimates may be less accurate due to limited individual prediction attempts. However, knowledge distillation techniques could potentially bridge the aforementioned gap, but assessing its feasibility was not a subject of this work.

From a feature importance standpoint, many studies attempted to solve the transparency issues of deep learning models [24, 25, 26, 27, 28]. While the Shap analysis represents a promising solution, the KernelShap method engaged in explaining the reasoning of deep neural networks uses a set of stochastic processes that simulate various coalitions of input features and their impact on the outcome. Due to potential biases stemming from the sampling strategy, these processes might represent a source of error. However, explanations of relatively simpler machine learning algorithms (e.g. RandomForest classifiers) are more reliable by nature as compared to sophisticated solutions. Therefore, to achieve the best transparency of autonomous systems one guiding principle should be the usage of minimal complex solutions for a certain task, when the performance metrics do not justify the opposite.

Moreover, he hypothesize that besides allowing clinicians to assess the reliability of predictions, explainable AI has the potential to identify possible unknown patterns in the data that could lead to a better disease understanding. Therefore, such explanations could trigger clinical trials designed to further explore the impact of various features on certain medical conditions, possibly leading to groundbreaking discoveries in medicine.

Nevertheless, it is worth mentioning that the benefits of employing DL-CAD systems in appropriate settings are not only limited at reducing clinical work burden, but also the overall diagnosis accuracy. Since diagnostic precision has been shown to be dependent on the radiologist experience level [11, 10, 9], adoption of such autonomous solutions in clinical routines could represent a tremendous asset in supporting clinicians, especially at the beginning of their careers. For instance, Winkel et. al. [11] reported a 4 points improvement in AUC when radiologists were accompanied by the DL-CAD system employed in section 4.2 as a baseline, while the median reading time decreased by over 20%.

Overall, explainability, interpretability and uncertainty awareness represent building blocks to trustworthy AI, which should nowadays become a standard practice to delivering autonomous solutions in various industries, especially where potential errors could have disastrous consequences to the end-beneficiary.

5. Final Conclusions

Final conclusions
Original contributions
Dissemination of Research results
Discussion

5.1 Conclusions

This PhD thesis demonstrates how major challenges in developing deep learning based solutions for various clinical needs can be addressed by employing paradigms such as self-supervision, synthetic data generation, uncertainty quantification and feature importance estimation. In spite of holding an outstanding potential in improving current clinical practices, and ultimately patient outcomes, immediate adoption of AI in healthcare routines is hindered by a set of concerns, including patient confidentiality, data labeling requirements and trustworthiness aspects.

Recent technological advancements in semiconductor industry and subsequently computational power have placed healthcare into a transformation phase, substantially changing current practices towards adoption of minimally invasive diagnosis techniques, image guided therapy planning, delivery, and health monitoring. While exhibiting tremendous benefits for patients, the current guidelines have markedly increased the workload of clinical practitioners such as radiologists, thus urging for appropriate software developments to support a reliable healthcare delivery while maintaining caregivers work burden at manageable levels. Deep learning based approaches have been shown to be the current goal standard in approaching various clinical problems, but nowadays the only feasible training paradigm is supervised learning. To that extent, a proper data labeling step represents a prerequisite for any DL model training, step that could possibly further increase workload of clinical practitioners. To this end, chapter 2 presents a self-supervised approach to train neural networks in extrapolating medical images, where input-output pairs are randomly simulated at training time, thus not requiring any manual annotation phase.

Moreover, among all types of personal information, healthcare data is one of the most restricted types as a consequence of being imposed to guarantee patient confidentiality. Although fundamentally correct and efficient, current constrains regarding healthcare data circulation have an adverse impact on enabling extensive AI exploitation in various clinical scenarios, where current practices could be improved. While a robust training of machine learning algorithms heavily depends on the availability of large scale, complete and qualitative datasets, gathering such databases in practice is often unfeasible. Chapter 3 addresses this obstacle by describing a methodology to create synthetic longitudinal patient data, free of any privacy related risks, that can be used to train neural networks in responding to diverse clinical needs. We herein exemplify this idea by approaching a highly relevant clinical use-case, namely prostate cancer patient stratification.

Nonetheless, AI models are susceptible to delivering incorrect predictions when presented with samples under represented in their underlying database. The inherent lack of transparency exhibited by deep learning algorithms prevents their end users from genuinely understanding the reasoning processes behind certain predictions, hence raising skepticism and reticence in adopting such

solutions their clinical routines, especially given the enormous negative impact an error could have on patients well being. Chapter 4 addresses this concern by enhancing DL models with explainability, interpretability and uncertainty awareness properties. All these enriched capabilities have been demonstrated to play a significant role in boosting the DL based systems robustness and stability, acting as safe-guards to ensure a secure use in clinical routines through an increased transparency. We herein demonstrated that model ensembles and Shap analysis can be successfully employed in providing uncertainty estimations and quantitative explanations of how a model inferred certain outcomes by following a couple of relevant use-cases, namely clinically significant prostate cancer prediction and non-small cell lung cancer classification

5.1.1 Self supervised learning for thin image extrapolation and registration

Chapter 2 presents a self-supervised learning approach to enhance the robustness of intraoperative CT imaging based guidance systems. Due to the relatively reduced field of view of real time acquisitions, and thus reduced context information, the registration step required to align these images with high quality preoperative scans is error prone. Therefore, we showed how DL can be successfully engaged in expanding the FOV of thin CT images by employing a generative adversarial framework. Results presented in this thesis demonstrate the tremendous positive impact DL can have on the overall system stability, reducing the median registration errors with an order of magnitude, ultimately yielding an improved and robust guidance for surgical interventions.

On the other hand, in the context of this study the role of self-supervised learning paradigm is essential. Given the nature of this use-case, gathering input-output data pairs for training ML algorithms in a supervised fashion is not only costly and resource exhaustive, but also unfeasible due to the need of perfectly registering the inputs (i.e. intraoperative images) with outputs (i.e. preoperative acquisitions). Therefore, experiments presented within this thesis demonstrate how self supervised learning paradigms could act as enablers for some clinical scenarios previously deemed as unrealistic, going beyond their mostly explored abilities of pretraining neural networks to facilitate a superior optimization on the downstream tasks.

5.1.2 Synthetic data generation for prostate cancer patient stratification

In response to the currently experienced challenges on gathering large scale, complete and qualitative longitudinal datasets, chapter 3 presents a systematic approach to create synthetic electronic health records reflecting the pathway of prostate cancer patients from diagnosis to treatment follow-up. By carefully designing disease modules in form of clinical care-maps combined with relevant disease incidence and statistics, this thesis demonstrated that reliable, consistent and realistic purely synthetic patient records could be created and ultimately employed in addressing various clinical needs.

From a qualitative perspective, we demonstrated that our synthetically generated cohorts exhibit similar distributional properties as real, but less complete datasets presented in literature. Noteworthy, the aforementioned conclusion not only refers to individual laboratory measurements or test results, but also to the longitudinal nature of the data: we herein demonstrated consistency across various diagnostic tests and procedures - including laboratory, imaging and biopsy -, treatment options and subsequent side effects.

As a result of this activity, the synthetic data generator can essentially produce an unlimited number of high-fidelity synthetic data samples that could be employed in a wide range of activities, including modeling, data analysis, prototype development, system testing, etc. We herein exemplified this by training a prostate cancer stratification model to infer patients TNM stage based on their entire EHRs. The nearly perfect classification accuracy obtained on the testing sets (where all stochastic processes were governed by different seeds as compared to the ones selected in the training data generation) serve as an extra proof of data consistency, providing a data quality check. Moreover, parts of these resulting predictors could act as foundational models that can be further fine-tuned

on relatively smaller datasets to accomplish a wide variety of clinically relevant tasks that could leverage the abstract feature representations learned from our synthetic data.

5.1.3 Trustworthy AI

Chapter 4 presents a couple of techniques that could be attached to any DL framework to increase the overall stability and transparency of the system. We have herein provided concrete examples of how feature importance estimations can significantly ameliorate the black-box nature of DL models, while uncertainty quantification can be employed in ensuring their safe operation in clinical routines.

Concretely, Shapely values were considered as a reliable proxy of how various features influenced the model to delivering certain predictions, providing quantitative estimations of the relative importance that each ingredient had to the final outcome. This type of analysis can be either performed at cohort level or at instance level, each having a distinct important role in providing trustworthy AI: while the latter aims at improving the safety related aspects of using such systems in practice by providing an explanation along with prediction, the earliest allows for an early stage model characterization and evaluation by furnishing insights into patterns being learned from the data. This thesis demonstrates the feasibility and utility of this analysis in (1) improving prediction of clinically significant prostate cancer by employing additional clinical and demographics features, and (2) differentiating non-small cell lung cancer lesions into sub-types based on sparse and heterogeneous genomics profiles: in both scenarios, Shap analysis provided clear insights into how different features steered the inference towards certain predictions, thus placing a spotlight on the reasoning process. Therefore, this type of information accompanying the model output substantially increases its trustworthiness, presumably paving the path of AI based solutions towards adoption in clinical routines.

Moreover, an ensemble based uncertainty quantification approach have been explored to identify inconclusive samples that may also lead to unreliable NSCLC sub-type predictions. Since neural networks tend to be overconfident in their inference, such techniques significantly contribute to delivering trustworthy AI, by providing models with a way to identify cases where its prediction might be error-prone, and hence deliver an "I don't know" answer rather than attempting at supplying a certain response. Section 4.3 provides a demonstration of how an ensemble-powered uncertainty measure could reliably identify inconclusive samples that cannot be differentiated based on the set of features considered as input. Overall, this added capability can be interpreted as a safe-guard, significantly improving the system stability and trustworthiness aspects while still maintaining clinicians workload at manageable levels: for instance, low-uncertainty cases can be handled autonomously while high-uncertainty predictions could trigger a flag requiring further assistance from caregivers.

5.2 Original contributions

To begin with, all contributions made in this thesis rely on an extensive documentation phase, where **a comprehensive list of challenges specific to employing deep learning based solutions in Healthcare industry was compiled**. In particular, (1) the substantial demands of training data are often unfeasible due to privacy constrains, data incompleteness and/or lack of annotations, and (2) reticence to the inherent non-comprehensiveness of deep neural networks often represents a roadblock in general adoption of such solutions in clinical routines. **Furthermore, a set of workarounds to these limitations were defined and explored through a series of clinically relevant use-cases, bringing significant contributions in the field** as further described.

Self supervised learning for thin image extrapolation and registration

The idea of engaging a DL-based extrapolation solution as a prior step to registering small-to-large field of view images represents an original approach that ultimately boosted the performance of interventional CT image guidance systems. The thin intra-operative image FOV ex-

pansion provided more contextual information for the subsequent registration step, which became more robust, yielding an improved stability demonstrated through a reduction of median errors with a magnitude factor.

The self-supervised learning paradigm was utilized to create input-output pairs that guided optimization of the extrapolation model through a generative adversarial training process. Two different approaches were explored: Firstly, we investigated an asymmetric extrapolation approach, **where an extra registration step was originally proposed to derive the spatial information of extrapolated images. Concurrently, a relatively simpler symmetric outpainting strategy was assessed, outlining the pros and cons of each method in the context of general clinical feasibility.**

Synthetic data generation for prostate cancer patient stratification

Inspired by Synthea framework, **high quality prostate cancer disease modules were designed to generate coherent longitudinal data that reflects various patient pathway phases, ranging from diagnosis to staging, treatment, and ultimately to the follow-up. By leveraging invaluable guidance from medical experts, comprehensive clinical care-maps were defined to integrate all relevant steps a patient goes through to diagnose, treat and monitor prostate cancer. Furthermore, to ensure realness and consistency across generated data, a comprehensive list of relevant publications was compiled to extract statistical properties of real prostate cancer cohorts and constrain all stochastic states of the care-maps accordingly.** Consequently, the quality and realness of purely synthetic data produced in this work was assessed to ensure statistical equivalence to the real cohorts. Moreover, while real data collected in various clinical trials follow a clear set of inclusion/exclusion criteria, usually being targeted on some specific disease phase or aggressiveness, the synthetic data generator herein presented does not lack diversity, being able to produce the entire spectrum of cases (e.g. from low risk to very high risk, from localized to regional and/or metastatic patients, from cases handled through an active surveillance regiment to the very advanced cases handled with palliative care, etc.).

Nevertheless, this work assessed the feasibility of using resulted synthetically generated longitudinal datasets in developing predictive models. **A natural language processing approach was originally employed in stratifying prostate cancer patients through the TNM staging system based on clinical code descriptions presented in the EHRs, with very promising preliminary results.**

Trustworthy AI

To obtain a superior discrimination of clinically significant prostate cancer a rectification step was proposed, where clinical and demographics information was originally associated with the output of a state of the art computer aided diagnosis system to improve its prediction accuracy. The results presented in this thesis clearly demonstrate the benefits of this approach not only through performance evaluation on testing sets, but also by employing feature importance estimations to enhance the overall system transparency. Moreover, insights derived from this analysis are consistent to widely published research, hence indicating the overall feasibility of embracing such solutions in clinical practices.

While genomics information is recognized to be associated with certain diseases predisposition and treatment response, the standard approach to developing prognostic or predictive models is to preselect a small subset of genes established to be associated with the target phenotype. Conversely, **this work originally envisioned a modeling framework that aims at finding interactions between a substantially larger number of genes that could be leveraged to create a fingerprint of patient's phenotype (e.g. foundational model).** In spite of NSCLC subtype classification herein employed as a pretext task, such genomic profiling models can be further fine-tuned and specialized for more clinically relevant scenarios, such as treatment response prediction or prognosis.

To account for the inherent heterogeneity of the mutational statuses, **a knowledge transfer technique was originally explored, where abstract phenotype representations created based on not**

routinely collected gene expression data were leveraged to regularize the training of genomic profiling models, demonstrating the overall feasibility through manifold visualizations. Moreover, with ambiguities in gene mutational profiles not being an uncommon situation, **this work presents an ensemble based-approach to quantify prediction uncertainty, and thus identify inconclusive samples that might require additional investigation. Results outlined in this thesis reveal a good association of the proposed uncertainty metric with vastly relevant outcomes, such as progression free status. Concurrently, feature contribution analysis further exposed contradicting mutational patterns that represent limiting factors of the classification performance, significantly ameliorating the black-box nature of such DL models, and thus, enhancing the overall system trustworthiness.**

5.2.1 Summary of contributions

Table 5.1: Summary of contributions and dissemination.

No. Order	Contribution	Chapter/ Section	Dissemination article
1	An extensive literature search was performed to define the challenges faced nowadays in developing AI-based solutions for diverse clinical needs, stemming from data scarcity and reticence to non-transparent solutions. Workarounds to these limitations were subsequently explored through various clinical scenarios.	1	Puiu, A., Vizitiu, A., Nita, C., Itu, L., Sharma, P., & Comaniciu, D. (2021). Privacy-Preserving and Explainable AI for Cardiovascular Imaging. Studies in Informatics and Control, 30(2), 21–32. https://doi.org/10.24846/v30i2y202102.
2	In the context of image-guided interventions, an original contribution stems from the improved alignment of intra-operative and pre-operative images, which is powered by a DL-based extrapolation phase that enhances the FOV of thin CT acquisitions as a prior step to registration.	2	Puiu, A., Reaungamornrat, S., Pheiffer, T., Itu, L. M., Suciu, C., Ghesu, F. C., & Mansi, T. (2022). Generative Adversarial CT Volume Extrapolation for Robust Small-to-Large Field of View Registration. Applied Sciences (Switzerland), 12(6). https://doi.org/10.3390/app12062944
3	Another original contribution stands in designing an extra registration step to derive spatial information of enhanced images, enabling the use of an asymmetric extrapolation approach.	2.3	

Continued on next page

Table 5.1 – continued from previous page

4	Synthea disease modules were originally designed to generate consistent longitudinal electronic health records for prostate cancer patients, not only limited to a certain pathway phase, but spanning from diagnosis to treatment and even to the follow-up. By properly constraining all stochastic states in these modules, the synthetically generated data exhibit the same statistical properties as real cohorts widely presented in literature.	3, 3.3.1	The results are not yet published
5	The usability of the resulting synthetic data in developing predictive models was explored through a relevant clinical use-case, namely prostate cancer patient stratification using the TNM staging system. A natural language processing approach was originally proposed by leveraging clinical code descriptions randomized in the EHRs.	3.3.2	The results are not yet published
6	In terms of identifying clinically significant prostate cancer, the addition of supplementary clinical and/or demographic features to a state-of-the-art DL-CAD solution has proven to play a significant role in improving the overall classification performance. Moreover, quantifying the relative impact of each input feature to the final prediction further enhanced the reliability and transparency of the proposed method.	4.2	The results are not yet published
7	Another original contribution in this thesis stands in the development of a non-small cell lung cancer subtype predictor based on a vast set of genes, with the ultimate goal of creating a deep genomic profiling model (as a proxy to patient phenotype) that can be further repurposed to address various clinical unmet problems, such as prognosis or treatment response prediction.	4.3	Puiu, A., Gómez Tapia, C., Weiss, M. E. R., Singh, V., Kamen, A., & Siebert, M. (2024). Prediction uncertainty estimates elucidate the limitation of current NSCLC subtype classification in representing mutational heterogeneity. Scientific Reports, 14(1), 6779. https://doi.org/10.1038/s41598-024-57057-3

Continued on next page

Table 5.1 – continued from previous page

8	Given the nature of genomics data where the inherent heterogeneity poses significant challenges for machine learning approaches, a knowledge transfer scheme was employed in enhancing the latent representation by utilizing the less frequently explored, but more informative gene expressions. Moreover, the usage of feature importance estimations along with prediction uncertainty estimates enhanced method trustworthiness by allowing identification of inconclusive samples, which appeared to be associated with prognosis.	4.3	Puiu, A., Gómez Tapia, C., Weiss, M. E. R., Singh, V., Kamen, A., & Siebert, M. (2024). Prediction uncertainty estimates elucidate the limitation of current NSCLC subtype classification in representing mutational heterogeneity. Scientific Reports, 14(1), 6779. https://doi.org/10.1038/s41598-024-57057-3

5.3 Dissemination of research results

As a result of the research conducted throughout this PhD program, 14 publications as author or co-author in various scientific journals were attained.

Three journal articles were published as first author during the PhD program:

- Puiu, A., Gómez Tapia, C., Weiss, M. E. R., Singh, V., Kamen, A., & Siebert, M. (2024). Prediction uncertainty estimates elucidate the limitation of current NSCLC subtype classification in representing mutational heterogeneity. Scientific Reports, 14(1), 6779. https://doi.org/10.1038/s41598-024-57057-3. (impact factor: 4.6, Q2)
- Puiu, A., Reaungamornrat, S., Pheiffer, T., Itu, L. M., Suciu, C., Ghesu, F. C., & Mansi, T. (2022). Generative Adversarial CT Volume Extrapolation for Robust Small-to-Large Field of View Registration. Applied Sciences, 12(6), 2944. https://doi.org/10.3390/app12062944. (impact factor: 2.7, Q3)
- Puiu, A., Vizitiu, A., Nita, C., Itu, L., Sharma, P., & Comaniciu, D. (2021). Privacy-Preserving and Explainable AI for Cardiovascular Imaging. Studies in Informatics and Control, 30(2), 21–32. https://doi.org/10.24846/v30i2y202102. (impact factor: 1.6, Q4)

Nine journal articles were published as co-author during the PhD program:

- Benedek, T., Ferent, I., Benedek, A., Cernica, D., Nita, C., Puiu, A., Itu, L., Rapaka, S., Puneet, S., & Benedek, I. S. (2020). P1434 Evolution of coronary wall shear stress following implantation of bioabsorbable vascular scaffolds—First results of a 1-year follow-up pilot study. European Heart Journal - Cardiovascular Imaging, 21(Supplement 1), https://doi.org/10.1093/ehjci/jez319.863. (impact factor: 6.20, Q1)
- Ciusdel, C., Turcea, A., Puiu, A., Itu, L., Calmac, L., Weiss, E., Margineanu, C., Badila, E., Berger, M., Redel, T., Passerini, T., Gulsun, M., & Sharma, P. (2020). Deep neural networks for ECG-free cardiac phase and end-diastolic frame detection on coronary angiographies. Computerized Medical Imaging and Graphics, 84, 101749. https://doi.org/10.1016/j.compmedimag.2020.101749. (impact factor: 5.70, Q1)
- Vizitiu, A., Nita, C. I., Puiu, A., Suciu, C., & Itu, L. M. (2020). Applying Deep Neural Networks over Homomorphic Encrypted Medical Data. Computational and Mathematical Methods in Medicine, 2020, 1–26. https://doi.org/10.1155/2020/3910250. (impact factor: 0.94, Q3)

- Nita, C.-I., Puiu, A., Bunescu, D., Mihai Itu, L., Mihalef, V., Chintalapani, G., Armstrong, A., Zampi, J., Benson, L., Sharma, P., & Rapaka, S. (2022). Personalized Pre- and Post-Operative Hemodynamic Assessment of Aortic Coarctation from 3D Rotational Angiography. Cardiovascular Engineering and Technology, 13(1), 14–40. https://doi.org/10.1007/s13239-021-00552-9. (impact factor: 1.80, Q3)

- Ploscaru, V., Popa-Fotea, N.-M., Calmac, L., Itu, L. M., Mihai, C., Bataila, V., Dragoescu, B., Puiu, A., Cojocaru, C., Costin, M. A., & Scafa-Udriste, A. (2022). Artificial intelligence and cloud based platform for fully automated PCI guidance from coronary angiography-study protocol. PLOS ONE, 17(9), e0274296. https://doi.org/10.1371/journal.pone.0274296. (impact factor: 3.70, Q2)

- Ogrezeanu, I., Vizitiu, A., Ciușdel, C., Puiu, A., Coman, S., Boldișor, C., Itu, A., Demeter, R., Moldoveanu, F., Suciu, C., & Itu, L. (2022). Privacy Preserving and Explainable AI in Industrial Applications. In Applied Sciences (Switzerland) (Vol. 12, Issue 13). MDPI. https://doi.org/10.3390/app12136395. (impact factor: 2.70, Q3)

- Hatfaludi, C. A., Tache, I. A., Ciușdel, C. F., Puiu, A., Stoian, D., Itu, L. M., Calmac, L., Popa-Fotea, N. M., Bataila, V., & Scafa-Udriste, A. (2022). Towards a Deep-Learning Approach for Prediction of Fractional Flow Reserve from Optical Coherence Tomography. Applied Sciences (Switzerland), 12(14). https://doi.org/10.3390/app12146964. (impact factor: 2.70, Q3)

- Tache, I. A., Hatfaludi, C. A., Puiu, A., Itu, L. M., Popa-Fotea, N. M., Calmac, L., & Scafa-Udriste, A. (2023). Assessment of the functional severity of coronary lesions from optical coherence tomography based on ensembled learning. BioMedical Engineering Online, 22(1). https://doi.org/10.1186/s12938-023-01192-x. (impact factor: 3.90, Q3)

- Scafa-Udriște, A., Itu, L., Puiu, A., Stoian, A., Moldovan, H., & Popa-Fotea, N.-M. (2023). In-stent restenosis in acute coronary syndrome—a classic and a machine learning approach. Frontiers in Cardiovascular Medicine, 10. https://doi.org/10.3389/fcvm.2023.1270986. (impact factor: 3.60, Q2)

One book chapter was published as co-author:

- Meister, F., Houle, H., Nita, C., Puiu, A., Itu, L. M., & Rapaka, S. (2020). Additional clinical applications. In Artificial Intelligence for Computational Modeling of the Heart (pp. 183–210). Elsevier. https://doi.org/10.1016/B978-0-12-817594-1.00017-6.

5.4 Discussion

While individual strengths and limitations were discussed for each use-case and approach presented within this thesis in the appropriate chapters, this section aims at emphasizing on the current work at a holistic level.

This PhD thesis exemplifies how several barriers in employing AI based solutions in solving various clinical needs could be surmounted by considering a set of paradigms such as self-supervised learning, synthetic data generation, feature importance estimation and model uncertainty quantification. All these approaches have been explored through a set of clinical use-cases related to different stages of cancer management, including diagnosis, characterization and treatment. However, the current work has several limitations as further described.

To begin with, although the introduction section provides a more comprehensive list of state of the art solutions to the identified challenges, some of them being concurrent to some extent in addressing certain topics, we would like to note that the scope of this PhD thesis was not to run a comparative study between those, but rather to empirically select the most suitable option for solving a specific problem. Therefore, one limitation of the current work stands in the lack of comparison

between multiple state of the art solutions applicable to some particular obstacle. While this represents a future direction of this work, usually in practice the specifics of problems addressed would give insights towards adoption of one methodology over the others. For example, choosing between semi-supervised and self-supervised learning paradigms would mainly depend upon the availability and quality of weakly labeled data, or the overall purpose of the model (i.e. contrasting a general and a problem specific desired behavior).

Secondly, although very promising, most of the results generated in this PhD program need further validation on large-scale representative testing datasets or real clinical scenarios. However, due to the unavailability of such datasets this step currently represents a future direction of this work.

Ultimately, with the extensive research being nowadays directed towards obtaining better and better artificial intelligence algorithms, a series of groundbreaking recent technologies were not explored during the PhD program. For example, the outstanding improvements recently obtained in the field of large language modeling could be leveraged in improving the realness of our prostate cancer synthetic data, or in overcoming modeling challenges owing to the inherent genomics heterogeneity, while their computer-vision derivatives could further improve the CT image extrapolation performance. Therefore, since all these novel algorithms hold many promises in further improving the overall performance, their adoption in the proposed methodologies remains to be explored in the future.

Overall, we strongly believe that the results obtained throughout this PhD program could have a great positive impact on the Healthcare transformation, bringing significant evidence that the major challenges and fears around the adoption of AI solutions in clinical routines could be efficiently addressed towards improving patients care, and ultimately their outcomes.

References

[1] World Health Organization. Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019. Tech. rep. 2020. URL: https://www.who.int/data/gho/data/themes/topics/indicator-groups/indicator-group-details/GHO/gho-ghe-global-health-estimates-life-tables (visited on 06/18/2021).

[2] Khanh Bao Tran et al. "The global burden of cancer attributable to risk factors, 2010–19: a systematic analysis for the Global Burden of Disease Study 2019". In: The Lancet 400 (10352 Aug. 2022), pp. 563–591. ISSN: 01406736. DOI: 10.1016/S0140-6736(22)01438-6. URL: https://linkinghub.elsevier.com/retrieve/pii/S0140673622014386.

[3] Robert E. Schoen et al. "Colorectal-Cancer Incidence and Mortality with Screening Flexible Sigmoidoscopy". In: New England Journal of Medicine 366 (25 June 2012), pp. 2345–2357. ISSN: 0028-4793. DOI: 10.1056/nejmoa1114635.

[4] Andrei Puiu et al. "Privacy-Preserving and Explainable AI for Cardiovascular Imaging". In: Studies in Informatics and Control 30 (2 June 2021), pp. 21–32. ISSN: 1841429X. DOI: 10.24846/v30i2y202102.

[5] R. J.M. Bruls and R. M. Kwee. "Workload for radiologists during on-call hours: dramatic increase in the past 15 years". In: Insights into Imaging 11 (1 Dec. 2020). How imaging studies increased and therefore clinicans workload.
. ISSN: 18694101. DOI: 10.1186/s13244-020-00925-z.

[6] Christine Dan Lantsman et al. "Trend in radiologist workload compared to number of admissions in the emergency department". In: European Journal of Radiology 149 (Apr. 2022), p. 110195. ISSN: 0720048X. DOI: 10.1016/j.ejrad.2022.110195.

[7] Tommaso Mansi, Tiziano Passerini, and Dorin Comaniciu, eds. Artificial intelligence for computational modeling of the heart. 1st ed. San Diego: Academic press in an imprint of Elsevier, 2019. ISBN: 978-0-12-817594-1.

[8] A. Fischer et al. "Deep Learning Based Automated Coronary Labeling For Structured Reporting Of Coronary CT Angiography In Accordance With SCCT Guidelines". en. In: Journal of Cardiovascular Computed Tomography 14.3 (July 2020), S21–S22. ISSN: 19345925. DOI: 10.1016/j.jcct.2020.06.019. URL: https://linkinghub.elsevier.com/retrieve/pii/S1934592520301921 (visited on 12/20/2021).

[9] Moritz Kasel-Seibert et al. "Assessment of PI-RADS v2 for the Detection of Prostate Cancer". In: European Journal of Radiology 85 (4 Apr. 2016), pp. 726–731. ISSN: 0720048X. DOI: 10.1016/j.ejrad.2016.01.011.

[10] Antonio C. Westphalen et al. "Variability of the Positive Predictive Value of PI-RADS for Prostate MRI across 26 Centers: Experience of the Society of Abdominal Radiology Prostate Cancer Disease-focused Panel". In: Radiology 296 (1 July 2020). Interuser variability in PI-RADS
, pp. 76–84. ISSN: 0033-8419. DOI: 10.1148/radiol.2020190646.

[11] David J. Winkel et al. "A Novel Deep Learning Based Computer-Aided Diagnosis System Improves the Accuracy and Efficiency of Radiologists in Reading Biparametric Magnetic Resonance Images of the Prostate: Results of a Multireader, Multicase Study". In: Investigative Radiology 56 (10 Oct. 2021), pp. 605–613. ISSN: 15360210. DOI: 10.1097/RLI.0000000000000780.

[12] Anamaria Vizitiu et al. "Applying Deep Neural Networks over Homomorphic Encrypted Medical Data". In: Computational and Mathematical Methods in Medicine 2020 (2020). ISSN: 17486718. DOI: 10.1155/2020/3910250.

[13] Shih Cheng Huang et al. "Self-supervised learning for medical image classification: a systematic review and implementation guidelines". In: npj Digital Medicine 6 (1 Dec. 2023). ISSN: 23986352. DOI: 10.1038/s41746-023-00811-0.

[14] Jakub Konečný et al. "Federated Learning: Strategies for Improving Communication Efficiency". In: (Oct. 2016). URL: http://arxiv.org/abs/1610.05492.

[15] Aviad Kipnis and Eliphaz Hibshoosh. "Efficient Methods for Practical Fully Homomorphic Symmetric-key Encrypton, Randomization and Verification". In: IACR Cryptol. ePrint Arch. 2012 (2012), p. 637. URL: https://api.semanticscholar.org/CorpusID:14250402.

[16] Ehsan Hesamifard, Hassan Takabi, and Mehdi Ghasemi. "CryptoDL: Deep Neural Networks over Encrypted Data". In: (Nov. 2017).

[17] Hervé Chabanne et al. Privacy-Preserving Classification on Deep Neural Network. Cryptology ePrint Archive, Paper 2017/035. https://eprint.iacr.org/2017/035. 2017. URL: https://eprint.iacr.org/2017/035.

[18] Charu C. Aggarwal and Philip S. Yu, eds. Privacy-Preserving Data Mining. Vol. 34. Springer US, 2008. ISBN: 978-0-387-70991-8. DOI: 10.1007/978-0-387-70992-5.

[19] Henry Surendra and S MohanH. "A Review Of Synthetic Data Generation Methods For Privacy Preserving Data Publishing". In: International Journal of Scientific & Technology Research 6 (2017), pp. 95–101. URL: https://api.semanticscholar.org/CorpusID:67051890.

[20] Aldren Gonzales, Guruprabha Guruswamy, and Scott R. Smith. "Synthetic data in health care: A narrative review". In: PLOS Digital Health 2 (1 Jan. 2023), e0000082. DOI: 10.1371/journal.pdig.0000082.

[21] Jacob Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019, pp. 4171–4186. URL: https://github.com/tensorflow/tensor2tensor.

[22] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. "Globally and Locally Consistent Image Completion". In: ACM Transactions on Graphics (Proc. of SIGGRAPH) 36.4 (2017), p. 107.

[23] Pengpeng Liu et al. Semantically Consistent Image Completion with Fine-grained Details. 2017. arXiv: 1711.09345 [cs.CV].

[24] Sebastian Bach et al. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation". In: PLOS ONE 10 (7 July 2015), e0130140. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0130140.

[25] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: (Feb. 2016).

[26] Avanti Shrikumar et al. "Not Just a Black Box: Learning Important Features Through Propagating Activation Differences". In: (May 2016).

[27] Scott Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: (May 2017). URL: http://arxiv.org/abs/1705.07874.

[28] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic Attribution for Deep Networks". In: (Mar. 2017).

[29] Florin C. Ghesu et al. "Quantifying and leveraging predictive uncertainty for medical image assessment". In: Medical Image Analysis 68 (2021), p. 101855. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2020.101855. URL: https://www.sciencedirect.com/science/article/pii/S136184152030219X.

[30] Costin Florian Ciușdel et al. "Normalizing Flows for Out-of-Distribution Detection: Application to Coronary Artery Segmentation". In: Applied Sciences 12.8 (2022). ISSN: 2076-3417. DOI: 10.3390/app12083839. URL: https://www.mdpi.com/2076-3417/12/8/3839.

[31] Natália Alves et al. "Prediction Variability to Identify Reduced AI Performance in Cancer Diagnosis at MRI and CT". In: Radiology 308 (3 Sept. 2023), e230275. ISSN: 15271315. DOI: 10.1148/radiol.230275.

[32] Joeran S. Bosma et al. "Semisupervised Learning with Report-guided Pseudo Labels for Deep Learning–based Prostate Cancer Detection Using Biparametric MRI". In: Radiology: Artificial Intelligence 5 (5 Sept. 2023). ISSN: 26386100. DOI: 10.1148/ryai.230031.

[33] Fabian Isensee et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: Nature Methods 18 (2 Feb. 2021), pp. 203–211. ISSN: 15487105. DOI: 10.1038/s41592-020-01008-z.

[34] Kamal M Ali and Michael J Pazzani. Error Reduction through Learning Multiple Descriptions. 1996, pp. 173–202.

[35] Janneke M T Hendriksen et al. "Clinical characteristics associated with diagnostic delay of pulmonary embolism in primary care: a retrospective observational study". In: BMJ Open 7 (3 Mar. 2017), e012789. ISSN: 2044-6055. DOI: 10.1136/bmjopen-2016-012789.

[36] Andrei Puiu et al. "Generative Adversarial CT Volume Extrapolation for Robust Small-to-Large Field of View Registration". In: Applied Sciences 12.6 (2022). ISSN: 2076-3417. DOI: 10.3390/app12062944. URL: https://www.mdpi.com/2076-3417/12/6/2944.

[37] Matthew A. Mauro et al. Mauro: image-guided interventions: expert radiology series. Third. Philadelphia: Elsevier, Inc, 2020. ISBN: 978-0-323-61204-3.

[38] Kevin Cleary and Terry M. Peters. "Image-Guided Interventions: Technology Review and Clinical Applications". In: Annual Review of Biomedical Engineering 12.1 (2010). PMID: 20415592, pp. 119–142. DOI: 10.1146/annurev-bioeng-070909-105249. eprint: https://doi.org/10.1146/annurev-bioeng-070909-105249. URL: https://doi.org/10.1146/annurev-bioeng-070909-105249.

[39] J. Modersitzki. Numerical Methods for Image Registration. Jan. 2004, pp. 27–44.

[40] R. Liao et al. "A Review of Recent Advances in Registration Techniques Applied to Minimally Invasive Therapy". In: IEEE Transactions on Multimedia 15.5 (2013), pp. 983–1000. ISSN: 1941-0077. DOI: 10.1109/TMM.2013.2244869.

[41] Barbara Zitová and Jan Flusser. "Image Registration Methods: A Survey". In: Image and Vision Computing 21 (Oct. 2003), pp. 977–1000. DOI: 10.1016/S0262-8856(03)00137-9.

[42] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever. "Mutual-information-based registration of medical images: a survey". In: IEEE Transactions on Medical Imaging 22.8 (2003), pp. 986–1004. ISSN: 1558-254X. DOI: 10.1109/TMI.2003.815867.

[43] Ian J. Goodfellow et al. Generative Adversarial Networks. 2014. arXiv: 1406.2661 [stat.ML].

[44] Tero Karras et al. Progressive Growing of GANs for Improved Quality, Stability, and Variation. 2017. arXiv: 1710.10196 [cs.NE].

[45] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. 2018. arXiv: 1812.04948 [cs.NE].

[46] Taesung Park et al. “Semantic Image Synthesis with Spatially-Adaptive Normalization”. In: arXiv:1903.07291 [cs] (Nov. 2019). arXiv: 1903.07291. URL: http://arxiv.org/abs/1903.07291 (visited on 12/20/2021).

[47] Yue Zhang et al. Task Driven Generative Modeling for Unsupervised Domain Adaptation: Application to X-ray Image Segmentation. 2018. arXiv: 1806.07201 [cs.CV].

[48] Chenyu You et al. “CT Super-resolution GAN Constrained by the Identical, Residual, and Cycle Learning Ensemble (GAN-CIRCLE)”. In: IEEE Transactions on Medical Imaging (2019), pp. 1–1. ISSN: 1558-254X. DOI: 10.1109/tmi.2019.2922960. URL: http://dx.doi.org/10.1109/TMI.2019.2922960.

[49] Q. Yang et al. “Low-Dose CT Image Denoising Using a Generative Adversarial Network With Wasserstein Distance and Perceptual Loss”. In: IEEE Transactions on Medical Imaging 37.6 (2018), pp. 1348–1357. ISSN: 1558-254X. DOI: 10.1109/TMI.2018.2827462.

[50] A. Vizitiu et al. “Data-Driven Adversarial Learning for Sinogram-Based Iterative Low-Dose CT Image Reconstruction”. In: 2019 23rd International Conference on System Theory, Control and Computing (ICSTCC). 2019, pp. 668–674. DOI: 10.1109/ICSTCC.2019.8885947.

[51] Zhu Jun-Yan et al. “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks”. In: Computer Vision (ICCV), 2017 IEEE International Conference on. 2017.

[52] Armanious Karim et al. “MedGAN: Medical Image Translation using GANs”. In: Computerized Medical Imaging and Graphics (2019), p. 101684. ISSN: 0895-6111. DOI: 10.1016/j.compmedimag.2019.101684. URL: http://dx.doi.org/10.1016/j.compmedimag.2019.101684.

[53] Chao Yang et al. High-Resolution Image Inpainting using Multi-Scale Neural Patch Synthesis. 2016. arXiv: 1611.09969 [cs.CV].

[54] Mark Sabini and Gili Rusak. Painting Outside the Box: Image Outpainting with GANs. 2018. arXiv: 1808.08483 [cs.CV].

[55] Yi Wang et al. “Wide-Context Semantic Image Extrapolation”. In: June 2019.

[56] Julius Surya Sumantri and In Kyu Park. 360 Panorama Synthesis from a Sparse Set of Images with Unknown Field of View. 2019. arXiv: 1904.03326 [cs.CV].

[57] Donald Ervin Knuth. The art of computer programming. 3rd ed. Reading, Mass: Addison-Wesley, 1997, p. 232. ISBN: 978-0-201-89683-1 978-0-201-89684-8 978-0-201-89685-5.

[58] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization. 2016. arXiv: 1607.08022 [cs.CV].

[59] Phillip Isola et al. Image-to-Image Translation with Conditional Adversarial Networks. 2016. arXiv: 1611.07004 [cs.CV].

[60] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. 2017. arXiv: 1701.07875 [stat.ML].

[61] Ishaan Gulrajani et al. “Improved Training of Wasserstein GANs”. In: CoRR abs/1704.00028 (2017). arXiv: 1704.00028. URL: http://arxiv.org/abs/1704.00028.

[62] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. 2016. arXiv: 1603.08155 [cs.CV].

[63] Fabian Isensee et al. “Brain Tumor Segmentation and Radiomics Survival Prediction: Contribution to the BRATS 2017 Challenge”. In: ArXiv abs/1802.10508 (2017).

[64] Florin C. Ghesu et al. “An Artificial Agent for Anatomical Landmark Detection in Medical Images”. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016. Ed. by Sebastien Ourselin et al. Cham: Springer International Publishing, 2016, pp. 229–237. ISBN: 978-3-319-46726-9.

[65] Florin-Cristian Ghesu et al. "Multi-Scale Deep Reinforcement Learning for Real-Time 3D-Landmark Detection in CT Scans". In: IEEE Transactions on Pattern Analysis and Machine Intelligence 41.1 (Jan. 2019), pp. 176–189. ISSN: 0162-8828, 2160-9292, 1939-3539. DOI: 10.1109/TPAMI.2017.2782687. URL: https://ieeexplore.ieee.org/document/8187667/ (visited on 12/20/2021).

[66] Yasmina Chenoune et al. "Rigid registration of Delayed-Enhancement and Cine Cardiac MR images using 3D Normalized Mutual Information". In: vol. 37. Oct. 2010, pp. 161 –164.

[67] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: (Oct. 2020).

[68] M. I. Jordan and T. M. Mitchell. "Machine learning: Trends, perspectives, and prospects". In: Science 349 (6245 July 2015), pp. 255–260. ISSN: 0036-8075. DOI: 10.1126/science.aaa8415.

[69] Jason Walonoski et al. "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record". In: Journal of the American Medical Informatics Association 25 (3 Mar. 2018), pp. 230–238. ISSN: 1527974X. DOI: 10.1093/jamia/ocx079.

[70] Jean-Francois Rajotte et al. "iScience Synthetic data as an enabler for machine learning applications in medicine". In: (). DOI: 10.1016/j.isci. URL: https://doi.org/10.1016/j.isci..

[71] Bogdan A. Gheorghiță et al. "Improving robustness of automatic cardiac function quantification from cine magnetic resonance imaging using synthetic image data". In: Scientific Reports 12 (1 Dec. 2022). ISSN: 20452322. DOI: 10.1038/s41598-022-06315-3.

[72] Shu Hui Hsu et al. "Synthetic CT generation for MRI-guided adaptive radiotherapy in prostate cancer". In: Frontiers in Oncology 12 (Sept. 2022). ISSN: 2234943X. DOI: 10.3389/fonc.2022.969463.

[73] Maram Mahmoud A. Monshi, Josiah Poon, and Vera Chung. "Deep learning in generating radiology reports: A survey". In: Artificial Intelligence in Medicine 106 (2020), p. 101878. ISSN: 0933-3657. DOI: https://doi.org/10.1016/j.artmed.2020.101878. URL: https://www.sciencedirect.com/science/article/pii/S0933365719302635.

[74] Yuxiang Liao, Hantao Liu, and Irena Spasić. "Deep learning approaches to automatic radiology report generation: A systematic review". In: Informatics in Medicine Unlocked 39 (2023), p. 101273. ISSN: 2352-9148. DOI: https://doi.org/10.1016/j.imu.2023.101273. URL: https://www.sciencedirect.com/science/article/pii/S235291482300117X.

[75] Edward Choi et al. "Generating Multi-label Discrete Patient Records using Generative Adversarial Networks". In: (Mar. 2017). URL: http://arxiv.org/abs/1703.06490.

[76] Omid Sayadi, Mohammad B. Shamsollahi, and Gari D. Clifford. "Synthetic ECG generation and Bayesian filtering using a Gaussian wave-based dynamical model". In: Physiological Measurement 31 (10 2010), pp. 1309–1329. ISSN: 13616579. DOI: 10.1088/0967-3334/31/10/002.

[77] Edmond Adib, Fatemeh Afghah, and John J. Prevost. "Synthetic ECG Signal Generation Using Generative Neural Networks". In: (Dec. 2021). URL: http://arxiv.org/abs/2112.03268.

[78] Burak Yelmen et al. "Creating artificial human genomes using generative neural networks". In: PLoS Genetics 17 (2 Feb. 2021). ISSN: 15537404. DOI: 10.1371/JOURNAL.PGEN.1009303.

[79] Sahel Shariati Samani et al. "Quantifying genomic privacy via inference attack with high-order SNV correlations". In: Institute of Electrical and Electronics Engineers Inc., July 2015, pp. 32–40. ISBN: 9781479999330. DOI: 10.1109/SPW.2015.21.

[80] Bristena Oprisanu, Georgi Ganev, and Emiliano De Cristofaro. "On Utility and Privacy in Synthetic Genomic Data". In: (Feb. 2021). URL: http://arxiv.org/abs/2102.03314.

[81] Karim Armanious et al. "MedGAN: Medical Image Translation using GANs". In: (June 2018). DOI: 10.1016/j.compmedimag.2019.101684. URL: http://arxiv.org/abs/1806.06397http://dx.doi.org/10.1016/j.compmedimag.2019.101684.

[82] Jun-Yan Zhu et al. "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks". In: (Mar. 2017). URL: http://arxiv.org/abs/1703.10593.

[83] Rebecca L. Siegel et al. "Cancer Statistics, 2021". In: CA: A Cancer Journal for Clinicians 71 (1 Jan. 2021), pp. 7–33. ISSN: 0007-9235. DOI: 10.3322/caac.21654.

[84] Harold Evelyn Taitt. Global Trends and Prostate Cancer: A Review of Incidence, Detection, and Mortality as Influenced by Race, Ethnicity, and Geographic Location. Nov. 2018. DOI: 10.1177/1557988318798279.

[85] Mary Beth B. Culp et al. "Recent Global Patterns in Prostate Cancer Incidence and Mortality Rates". In: European Urology 77 (1 Jan. 2020), pp. 38–52. ISSN: 0302-2838. DOI: 10.1016/J.EURURO.2019.08.005.

[86] Stephen B. Edge and American Joint Committee on Cancer. AJCC cancer staging manual, p. 650. ISBN: 9780387884400.

[87] James D Brierley, Mary K Gospodarowicz, and Christian Wittekind. TNM classification of malignant tumours. John Wiley & Sons, 2017. ISBN: 9781119263548.

[88] Jeremy L. Warner, Mia A. Levy, and Michael N. Neuss. "Feasibility and accuracy of extracting cancer stage information from narrative electronic health record data". In: Journal of Oncology Practice 12 (2 Feb. 2016), e169–e179. ISSN: 1935469X. DOI: 10.1200/JOP.2015.004622.

[89] Georgina Cosma et al. "Prediction of pathological stage in patients with prostate cancer: A neuro-fuzzy model". In: PLoS ONE 11 (6 June 2016). ISSN: 19326203. DOI: 10.1371/journal.pone.0155856.

[90] Stacy Loeb et al. "Distribution of PSA velocity by total PSA levels: Data from the Baltimore Longitudinal study of aging". In: Urology 77 (1 Jan. 2011), pp. 143–147. ISSN: 00904295. DOI: 10.1016/j.urology.2010.04.068.

[91] Thiago N. Valette et al. "Probability of extraprostatic disease according to the percentage of positive biopsy cores in clinically localized prostate cancer". In: International Braz J Urol 41 (3 2015), pp. 449–454. ISSN: 16776119. DOI: 10.1590/S1677-5538.IBJU.2014.0223.

[92] Jure Murgic et al. "The role of the maximum involvement of biopsy core in predicting outcome for patients treated with dose-escalated radiation therapy for prostate cancer". In: Radiation Oncology 7 (1 Aug. 2012). ISSN: 1748717X. DOI: 10.1186/1748-717X-7-127.

[93] John B. Eifler et al. "An updated prostate cancer staging nomogram (Partin tables) based on cases from 2006 to 2011". In: BJU International 111 (1 2013), pp. 22–29. ISSN: 1464410X. DOI: 10.1111/j.1464-410X.2012.11324.x.

[94] Hirohiko Kamiyama et al. Unusual False-Positive Mesenteric Lymph Nodes Detected by PET/CT in a Metastatic Survey of Lung Cancer. 2016. DOI: 10.1159/000446579.

[95] Mingxia Feng et al. "Retrospective analysis for the false positive diagnosis of PET-CT scan in lung cancer patients". In: Medicine (United States) 96 (42 Oct. 2017). ISSN: 15365964. DOI: 10.1097/MD.0000000000007415.

[96] James L. Mohler et al. "Prostate cancer, version 2.2019". In: JNCCN Journal of the National Comprehensive Cancer Network 17 (5 2019), pp. 479–505. ISSN: 15401413. DOI: 10.6004/jnccn.2019.0023.

[97] Aaron C. Spalding et al. "Percent Positive Biopsy Cores as a Prognostic Factor for Prostate Cancer Treated with External Beam Radiation". In: Urology 69.5 (2007), pp. 936–940. ISSN: 0090-4295. DOI: https://doi.org/10.1016/j.urology.2007.01.066. URL: https://www.sciencedirect.com/science/article/pii/S009042950700132X.

[98] Ji Won Seo et al. "PI-RADS version 2: Detection of clinically significant cancer in patients with biopsy gleason score 6 prostate cancer". In: American Journal of Roentgenology 209 (1 July 2017), W1–W9. ISSN: 15463141. DOI: 10.2214/AJR.16.16981.

[99] Kouji Izumi et al. "The Relationship Between Prostate-Specific Antigen and TNM Classification or Gleason Score in Prostate Cancer Patients With Low Prostate-Specific Antigen Levels". In: The Prostate 75 (Mar. 2015). DOI: 10.1002/pros.22985.

[100] Daniel Jones et al. The diagnostic test accuracy of rectal examination for prostate cancer diagnosis in symptomatic patients: A systematic review. June 2018. DOI: 10.1186/s12875-018-0765-y.

[101] Anthony V D et al. Preoperative PSA Velocity and the Risk of Death from Prostate Cancer after Radical Prostatectomy. 2004. URL: www.nejm.org.

[102] Stacy Loeb et al. "PSA Doubling Time Versus PSA Velocity to Predict High-Risk Prostate Cancer: Data from the Baltimore Longitudinal Study of Aging". In: European Urology 54 (5 Nov. 2008), pp. 1073–1080. ISSN: 03022838. DOI: 10.1016/j.eururo.2008.06.076.

[103] David E. Neal et al. "Ten-year Mortality, Disease Progression, and Treatment-related Side Effects in Men with Localised Prostate Cancer from the ProtecT Randomised Controlled Trial According to Treatment Received". In: European Urology 77.3 (2020), pp. 320–330. ISSN: 0302-2838. DOI: https://doi.org/10.1016/j.eururo.2019.10.030. URL: https://www.sciencedirect.com/science/article/pii/S0302283819308371.

[104] Jenny L. Donovan et al. "Patient-Reported Outcomes after Monitoring, Surgery, or Radiotherapy for Prostate Cancer". In: New England Journal of Medicine 375 (15 Oct. 2016), pp. 1425–1437. ISSN: 0028-4793. DOI: 10.1056/nejmoa1606221.

[105] Anthony V. D'Amico et al. "Biochemical Outcome After Radical Prostatectomy, External Beam Radiation Therapy, or Interstitial Radiation Therapy for Clinically Localized Prostate Cancer". In: JAMA 280.11 (Sept. 1998), pp. 969–974. DOI: 10.1001/jama.280.11.969. eprint: https://jamanetwork.com/journals/jama/articlepdf/187980/joc80111.pdf. URL: https://doi.org/10.1001/jama.280.11.969.

[106] Prasanna Sooriakumaran et al. "Comparative effectiveness of radical prostatectomy and radiotherapy in prostate cancer: Observational study of mortality outcomes". In: BMJ (Online) 348 (Feb. 2014). ISSN: 17561833. DOI: 10.1136/bmj.g1502.

[107] John K. Gohagan et al. "Prostate Cancer Screening in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial of the National Cancer Institute". In: The Journal of Urology 152.5, Part 2 (1994), pp. 1905–1909. ISSN: 0022-5347. DOI: https://doi.org/10.1016/S0022-5347(17)32412-6. URL: https://www.sciencedirect.com/science/article/pii/S0022534717324126.

[108] Tim Hulsen. "An overview of publicly available patient-centered prostate cancer datasets". In: Translational Andrology and Urology 8 (S1 Mar. 2019), S64–S77. ISSN: 22234683. DOI: 10.21037/tau.2019.03.01.

[109] Rupam Deori, Bijoyananda Das, and Mustafa Abdur Rahman. "A Study of Relationship of Prostate Volume, Prostate Specific Antigen and age in Benign Prostatic Hyperplasia". In: International Journal of Contemporary Medical Research 4 (2017). ISSN: 2454-7379. URL: www.ijcmr.com.

[110] Emily Alsentzer et al. Publicly Available Clinical BERT Embeddings. 2019, pp. 72–78. URL: https://www.ncbi.nlm.nih.gov/pmc/.

[111] David J. Winkel et al. "Autonomous detection and classification of pi-rads lesions in an mri screening population incorporating multicenter-labeled deep learning and biparametric imaging: Proof of concept". In: Diagnostics 10 (11 Nov. 2020). ISSN: 20754418. DOI: 10.3390/diagnostics10110951.

[112] Pritesh Mehta et al. "Computer-aided diagnosis of prostate cancer using multiparametric MRI and clinical features: A patient-level classification framework". In: Medical Image Analysis 73 (2021), p. 102153. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2021.102153. URL: https://www.sciencedirect.com/science/article/pii/S1361841521001997.

[113] Hashim U. Ahmed et al. "Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study". In: The Lancet 389 (10071 Feb. 2017), pp. 815–822. ISSN: 1474547X. DOI: 10.1016/S0140-6736(16)32401-1.

[114] Baris Turkbey et al. "Prostate Imaging Reporting and Data System Version 2.1: 2019 Update of Prostate Imaging Reporting and Data System Version 2". In: European Urology 76.3 (2019), pp. 340–351. ISSN: 0302-2838. DOI: https://doi.org/10.1016/j.eururo.2019.02.033. URL: https://www.sciencedirect.com/science/article/pii/S0302283819301800.

[115] Shijun Wang et al. Computer aided-diagnosis of prostate cancer on multiparametric MRI: A technical review of current research. 2014. DOI: 10.1155/2014/789561.

[116] Kimia Kohestani et al. "Performance and inter-observer variability of prostate MRI (PI-RADS version 2) outside high-volume centres". In: Scandinavian Journal of Urology 53 (5 Sept. 2019), pp. 304–311. ISSN: 21681813. DOI: 10.1080/21681805.2019.1675757.

[117] Xin Yu et al. "False Positive Reduction Using Multiscale Contextual Features for Prostate Cancer Detection in Multi-Parametric MRI Scans". In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). 2020, pp. 1355–1359. DOI: 10.1109/ISBI45749.2020.9098338.

[118] Satoshi Washino et al. "Combination of prostate imaging reporting and data system (PI-RADS) score and prostate-specific antigen (PSA) density predicts biopsy outcome in prostate biopsy naïve patients". In: BJU International 119 (2 Feb. 2017), pp. 225–233. ISSN: 1464410X. DOI: 10.1111/bju.13465.

[119] Tobias Nordström et al. "Prostate-specific antigen (PSA) density in the diagnostic algorithm of prostate cancer". In: Prostate Cancer and Prostatic Diseases 21.1 (2018), pp. 57–63. ISSN: 1476-5608. DOI: 10.1038/s41391-017-0024-7. URL: https://doi.org/10.1038/s41391-017-0024-7.

[120] Yuan Fei Lu et al. "Optimizing prostate cancer accumulating model: Combined PI-RADS v2 with prostate specific antigen and its derivative data". In: Cancer Imaging 19 (1 May 2019). ISSN: 14707330. DOI: 10.1186/s40644-019-0208-6.

[121] Dong Yang et al. "Automatic Liver Segmentation Using an Adversarial Image-to-Image Network". In: (July 2017). URL: http://arxiv.org/abs/1707.08037.

[122] Andrei Puiu et al. "Prediction uncertainty estimates elucidate the limitation of current NSCLC subtype classification in representing mutational heterogeneity". In: Scientific Reports 14 (1 Mar. 2024), p. 6779. ISSN: 2045-2322. DOI: 10.1038/s41598-024-57057-3. URL: https://www.nature.com/articles/s41598-024-57057-3.

[123] Ayal B Gussow, Eugene V Koonin, and Noam Auslander. "Identification of combinations of somatic mutations that predict cancer survival and immunotherapy benefit". en. In: NAR Cancer 3.2 (May 2021), zcab017.

[124] Junyu Long et al. "A mutation-based gene set predicts survival benefit after immunotherapy across multiple cancers and reveals the immune response landscape". en. In: Genome Med 14.1 (Feb. 2022), p. 20.

[125] Robert Clarke et al. "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data". en. In: Nat Rev Cancer 8.1 (Jan. 2008), pp. 37–49.

[126] Mohan Babu and Michael Snyder. "Multi-omics profiling for health". en. In: Mol Cell Proteomics 22.6 (Apr. 2023), p. 100561.

[127] Kathryn A Phillips et al. “Availability and funding of clinical genomic sequencing globally”. en. In: BMJ Glob Health 6.2 (Feb. 2021).

[128] A Bayle et al. “ESMO study on the availability and accessibility of biomolecular technologies in oncology in Europe”. en. In: Ann Oncol 34.10 (July 2023), pp. 934–945.

[129] Gemma L D’Adamo, James T Widdop, and Edward M Giles. “The future is now? Clinical and translational aspects of “omics” technologies”. en. In: Immunol Cell Biol 99.2 (Oct. 2020), pp. 168–176.

[130] Valeria Relli et al. “Abandoning the notion of non-small cell lung cancer”. en. In: Trends Mol Med 25.7 (May 2019), pp. 585–594.

[131] Katherine A Hoadley et al. “Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer”. en. In: Cell 173.2 (Apr. 2018), 291–304.e6.

[132] Valeria Relli et al. “Distinct lung cancer subtypes associate to distinct drivers of tumor progression”. en. In: Oncotarget 9.85 (Oct. 2018), pp. 35528–35540.

[133] Shih-Hsin Hsiao et al. “Comparative survival analysis of platinum-based adjuvant chemotherapy for early-stage squamous cell carcinoma and adenocarcinoma of the lung”. en. In: Cancer Med 11.10 (Mar. 2022), pp. 2067–2078.

[134] Giorgio Scagliotti et al. “Treatment-by-histology interaction analyses in three phase III trials show superiority of pemetrexed in nonsquamous non-small cell lung cancer”. en. In: J Thorac Oncol 6.1 (Jan. 2011), pp. 64–70.

[135] Navneet Singh et al. “Therapy for stage IV non-small-cell lung cancer with driver alterations: ASCO living guideline”. en. In: J Clin Oncol 40.28 (July 2022), pp. 3310–3322.

[136] Navneet Singh et al. “Therapy for stage IV non-small-cell lung cancer without driver alterations: ASCO living guideline”. en. In: J Clin Oncol 40.28 (July 2022), pp. 3323–3343.

[137] Caicun Zhou et al. “Interim survival analysis of the randomized phase III GEMSTONE-302 trial: sugemalimab or placebo plus chemotherapy as first-line treatment for metastatic NSCLC”. en. In: Nat Cancer 4.6 (June 2023), pp. 860–871.

[138] Neil M Woody et al. “A histologic basis for the efficacy of SBRT to the lung”. en. In: J Thorac Oncol 12.3 (Dec. 2016), pp. 510–519.

[139] Nozomi Kita et al. “Comparison of recurrence patterns between adenocarcinoma and squamous cell carcinoma after stereotactic body radiotherapy for early-stage lung cancer”. en. In: Cancers (Basel) 15.3 (Jan. 2023).

[140] Jang Hyun Cho and Bharath Hariharan. “On the efficacy of knowledge distillation”. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019, pp. 4793–4801.

[141] Ferdinandos Skoulidis and John V Heymach. “Co-occurring genomic alterations in non-small-cell lung cancer biology and therapy”. en. In: Nat Rev Cancer 19.9 (Aug. 2019), pp. 495–509.

[142] Fan Zhang et al. “Co-occurring genomic alterations and immunotherapy efficacy in NSCLC”. en. In: NPJ Precis Oncol 6.1 (Jan. 2022), p. 4.

[143] Xiang Ge Luo, Jack Kuipers, and Niko Beerenwinkel. “Joint inference of exclusivity patterns and recurrent trajectories from tumor mutation trees”. en. In: Nat Commun 14.1 (June 2023), p. 3676.

[144] Catherine Labbé et al. “Prognostic and predictive effects of TP53 co-mutation in patients with EGFR-mutated non-small cell lung cancer (NSCLC)”. en. In: Lung Cancer 111 (June 2017), pp. 23–29.

[145] Zhong-Yi Dong et al. “Potential Predictive Value of TP53 and KRAS Mutation Status for Response to PD-1 Blockade Immunotherapy in Lung Adenocarcinoma”. en. In: Clin Cancer Res 23.12 (Dec. 2016), pp. 3012–3024.

[146] William McLaren et al. “The Ensembl variant effect predictor”. en. In: Genome Biol 17.1 (June 2016), p. 122.

[147] Lei Zhang et al. “AutoGGN: A gene graph network AutoML tool for multi-omics research”. In: Artificial Intelligence in the Life Sciences 1 (Dec. 2021), p. 100019.

[148] Haitham A Elmarakeby et al. “Biologically informed deep neural network for prostate cancer discovery”. en. In: Nature 598.7880 (Sept. 2021), pp. 348–352.

[149] Marylyn D Ritchie et al. “Methods of integrating data to uncover genotype-phenotype interactions”. en. In: Nat Rev Genet 16.2 (Jan. 2015), pp. 85–97.

[150] Ethan Cerami et al. “The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data”. en. In: Cancer Discov 2.5 (May 2012), pp. 401–404.

[151] John G Tate et al. “COSMIC: the Catalogue Of Somatic Mutations In Cancer”. en. In: Nucleic Acids Res 47.D1 (Jan. 2019), pp. D941–D947.

[152] Pavlos S. Efraimidis and Paul G. Spirakis. “Weighted random sampling with a reservoir”. In: Information Processing Letters 97.5 (2006), pp. 181–185.

[153] Matthew H Bailey et al. “Comprehensive Characterization of Cancer Driver Genes and Mutations”. en. In: Cell 173.2 (Apr. 2018), 371–385.e18.

Abstract

In the context of Artificial Intelligence (AI) that is nowadays driving the healthcare transformation, this PhD thesis delves into the specific challenges practitioners face in designing cutting edge technology to improve current clinical practices. Focusing on various medical scenarios related to one of the most devastating diseases of our time, cancer, this work explores several strategies to overcoming these challenges with the ultimate goal of bridging current gaps, and thus, paving the road of Deep Learning (DL) based solutions towards large scale adoption. Therefore, the aim of this PhD thesis is to unblock the full potential of AI in streamlining clinical routines, and ultimately, in improving patient care.

Organized into comprehensive sections, this PhD thesis features the tremendous potential AI has in supporting healthcare transformation, highlights typical challenges that nowadays obstructs its journey in addressing various clinical needs and ultimately identifies several strategies to overcoming these obstacles.

With data scarcity being a common problem in developing data-driven solutions for healthcare, statistically equivalent synthetic counterparts were created and further leveraged to enable innovative approaches for solving various clinical needs. Conversely, when the inherent complexity of medical domain prevents the accumulation of large scale annotated datasets to drive Machine Learning (ML) algorithms, training paradigms such as self-supervised learning exhibit a great potential in enabling data-driven developments, and hence, in crafting the next generation of healthcare. Nevertheless, integration of AI in clinical routines must be rigorously accomplished with emphasis on trustworthiness aspects. Exhibiting a direct impact on patients well being, achieving a safe functionality in practice is a crucial requirement of any medical software. Therefore, this work considered recent advancements in trustworthy AI to enhance transparency and reliability of DL models.

To demonstrate the overall feasibility of these approaches, a series of practical scenarios around innovative cancer care were pursued. In particular, this thesis demonstrates AI's impact on various disease management stages, including diagnosis of clinically significant prostate lesions, stratification of non-small cell lung and prostate cancer for therapy planning, and image extrapolation for guiding treatment delivery.

Finally, all results and insights derived from the activities performed during this PhD program are highlighted to offer a complete landscape of the thesis. Overall, our findings provide compelling evidence that the primary concerns and uncertainties regarding the integration of AI solutions into clinical practices can be effectively managed, leading to enhanced patient care and ultimately improved outcomes.