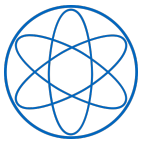
HELMHOLTZ MUNICH

Deep learning based analysis of medical imaging data in oncology with a focus on radiation therapy

Daniel M. Lang

Dissertation

Technical University of Munich and Helmholtz Center Munich



This thesis was created under the supervision of Prof. Dr. Jan J. Wilkens and Dr. Stefan Bartzsch.

Technische Universität München
TUM School of Natural Sciences

Deep learning based analysis of medical imaging data in oncology with a focus on radiation therapy

Daniel M. Lang

Vollständiger Abdruck der von der TUM School of Natural Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
genehmigten Dissertation.

Vorsitz: Prof. Dr. Björn Garbrecht

Prüfer*innen der Dissertation:

1. Prof. Dr. Jan J. Wilkens
2. Prof. Dr. Franz Pfeiffer

Die Dissertation wurde am 09.11.2022 bei der Technischen Universität München eingereicht und durch die TUM School of Natural Sciences am 28.06.2023 angenommen.

Abstract

Medical imaging data depicts a main information source for assessment of tumor characteristics in clinical oncology. Despite advances in image acquisition techniques, allowing for delineation of more and more detailed structures, clinical evaluation of imaging features is performed based on simplistic scores, like largest tumor diameter. Deep learning models have proven their power for image analysis in a variety of settings, and also the medical domain is shifting towards an application of the technique. However, existing models often rely on architectures engineered to be used on natural images, not taking into account the features specific to medical imaging data, or are limited to the area of image segmentation. Different deep learning based classification models customized to the medical problem settings at hand were developed in this thesis.

First, a 3D convolutional neural network was trained on computed tomography (CT) and clinical data for classification of risk scores in renal mass patients. Element-wise fusion of features extracted from both kidneys, utilizing a Siamese network, was established. The model was able to differentiate clinically relevant groups with an area under the receiver operating characteristics curve (AUC) of 0.814 on the test set.

Next, CT based detection of oropharynx cancer cases caused by an infection with a human papilloma virus (HPV) was analyzed. A transfer learning approach relying on sports video clips, featuring two spatial and one temporal dimensions, was utilized to be able to handle 3D medical imaging data in the downstream task. On the external test set, the model was able to discriminate between HPV positive and HPV negative cases with an AUC of 0.814. For in-domain pretraining, a masked autoencoder, utilizing modern transformer layers, was customized, and the dataset was modified to include unlabeled cases. The self-supervised model achieved an AUC of 0.723, while the transfer learning approach was able to distinguish cases with an AUC of 0.710 on the modified test set.

Time dependent endpoints, like overall survival, are of essential importance in oncology. Therefore, ability of deep learning architectures for modeling of survival data was studied. However, handling of time dependent data requires special architectures. For prediction of progression free survival in head and neck cancer patients, a discrete time survival model in conjunction with the video clip based transfer learning approach developed before was established. The model was trained on positron emission tomography (PET)/CT images and

clinical data and was able to asses patients' risk for cancer progression with a concordance index of 0.668.

Enabling reproducible research, all studies were either preformed on open datasets or developed in the setting of publicly held biomedical imaging challenges. Further studies on larger patient cohorts are needed. In general, medical imaging datasets contain a relatively small number of cases. Transfer and self-supervised models developed in this thesis have the power to be trained as representation learning approaches on larger datasets, to be then finetuned on the specific task at hand. Survival models could be used to establish novel scores, that feature the power to be utilized for decision making in therapy and spare patients from overdosage.

Zusammenfassung

Medizinische Bilddaten stellen eine wichtige Informationsquelle in der Onkologie dar. Essentielle Tumorcharakteristika können durch bildgebende Verfahren, wie etwa der Computertomographie (CT), erfasst werden um in der Auswahl des Therapieverfahrens Berücksichtigung zu finden. Während jedoch technische Weiterentwicklungen in der Bildaufnahme ein Auflösen immer kleinerer Details ermöglicht erfolgt die Bildauswertung basierend auf primitiven Kenngrößen, wie etwa des größten Tumordurchmessers. Algorithmen des tiefen Lernens (engl. deep learning) konnte in einer Vielzahl von Studien ihre Leistungsfähigkeit im Bereich der Bildauswertung nachweisen. Auch im medizinischen Bereich findet das Verfahren eine zunehmende Verbreitung. Modelle beruhen jedoch zumeist auf Techniken entwickelt zur Anwendung auf herkömmlichen Bilddaten, ohne dabei auf die speziellen Eigenschaften der medizinischen Bilder einzugehen, oder beschränken sich auf den Bereich der Bildsegmentierung. In der vorliegende Arbeit wurden verschiedene Deep Learning Algorithmen zur Analyse von Tumorcharakteristika entwickelt, wobei speziell auf die Gegebenheiten medizinischer Bilddaten eingegangen wurde.

In einem ersten Schritt wurde hierzu ein faltendes neuronales Netzwerk (engl. convolutional neural network) zur Risikoabschätzung in Nierentumoren anhand von CT Bildern und klinischen Daten entwickelt. Durch ein siamesisches Netzwerk wurden Muster beider Nieren extrahiert und anschließend elementweise zusammengefügt. Das Modell konnte klinisch relevante Subgruppen im Testdatensatz mit eine Fläche unter der Grenzwertoptimierungskurve (AUC) von 0.814 unterscheiden.

Im Anschluss wurde ein CT-basiertes Modell zu Detektion einer Infektion mit Humanen Papillomviren (HPV) in Oropharynxkarzinompatienten erstellt. Um der Herausforderung kleiner Datensätze zu begegnen, wurde ein Transfer-Lernansatz beruhend auf Sportvideos in der Vorlernphase erstellt. Videodaten besitzen eine dreidimensionale Struktur, mit zwei räumlichen und einer zeitlichen Dimension, dies ermöglicht ein Weiterlernen auf den 3D CT Bilddaten. Das so entwickelte Netzwerk war fähig HPV positive von HPV negativen Tumore mit einer AUC von 0.814 zu unterscheiden. Zur Entwicklung einer selbstlernenden Methode, die ein Vortrainieren in der selben Domäne ermöglicht, wurde ein Maskierender-Autokodierer an die Daten angepasst, dabei wurden moderne Transformierschichten verwendet. Der Datensatz wurde modifiziert und

um Fälle für die kein HPV Testergebnis vorlag erweitert. Der Maskierende-Autokodierer Ansatz erreichte dabei eine AUC von 0.723, der zuvor entwickelte Transfer-Lernansatz erzielte eine AUC von 0.710 auf dem erweiterten Datensatz.

Zeitabhängige klinische Endpunkte, wie etwa das Gesamtüberleben, spielen in der Onkologie eine wichtige Rolle. Die Fähigkeit von Deep Learning Netzwerken zur Modellierung von temporären Daten wurde untersucht. Dies erfordert die Verwendung spezieller Architekturen. Zur Prognose des progressionsfreien Überlebens in Kopf- und Halskarzinompatienten wurde ein Intervallzeitmodell mit dem zuvor entwickelten Transfer-Lernansatz kombiniert um anhand von Positron-Emissions-Tomographie (PET)/CT Bildern und klinischen Daten trainiert werden zu können. Dabei konnte auf dem Testdatensatz ein Übereinstimmungskoeffizient (engl. concordance index) von 0.668 erreicht werden.

Um eine Reproduzierbarkeit der hier durchgeführten Studien zu erreichen wurden alle Modelle anhand von öffentlich zugänglichen Daten oder im Rahmen von öffentlichen Wettbewerben entwickelt. Weiterführende Studien an größeren Datensätzen werden jedoch benötigt. Das Fehlen eben jener großen Datensätze stellt eines der Hauptprobleme im Bereicher der Deep Learning basierten Auswertung von medizinischen Bilddaten dar. Der vorgestellte Transfer-Lernansatz und die entwickelte selbstlernende Methode bilden eine Basis um Modelle an größeren Datensätzen vorzutrainieren und an spezifischen kleineren Datensätzen weiterzuentwickeln. Überlebensmodelle besitzen das Potential zur Entwicklung neuartiger Risikoabschätzungen. Diese können zur Auswahl der jeweiligen Therapie miteinbezogen werden um so etwa eine Überdosierung zu verhindern.

Publications

List of peer reviewed publications developed during the course of this thesis:

- Daniel M. Lang, Jan C. Peeken, Stephanie E. Combs, Jan J. Wilkens and Stefan Bartzsch. "Deep learning based hpv status prediction for oropharyngeal cancer patients". *Cancers* 13.4 (2021)
- Daniel M. Lang, Jan C. Peeken, Stephanie E. Combs, Jan J. Wilkens and Stefan Bartzsch. "Deep learning based GTV delineation and progression free survival risk score prediction for head and neck cancer patients". *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer. 2021
- Daniel M. Lang, Jan C. Peeken, Stephanie E. Combs, Jan J. Wilkens and Stefan Bartzsch. "Risk Score Classification of Renal Masses on CT Imaging Data Using a Convolutional Neural Network". *2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*. IEEE. 2022
- Daniel M. Lang, Jan C. Peeken, Stephanie E. Combs, Jan J. Wilkens and Stefan Bartzsch. "A Video Data Based Transfer Learning Approach for Classification of MGMT Status in Brain Tumor MR Images". *International MICCAI Brainlesion Workshop*. Springer. 2022

Contents

Chapter 1

Introduction

Cancer remains a major health risk in every country in the world, with 19.3 million new cases and almost 10.0 million deaths determined worldwide for the year 2020 [1]. Significant advances in therapy have been achieved in the past decades, leading to improved cure rates and therefore higher survival probabilities. However, early detection and precise diagnosis is crucial for successful treatment. Solid cancers appear in the form of a tumor, which is the end result of a whole series of changes, possibly developed over several years [2]. But what mechanisms are causing such a development? The underlying biological characteristics were most prominently summarized by Hanahan and Weinberg [3, 4] as the *hallmarks of cancer*, including the ten properties depicted in Figure 1.1. The combination of all of those hallmarks leads to an abnormal growth of cells, ultimately resulting in macroscopic formation of a tumor able to invade surrounding and distant tissue. However, underlying mechanisms of those characteristics are genomic ones and actual tumor formation features a broad richness in diversity. Cancers are no uniform entities but depict heterogeneous diseases with different courses and response to treatment.

For example, tumors collectively summarized under the term *head and neck cancer* include different epithelial malignancies that can be separated into five basic subregions with some of them featuring a even more precise differentiation [5]. Known risk factors for development of such tumors include: tobacco use, alcohol consumption, infection with the Epstein-Barr virus or a high-risk human papilloma virus (HPV) [6]. Head and neck cancers in the oropharynx, a subpart of the throat, are more likely to be caused by an infection with HPV, but HPV positive cases are associated with a better radiosensitivity than cases caused by alcohol or tobacco consumption [7].

Apart from such a heterogeneity in appearance, tumors are also detected at different stages. In later stages the cancer can have already spread to nearby lymph nodes or even distant organs, while at an earlier stage the tumor may be of small local extent. In addition, not only do cancers feature large heterogeneity but also patients differ fundamentally in various properties like age or vital status.

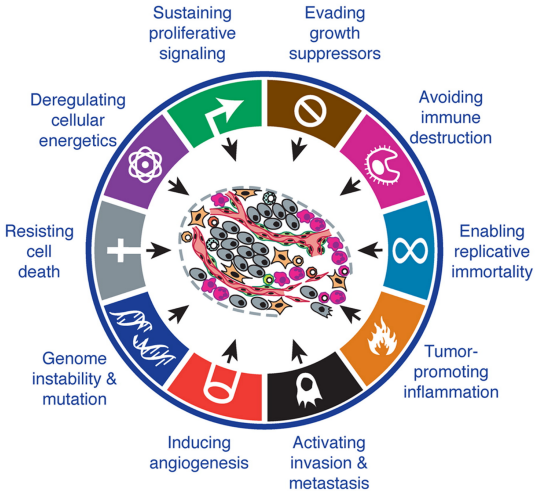


Figure 1.1: Hallmarks of cancer, taken from Hanahan and Weinberg [4]. The six hallmarks initially formulated in [3] were extended by two emerging hallmarks and two enabling characteristics leading to a total of 10 features shown here.

All of those characteristics have to be taken into account for the decision on the right treatment regime. Surgery, chemotherapy and radiation therapy are the most common cancer treatment options, but also hormone therapy, immuotherapy, and hyperthermia are available. Most often a combination of various methods is used. Assessment of the treatment regime does not only involve selection of the best therapy but also requires detailed determination of characteristics like dose and follow-up care.

This diversity in cancer development and extent, as well as the richness of treatment options, demands precise diagnosis such that an optimal therapy can be selected. Different tools have been developed to facilitate high diagnostic precision. Lab tests are able to identify biomarkers in body fluid samples, cells extracted via biopsies can be analyzed pathologically, genetic testing permits detection of DNA changes, and medical imaging allows for detailed rendering of tissue.

In general, imaging depicts an essential part of management for every tumor patient. Development of techniques like computed tomography (CT), magnetic resonance imaging (MRI) and positron emission tomography (PET) allow for delineation of macroscopic characteristics, caused by the inherent hallmarks of cancer, that can be used for evaluation of disease specificity. Technological advances have had huge impact on cancer detection and management and further improvements in technologies like phase contrast and dark-field imaging have the power for even more detailed resolution. However, in contrast to all of those technical enhancements, evaluation of images still relies on very basic anatomi-

cal features, like largest tumor diameter, or visual assessment by highly trained medical experts.

Today, artificial intelligence (AI) is present in our everyday life. Algorithms suggest music or movies to us, capture which mails are considered spam, and allow cars to drive on their own. The success of AI was mainly driven by the development of deep learning, a technique whose superiority in terms of image classification was proven by Krizhevsky et al. [8] in 2012 when they were able to win the ImageNet Large Scale Visual Recognition Challenge [9] with considerable distance by construction of a convolutional neural network (CNN). Deep learning is now considered state of the art for a large variety of tasks involving imaging data.

However, utilization of CNNs for examination of medical imaging data still lacks widespread application. Availability of large, publicly available datasets like ImageNet, MNIST [10] or CIFAR [11] was one key factor for the success of deep learning in the natural imaging domain. Medical cohorts differ significantly from such resources. Curation of medical data is cumbersome, causing cohorts to be several magnitudes smaller than those usually encountered in conventional settings. Features like dimensional extend of images, uniformity in perspective, proportions, and intensity require customized techniques and model building tailored for an application in the medical domain. Characteristics between different classes are often not as distinct as for tasks on natural imaging data, or can even be completely unknown. Moreover, endpoints like overall survival or long-term side effects, require models able to handle time dependent data, a type of model output not so commonly encountered in deep learning.

Therefore, deep learning techniques engineered for an application on standardized, well studied natural imaging datasets have to be customized to work in the medical domain and new approaches and models for problem settings solely related to medical imaging have to be developed.

Chapter 2

Aim of this thesis

Medical imaging data in oncology contains rich information about the tumor as well as surrounding tissue and nearby organs at risk. Imaging patters are visually assessed by medical experts, i.e. radiologists, for determination of properties essential for treatment. Three main image based radiological tasks can be differentiated: detection, characterization, and monitoring, with characterization involving the subtasks: segmentation, diagnosis and staging. However, human based assessment is time consuming and costly and relies on education and experience of the annotator. [12] Deep learning algorithms have proven their power for automated, high throughput image analysis in several settings outside of medicine over the past years, leading multiple studies to investigate the capability of such networks in the field of medical imaging. Most progress has been made in the area of image segmentation, mainly due to development of the U-net model of Ronneberger et al. [13], but also classification tasks can be solved utilizing deep learning algorithms. For example, Kirienko et al. [14] trained a PET/CT based convolutional neural network for prediction of tumor extend in lung cancer patients, and Lee et al. [15] developed a network for CT based detection of cervical lymph node metastasis in patients with thyroid cancer. However, not only can algorithms be trained for automation of tasks usually performed by human readers but also for prediction of novel scores, currently not assessed with the help of medical imaging. Chaunzwa et al. [16] build a model for CT based classification of non-small cell lung cancer tumor histology, usually determined pathologically. Medical imaging based detection of such clinical parameters has the advantage to be fast, non-invasive and comes without any extra costs if performed on routinely acquired images. Those characteristics are of particular interest in the field of radiation oncology, where treatment features a non-invasive character itself and imaging data depicts one of the main information sources utilized for therapy decisions. However, despite the advances made in deep learning based medial imaging analysis, there exist several limitations that the field has to overcome for successful widespread clinical application.

Medical imaging data differs significantly from natural images, commonly

encountered in the domain of deep learning. Instead of 2D RGB-color images, a grayscale 3D delineation of tissue is depicted, with intensity values in CT reflecting a quantitative measure. Those changes in imaging features have to be taken into account during model design. Most often, studies performed in the medical domain simply apply architectures developed on natural images, without adapting to the changed conditions. This, for example, results in models expecting input to have a 2D structure featuring three color channels. Allowing only slices of three-dimensional medical images, like CT or MR, to be input into the model, but no complete 3D images. Even though models can be trained successfully using 2D networks, see e.g. [17, 18], disease patterns in patients' body are three dimensional and algorithms aiming for full exploitation of tissue information have to incorporate this three-dimensional structure. Starke et al. [19] compared 2D and 3D convolutional neural networks for outcome modeling in head and neck squamous cell carcinoma patients and found a superiority of the 3D approach. Therefore, all the networks developed within this thesis are customized to the characteristics of medical imaging. Architectures employ a full three-dimensional structure. Furthermore, respective tumor expansions are taken into account, which is achieved by identification of the respective regions of interest and development of architectures able to take only those regions as an input. The model of Chapter 5, aiming for risk evaluation in kidney masses, accomplishes this by utilization of a Siamese network. Two patches, one for each kidney region, are cropped from the complete CT and used as an input, respective feature vectors are merged taking the element-wise maximum. Later chapters adapt techniques developed in the natural imaging domain, i.e. transfer and self-supervised learning methods, for an application on medical imaging.

Development of deep learning architectures is often performed on relatively well prepared large datasets, like the ImageNet database. As argued above, imaging data of this format differs from oncological imaging data in several aspects, requiring special handling. One of the most significant differences is given by the small size of medical cohorts, making model overfitting a substantial problems of the field. Different techniques have been developed to facilitate training on sparse data but must also be adapted to the needs specific to medical imaging data. Transfer learning and self-supervised learning approaches pretrain architectures on a related task before training of the actual task is started, such that general prior knowledge is inject into the model. Most often pretrain is performed on large 2D datasets, which forbids handling of three-dimensional data in the downstream task. In this thesis transfer and self-supervised learning techniques allowing full three-dimensional exploitation in the downstream task are investigated. In Chapter 6, a transfer learning approach pretrained on sports video clips, featuring two spatial and one temporal dimensions, is studied. Furthermore, the self-supervised masked autoencoder of He et al. [20] is remodeled to be able to handle three-dimensional medical imaging data.

Disease characteristics in medicine are often time dependent. Overall survival depicts the gold standard endpoint of clinical cancer trials [21], but also other endpoints, like long-term side effects, feature a temporal dependency. Frequently, studies circumvent this problem by prediction of endpoints at a given

time point, e.g. 2-year survival. However, utilization of such an approach does not allow for detailed risk assessment. Moreover, no appropriate handling of censored cases is possible, such that patient lost to follow-up have to be removed from the dataset. Therefore, for proper handling of temporal dependency models able to incorporated time-to-event data have to be employed. Ability of such models is studied within this thesis. Namely, in Chapter 7 the discrete survival model of Gensheimer and Narasimhan [22] is combined with the transfer learning approach developed in Chapter 6.2.1 for prediction of progression free survival in head and neck cancer patients.

In 2017, *The Economist* published an article claiming that "the world's most valuable resource is no longer oil, but data" [23], highlighting the rising value of data owned by big tech companies. Data is not only a valuable resource for tech giants but also in the scientific domain, especially in the medical field. Curation costs caused by labor intensive work of experts puts high value on medical cohorts. Access to such datasets allows research groups to perform studies that cannot be conducted by others, which increases their scientific impact. Additionally, data protection issues limit public distribution. This leads to the fact that studies are often performed on private datasets, not shared with the community. However, replication is one of the most essential factors in science, depicting the ultimate standard by which scientific claims have to be judged [24] and development of models on hidden, private datasets hinders replication. Moreover, in the small data domain of medicine, model performance can heavily depend on the cohorts used during training and evaluation [25]. Therefore, architectures developed on different datasets cannot be compared, which impedes technical advances but also reproducibility. Hence, data sharing and model development on public datasets should be aspired. This can be achieved by the utilization of public datasets during model construction, or by competition in publicly held challenges, which has become the standard for validation of biomedical image analysis methods [26]. Both approaches are investigated in this thesis. The studies of Chapter 6 were performed on open access data mined from The Cancer Imaging Archive (TCIA) [27], a public archive including different oncological cohorts. Model development and evaluation in Chapter 5 was conducted in the setting of the pubic KNIGHT challenge [28] held at the 2022 IEEE International Symposium on Biomedical Imaging (ISBI) and the studies in Chapter 7 were performed in conjunction with the 2021 MICCAI HEad and neCK TumOR (HECKTOR) challenge [29].

Endpoints investigated in this thesis involved histological test results and novel risk scores commonly not determined on medical imaging data. In Chapter 5 a network is presented that was trained for prediction of pathological risk scores in renal masses based on preoperative CT imaging data. Advances in non-invasive therapy options, like radiation therapy, increase the demand for such non-invasive testing. Approaches capable of detection of head and neck cancer cases caused by a infection with the human papilloma virus are discussed in Chapter 6. HPV depicts a essential information source in radiation oncology that is conventionally tested histopathologically. The network presented in Chapter 7 was trained on prediction of tumor progression in head and

neck cancer patients. Development of a risk score predicting tumor progression allows for identification of high and low risk cases and therefore for adoption of treatment dose.

Chapter 3 introduces the main background of medical imaging and deep learning utilized in later sections, while Chapter 4 discusses the problem settings specially related to an application of deep learning architectures in the medical imaging domain. A summary and general discussion about the application of deep learning models in the domain of oncological imaging data is presented in Chapter 8. Finally, a conclusion in given in Chapter 9.

Chapter 3

Background

3.1 Radiation oncology and medical imaging

Imaging is essential for treatment of all patients with solid tumors and is involved in nearly every part of cancer management, from initial discovery, to image guided biopsy and surgery, to cancer staging, and treatment planning in radiation therapy [30].

Medical imaging techniques most commonly used in oncology, especially for treatment planning in radiation oncology, are computed tomography (CT) and magnetic resonance imaging (MRI), but also nuclear imaging techniques, like positron emission tomography (PET) and single photon emission computed tomography (SPECT), are utilized.

3.1.1 Radiation therapy

Apart from surgery and chemotherapy, radiation therapy (RT) depicts one of the main treatment options in oncology. Today, more of half of adult tumor patients can definitely be cured, with half of them receiving RT [31]. Radiation therapy utilizes high energy particles of ionizing radiation to destroy cancerous tissue and stop tumors from further development. Usually, the total radiation dose is split into multiple *fractions*, delivered over the course of several weeks. Reason for this are repair mechanisms setting in between radiation sessions, that are able to restore healthy tissue but not malignant cells. Hence, damage to healthy tissue is reduced while tumor control can be maintained. [32] Most commonly, beams of photons are utilized, but also electrons, protons and other particles can be used [33]. Curative intended RT can be administered

- on its own, as a primary treatment option,
- as a neoadjuvant therapy option before surgery,
- as an adjuvant therapy option after surgery,

- in combination with other therapy methods, i.e. chemotherapy.

External beam radiation therapy (EBRT) delivers radiation from a source outside of patient's body to the tumor region, while for brachytherapy a radiation source is placed inside of patient's body. In the following, only EBRT will be considered.

Any cancer therapy faces the challenge to achieve the highest probability of cure with the least damage to healthy tissue [34]. Treatment planning aims for determination of parameters considered optimal in management of patients disease. A fundamental part of treatment planning in RT is given by delineation of a *target volume*, including the tumor and its spread to surrounding tissue and lymphatics, and also identification of nearby organs at risk. Medical imaging, discussed in Section 3.1.2, depicts the most essential information source for identification of those volumes. Other parameters to be optimized include: dose prescription, dose fractionation and dose distribution. [35]

Over the past decades, progress in image acquisition allowed for advances in tissue delineation, and technical developments in radiation devices and dose planning improved precision of dose delivery. Leading RT to evolve from a non-site-specific technique towards specialized planning based on three-dimensional reconstruction of images and optimization algorithms [34]. Technical advances in dose delivery include the development of intensity modulated radiotherapy (IMRT). Multiple beams of varying intensities are utilized to deliver high radiation dose to the target and minimum dose to nearby organs at risk [36]. Today, IMRT depicts the most commonly used treatment option in RT [37].

Further developments include the introduction of volumetric modulated arc therapy (VMAT) and stereotactic body radiation therapy (SBRT). VMAT delivers radiation in a continuous manner while rotating around the patient. This allows for superior conformity of the delivered dose with the target volume. However, treatment planning in VMAT requires higher quality and is also more time consuming than for IMRT. [32] SBRT, delivers a small number of ultra-high doses to the target volume with high precision, such that sparring of surrounding healthy tissue is improved [38]. But, utilization of high dose rates requires exact dose delivery, such that not surrounding normal tissue will be exposed. This can only be secured for small target volumes, limiting the applicability of the approach. [35]

Finally, adaptive radiation therapy (ART) [39, 40] aims for incorporation of any changes in patients' body during treatment. Such changes can be given by organ movement or anatomical modifications during treatment. Offline-ART measures changes prior to or during treatment and applies changes in the next treatment session, while online-ART aims for continuous monitoring and immediate adaption. [32]

3.1.2 Imaging techniques

Computed Tomography

Computed tomography (CT) images display a quantitative density measurement of tissue. X-rays are aimed on the patient using a fan beam. Attenuation coefficients of those x-rays are measured by a detector. The x-ray source and detector are rotated around the patient to probe the tissue from numerous angles. In this way a cross sectional image of the patient, with intensity values reflecting quantitative characteristics of probed tissue, is constructed.

The measured attenuation coefficients μ are rescaled into Hounsfield units (HU) using

$$\mathrm{HU} = 1000 \times \frac{\mu - \mu_{\mathrm{water}}}{\mu_{\mathrm{water}} - \mu_{\mathrm{air}}}, \tag{3.1}$$

assigning a value of 0 HU to water and -1000 HU to air. Imaging is done in a slice wise fashion and patients, lying on a bench, are moved trough the scan plane for construction of a three dimensional image. Interpolation is applied between slices for reconstruction of a artifact-free image.

CT comes with the drawback of radiation exposure for patients, caused by the x-rays. This pushed the development of scanner hardware and reconstruction filters, allowing for probing with lower dose. However, even though advances could be made for such low dose CT (LDCT), the technique still suffers from a increase of noise in the images.

Another variation of CT is given by cone beam imaging. Cone shaped x-rays are used and attenuation coefficients are measured with a two dimensional detector. This results in a decreased examination time but increases artifacts induced by scatter radiation.

Contrast media, influencing tissue attenuation coefficients by induction of media containing atoms of higher atomic number, is regularly used for improvement of contrast in soft tissue.

Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) utilizes high external magnetic fields for image acquisition. Application of the external field yields hydrogen nuclei in patient's body to partially align and accumulate a net magnetization by spinning around the direction of the external field with a frequency proportional to the field strength. Most scanners directing a field of 1.5 or 3 Tesla on the patient.

This equilibrium state of partially aligned nuclei is then temporarily excited by a radio frequency (RF) pulse. The time it takes the nuclei to return to the equilibrium after the RF pulse is switched off is measured in longitudinal (T1 relaxation) and transversal (T2 relaxation) direction, with reference to the external field. Both of those times depend on the surrounding tissue and chemical composition of the nuclei and therefore allow for determination of contrast between different tissues.

Changes in the protocol allow for emphasis on contrast between different types of tissue. In general, MR has better soft tissue contrast than CT. How-

ever, intensity values do not display a quantitative measure, limiting the comparability between different scanners and protocols. More advanced techniques involve the injection of contrast agent, like gadolinium.

Positron Emission Tomography

For PET imaging, a radionuclide bound to a drug is injected into the body. The metabolic activity caused by the drug can then be imaged by particles emitted from the radioisotope, which allows for detection of biochemical abnormalities in organs or tissue.

For tumor imaging usually ^{18}F-fluorodeoxyglucose (^{18}F-FDG) is used. FDG is absorbed by the tumor while the ^{18}F decays and emits a neutrino-positron pair. The emitted positrons collide with electrons in the body and annihilate into pairs of two photons that can be measured by detectors. This signaling allows for reconstruction of physiological structures in the body.

For semi-quantitative comparison, PET intensity values are usually transformed into standardized uptake values (SUV), taking injection dose, time between injection and imaging and patient weight into account

$$SUV = \frac{\lambda(t)}{D_{\text{inj}}(t)/w}, \tag{3.2}$$

with $\lambda(t)$ the measured radioactivity, D_{inj} the injected dose and w the body weight of the patient [32]. SUV is widely used as a functional biomarker able to stratify patients into subgroups and allowing for prediction of clinical outcomes like survival [41]. PET imaging can also be combined with CT in one sequential scanner able to generate a registered combination of both modalities known as PET/CT.

Other imaging techniques involved in clinical cancer management include: x-ray radiography, ultrasound and single-photon emission computed tomography (SPECT).

3.1.3 Utilization of medical imaging for diagnosis and therapy

Medical images incorporate rich information about tissue, which can be utilized for clinical decision making. In clinical practice, images are mainly visually interpreted by highly trained medical experts, i.e. radiologists, with standardized guidelines developed to improve decision processes in cancer management.

In terms of diagnosis, imaging reporting and data system (I-RADS) guidelines have been established for standardization of visual findings. Such guidelines exist for estimation of risk in patients suspicious of prostate (PI-RADS), breast (BI-RADS) and liver (LI-RADS) cancer [30]. Once presence of cancerous tissue is proven, cancer staging is applied for characterization of extent and spreading of the tumor. Clinical staging is performed prior to treatment by

application of a certain system. TNM [42] is the most widely applied system for anatomical staging of solid cancers, with tumors being classified by:

- T - primary tumor extent,
- N - involvement of regional lymph nodes,
- M - involvement of distant metastases.

However, medical imaging can also be used to assess response to therapy. The WHO [43] and RECIST criteria [44] classify patients into four response groups based on change in tumor diameter over time. But, for irregular shapes both criteria fail to capture changes adequately [41]. Moreover, do those criteria not provide any guideline on how to use their information for readjustment of therapy, and even if standardized response guidelines could be formulated, the methods are only able to catch changes during treatment. While a method capable of risk stratification and prognostication prior to treatment would allow for adaption of treatment dose and therefore possibly spare patients from overdosage even before therapy is started.

Functional imaging like ^{18}F-FDG PET has the potential for such pretreatment stratification. Markers like maximum standardized uptake values (SUV_{max}) and metabolic tumor volume (MTV) of ^{18}F-FDG PET can be associated with endpoints like overall survival in non-small cell lung cancer (NSCLC) [45, 46] or head and neck cancer [47].

However, medical imaging data contains rich information about the tumor and its environment, nodal and metastatic involvement, and general disease patterns, that can simply not be captured by such simplistic markers.

3.2 Radiomics

Advances in digitization in the 1980s and 1990s, involving the development of picture archiving and communication systems (PACS) and standards like digital imaging and communications in medicine (DICOM), resulted in first developments of computer aided diagnosis (CADx) systems in radiology [48]. Those early approaches applied a very limited number of filters for image processing and feature extraction and were developed to facilitate tasks like detection of breast lesions and microcalcifications on mammograms [49, 50] and lung nodules on CT images [51]. Over time, further improvements in terms of algorithms and filters established a research field meanwhile termed *radiomics* [1] [52, 53].

Radiomics extracts a large number of quantitative features from a given region of interest (ROI) by application of predefined, handcrafted filters. Typically, the gross tumor volume (GTV), segmented by a human annotator, will be used as the ROI, i.e. resulting features provide only information about this

[1] Terminology of the expression *radiomics* is not uniquely defined. In this work, the term will be used in the context of feature extraction from medical imaging data relying on predefined filters, but not in correlation with deep learning models.

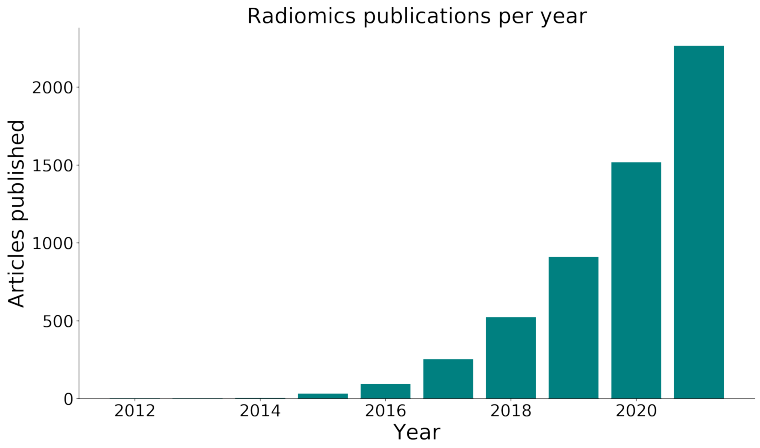


Figure 3.1: Radiomics publications per year. Data was extracted from https://pubmed.ncbi.nlm.nih.gov using the search term "radiomics".

specific region. Statistical tests exclude redundant features. The remaining features will then be used for training of classical machine learning algorithms, like logistic regression models, random forests classifiers, support vector machines and Cox proportional hazards models [54]. The principle workflow of radiomics is shown in Figure 3.2a.

Since introduction of the terminology of *radiomics* by Lambin et al. [52] beginning of 2012, the research field is constantly growing, Figure 3.1.

An example application is given by Aerts et al. [55], who extracted 440 features from CT images of head and neck and lung tumor patients to train a Cox model. They were able to prove significant performance on survival prediction and could associate predictions with gene-expression patterns. Bogowicz et al. [56] calculated 317 CT radiomics features for head and neck squamous cell carcinoma (HNSCC) patients. A logistic regression model was trained for detection of infections with the human papillomavirus (HPV). Four features could be associated with patients HPV status, achieving an area under the receiver operating characteristics curve (AUC) of 0.78 on the validation set.

However, all this work has not led to clinical application yet. A significant proportion of studies lack sufficient quality, rely on gray level patterns in the ROI, and depend on the scanning devices present in the training set [57]. Leading studies to seek standardization of features and frameworks [58, 59].

But, even though those standards will possibly lead to improvements in terms of reproducibility and robustness, radiomics still depends on delineation of the tumor region for feature extraction. Not only is tumor delineation on medical images labor intensive, requiring expert knowledge, but also known to suffer from inter- and intra-observer variability [60]. Changes in volume have

significant influence on radiomics features [61–63]. Hence, radiomics algorithms trained on volumes defined by one annotator can lead to different output for annotations performed by a second annotator, or even between different annotations of the same radiologist. Studies have been conducted to search for features that are stable under such shifts in tumor volume [64–66]. However, even if the issue could be solved and robust features may be established, limitation to a certain region of interest introduces another problem.

The TNM system includes nodal and metastatic information for staging, highlighting the importance of input coming from outside of the main tumor volume. The possibility to included those regions in the radiomics workflow exists, but cancer is known to be a lot more complex than just being a solid encapsulated mass. In order to sustain progression, tumor builds up a surrounding called the tumor microenvironment (TME). The TME is a cellular environment encompassing the surrounding immune cells, blood vessels, extracellular matrix (ECM), fibroblasts, lymphocytes, bone marrow-derived inflammatory cells, and signaling molecules [67]. Medical imaging is not (yet) able to resolve interactions on a molecular level, but growth of blood vessels, induced by the cancer hallmark angiogenesis, manifests in macroscopic characteristics. Also, the two other hallmarks: activating invasion and metastasis and tumor-promoting inflammation, lead to physiological changes in the TME. Such structures in the TME play a fundamental role in tumor progression and metastasis and can be utilized for evaluation of tumor aggressiveness [68]. Hence, non-invasive imaging of structures in the TME can provide information of cancer aggressiveness, metastasis, and help to determine early response to treatment [69]. Future innovations will most probably allow for even more detailed resolution. Therefore, focusing on a encapsulated volume, like the GTV, removes valuable information from imaging data, limiting the capability of approaches relying on it.

Deep learning algorithms may be better suited to be used for extraction of information from medical imaging data. Deep learning techniques do not rely on definition of a volume for feature extraction and construction of predefined handcrafted filters, but work directly on the input images. Workflows of both techniques, radiomics and deep learning, are shown in Figure 3.2.

3.3 Deep learning

Today, *artificial intelligence (AI)* is present in your everyday life. Cell phones automatically detect objects in images, react to spoken words, know which topics we are (supposedly) interested in and which mails we do consider as spam. But, which part makes an algorithm *intelligent*? And which part of this intelligence is *artificial* and which *human*? Moreover, what is the difference between artificial and human intelligence. When does AI perform *human-like*? And what even makes human performance *human-like*?

A significant part of the answer to those questions is of philosophical nature, laying outside of the scope of this thesis. However, for better understanding, one must know how the term *artificial intelligence* evolved. Starting in the 1940s

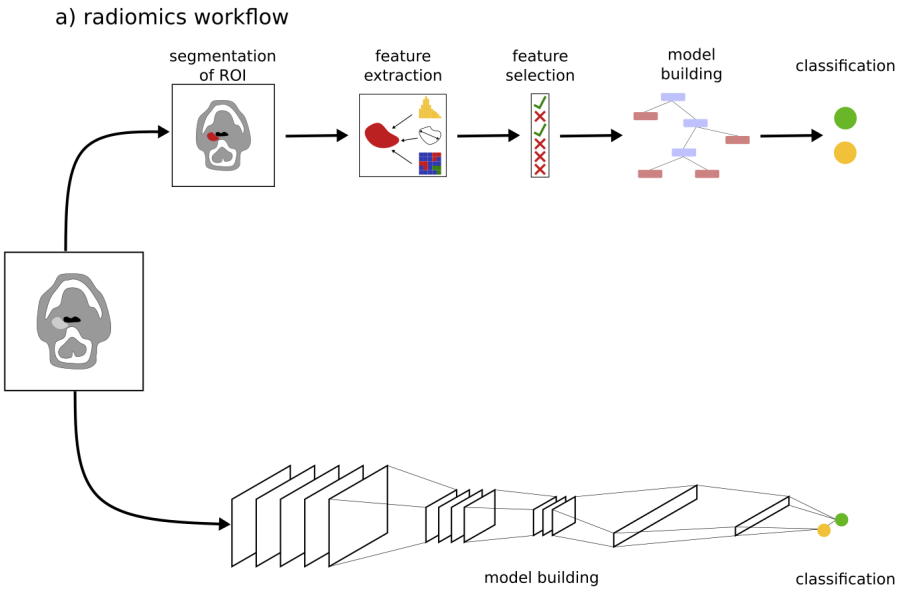


Figure 3.2: Workflow for radiomics and deep learning models, following Hosny et al. [12]. For the radiomics workflow, shown in a), a volume of interest has to be defined. Features are extracted from this volume and redundant ones are neglected. The remaining features are then used to train a more classical machine learning model, like a decision tree. Deep learning models, shown in b), work directly on the input images and do not require volume delineation or feature selection.

and 1950s, researchers tried to formulate models of the human brain. McCulloch and Pitts [70] and Rosenblatt [71] introduced the perceptron, a mathematical model of the smallest processing unit in the human brain, the neuron. In this time the terminology of *artificial intelligence* was introduced. The existence of a formal description of human thinking was assumed, that researchers were seeking to model. However, over the course of the twentieth century the idea was abandoned. [72] Today, researchers aim for development of algorithms that perform well in complex domains by any means and not by human-like standards [73]. Dick [72] states that "the most powerful and profitable artificial intelligences we have produced [...] exhibit a rather limited range of intelligent behavior" and argues that models are only trained for one task, that is to make accurate predictions.

Therefore, there is little to be said against an suspension of the terminology of *artificial intelligence*. This thesis will mainly deal with the research field of *deep learning*, commonly considered a sub-field of AI. Therefore, algorithms will be referred to as deep learning models or architectures. However, in reference to the very broad research field, taking into account deep learning architectures but also more classical machine learning models, the common notation of AI can not be abolished and will be used.

3.3.1 Deep neural networks

The very basic concept behind deep learning based algorithms stems from the idea to develop artificial neural networks, modeling interactions in the human brain. This is also reflected in the title of Rosenblatt's 1958 paper "The perceptron: a probabilistic model for information storage and organization in the brain" [71]. The underlying signaling unit of the human brain is given by the neuron. Single neurons build connections with other neurons to form a biological network. Connections between neurons are strengthened or weakened based on how often they are used. All incoming signaling to a neuron is summed in the cell body, transformed into a output signal, and then propagated to other connected neurons. [74] However, even though signaling processes in the brain, conducted by neurons, had huge influence on the initial development of deep learning, researchers realized over time that sticking to closely onto the concept of an exact brain model was misleading. For example, only activation functions found to be similar to the biological processing of neurons were used in the beginning, even though other mathematical functions are now known to work better. Therefore, artificial neural networks have become quite different from their biological counterparts over time [75], but the core concept of connected neurons remains.

An artificial neuron processes its input $\mathbf{x}$ by weighting it with the adjustable coefficients $\mathbf{w}$ for construction of a output signal y. A schematic of this is shown in Figure 3.3. The input vector $\mathbf{x} = (x_1\ x_2\ x_2\ ...\ x_n)$ is multiplied with the weight vector $\mathbf{w} = (w_1\ w_2\ w_2\ ...\ w_n)$ to form a scalar signal, which is usually accompanied with some bias term b. This *logit* is then passed to a *activation*

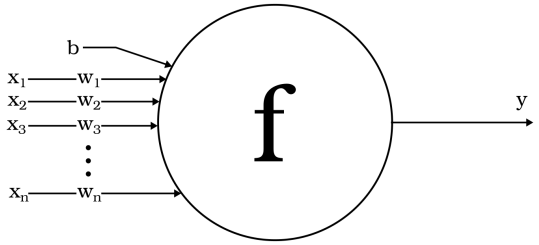


Figure 3.3: Schematic of a artificial neuron. The input to the neuron $\mathbf{x} = (x_1\ x_2\ x_2\ ...\ x_n)$ is weighted by $\mathbf{w} = (w_1\ w_2\ w_2\ ...\ w_n)$ to be summed with the bias term b in the cell body. This logit is then passed through an activation function to generate the scalar output y.

function f to produce the output [74]

$$y(\mathbf{x}; \mathbf{w}, b) = f(\mathbf{x} \cdot \mathbf{w} + b). \tag{3.3}$$

Rosenblatt's model of such an artificial neuron, termed *perceptron*, used the Heaviside step function as an activation function and a negative bias term to model scalar output, such that the model read

$$y = \begin{cases} 1 & \text{if} \quad \sum_j x_j \cdot w_j > b \\ 0 & \text{if} \quad \sum_j x_j \cdot w_j \leq b. \end{cases} \tag{3.4}$$

Hence, the perceptron is able to combine different inputs in a weighted manner to form a binary output decision. An example for this could be the decision whether or not a cancer patient should receive adjuvant radiation therapy, with the input variables x_j being clinical parameters like: age, sex, TNM-stage and cancer grading.

Complex modeling problems, of course, require a more complex model than a single perceptron, and as neurons in the human brain form a network, perceptrons can be combined into a network called multilayer perceptron (MLP). An example network for this can be seen in Figure 3.4. The architecture persists of four input neurons x_i, directly connected to four neurons in the first layer, followed by three neurons in the second layer and one output neuron. Layers not directly connected to the input or output are called hidden layers, models featuring two or more hidden layers are called deep neural networks [75]. Neurons in the first hidden layer are directly working on input features, while later layers have access to patterns extracted by earlier layers, which allows for more complex decision making. The structural organization of multilayer perceptrons utilizes neurons in such a way that each neuron in a given layer is connected to all neurons in the prior and in the subsequent layer, Figure 3.4. Such type of layer is also called dense or fully connected layer.

For the network to learn the right set of weights, an algorithm will be applied that changes the weights in a stepwise fashion such that model output will

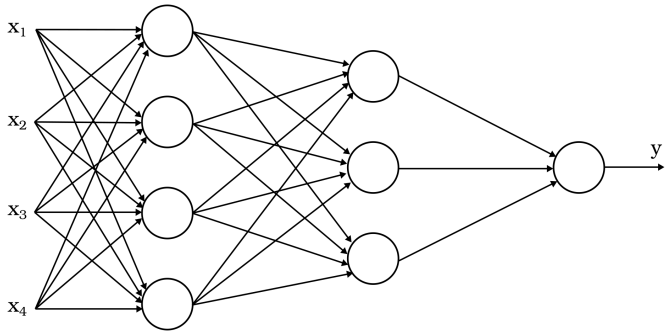


Figure 3.4: Exemplary multilayer perceptron (MLP). The model features four input neurons x_i and one output neuron y. Two hidden layers are present, with four neurons in the first hidden layer and three neurons in the second hidden layer.

be more similar to the desired label, i.e. small changes in the weights should cause small changes in the output [76]. This is not the case if the Heaviside step function is used as a activation function. Therefore, dense layers are trained using differentiable continuous activation functions. Notably, any kind of differentiable activation function can be used, but if only linear activations are employed the whole network reduces to a simple linear model [77]. Therefore, hidden layers feature non-linear activations. The most frequently used ones are depicted in Figure 3.5.

For hidden neurons a typical activation function is given by the rectified linear unit (ReLU)

$$\phi(x) = \max(0, x). \tag{3.5}$$

Selection of the activation function in the output layer depends on the modeling problem at hand. In case of a single output neuron for binary classification, typically the sigmoid activation

$$\sigma(z) = \frac{1}{1 + \exp(-z)} \tag{3.6}$$

will be used. While for classification problems involving K different classes commonly the softmax function

$$\sigma(\mathbf{z})_i = \frac{\exp(z_i)}{\sum_{j=1}^{K} \exp(z_j)} \tag{3.7}$$

will be applied.

3.3.2 Backpropagation and gradient descent

As for more classical machine learning techniques, deep neural networks are trained by minimization of a loss function, fitting the model to the data. A very

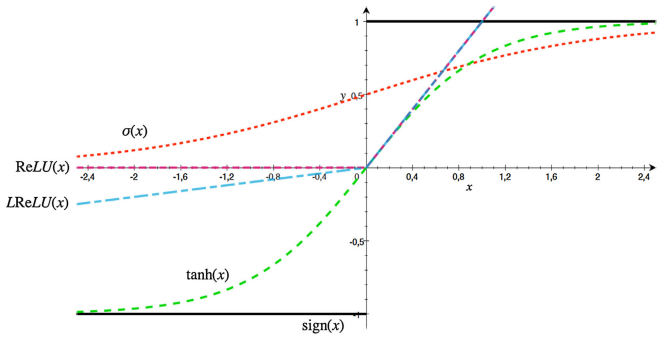


Figure 3.5: Most commonly used activation functions, taken from Maier et al. [78]. The initially used Heaviside step function ($sign(x)$) is shown in black. Modern activation functions are given by the hyperbolic tangent ($tanh(x)$, green), the sigmoid activation ($\sigma(x)$, red), the rectified linear unit ($ReLU(x)$, pink) and the leaky rectified linear unit ($LReLU(x)$, blue).

basic loss function is given by mean squared error (MSE)

$$L_{\text{MSE}} = \frac{1}{2n} \sum_{x} \|y(x) - \tilde{y}\|^2, \tag{3.8}$$

with n training examples $\mathbf{x}$ with corresponding labels $\tilde{y}$ and $y(x)$ the network output.

After network construction, weights are randomly initialized, e.g. by using the method of Glorot and Bengio [79], and then iteratively modified for the model to fit to the training data. Hence, a way to minimize the loss by tweaking of trainable network weights has to be found. For a long time, it was unclear to researches how to optimize the trainable parameters of MLPs, but the problem was solved by introduction of the backpropagation algorithm [80, 81].

Gradient descent, introduced by Cauchy et al. [82] in 1847, depicts a well known iterative optimization algorithm for identification of local minima. The algorithm takes small steps in the opposite direction of objective's gradient, with the size of those steps controlled by the learning rate. So, the change in trainable network parameters θ can be computed by

$$\Delta\theta \equiv -\epsilon \, \nabla_\theta L(\theta), \tag{3.9}$$

with ϵ, the learning rate. In each optimization step the network parameters can then be updated by [75]

$$\theta \leftarrow \theta - \epsilon \nabla_\theta L(\theta). \tag{3.10}$$

With the shift in network parameters given by

$$\Delta\theta = -\epsilon \, \nabla_\theta \left(\frac{1}{2n} \sum_{x} \|y(x) - \tilde{y}\|^2 \right) \tag{3.11}$$

if the MSE loss of eq. 3.8 is used.

Therefore, with the sum in eq. 3.11 running over all training examples, each input volume has to be passed trough the network for computation of all network outputs $y(x)$, before the weights can be tweaked using gradient descent. This is exactly what the backpropagation algorithm is doing: "for each training instance the backpropagation algorithm first makes a prediction (forward pass), measures the error, then goes through each layer in reverse to measure the error contribution from each connection (reverse pass), and finally slightly tweaks the connection weights to reduce the error (Gradient Descent step)"[75].

This procedure is also known as batch gradient descent. For computation of the next gradient descent step the whole dataset is used. As stated above, the algorithm searches for local minima, which in general will be different from the global minimum for non-trivial problem settings [74]. In order to find the best fit to the data, one is of course interested in identification of the global minimum of the loss function. Therefore, a modified version called minibatch gradient descent is typically used for training of neural networks. Minibatch gradient descent only uses a subset of the complete training set for computation of the error function before optimizing the weights. As the examples utilized for loss calculation change during optimization, the surface of the loss changes, possibly leading the optimization algorithm to not get stuck in local minima. The number of examples presented to the algorithms before each optimization step, called the batch size, influences the optimization performance. However, not only is the application of minibatches influencing the optimization process, but it also reduces the computational burden, as only part of the input data has to be fitted into memory during evaluation.

Modern optimization algorithms also add a *momentum term* to the gradient descent step in order to account for previous steps [75]

$$\begin{aligned} \mathbf{m} &\leftarrow \beta\mathbf{m} - \epsilon\nabla_\theta L(\theta) \\ \theta &\leftarrow \theta + \mathbf{m}, \end{aligned} \tag{3.12}$$

with $\mathbf{m}$ being the momentum term and β given between 0 and 1. A frequently utilized method making use of such a momentum term is given by the *Adam* optimizer [83].

3.3.3 Model parametrization

By minimization of the loss function on the training data it is assumed that general patterns, associated with the underlying problem, will be learned, and that the model will therefore be able to generalize well on data not used during optimization. The performance of the model on previously unobserved inputs is called generalization, a low generalization error is aspired [84].

Neural networks typically feature a large number of trainable weights. For example, AlexNet [8], one of the first successful neural networks, already consisted of 650,000 neurons and 60 million trainable parameters. On one hand, networks with too many degrees of freedom will not learn a general representation but only fit very precisely to the training data by remembering single data

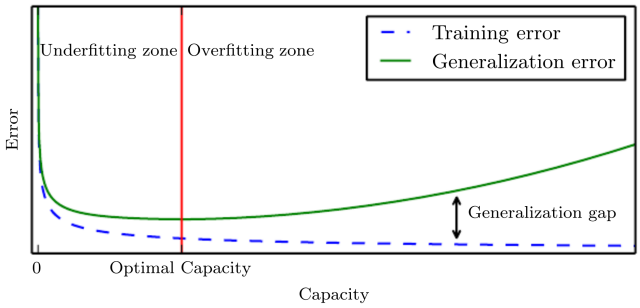


Figure 3.6: Underfitting and overfitting zones based on model capacity, taken from Goodfellow et al. [84]. A model not powerful enough will not be able to learn the task reasonably well. Hence, even the training loss will not reach a desirable minimum. A model too large will overfit on the data by remembering single data points instead of recognizing general patterns of the underlying problem. This leads to a large gap between training and generalization error.

points, a process called overfitting. On the other, models need to be powerful enough to perform well on non-trivial tasks. Image classification for a diverse set of classes requires a large architecture, like AlexNet. If a model is too small, it will not be able to reach sufficient performance, a phenomena entitled underfitting.

So, choosing a model's capacity involves a tradeoff between improving performance on the training dataset but also minimizing the generalization error, as shown in Figure 3.6.

In order to keep track of the generalization error during training, a subset of the training data will be used as validation set. This independent set, not involving any training samples, is used to calculate a unbiased loss, and possibly other metrics, at every optimization step/training epoch, in order to keep track of the generalization gap. Differences in training and validation performance can indicate over- or underfitting, Figure 3.7 shows an example of overfitting. Therefore, validation set performance can be used as a measure for optimization of hyperparameters like number of layers or size of the learning rate. During inference a third, independent set is then used for computation of a final unbiased performance measure. Hence, training of deep learning models involves three datasets: a training set, used for weight optimization, a validation set, used to keep track of the generalization error during training, and a completely independent test set, used for final model evaluation.

3.3.4 Problem formulation and performance measure

Choice of the right loss function to be optimized during training is essential. A diverse set of loss functions tailored for specific settings exists. The mean

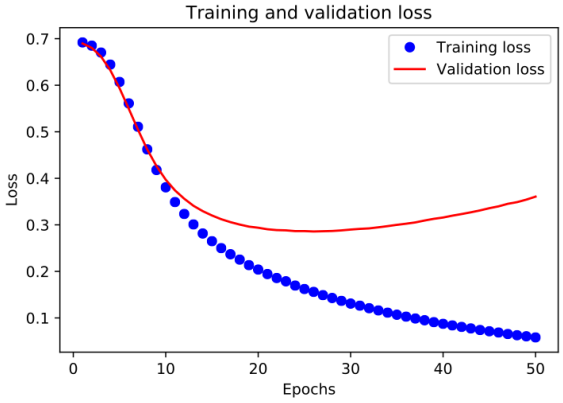


Figure 3.7: Example learning curve showing training and validation performance, taken from Murphy [77]. A declining validation set performance, setting in at around epoch 25, indicates model overfitting.

squared error loss of eq. (3.8) is mostly used in regression problem settings, modeling scalar output. While classification models are usually trained using the categorical cross entropy loss

$$L_{\text{CE}} = -\sum_{i=1}^{n} y_i \log(p_i) \tag{3.13}$$

with y_i the true label, p_i the softmax probability for class i and n the total number of classes. The cross entropy loss penalizes false predictions more heavily than MSE and is therefore better suited for classification tasks.

Categorical cross entropy eq. (3.13) and mean squared error eq. (3.8) in combination with the Dice loss, that will be introduce in Section 3.3.7, are the most essential losses used to train deep learning models. However, other loss functions exist and selection of the right loss is an important step in model formulation.

For imbalanced datasets with a unequal proportion of cases per class, introduction of a weighting term in the loss can help prevent the optimization process to solely focus on overrepresented classes. In case of the categorical cross entropy eq. (3.13), this would be done by multiplication of each summand with a weighting w_i, giving higher emphasis to underrepresented classes.

For evaluation of model performance during validation and testing, selection of suitable metrics is also essential. A common metric used for classification is given by accuracy, i.e. the number of correct classified cases divided by the total number of cases. However, for imbalanced distributions of classes this can cause serious problems [85].

Clinical tests often involve binary classification tasks, i.e. distinguishing between patients having a disease and patients without the disease. In this case,

sensitivity and specificity can be used to measure model performance. Sensitivity represents the ability of a test to identify cases with a positive label, i.e. having the disease, while specificity measures the ability to identify cases with a negative label, i.e. without the disease [86]. A graphical illustration of binary classification performance is given by the receiver operating characteristics (ROC) curve. The area under the receiver operating characteristics curve (AUC) represents the probability that a randomly chosen case with a positive ground truth label is ranked with greater suspicion than a randomly chosen ground truth negative case [87].

3.3.5 Regularization and augmentation techniques

As mentioned above, acquisition of large datasets is not possible for all settings. However, non-trivial tasks require large enough networks, increasing the risk for model overfitting. Two approaches designed to prevent networks from overfitting, while at the same time allowing for reasonable model sizes, are given by *regularization* and *augmentation*.

"Regularization techniques are a set of best practices that actively impede the model's ability to fit perfectly to the training data, with the goal of making the model perform better during validation" [88]. One such technique is given by weight regularization. A term sanctioning large values for trainable model parameters is added to the loss. In this way, the space of possible weights for the model to choose from is reduced, limiting the ability of the model to exactly fit to the training data. The weights are either added proportionally to their absolute value, called $L1$ *regularization*, or by using the squared value, called $L2$ *regularization* or weight decay

$$\begin{aligned} Loss &\to Loss + \lambda \sum_j |\theta_i| \qquad L1 \text{ regularization}, \\ Loss &\to Loss + \lambda \sum_j \theta_i^2 \qquad L2 \text{ regularization}, \end{aligned} \tag{3.14}$$

with the trainable network parameters θ_i and λ controlling the impact of the regularization term.

However, the most essential regularization technique used in neural networks is given by dropout [89]. The technique set a fraction of randomly chosen neurons for specific layers to zero, such that they are excluded during optimization. In this way, ability of single neurons to memorize specific data points is reduced, preventing the model from overfitting.

Data augmentation performs modifications on the training data to artificially increase dataset size. For imaging data, such augmentation techniques include: random rotations of the images, flipping/mirroring on a certain axis, random zooming, random cropping, addition of random noise, scaling of image brightness and modification of contrast and saturation.

Data augmentation often leads to significant improvements in terms of performance and robustness, since valid augmentation mechanisms algorithmically

inject prior knowledge into the model [77]. Especially, for image classification tasks augmentation has been proven to be effective, even simple transformations like scaling or rotations can lead to significant performance gains [84]. For problem settings involving only a few training samples, the approach is essential to teach the network certain invariance and robustness properties [13]. Shorten and Khoshgoftaar [90] concluded that augmentation modifies limited datasets in such a way that the characteristics of big data are acquired.

3.3.6 Convolutional neural networks

Images are usually given in the form of 2D arrays, but MLPs, introduced in Section 3.3.1, require one dimensional input. Therefore, pixel values have to be rearranged into a 1D vector to fulfill the desired input format of the network. This results in large input sizes and, as neurons are connected densely, huge networks with lots of trainable parameters. Moreover, dense networks trained on imaging data are not translational invariant to shifts in the input images, as each input neuron will be associated with a given location in the image. If a dense model is trained on cancer images with the cancer always present in the lower right corner of the image, it will not be able to perform well on images with the cancer in the upper left corner.

A type of deep learning architecture able to solve those problems is given by convolutional neural networks (CNNs). CNNs are again inspired by models of the brain, especially the visual cortex. Studies performed in the 1950s by Hubel and Wiesel [91] found that the visual cortex consists of layers with a hierarchical structure, deeper layers build upon features detected by previous layers. Initial layers detect lines and edges and deeper layers combine those patterns to identify contours and shapes and finally entire objects [74], Figure 3.8. Those findings first lead to the application of hand-crafted filters for feature extraction from imaging data and finally to development of CNNs.

As the name already reveals, the most essential component in a CNN is given by a convolutional operation. The idea of using filters for feature extraction remains, but for CNNs the filters, also called kernels, are not predefined but learned during training. Weights in the kernels are again randomly initialized and then tweaked for optimization, in the same manner as for dense neural networks.

A basic example of a convolution operation is shown in Figure 3.9. Values in the source layer that are in view of the kernel are multiplied with their kernel counterparts to then be summed up for generation of a value in the destination layer. In order to filter the complete input, the kernel is slid over the input array with a given *stride*. As array elements on the edges of the input will be involved less frequently in this filtering process, padding can be applied. Most commonly, array elements with a value of zero are padded to the edges, such that the spatial dimension of the input volume will be preserved. An example process can be seen in Figure 3.10, a convolutional kernel of size 3×3 is used, stride in horizontal direction is given by 1, and a padding of size one is applied.

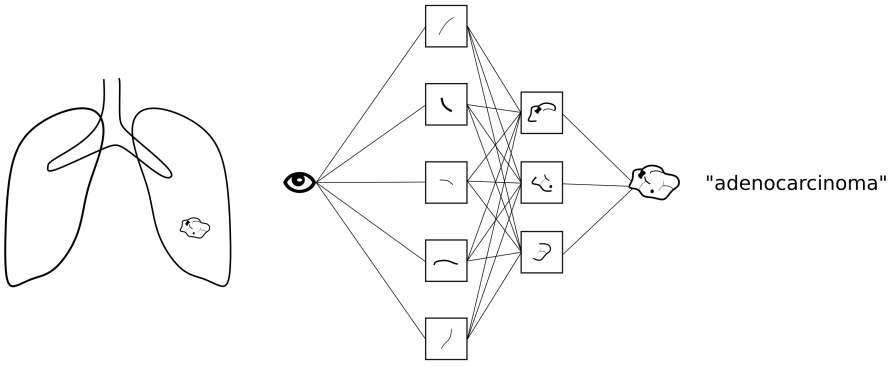


Figure 3.8: Schematic of the hierarchical structure in the visual cortex. Initial layers recognize low level features like lines and edges. Deeper layers combine patterns found in lower layers for detection of more complex patterns like contours and shapes, with final layers recognizing complete objects.

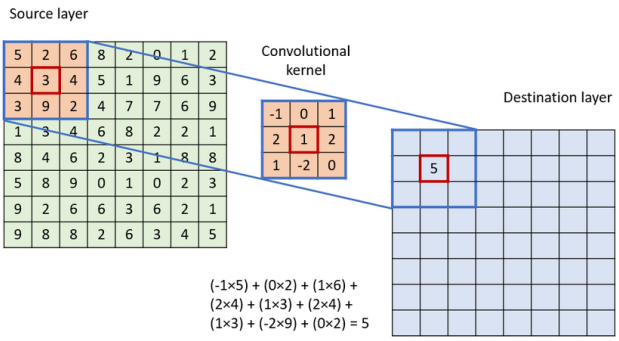


Figure 3.9: Convolutional operation, taken from Podareanu et al. [92]. Values in the source layer are multiplied with their counterparts in the convolutional kernel, featuring here a size of 3×3. Those input-kernel pairs are then summed up for generation of the respective value in the destination layer.

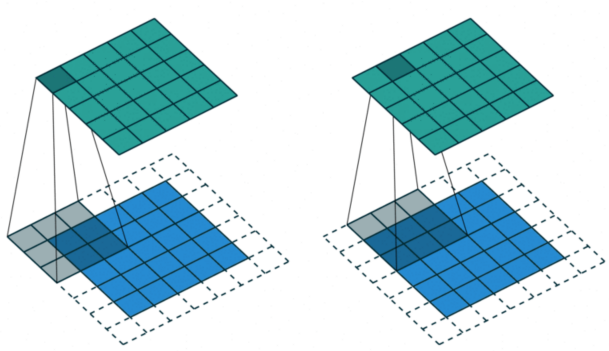


Figure 3.10: Convolution operation, taken from Dumoulin and Visin [93]. Input values (blue) are convolved with the kernel of size 3×3 (shaded) for extraction of the respective value in the feature map (cyan). In the next step, the kernel is moved with a stride of 1 in horizontal direction for generation of the next feature map value. For preservation of spatial dimensionality the input is padded.

In this way, a *feature map* is generated, that can again be filtered by kernels. At every network depth, different filters are probed for feature extraction, leading a variety of feature maps.

Additionally to convolutional layers, pooling layers are usually employed in CNNs. Again, filters are slid over the input of the layer, however, no weighted combination of input values in view of the kernel is computed but static merging is performed. For *max pooling* layers, the maximum of the input values is selected, Figure 3.11. Other pooling operations are given by taking the mean or average.

Pooling layers are usually employed for reduction of spatial dimension, while convolutional layers typically preserve it. Most convolutional networks feature a

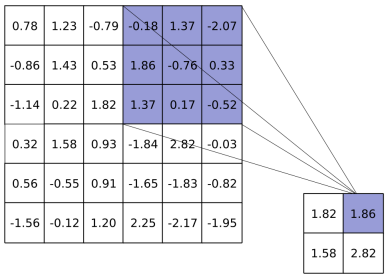


Figure 3.11: Max pooling operation with a kernel and stride of size 3×3. For each max pooling operation the maximum of the current field of view of the kernel is taken.

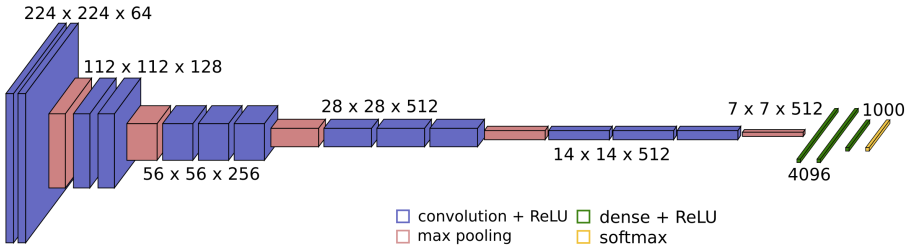


Figure 3.12: VGG network architecture of Simonyan and Zisserman [94]. The model was trained for classification of natural images from the ImageNet dataset [9], featuring 1000 classes. The convolutional part of the model, consisting of convolutional layers (blue) and max pooling layers (red), is utilized for feature extraction. Classification is performed by fully connected layers (green) ending in an output layer with 1000 neurons normalized by the softmax function (yellow), reflecting class probabilities.

pyramid-like structure. The number of filters grows with network depth, while the size of the feature maps shrinks accordingly [88]. A typical CNN structure is given by the VGG16 model of Simonyan and Zisserman [94], shown in Figure 3.12. The input size is given by colored images of size 224 $\times$224 $\times$3, with 3 reflecting the number of red-green-blue (RGB) color channels. The first convolutional layer utilizes a feature map size of 64. Those feature maps are further filtered for patterns by another convolutional layer of the same size. Next, a max pooling layer is applied for reduction of spatial dimensionality. In total, five such combinations of convolutional and max pooling layers are applied for feature extraction, with the two first blocks utilizing two convolutional layers followed by max pooling and the remaining blocks using three convolutional layers before max pooling. The architecture was trained for classification of the ImageNet dataset [9], featuring 1000 categories. Therefore, after the convolutional part, used for feature extraction, fully connected layers are applied for classification, ending in a output layer with 1000 neurons normalized by application of the softmax function.

Convolutional and pooling operations can not only be performed in a two dimensional way. Video data is given in three dimensions, with the third dimension being depicted by the time axis, for classification of sports video data Tran et al. [95] used 3D convolutions and max pooling for feature extraction. Medical imaging data, like CT or MR, is also three dimensional and proper application of CNN models demands the application of three dimensional feature extraction layers.

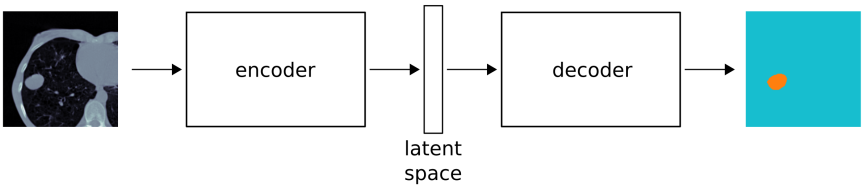


Figure 3.13: Autoencoder architecture, the basic building blocks are given by a decoder and a encoder part. The encoder transforms the input into a smaller dimensional space, the latent space. While the decoder restores the dimensionality of the input data from the latent space. Here, segmentation of the gross tumor volume in a lung cancer CT image, is depicted as an example task.

3.3.7 Segmentation networks

Apart form classification, deep learning models are regularly trained on image segmentation tasks. Segmentation can be categorized into two types, semantic segmentation and instance segmentation. For semantic segmentation the learning task is given by a pixel wise classification into different categories, with different objects of the same class not being differentiated. While instance segmentation learns to differentiate objects of the same class. For segmentation of kidneys this would for example mean that semantic segmentation only differentiates between kidney and other tissue while for instance segmentation the task could be given by segmenting the left kidney, the right kidney and other tissue.

The VGG model depicted in Figure 3.12 was trained for image classification, featuring a pyramid-like structure. For segmentation tasks, network output usually is of the same dimensionality as the input, requiring another type of structure. A architecture featuring such kind of structure is given by autoencoders, Figure 3.13. Autoencoders consist of a encoder and a decoder part. The encoder transforms the input into a latent representation, while the decoder restores the dimensionality of the input from this latent representation. Autoencoders trained on imaging data typically feature convolutional layers only. The encoder can, for example, be given by the convolutional part of the VGG16 model in Figure 3.12, while the encoder takes the same form but will be mirrored. For recovery of the input image's dimensionality, transpose convolutions, Figure 3.14, are used in the decoder.

Autoencoders can be applied for image denoising, with the noisy image as an input and the denoised images as output, but also for anomaly detection, dimensionality reduction and last but not least image segmentation.

However, autoencoders require feature information to flow through every layer in the architecture, which can cause information extracted in initial layers to be lost in deeper layers. A solution to this problem is given by the introduction of connections feeding information of earlier layers directly to deeper layers, by skipping layers in between [96, 97]. Ronneberger et al. [13] devel-

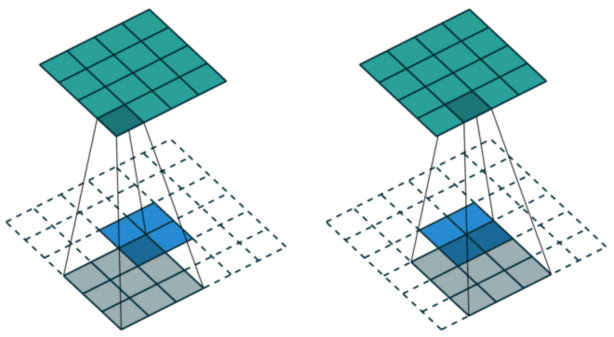


Figure 3.14: Transpose convolution operation, taken from Dumoulin and Visin [93]. The input (blue) is padded such that convolution with the kernel (shaded) results in an increase in spatial dimension for the feature map (cyan).

oped the U-Net, a fully convolutional segmentation model following the basic encoder-decoder principle of autoencoders, but at the same time making use of such *skip connections*.

The network structure can be seen in Figure 3.15. In the encoder path, convolutional and max pooling layers are used to transform the input image into the latent space representation, given by the feature map with 32×32 pixels on the bottom. In the decoder path, transpose convolutions, here named *up-conv*, are used to increase spatial dimensionality. Skip connections, copying feature maps from the encoder to the decoder, are introduced. In this way, global information, extracted by initial layers, is available at all stages in the decoder path.

The U-Net model was able to win the ISBI 2015 cell tracking challenge [13] and has established as the basic architecture in biomedical image segmentation since then. Isensee et al. [98] were able to enhanced the model even further by introduction of predefined preprocessing and training strategies, but the basic U-Net like structure remains.

As discussed in Section 3.3.4, formulation of the right optimization problem plays a key role. Even though it is possible to train segmentation models using the MSE loss eq. (3.8) or the cross entropy loss eq. (3.13), typically the Dice loss will be utilized. The Dice coefficient

$$DSC(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{3.15}$$

measures the similarity between a set of points A and B. Values for the Dice coefficient range between zero, reflecting no agreement between both sets, and one, reflecting perfect agreement. The Dice loss, to be minimized, is defined as

$$L_{DSC} = 1 - DSC(A, B). \tag{3.16}$$

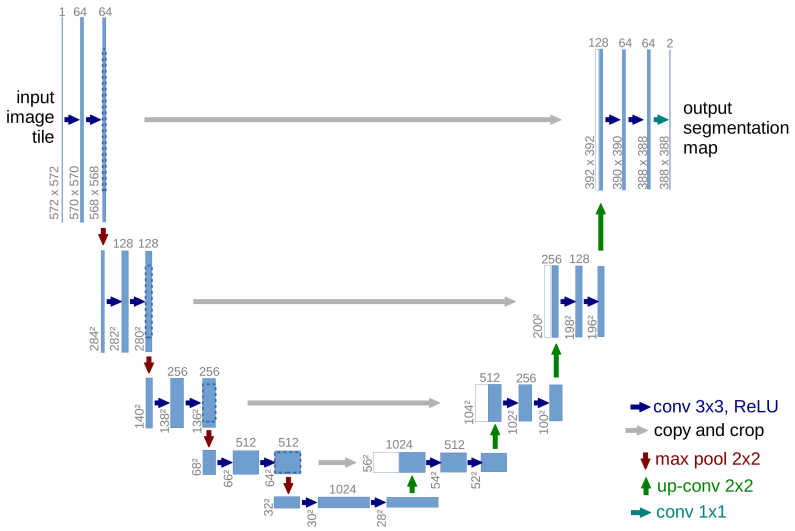


Figure 3.15: U-Net architecture, taken from Ronneberger et al. [13]. The model follows the basic encoder-decoder structure of autoencoders, but employs skip connections, called *copy and crop* here. Downsampling in the encoder path is done by convolutional and max pooling layers and upsampling by convolutional and transpose convolutional layers, called *up-conv* here. Convolutional layers in the encoder feature no padding and therefore introduce a reduction in spatial dimensionality. Hence, feature maps in the encoder have to be cropped before they can be copied to the decoder. By introduction of skip connections, global information, extracted by initial layers, is available at every stage of the decoder path and does not have to be passed through the whole architecture.

The Hausdorff distance measures the extend to which a set of points A is displaced from another set of points B, using [99]

$$HD(A,B) = \max\left\{\sup_{a\in A}\inf_{b\in B} \mathrm{d}(a,b), \sup_{b\in B}\inf_{a\in A} \mathrm{d}(a,b)\right\}, \tag{3.17}$$

with the Euclidean distance $\mathrm{d}(a,b)$ between point a and b and sup and inf the supremum and infimum. The Hausdorff distance is also frequently used as a performance measure in segmentation problems.

3.3.8 Transfer learning

Transfer learning utilizes knowledge learned in an initial task P_1 to improve performance on a downstream task P_2. It is assumed that variations learned in P_1 are relevant for variations that have to be identified in P_2 [84]. So, the main idea is it to use the knowledge gained from solving one problem setting and transfer it to solve another, related problem. With the initial task P_1 and its related source domain being independent from the downstream target domain of task P_2.

As discussed in Section 3.3.5, gathering of large datasets in order to train a complex model is not always possible. *Pretraining* a model on a larger dataset, that is somehow related to the actual task at hand, induces prior knowledge before training of the actual task is started. Transfer learning assumes that induction of this prior knowledge will enable the model to learn on less data, compared to a training strategy with randomly initialized weights.

An example for this would be the usage of a model pretrained on ImageNet, with several million images present, and transfer parts of it to train a model for detection of diseases in lung x-ray images, featuring a dataset with only a few hundred patients. By pretraining on the natural images of ImageNet the model will learn to detect very basic patterns in its initial layers like edges, lines and corners. Such basic knowledge about vision is possibly also essential for the classification of x-ray images and therefore allows the model to then be trained on this task with fewer examples present.

For CNN architectures, the knowledge transfer is usually achieved by copying of weights from the pretrained model to the downstream model, Figure 3.16. Fully connected layers are (partially) replaced with new layers of customized size, suitable for the problem setting of the downstream task. The network is then fine tuned, with parts of the initial weights kept fix. Hence, the transfer of knowledge is performed by copying pretrained filters learned on the initial task to the downstream model.

3.3.9 Self-supervised learning

Deep learning algorithms discussed in detail so far were trained to associate input with a given output value, both present during training. This type of learning procedures is called supervised learning. In contrast to that, unsupervised learning algorithms are trained on data for which no output label is

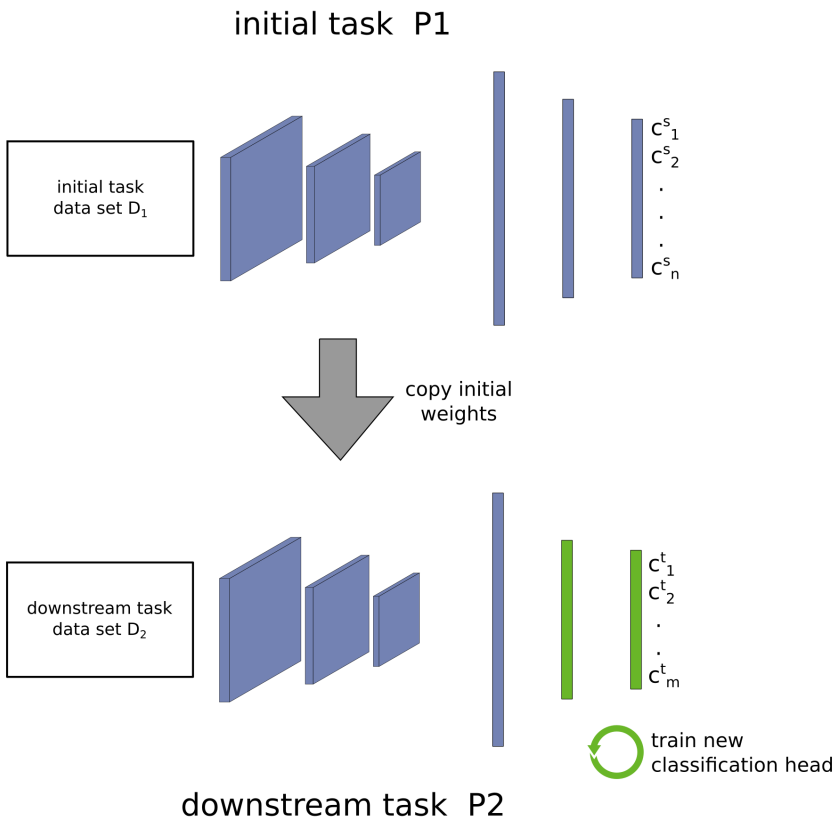


Figure 3.16: Transfer Learning. For convolutional neural networks the model is first trained on task P_1, featuring n classes c_i^s. After finishing training of the initial task, part of the weights (blue) are copied to the model of the downstream task, which will then be trained on task P_2, with m classes c_i^t. Notably, the classes in the initial and the downstream task can be completely independent. Copying of model weights transfers knowledge gained about patterns in task P_1 to the downstream model. The classification head of the downstream model (green) is then customized to fulfill the structural demands of task P_2.

present. More precisely: "unsupervised learning involves observing several examples of a random vector x, and attempting to implicitly or explicitly learn the probability distribution $p(x)$, or some interesting properties of that distribution, while supervised learning involves observing several examples of a random vector x and an associated value or vector y, and learning to predict y from x, usually by estimating $p(y|x)$" [84]

An example for unsupervised learning is given by the denoising autoencoder, briefly mentioned in Section 3.3.7. For training of a image based denoising autoencoder, Gaussian or any other kind of noise can be added to the input images, while network output is given by the uncorrupted images. Hence, only a set of input data is required for training but no associated labels. If the noise added during training will be chosen in the right manner, the approach will result in a model able to enhance image quality.

However, autoencoders can also be trained to simply restore the input image without any corruption happening. The latent space is usually chosen such that it is of lower dimensionality then model input. Therefore, the network has to learn a transformation into a lower dimensional space. As the decoder is trained to restore the input images from the latent space, the transformation has to be meaningful, such that no essential information will be lost.

This application is also an example for *representation learning*, which in general studies the process of developing algorithms able to map their input into a lower dimensional embedding while at the same time sustaining all the essential information. Pretraining CNNs on larger datasets can also be seen as a form of representation learning method. The convolutional part maps the input image into a lower dimensional embedding that has to be meaningful in order to succeed on the initial task.

However, the application of transfer learning requires the initial and the downstream domain to be somehow similar in the patterns essential for their respective tasks. Otherwise, the embedding, learned in the initial task, will be useless in the downstream task, resulting in a insufficient model. For example, the usefulness of pretraining on natural imaging data like the ImageNet for a downstream applications in the medical imaging domain was challenged by Raghu et al. [100]. Limitations of the approach will be discussed later on.

Most of the time, the limiting factor for generation of large datasets is not given by a lack of input data but of corresponding labels. Therefore, datasets frequently involve a subset of unlabeled cases. *Self-supervised learning* aims for inclusion of unlabeled data in the training procedure. Models are pretrained in a unsupervised fashion to then be finetuned on the subset of labeled data points, or a related dataset from the same domain.

Vincent et al. [101] trained a denoising autoencoder to recover corrupted images, in a fashion similar to that described above, for classification of images from the MNIST database, containing images of handwritten digits [10]. The autoencoder was first trained for image denoising and then finetuned for classification, using the same dataset for both tasks. Finetuning can be performed in such a way that weights of the encoder are copied to the downstream network and a new, randomly initialized classification head is attached, similar to the

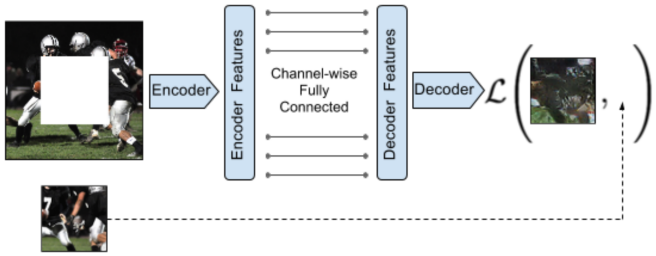


Figure 3.17: Image inpainting as a self-supervised learning approach, taken from Pathak et al. [102]. A patch of the input image is removed and feed to a autoencoder-like encoder-decoder network. The encoder learns to map the input into the lower dimensional latent space, entitled *Encoder Features* here. A channel wise fully connected layer is used for generation of features provided to the decoder, which aims for reconstruction of the removed part. L_2 is used as a loss, measuring the difference between the removed and the reconstructed patch.

approach discussed in Section 3.3.8 and shown in Figure 3.16.

Another example for a self-supervised approach using a autoencoder for pretraining was developed by Pathak et al. [102], called image inpainting, Figure 3.17. The autoencoder architecture was slightly modified, with a channel-wise fully connected layer in the latent space, but the basic encoder-decoder structure remains. For corruption, a whole patch is removed from the input image. The model is then trained for reconstruction of this missing patch. Again, the pretrained encoder can then be utilized as a feature extractor in a downstream task.

Gidaris et al. [103] rotated images by a multiple of 90° to pretrain a network on prediction of the rotation angle. Doersch et al. [104] cropped two patches from the input image and let the network predict their relative position in the pretraining step. While Noroozi and Favaro [105] constructed a jigsaw puzzle, cropping several patches from the images, to then shuffle them randomly and let the network reconstruct their initial order.

Another self-supervised approach applicable on imaging data is given by the masked autoencoder model of He et al. [20], depicted in Figure 3.18. The architecture utilizes transformer layers, that will be introduced in the next section. Images were divided into patches, then a fraction of those patches was removed/masked from the input images. The autoencoder architecture is pretrained for reconstruction of the unmasked image. Again, the trained encoder can be utilized in a given downstream task. Inspiration for the masking approach was taken from the BERT architecture of Devlin et al. [106], a natural language processing (NLP) model.

Recently, NLP networks have seen a huge increase in performance. The GTP-3 model [107] is able to generate text so true to human produced one that

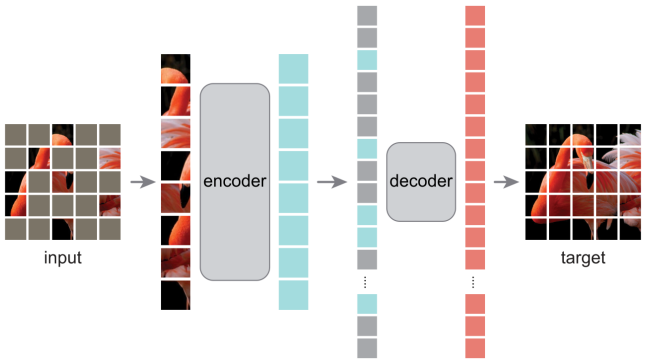


Figure 3.18: Masked autoencoder model, taken from He et al. [20]. Images are divided into patches. A fraction of those patches is removed, i.e. *masked*, from model input. The architecture is trained for reconstruction of the unaltered full image.

the authors felt the need to incorporate a whole section about possible misuse and dangers in the publication. Training of the model was made possible by pretraining in a unsupervised way on a huge text dataset crawled from the web that involved 300 billion tokens.

Another ingredient to the success of NLP architectures was the development of transformer models, discussed in the next section.

3.3.10 Transformer models

When our brain is inspecting a visual scene it receives more information than it is able to process, therefore, it has learned to neglect parts of the input and only pay *attention* to a fraction of it. If we do not focus on anything in particular, attention can *involuntarily* be focused on bright colors or moving objects. While *volitional* attention can suppress input from larger objects if we are searching for something small. [108] Machine learning researches took inspiration from this mechanism to develop certain type of layers, called attention layers, which are extensively used in *transformer* models.

For all the deep learning layers discussed so far, output features $\mathbf{h}$ in hidden layers are obtained by linear combination of input features $\mathbf{x} \in \mathbb{R}^v$ with layers' trainable parameters $\mathbf{W} \in \mathbb{R}^{v' \times v}$ followed by a activation function f

$$\mathbf{h} = f(\mathbf{W}\mathbf{x}), \tag{3.18}$$

with v the number of input neurons and v' the number of output neurons, in case of a dense layer. The idea behind attention layers is the application of a flexible set of m feature vectors in this case called *values* $\mathbf{V} \in \mathbb{R}^{m \times v}$. Based on input, the model pays different attention and uses a different set of those vectors

for feature extraction. This is achieved by measuring the similarity between the input *query* vector $\mathbf{q} \in \mathbb{R}^q$ and a set of m *keys* $\mathbf{K} \in \mathbb{R}^{m \times k}$, if $\mathbf{q}$ is most similar to key i value $\mathbf{v}_i$ will be used. [77]

Attention can be thought of as a dictionary lookup, but in order to make the operation differentiable, not a single value is retrieved per query but a weighted combination of all possible values,

$$\begin{aligned} \text{Attention}(\mathbf{q}, (\mathbf{k}_1, \mathbf{v}_1), ..., (\mathbf{k}_m, \mathbf{v}_m)) &= \text{Attention}(\mathbf{q}, (\mathbf{k}_{1:m}, \mathbf{v}_{1:m})) \\ &= \sum_{i=1}^{m} \alpha_i(\mathbf{q}, \mathbf{k}_{1:m}) \mathbf{v}_i, \end{aligned} \tag{3.19}$$

with the attention weights $\alpha_i(\mathbf{q}, \mathbf{k}_{1:m})$ satisfying $0 \leq \alpha_i(\mathbf{q}, \mathbf{k}_{1:m}) \leq 1 \; \forall \; i$ and $\sum_i \alpha_i(\mathbf{q}, \mathbf{k}_{1:m}) = 1$. A attention score function $a(\mathbf{q}, \mathbf{k}_i) \in \mathbb{R}$ measuring the similarity between q and k_i can be used in combination with the softmax function to compute the attention weights [77],

$$\alpha_i(\mathbf{q}, \mathbf{k}_{1:m}) = \frac{\exp(a(\mathbf{q}, \mathbf{k}_i))}{\sum_{j=1}^{m} \exp(a(\mathbf{q}, \mathbf{k}_j))}. \tag{3.20}$$

A certain type of attention score function is given by *scaled dot-product attention*,

$$a(\mathbf{q}, \mathbf{k}) = \mathbf{q}^\mathbf{T} \mathbf{k} / \sqrt{d}, \tag{3.21}$$

which requires $\mathbf{q}$ and $\mathbf{k}$ to have the same length d, Figure 3.19. Finally, for the case of mini batches, queries, keys and values are given by matrices, leading

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}. \tag{3.22}$$

A special kind of attention mechanism is given by self-attention. The input to layer $\mathbf{X}$ is projected using the weights matrices $\mathbf{W}_Q$, $\mathbf{W}_K$ and $\mathbf{W}_V$ to construct queries, keys and values,

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_V. \tag{3.23}$$

By optimization of those weight matrices self-attention is able to learn relations between patterns in the input. The mechanism is heavily employed in natural language processing models, where the relation between different words in the input sentences plays a major role.

For an application in transformer models, multi-headed self attention, allowing for a multitude of patterns to be recognized, is used

$$\begin{aligned} \text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= \text{Concat}(\mathbf{h}_1, ..., \mathbf{h}_n)\mathbf{W}^O \\ \text{with} \quad \mathbf{h}_i &= \text{Attention}(\mathbf{X}\mathbf{W}_Q^i, \mathbf{X}\mathbf{W}_K^i, \mathbf{X}\mathbf{W}_V^i). \end{aligned} \tag{3.24}$$

Multi-headed attention layers are combined with normalization layers and multilayer perceptrons (MLPs) in order to form transformation blocks, shown on the right hand side of Figure 3.19.

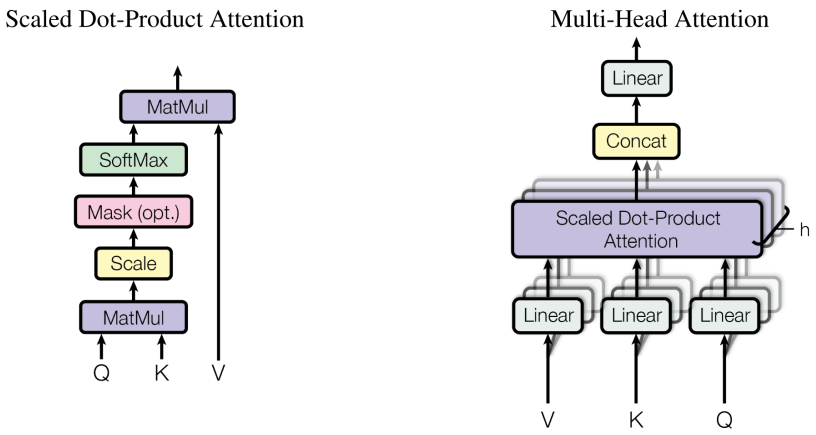


Figure 3.19: Scaled dot-product attention and multi-head attention, taken from Vaswani et al. [109]. For the scaled dot-product, queries Q are multiplied with keys K, to then be scaled by $\sqrt{d_k}$ and processed by the softmax function, to be finally multiplied with the values V. For multi-head attention, h of those weighted dot-product attentions are generated and concatenated in order to form the output.

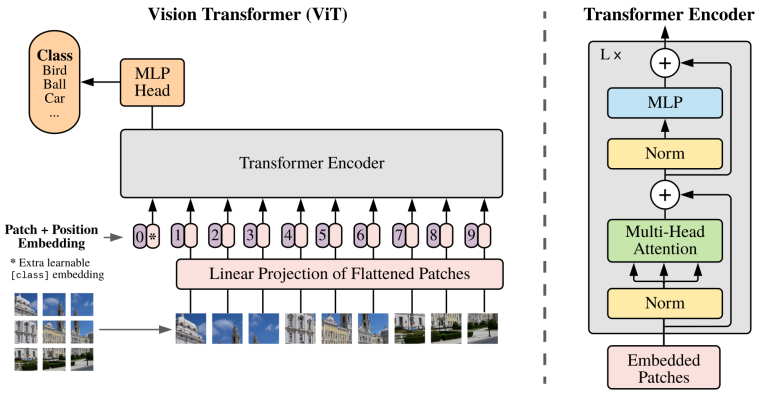


Figure 3.20: Vision transformer model, taken from Dosovitskiy et al. [110]. Input images are cut into patches, which are then transformed into an embedding space, using a linear projection, and merged with positional embeddings, as no inherent spatial knowledge is present in the model. This embedding is then processed by transformer blocks, shown on the right hand side, ending in a MLP head utilized for classification.

The superiority of transformer models over other architectures for an application in NLP tasks has been proven by a variety of architectures in recent years, see e.g. [106, 107]. In order to be mathematically processible, single words in the input are transformed into an embedding space. This input embedding is then merged with a possitional encoding, as no inherent spatial knowledge about the input is build into the model, like for example in CNNs where spatial kernels are used.

Transformer models are not limited to be used in NLP tasks only but were also adopted as vision transformers (ViT) to be applied in computer vision [110]. ViTs are working on image patches that are embedded by a linear projection layer. This projection is again merged with a positional encoding to be processed by transformer layers, Figure 3.19.

Chapter 4

Deep Learning on medical imaging data

One of the problem settings studied most intensively in the area of deep learning is given by the classification task of ImageNet [9], a database with more than 14 million images incorporating classes like animals, food and vehicles. Such images include a large variety in features like perspective, proportions, color and imaging equipment used for acquisition.

Medical imaging cohorts differ from such data in various ways. The image acquisition techniques mentioned in Section 3.1.2 all result in three dimensional grayscale data, that is not just stored as voxel arrays but contains information about physical space. Most importantly, information about voxel spacing and image orientation allows for standardization in terms of proportion and perspective.

Medical cohorts usually involve a very limited number of institutions, often even data from one center only, with each center exhibiting only few imaging scanners. Thus, medical cohorts commonly include data from few image acquisition devices. Another, very significant, feature of medical datasets is their small size. Medical cohorts are generally several magnitudes smaller than dataset sizes of standard deep learning problem settings, involving a few thousand or even hundred cases. Reasons for this are privacy protection regulations, limiting the ability to share and distribute medical information, and the fact that annotation requires labor intensive work of medical experts, which makes the generation of cohorts expensive and time consuming.

Therefore, medical imaging cohorts involve more standardized images than natural imaging datasets, but at the same time dataset sizes are smaller, with images given by three dimensional grayscale arrays. Due to such differences to standard deep learning problem settings, the techniques used before, during, and after training also differ from standard approaches.

4.1 Data augmentation techniques

As mentioned above, collection of large medical datasets is cumbersome. However, augmentation techniques can be used to artificially increase dataset sizes.

On one hand, modifications have to be applied such that no information in the images, essential for the classification task, will be destroyed. Characteristics in medical imaging data, specifying affiliation to a certain class, can be more subtle than for natural images. Therefore, image corruption methods have to be chosen such that this information will not be annihilated.

On the other, reasonable augmentation techniques can introduce robustness and invariance in the model. For example, rotations and flipping/mirroring during training introduce invariance to those transformations during inference. Slight zooming of the images still gives a physically meaningful copy and prevents predictions relying solely on the size of the object at hand, i.e. the tumor.

Also, physically motivated augmentation techniques can be applied. Deformation techniques make use of random displacement vectors to shift pixel/voxel values by a certain range. Idea behind this is the generation of physically plausible variations, mimicking deformations caused by patient movement. For elastic deformations, displacement vectors are sampled from a Gaussian distribution [13, 111]. More advanced techniques make use of deep learning architectures, like generative adversarial networks (GANs), themselves to generate augmented samples [112].

4.2 Transfer learning

Even though the domain of natural images differs significantly from that of medical images, several studies reported improved performances for models that were pretrained on ImageNet-like data to be then finetuned on a medical task [17, 18, 113]. However, real benefit of this approach remains controversial, with different studies leading to different conclusions [113, 114]. Raghu et al. [100] accounted improvements to the utilization of over-parameterized models and inferred no actual gain from transferred knowledge.

Apart from the questionable advantage of transfer learning on ImageNet, utilization of the approach introduces fundamental limitations. Natural images are two dimensional, and networks pretrained on such data also require input in the downstream task to be of the same dimensionality. This forbids direct processing of three dimensional medical imaging data. Workarounds involve approaches feeding three slices of the 3D grayscale images to the color channels of the network, calling it a 2.5D approach. With some of them using one slice for each dimension in the image, see e.g. Saint-Esteven et al. [115]. However, 2D convolutions result in two dimensional feature maps, destroying any higher dimensional information from the input image [95]. Therefore, in order to investigate the real benefit of transfer learning in the medical imaging domain, a 3D approach has to be developed.

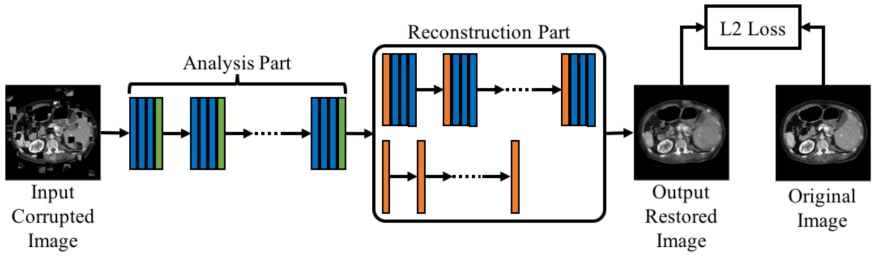


Figure 4.1: Self-supervised context restoration approach in medical imaging, taken from Chen et al. [118]. For image corruption, some patches in the input image are randomly shuffled. The autoencoder architecture, consisting of the encoder (called *Analysis Part*), and the decoder (called *Reconstruction Part*), is trained for restoration of the uncorrupted image. Loss is given by $L2$. Depending on the respective downstream task, size of the decoder is varied. For a segmentation downstream task a larger decoder is used, shown in the upper part, while for a classification downstream task a smaller decoder will be trained.

4.3 Self-supervised learning

As already mentioned in Section 3.3.9, the limiting factor for generation of huge datasets is most often given by a lack of labels and not of input data. This is also the case in the medical domain, where labels often have to be identified by expensive laboratory tests or highly trained medical experts. In contrast to that, unlabeled imaging data can be mined relatively easily from local data storage, like the picture archiving and communication system (PACS).

Inclusion of such unlabeled data by application of self-supervised learning was proven to be able to outperform supervised approaches in the natural imaging domain [116, 117]. Application of self-supervised learning has also been studied in the medical domain. Chen et al. [118] formulated a context restoration pretraining task by shuffling a given number of patches in the image and let the model learn to reconstruct the original image, Figure 4.1. While Azizi et al. [119] trained a contrastive approach, showing the model two augmented views of the same image with the objective to maximize agreement for both projections in a lower dimensional embedding space. However, further studies to test a variety of approaches and their performance on larger cohorts are needed.

4.4 Survival analysis

Typical problems in deep learning, especially in the area of image classification, involve tasks that require prediction of a fixed label, like training a model for classification of ImageNet. Such problem settings are also common in the medical domain. Grading of soft tissue sarcomas based on CT imaging would be an example. But, network output can also be continuous, e.g. for a model trying

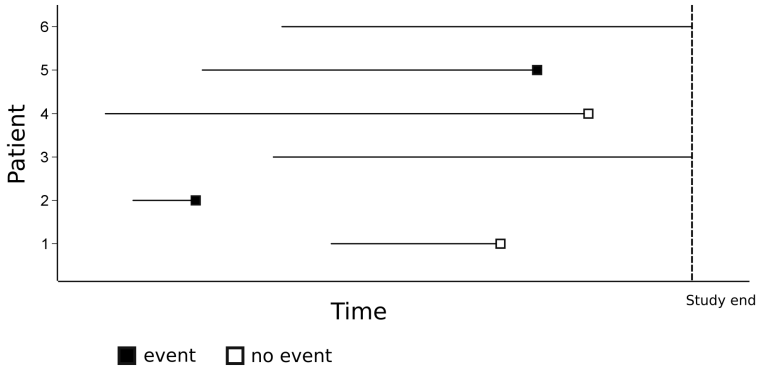


Figure 4.2: Example for *right censored* data. Patients 2 and 5 receive the event, e.g. death, during course of the study, while patients 1 and 4 are lost to follow up. Patients 3 and 6 did not receive the event before the study ended.

to predict exact tumor volume.

However, another type of modeling problem, encountered not so regularly in typical deep learning settings, is given by prediction of time dependent data. If the success of different treatment regimes should be analyzed, *time to failure* plays a critical role. It makes a fundamental difference if a cancer patient dies 5 years or 5 month after therapy. Therefore, binary modeling of survival vs. death would neglect the most essential information.

Real world survival data involves *censored* cases that cannot be handled by simple regression models. Censoring describes the involvement of data for which certain time spans are missing. In the medical domain, one usually has to deal with *right censored* data, Figure 4.2. That means that a last time point exists for which it is known that the *event* under consideration did not occur so far. However, after that time point no information about the event status is available. Example reasons for this are: end of the study or patients that drop out of the program, but also a lung cancer patient for whom progression free survival is studied, that dies on a heart stroke. Typical events of interested are, overall survival or cancer recurrence after therapy.

This type of data requires special models for handling. The two most commonly used quantities to describe such time to event problems are given by the survival and the hazard function. The survival function

$$S(t) = \Pr(T > t), \tag{4.1}$$

models the probability of an individual surviving beyond time t, with T representing the time until the event occurs, while the hazard function

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t}, \tag{4.2}$$

describes the momentary rate of occurrence for the event at time t, given that it has not occurred before [120].

A simple approach for estimation of the underlying survival function was developed by Kaplan and Meier [121]. The non-parametric approach assumes the survival function to take the form

$$S_{KM}(t) = \prod_{i:t_i \leq t} (1 - d_i/n_i), \tag{4.3}$$

with t_i the (ordered) observed event times, d_i the number of failures at t_i and n_i the number of subjects known to not have received the event before t_i [122]. However, the Kaplan-Meier estimator can only be applied in a descriptive manner, delineating the survival function based on observed data, but has no predictive power.

A widely applied approach able to be used in a predictive manner is given by the proportional hazards model of Cox [54]. The model assumes the hazard function for an individual j with covariates $\mathbf{x}_j$ to take the form

$$h_j(t|\mathbf{x_j}) = \lambda_0(t) \exp\{\mathbf{x}_j\beta\}, \tag{4.4}$$

with $\lambda_0(t)$ the baseline hazard function and $\exp\{\mathbf{x}_j\beta\}$ representing the relative risk associated with $\mathbf{x}_j$. The model does not take any assumption about the specific baseline hazard function $\lambda_0(t)$, but assumes a proportionality for the complete hazards h_j of all cases involved. It is semiparametric in the way that it makes parametric assumptions of the effect of covariates on the hazard function but no assumptions about the hazards function itself. Cox showed that a partial likelihood for the coefficients β is given by

$$L(\beta) = \prod_{T_i \text{ uncensored}} \frac{\exp\{\mathbf{x}_i\beta\}}{\sum_{T_j \geq T_i} \exp\{\mathbf{x}_j\beta\}}, \tag{4.5}$$

that can be treated as an ordinary log likelihood to derive valid maximum likelihood estimates of β. [122]

With eq. 4.5 being completely independent of $\lambda_0(t)$, Cox proportional hazards model allows for comparison of survival between different individuals without the need to determine the underlying baseline hazard function.

Several authors adopted this approach to be used in combination with deep neural networks (e.g. Katzman et al. [123] and Ching et al. [124]). However, the partial likelihood 4.5 depends on the order of all cases involved in the dataset. Therefore, the loss for such models can not be computed for single data points but depends on all instances in the dataset, which impedes usage of batches during optimization, a favorable procedure as described in Section 3.3.2.

A survival model approach fitting way more naturally to the structure and essence of deep neural networks was developed by Gensheimer and Narasimhan [22], Figure 4.3. They developed a discrete time survival model, with every output neuron of the network corresponding to the conditional probability of

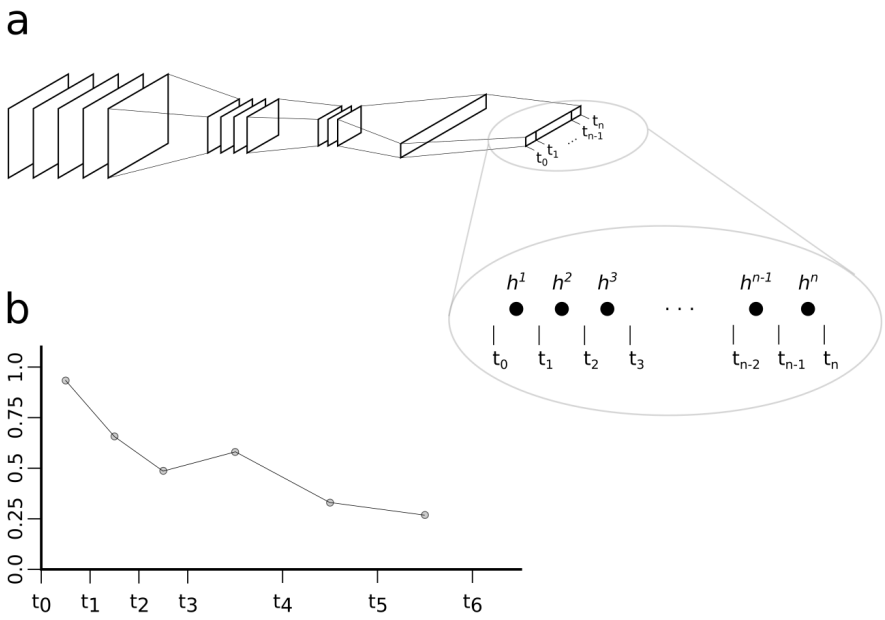


Figure 4.3: Nnet-survival model. The principle architecture for a CNN model is shown in a), neurons in the output layer reflect hazard probabilities for certain time intervals. Number and size of the intervals have to be chosen as a hyperparameter. In b) example output for a model with six neuron in the output layer is shown.

surviving a discrete time interval. Loss for time interval j of the model is then given by the negative log likelihood function

$$\sum_{i=1}^{d_j} \ln(h_j^i) + \sum_{i=d_j+1}^{r_j} \ln(1 - h_j^i), \tag{4.6}$$

with h_j^i the hazard probability for individual i during j and r_j the number of individuals not having experienced failure or censoring before the interval, with d_j of them suffering failure during the interval. The overall loss is given by the sum of the losses for all time intervals. Hence, the approach allows for loss computation on a individual level and therefore for an application of batches during training.

One of the most commonly used performance measures for survival models is given by the concordance index (c-index) [122]. Due to the involvement of censored cases, no order can be established between all cases, Figure 4.4. The c-index is a generalization of the area under the ROC curve that can be applied to censored output variables. It can be interpreted as the fraction of all pairs of subjects whose predicted survival times are correctly ordered among all subjects

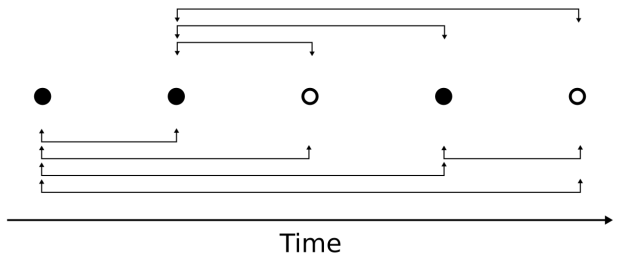


Figure 4.4: Ordering of censored data, following Steck et al. [125]. Empty circles represent right censored datapoints, and arrows the existence of a order between two datapoints. For censored datapoints, a order can only be established in relation to uncensored cases featuring a shorter survival time.

that can actually be ordered [125].

Chapter 5

Renal mass risk score prediction

Publicly held challenges allow for fair comparison between different algorithms and models and have now become the standard for validation of approaches in the biomedical image analysis domain [26]. Several biomedical imaging challenges were held in conjunction with the 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) [126]. Aim of the KNIGHT challenge was renal mass risk score prediction based on CT imaging and other clinical data [28]. Within this thesis, the performance of CNN models was tested in this competitive setting. Part of the studies presented in this Section were published in the challenge proceedings: "Risk Score Classification of Renal Masses on CT Imaging Data Using a Convolutional Neural Network", Lang et al. [127].

5.1 Introduction

Advances in medical imaging modalities like CT and MRI have led to an increase in detection of renal masses incidentally found during other workups. Renal masses depict an abnormal growth in the kidney, and are either solid or cystic. [128] Occurrences feature a large variety of behavior, ranging from benign lesions to aggressive carcinomas. Treatment options for management of clinically localized renal masses are diverse, including active surveillance, thermal ablation and radical or partial nephrectomy [129], but also radiation therapy in the form of SBRT is actively investigated in the management of renal masses [130, 131]. This diversity in characteristics and treatment options requires precise diagnostics, for selection of the right therapy regime.

Percutaneous biopsies can be utilized for determination of renal mass subtypes [132], but precision of the method remains controversial, especially for an application in small renal masses [133]. Moreover, advances in non-invasive methods like thermal ablation and SBRT are increasing the demand for non-invasive testing. Elderly patients with medical comorbidities are often admin-

istered to be kept under active surveillance, to be only treated after tumor progression [134]. Medical imaging based testing, able to detect aggressive cases, would allow non-invasive determination of patients that cannot be spared from treatment, without the need to wait for disease progression.

Findings of cystic renal masses on CT imaging can be categorized using the Bosniak classification system [135]. The risk for involvement of cancerous tissue in lesions is evaluated and patients are stratified into treatment regimes. However, the value of this system has been questioned [136, 137]. Leading Silverman et al. [138] to propose modifications in 2019 that are also expanding it to include findings on MRI data, but use in clinical practice is not widespread yet.

For solid renal masses, CT and MRI can be utilized for characterization based on lesion appearance. However, solid renal masses are often visually indistinguishable [139]. Roughly 15% of benign renal tumors that are smaller than 4 cm in size are classified as malignant, based on preoperative CT imaging [140]. The three nephrectomy scoring systems RENAL [141], PADUA [142] and centrality index [143] have been developed for standardized pre-operative assessment based on anatomical characterization, but come with the cost of labor intensive processing by medical experts.

Misclassification introduced by such testing methods can lead to incorrect decisions in patient management. Welch and Black [144] concluded that a substantial proportion of renal tumors identified as cancerous are *overdiagnosed*, either because they do not grow at all or because their growth is too slow for the tumor to cause symptoms before the patient dies of other causes. Therefore, a standardized well defined classification method, taking into account advances in non-extirpative treatment methods, like SBRT, is needed for renal mass staging.

Post-operative pathological classification is performed by the guidelines developed by the American Urological Association (AUA). The guidelines focus on the evaluation and management of clinically localized sporadic renal masses suspicious for renal cell carcinoma (RCC) in adults [130]. Classification involves the five risk groups: benign (B), low risk (LR), intermediate risk (IR), high risk (HR), and very high risk (VHR). Patients classified with high risk or very high risk scores should be treated with adjuvant therapy. Therefore, scores are grouped into the no adjuvant therapy (NoAT) class, involving benign, low risk and intermediate risk scores, and candidate for adjuvant therapy (CanAT) class, with high risk and very high risk cases, Figure 5.1.

A model able to predict those post-operative pathologically confirmed risk scores based on pre-operative imaging data would allow for precise testing that is non-invasive.

Task of the 2022 IEEE ISBI KNIGHT challenge was development of AI models able to identify patients' AUA risk score class based on pre-operative CT imaging and clinical data [28]. As presented in Section 3.3.6, convolutional neural networks depict a valuable approach to be used for analysis of imaging data. Within this thesis, ability of CNNs for AUA risk score classification based on CT imaging and additional clinical data has been tested and a segmentation

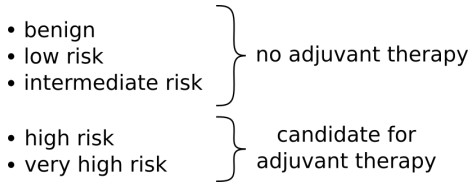


Figure 5.1: Risk scores for post-operative pathological classification of renal masses. For the three lowest classes adjuvant therapy is not recommended, while the two high risk classes mark candidates for adjuvant therapy.

model has been trained for identification of the rough ROI.

5.2 Material and methods

The challenge dataset involved 300 cases available for training and 103 cases for testing. Each case included an abdominal CT scan and clinical information about:

- age at nephrectomy,
- gender,
- body mass index,
- smoking history,
- age when quit smoking,
- pack years,
- chewing tobacco use,
- alcohol use,
- comorbidities,
- glomerular filtration rate,
- radiographic size,
- voxel spacing.

Training set cases also involved a respective AUA risk score classification label. Furthermore, as cases in the training set were already present in the MICCAI 2021 Kidney and Kidney Tumor Segmentation challenge (KiTS21) [145], segmentation maps for cystic, tumor and kidney tissue from different annotators were available for those instances.

Training set cases were divided into a train and a validation set using a stratified shuffle split, preserving the ratio of age, tumor size, voxel spacing in longitudinal direction, and AUA risk score in both data sets. The `model_selection.StratifiedShuffleSplit` function of the `sklearn` python module [146] was used to do so. Number of AUA risk score cases in the train and validation set are depicted in Table 5.1.

5.2.1 Segmentation of region of interest

Usually, CT images do not only involve information about respective regions of interest, essential for a specific task, but also depict surrounding tissue and

number of cases	benign	low risk	intermediate risk	high risk	very high risk
train set	19	119	40	32	40
validation set	6	23	8	8	5

Table 5.1: Number of AUA risk score cases in the train and validation set

body parts, up to the complete patient for whole body CTs. Removal of imaging parts not essential for the task of interest yields two advantages for deep learning models. First, graphics processing units (GPUs), used for training of deep neural networks, feature a finite amount of memory. This limits the number of trainable parameters that can be used, and therefore also the size of the network. Input size directly affects the number of parameters in a model. Hence, models with very large input volumes cannot be trained in a practical setting. Second, preprocessing of input data can depict a essential step in the training process. A machine learning model trained on a large input volume, with essential information only present in a small subvolume, has to learn to recognize the respective region of interest first, before it can learn the actual task of interest. So, cropping non-essential parts of the input volume helps to speed up learning and has the ability to improve models' performance. Of course, it has to be secured that no information essential for the learning task will be lost by the cropping process.

Therefore, abdominal CT images in the training and test set were cropped to only involve the region of interest, i.e. both kidneys and their surrounding tissue. However, delineations of the kidneys had to be known to do so. For the training set, segmentation maps of the KiTS21 challenge provided information about tissue of interest, but the test set was missing such labels. For generation of delineations in the test set, a U-Net model was trained. The KiTS21 data involved segmentation labels for cystic, tumor and kidney tissue and also delineations from different annotators. For construction of binary region of interest maps, segmentations labeling different tissue were merged taking the union, while segmentations from different annotators were combined using intersections.

Standardization and normalization of input features is a common practice in machine learning, facilitating model fitting and inference [77]. Often, image intensity values are rescaled to range between 0 and 1, such that randomly initialized weights of network layers and values in the input feature a common scale. Hounsfield units of kidney tissue are roughly given in a range from 20 HU to 40 HU, while fatty tissue ranges between -100 HU and -70 HU [32]. Taking this into account, intensity values of the images were cropped at -250 HU and 250 HU and then linearly rescaled to range from 0 to 1 for standardization.

Furthermore, all CT images were resampled to the same voxel spacing. The mean value of axial and longitudinal voxel spacing in the training set was given by 0.79 mm and 3.18 mm with respective 5-th and 95-th percentile values of 0.65 mm and 0.98 mm, and 0.5 mm and 5.0 mm. As voxel spacing is inversely

proportional to the dimensional extend of the images, a spacing of 1.5 mm × 1.5 mm × 3.0 mm has been chosen to keep input volumes small.

However, 3D imaging data still depicts a large input volume to be processed by U-Net like architectures. Intention of training of the segmentation model was rough orientation of the region of interest, but not detailed segmentation of different tumor tissues. Therefore, to keep model input small, a lightweight 2D architecture has been chosen for training of the segmentation model. Furthermore, CT and binary segmentation map slices have been center cropped to a patch size of 390 mm × 390 mm or respectively 260 × 260 pixels. During training, those input patches were then randomly cropped to a size of 256 × 256 pixels for data augmentation. Furthermore, from the augmentation techniques presented in Section 3.3.5, flipping on the coronal and sagittal plane and rotations by a multiple of 90° were performed.

The developed architecture followed the basic U-Net structure of Ronneberger et al. [13], depicted in Figure 3.15. However, zero padding was applied in all convolutional layers, to preserve spatial dimensionality. Furthermore, the number of down and up-sampling blocks and feature maps was reduced to prevent overfitting. The best performing model consisted of three down and respective up-sampling blocks of feature map size 32, 64, and 128. Optimization of model weights was performed using the Adam optimizer introduced in Section 3.3.2 with a learning rate of 1×10^{-4}. As a objective function the Dice loss, introduced in Section 3.3.7, was used. A fixed batch size of 12 was applied during optimization and all models were trained for 100 epochs.

During inference, all 2D slice level predictions were appended for construction of 3D segmentation maps.

The nature of CT imaging data is a three dimensional one, therefore, a model aiming for the best segmentation approach should employ 3D convolutional layers. The superiority of such an approach was proven by Isensee and Maier-Hein [147] in the KiTS19 challenge. They utilized of a residual 3D U-Net that achieved a Dice score of 0.974 and 0.851 for segmentation of kidney and tumor tissue. However, aim of the approach presented here was development of a lightweight segmentation model, able to roughly identify both kidneys. As construction was limited by the GPU available for training, given by a NVIDIA M60 card.

5.2.2 AUA risk score classification

Preprocessing of CT data for the AUA risk score classification model was performed by resampling all cases to a uniform voxel spacing of 0.75 mm × 0.75 mm × 2.5 mm, while rescaling of intensity values was performed in the same way as for the segmentation approach. Segmentations were utilized to identify center points of both kidneys in the CT images. For the train and validation set, the binary segmentation maps provided were used to do so, while for the test set segmentations generated with the U-Net model were utilized. Two patches of both kidneys featuring a size of 120 mm × 120 mm × 140 mm or respectively 160 × 160 × 56 pixels, were cropped from the CT images using those center

points for orientation.

Desired model output was given by AUA risk scores. Which are not completely independent from each other, but depicted a increase in severity. A ground truth *very high risk* case, classified as *benign*, represents a bigger problem than the same case being classified as *high risk*. Such a relation between labels is not reflected in the crossentropy loss, which depicts the standard loss utilized for multi-class classification problems. Therefore, the problem was handled in a regression manner. AUA risk scores were transferred into a stepwise continuous score ranging from 0 to 1 in steps of 0.25, which in terms of the vector $\mathbf{L}$ can be expressed by

$$\mathbf{L} = \begin{pmatrix} L_{\text{benign}} \\ L_{\text{low risk}} \\ L_{\text{intermediate risk}} \\ L_{\text{high risk}} \\ L_{\text{very high risk}} \end{pmatrix} = \begin{pmatrix} L_0 \\ L_1 \\ L_2 \\ L_3 \\ L_4 \end{pmatrix} = \begin{pmatrix} 0.0 \\ 0.25 \\ 0.50 \\ 0.75 \\ 1.0 \end{pmatrix}. \tag{5.1}$$

Clinical features essential for the task were identified by training of a random forest model and computation of mean decrease in impurity (MDI) [148]. MDI, also referred to as Gini index, provides a measure to calculate feature importance in random forest models [149]. The classifier consisted of 100 trees with a maximal depth of 4 layers. All variables with a Gini importance larger than 1% were selected to be included in the deep learning model. That were: *radiographic size*, *body mass index*, *age a nephrectomy*, *metastatic solid tumor* and *chronic kidney disease*. With *metastatic solid tumor* and *chronic kidney disease* given as binary scores.

The deep learning architecture developed for AUA risk score classification followed the principle structure of a Siamese neural network, with both kidney patches as an input. Siamese neural networks are usually employed for similarity computation between features of two inputs, which are retrieved using the same extractor [150]. However, for the two kidney patches no similarity was computed but the Siamese structure was used for construction of a model featuring a relatively small amount of trainable parameters, preventing the architecture from overfitting. A delineation of the architecture is depicted in Figure 5.2.

For feature extraction, a 3D convolutional part has been employed, featuring convolutional and max pooling layers. Kernels of size $3 \times 3 \times 3$ with a stride of $1 \times 1 \times 1$ were used in convolutional layers, feature maps were zero padded by one pixel in every dimension. The ReLU function, eq. (3.5), was used as an activation. For max pooling layers, again kernels of size $3 \times 3 \times 3$ were used, and stride was given by $2 \times 2 \times 2$, reducing spatial dimensionality by a factor of 2. Following the Siamese structure, two input paths with shared weights were used to process both kidney patches. Resulting feature vectors were merged using the element-wise maximum.

The decision for taking the element-wise maximum as a fusion method was made based on the fact that for classification it only matters if a given pattern is present, but not in which patch it is contained. Another merging method would

be given by stacking of feature vectors. However, this would increase the output size of the convolutional part by a factor of two, resulting in a larger network, more prone to overfit. After merging of feature vectors, dense layers were used for classification. Clinical features, identified as essential, were merged with deep features. The architecture ended in one output neuron modeling the risk score developed before.

Dropout, introduced in Section 3.3.5, and ReLU activations were applied after each hidden dense layer. Due to the decision for a regression-like handling, mean squared error was used as an objective function to be minimized using the Adam optimizer [83].

Data augmentation has been applied in order to prevent the model from overfitting. Utilized augmentations were similar to the ones used for training of the segmentation model, i.e. patches were randomly cropped to a size of $128 \times 128 \times 48$ voxels, flipped on the coronal and sagittal plane and rotated by a multiple of $90°$ on the axial axis.

Validation set loss was used as a measure for optimization of hyperparameters, which were tuned manually in a grid-like manner. Hyperparameters modified during this process included: size of convolutional feature maps, number and size of fully connected layers, layer at which clinical parameters are added, dropout rate, learning rate and batch size. All models were trained for 500 epoch. The model found to perform best featured three convolutional layers of size 16, 32, and 64, and four dense layers with output sizes of 256, 128, 128, and 1. Clinical features were added after the second dense layer. Dropout rate for all hidden dense layers was given by 0.25. Batches involved 8 cases and the learning rate was given by 1×10^{-4}.

As described above, to account for the increase in severity of AUA classes, output of the model was chosen to be scalar valued. However, for final classification the distance z_j between model output y and each label L_j has been computed

$$z_j = \|L_j - y\|, \tag{5.2}$$

with $\|\cdot\|$ depicting the euclidean norm. The softmax function was then used for construction of normalized scores

$$p_j = \sigma(-\mathbf{z})_j, \tag{5.3}$$

reflecting the probability of each case to belong to a certain class.

The AUA risk score classification model received input from CT and clinical data. Contribution of both inputs to the final output was studied by retraining the model based on one input modality only, i.e. taking only CT or only clinical data as an input. In case of the model taking only input from imaging data, input from the clinical features was simply removed from the model. While for the model taking only clinical data as an input, the architecture was kept the same, but CT input patches were randomly shuffled, such that no valuable information could come from the imaging path. As ground truth class labels were only available for the train and validation set, influence of input sources could only be determined for those two datasets, but not for the test set.

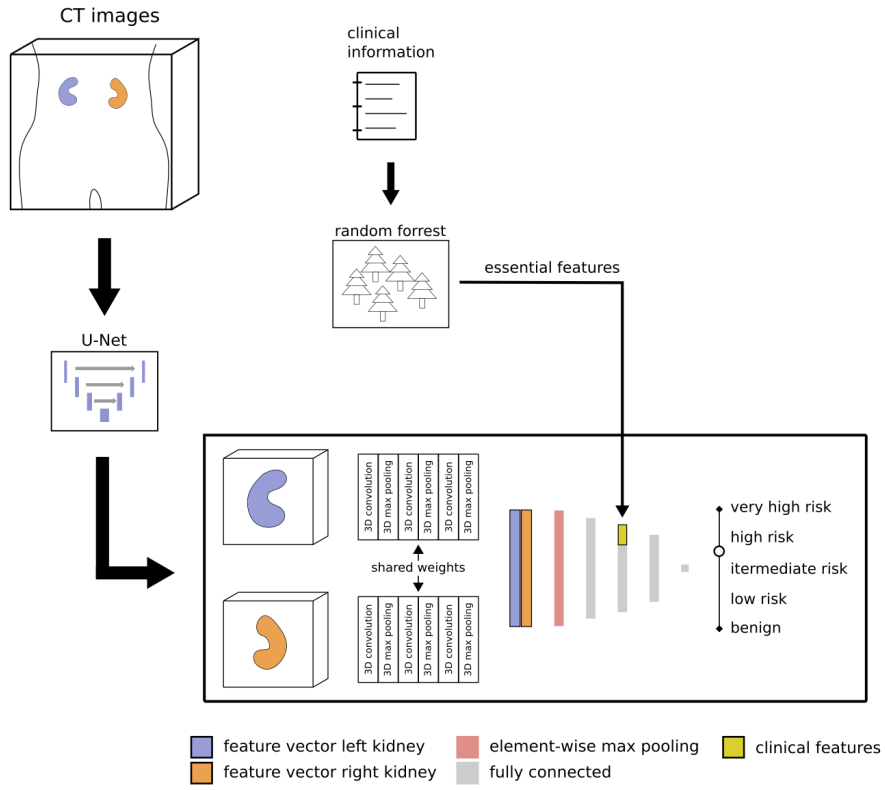


Figure 5.2: Architecture developed for AUA risk score classification in renal masses, adopted from Lang et al. [127]. Regions of interest for both kidneys were identified by training of a U-Net like architecture. Two patches, one for each kidney, were then cropped from the abdominal CT images. Those patches were then processed by a model featuring a Siamese like structure, i.e. two input path with shared weights. Resulting feature vectors were merged using the element-wise maximum and further processed by dense layers. Importance of clinical features were measured by training of a random forest classifier and calculation of mean decrease in impurity. Essential clinical features were merge with imaging features before construction of the output score.

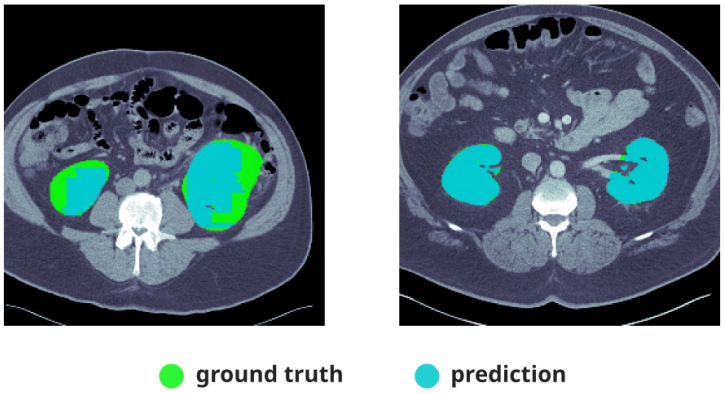


Figure 5.3: Example segmentations from the validation set. The left image shows a slice of the worst performing case in the validation set featuring a Dice score of 0.379, while on the right hand side a slice of the best performing example, with a Dice score of 0.977, is depicted.

5.3 Results

The segmentation model achieved a mean Dice score of 0.951 and 0.903 on the train and validation set. Performance on the test set could not be evaluated due to missing labels. Example predictions for the best and worst performing validation set cases are shown in Figure 5.3.

Performance on the binary discrimination between NoAT and CanAT cases was used as a measure to rank participants of the KNIGHT challenge, while detailed AUA score prediction performance, discriminating between the five risk classes, was only captured but not utilized. AUC has also been chosen as a metric for AUA score performance, comparing one risk class with all the others, e.g. discriminating between *low risk* and all other risk scores. This resulted in five AUC scores, one for each class. Those five scores were than merged using the mean, for generation of a single metric.

For the NoAT vs. CanAT task the model was able to achieve a training, validation and test set AUC score of 0.896, 0.865 and 0.814. Results for the detailed AUA risk score prediction can be seen in Table 5.2. Mean AUC scores on train, validation and test set were given by 0.836, 0.752 and 0.676. The developed model was placed second in the final challenge ranking. Top 4 results can be see in Table 5.3.

Performance of the developed model for input coming from different resources can be seen in Table 5.4, while validation set learning curves are shown in Figure 5.4.

AUC-scores	benign	low risk	intermediate risk	high risk	very high risk	mean
training	0.923	0.828	0.703	0.820	0.907	0.836
validation	0.674	0.802	0.640	0.753	0.889	0.752
test	0.606	0.763	0.524	0.656	0.830	0.676

Table 5.2: AUC scores for all risk classes. Scores are computed based on binary detection of one class in contrary to all other classes.

Rank	Team	AUC task 1	AUC task 2
1	agentili	**0.841**	0.529
2	*Helmholtz_IRM*	0.814	**0.676**
3	Taiyuan_University_lab_713 [151]	0.813	0.626
4	Medal [152]	0.808	0.646

Table 5.3: Ranking of the KNIGHT challenge. The developed model, submitted under the team name *Helmholtz_IRM*, was placed second, based on NoAT vs CanAT AUC score performance.

	AUC NoAT vs. CanAT		mean AUC risk score classification	
	training	validation	training	validation
full input	0.896	0.865	0.836	0.752
image only	0.739	0.688	0.650	0.607
clinical only	0.864	0.790	0.747	0.698

Table 5.4: AUC scores for NoAT vs. CanAT classification and mean AUC scores for AUA risk score classification with input coming from different resources.

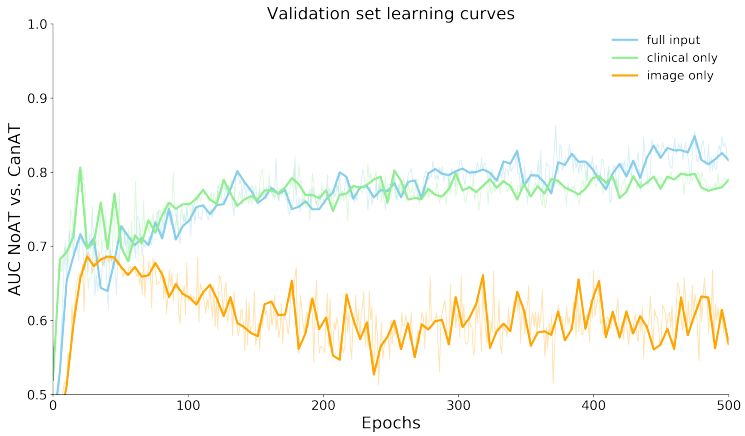


Figure 5.4: Smoothed validation set learning curves for input coming from different resources, taken from Lang et al. [127].

5.4 Discussion

Performance of the segmentation model could not be evaluated on the test set. However, the generalization gap of segmentation models is generally small. Isensee and Maier-Hein [147] reported the same validation and test set Dice score of 0.97 for delineation of kidney tissue in the KiTS19 challenge. Therefore, assumption of a low generalization error for the developed U-Net architecture is reasonable. Objective of the segmentation model was identification of the region of interest, but no exact pixel-wise differentiation between tissue. Accordingly, validation set performance given by a Dice score of 0.903 was valued sufficient to be used to crop patches of both kidneys from the CT images.

For binary discrimination between NoAT and CanAT cases the model was able to achieve a significant test AUC score of 0.814. Performance was relatively stable with respect to train and validation results. In contrast to that, detailed AUA risk score classification performance experienced a considerable drop when probed on the test set. Different mechanisms can lead to such behavior.

In general, the AUC metric is invariant to any global scaling or offset in the data. However, due to utilization of the AUC metric in the non-binary setting, comparing each class with all the other classes, output scores have to be associated with a given class. The developed model does this by computation of the distance between output scores and the label vector $\mathbf{L}$, eq. (5.2). Dependence on $\mathbf{L}$ breaks the global shift invariance of the AUC score, which is usually only used for evaluation of pure binary classification tasks. Deep learning models are known to be sensitive to distribution shifts between two datasets. Differences in scanners or scan protocols, but also discrepancy in pathologies and anatomies, can cause such shifts [25], which in turn can lead to global offsets in the output,

causing a decline in performance.

Another factor limiting the generalization capability of the model is given by the small dataset size. Only 300 cases were available for training, which had to be further separated into a train and validation set. This led to involvement of less than 10 cases for all classes, except for the low risk class, in the validation set. For such a low number of cases, the validation set only reflects the underlying data distribution to a limited amount. Hence, generalization error is not properly reflected by the set, which increases the possibility for selection of a suboptimal final model, resulting in a drop in performance.

Task of the KNIGHT challenge was prediction of risk classes based on "clinical Computed Tomography (CT) imaging of the kidneys" [28], but next to imaging data also clinical information was available for training. The influence of both modalities was tested by training of separate models. Performance of the model relying solely on imaging was inferior to the model trained with both inputs. The *image only* validation curve in Figure 5.4 shows signs of overfitting after 50 epochs. However, the curve also depicts a increase in AUC performance after training is started and reaches a score of 0.688. This suggests inclusion of information essential for the prediction task in imaging data. The model trained on clinical information only showed similar initial behavior than the complete model, but superior performance of the full model can be recognized after a few hundred epochs. This superiority in terms of validation performance is also present in both AUC measures shown in Table 5.4. Therefore, those findings suggest main contribution of clinical input data for formation of the output score, but a non-neglectable input coming from imaging. However, more studies on larger unbiased datasets are needed for evaluation.

The developed model was placed second based on the NoAT vs CanAT AUC score ranking of the challenge. For the more detailed task of AUA risk score classification, better performance than other participants could be achieved. The winning contribution reported no utilization of imaging data for their model, but no details of the developed model have been published. The approach ranked 3^{rd}, developed by Chaudhary et al. [151], also reported no advantage from inclusion of CT data. They tested the ability of different CNN architectures for extraction of imaging features, that have been fused with clinical information for prediction. However, their final model was based on clinical features only, employing a network specifically designed to handle tabular data, by application of attention layers [153]. Notably, they did input whole CT images, without cropping around the region of interest, for their imaging models. Varsha et al. [152], placed 4^{th}, trained a nnU-Net model for segmentation of detailed tissue delineations, provided by the KiTS19 dataset. Latent space features from a U-Net, included in the model, were utilized to extract information from the CT images. Those features were reduced using principle component analysis (PCA) and then fused with clinical features to be processed by an classification head featuring an attention layer. Influence of clinical and imaging features on the prediction were not investigated.

With other approaches reporting no influence of imaging data on model performance the need for further studies to investigate influence of both input

modalities is reaffirmed. However, Chaudhary et al. [151] used the whole CT volume as an input, which limits model's ability, as pointed out above.

Dataset size was very limited containing 300 training cases only. Further studies should be performed on larger cohort sizes, investigating the influence of different scanner types and the contribution coming from imaging and clinical input in more detail. Merging of both input patches was performed by computation of the element-wise maximum, but also other methods like taking the average or mean instead of the maximum are possible. These methods should also be further investigated.

5.5 Conclusion

A 3D CNN architecture for classification of AUA risk scores based on CT imaging and clinical data was presented. A Siamese network in combination with a element-wise fusion layer was developed. The architecture allows for reduction of trainable parameters and therefore prevents the model from overfitting. The network was tested in a competitive setting by participation in a public challenge. Significant model performance for a discrimination between relevant therapy classes was achieved. Further studies for investigation of influence of different input modalities are needed.

Chapter 6

HPV status in oropharynx cancer patients

6.1 Introduction

Human papilloma viruses (HPVs) are small DNA viruses. Over 120 different subtypes have been discovered with approximately one-third of them infecting the squamous epithelia of the genital tract [154]. Distinct high risk subtypes have been identified that were proven to be causative agents of cancer [155]. This carcinogenic character could first be demonstrated by Harald zur Hausen [156, 157], who was awarded with the Nobel Price in medicine in 2008 "for his discovery of human papilloma viruses causing cervical cancer" [158]. The subtypes HPV-16 and HPV-18 are the two most carcinogenic ones, responsible for about 70% of cervical cancers [159].

Subsequent studies were able to associate high risk HPVs with many penile, vulva and anal carcinomas and also oral cancers [155]. Infection with HPV-16 was identified as a risk factor for head and neck squamous cell carcinoma (HNSCC) [160]. HNSCCs feature a large variability and can be categorized into subtypes that differ with respect to risk factors, pathogenesis, and clinical behavior [161]. Squamous cell carcinomas originating in the oropharynx (Figure 6.1), a part of the pharynx which in turn is part of the throat, depict one such subtype. Gillison et al. [163] were able to show that HPV positive oropharyngeal cancers comprise a distinct molecular, clinical, and pathological disease entity that has a markedly improved prognosis. Incidences of HPV positive oropharynx cancer cases are on the rise [164], in 2016 about 30% of oropharynx cancer (OPC) cases were caused by an infection with HPV [165]. However, HPV positive OPCs feature a better responsiveness to chemotherapy and radiation treatment [166]. Therefore, testing for an infection with HPV depicts a essential diagnostic factor in OPC patients. Most frequently, immunohistochemistry staining for p16 is used as a surrogate marker of HPV, featuring a sensitivity of >90% and a specificity of >80% [167, 168]. Detection of HPV E6 and E7 messenger RNA

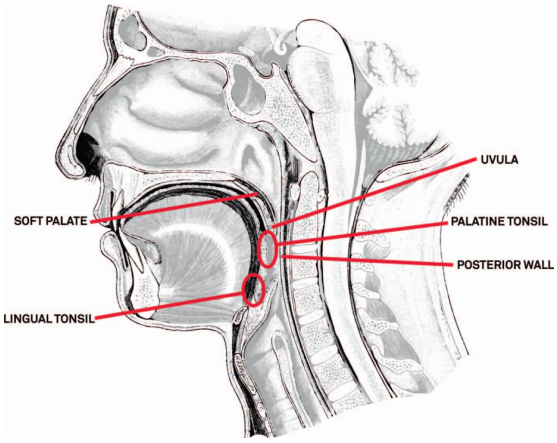


Figure 6.1: Anatomy of the oropharynx, taken form Lewis Jr et al. [162].

by in situ hybridization is often considered as the gold standard, but is difficult to be utilized in clinical settings. Hence, application of the best testing method remains controversial. [162]

Different radiomics studies have proven the ability of machine learning algorithms for detection of patients HPV status based on CT imaging data. Huang et al. [169] trained a logistic regression classifier on 113 HNSCC patients from The Cancer Genome Atlas, 540 quantitative image features were extracted from which five were selected for final model construction. When tested on 53 patients of an external data set, the model achieved an AUC of 0.76. Bogowicz et al. [170] utilized a logistic regression classifier in a privacy preserving way on 975 patients from 6 different cohorts. Six different models were respectively trained on data of five centers, while testing was done on the remaining center. In total, 981 radiomics features were extracted. Feature selection was performed in a centralized manner and in a privacy preserving distributed manner. Final models relied on input from 26-30 radiomics features, reaching AUC scores ranging between 0.69 and 0.82. Fujima et al. [171] finetuned a CNN model, pretrained on ImageNet, for classification of patients HPV status based on PET/CT imaging data, i.e. using a 2D approach. The model was trained on 2160 input slices and achieved an AUC score of 0.83 on the test set, but no external testing was performed and slices containing artifacts or tumors with diameters smaller than 1.5 cm in size were excluded from the dataset.

However, Starke et al. [19] have proven the superiority of 3D over 2D CNN approaches for outcome modeling in HNSCC. Therefore, investigating the potential of deep learning architectures, 3D models for prediction of patients HPV status based on CT imaging data were developed within this thesis.

First, a transfer learning approach relying on video classification for pretraining, and therefore allowing for three dimensional input in the downstream task, was developed. To do so, the convolutional part of the C3D model of

Tran et al. [95] was utilized as a feature extractor. C3D was trained for sports video clip classification on the Sports-1M data set [172]. For comparison, a 2D transfer learning approach using the aforementioned ImageNet pretraining strategy was tested. Hussein et al. [173] utilized the C3D model for lung nodule risk stratification based on CT imaging, but their network relied on additional radiographic information like tumor sphericity and texture while the network developed in this thesis was trained end-to-end, i.e. only relying on imaging input.

Next to transfer learning, self-supervised learning can be used to enable models to be trained on small datasets, cf. Section 3.3.9. Studying the ability of such self-supervised approaches in the domain of medical imaging data, the masked autoencoder (MAE) model of He et al. [20] was modified to be able to process 3D data. The architecture relied on the modern concept of transformers, introduced in Section 3.3.10. Zhou et al. [174] studied the ability of a MAE based pretraining approach to be applied on X-Ray, CT and MRI data. However, studies on 3D medical imaging data, i.e. CT and MRI, only involved segmentation problems, but no classification downstream task has been evaluated.

All investigations were performed on publicly available datasets in order to allow for reproducible research. Code of the transfer learning approach was published online [1]. Part of the studies performed on transfer learning were published in: "Deep learning based hpv status prediction for oropharyngeal cancer patients", Lang et al. [175].

6.2 Material and methods

6.2.1 Transfer learning

The radiomics and deep learning studies presented above, aiming for HPV prediction based on CT imaging data, relied on datasets not available to the public. As discussed in Section 2, application of such private datasets hinders reproducible research. Therefore, the publicly accessible [2] "The Cancer Imaging Archive (TCIA)" [27] was mined for appropriate cohorts containing CT images of head and neck cancer cases. TCIA is an online repository that comprises publicly shared cancer imaging data.

Four head and neck cancer cohorts could be identified including ground truth information about patients' HPV status, namely the *OPC-Radiomics* [176, 177], *HNSCC* [178, 179], *Head-Neck-PET-CT* [180, 181] and *Head-Neck-Radiomics-HN1* [55, 182] cohorts. Those cohorts were scanned for appropriate cases applying the inclusion criteria: oropharyngeal subtype, existence of a pre-treatment

[1] https://github.com/LangDaniel/hpv_status

[2] At the point of this writing head and neck cancer cohorts of the TCIA archive were freely available. In the meantime, due to privacy protection concerns a restricted license agreement has to be approved by the provider, as head and neck CT data could potentially be used for reconstruct of the human face.

	training set		validation set	test set
	OPC	**HNSCC**	**HN PET-CT**	**HN1**
Patients	412	263	90	80
HPV: pos/neg	290/122	223/40	71/19	23/57
HPV status				
Age				
pos	58.81 (52.00-64.75)	57.87 (52.00-64.00)	62.32 (58.00-66.00)	57.52 (52.00-62.50)
neg	64.82 (58.00-72.75)	60.02 (54.50-67.25)	59.11 (49.50-69.50)	60.91 (56.00-66.00)
Sex: Female/Male				
pos	47/243	32/191	14/56	5/18
neg	34/88	15/25	4/15	12/45
T-stage: T1/T2/T3/T4				
pos	46/93/94/57	60/93/41/29	10/37/15/9	4/8/9/8
neg	9/35/43/35	6/12/12/10	3/4/8/4	9/16/9/23
N-stage: N0/N1/N2/N3				
pos	33/22/215/20	19/30/170/4	11/10/47/3	6/2/15/0
neg	36/16/62/8	5/2/31/2	2/1/13/3	14/10/31/2
Tumor size $[cm^3]$				
pos	29.35 (10.52-37.78)	11.78 (3.94-14.04)	34.63 (14.91-41.77)	23.00 (10.83-34.29)
neg	36.99 (15.72-45.35)	23.57 (5.80-22.85)	35.09 (17.32-47.82)	40.19 (11.77-54.42)
transversal voxel spacing $[mm]$	0.97 (0.98-0.98)	0.59 (0.49-0.51)	1.06 (0.98-1.17)	0.98 (0.98-0.98)
longitudinal voxel spacing $[mm]$	2.00 (2.00-2.00)	1.53 (1.00-2.50)	2.89 (3.00-3.27)	2.99 (3.00-3.00)
manufacturer				
GE Med. Sys.	272	238	45	0
Toshiba	138	3	0	0
Philips	2	12	45	0
CMS Inc.	0	0	0	43
Siemens	0	4	0	37
other	0	6	0	0

Table 6.1: Cohorts used for training of the deep learning architectures, taken from Lang et al. [183]. Error margins depict 25^{th} and 75^{th} percentiles.

CT image, availability of a GTV segmentation, and tested HPV status. This led to a data set size of 850 individual patients, depicted in Table 6.1.

HPV testing was performed by IHC staining of p16 in the *OPC-Radiomics* and *Head-Neck-Radiomics-HN1* cohorts, while a combination of IHC and HPV DNA in situ hybridization was performed in the *HNSCC* cohort. Testing methods for the *Head-Neck-PET-CT* were not reported.

As for the risk score prediction approach in Section 5, CT images were cropped to patches of smaller size using the center of the ROI, given here by the GTV. All CT images were resampled to a uniform voxel size of 1.0 mm $\times$ 1.0 mm $\times$ 1.0 mm. Intensity values, given in Hounsfield units, were clipped at -250 HU and 250 HU, extending the typical clinical larynx Hounsfield-window given by a center of 50 HU and a width of 250 HU [32]. Clipped values were than linearly rescaled to range from 0 to 255.

The two largest cohorts of the data set, i.e. *OPC-Radiomics* and *HNSCC* were employed for training, while *Head-Neck-PET-CT* was utilized as validation set and *Head-Neck-Radiomics-HN1* as test set.

Hence, 675 cases were available to train the model. As illustrated in Chapter 3.3, deep learning models commonly require datasets of larger size, but actions to overcome this need and prevent model overfitting can be applied. Next to data augmentation and regularization methods, transfer learning can be applied to enable training on sparse data. Often, architectures pretrained on ImageNet are finetuned on the desired task, as done by Fujima et al. [171] for HPV classification on PET/CT imaging data. However, as ImageNet data is two dimensional the approach only allows for 2D input in the downstream

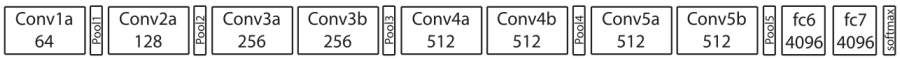


Figure 6.2: C3D model, taken from Tran et al. [95]. The architecture utilizes eight 3D convolutional and five 3D max pooling layers in the feature extraction part. While two fully connected hidden layers, followed by a softmax output layer are employed for classification. Convolutional layers feature kernels of size $3 \times 3 \times 3$ and a stride of 1 in all dimensions. Max pooling layer *pool1* employs a kernel size and stride of $1 \times 2 \times 2$, the two other pooling layers feature kernels and strides of size $2 \times 2 \times 2$. The output layer comprises 487 neurons, number of feature map sizes and neurons in dense hidden layers are denoted in the Figure.

task, i.e. slices of PET/CT images. In contrast to that, a 3D classification task utilized in pretraining would allow for full exploitation of 3D information in the downstream task. One such task is given by classification of video data. Video data features 2D images from different timepoints, and has therefore two spatial and one temporal dimensions. Video clip classification depicts a task commonly encountered in the area of deep learning, with different dataset of appropriate size available [172]. Hence, classification of video data has been chosen as a pretraining task in this thesis.

Training architectures on large 3D datasets is time-consuming and requires appropriate computing resources. Therefore, instead of training the video classification model from scratch, weights of a already trained model were used. Next to a reduction of computational workload, this approach also comes with the advantage of utilization of a reproducible starting point.

A suitable pretrained model was given by C3D [95]. C3D processes its input data in a 3D convolutional manner, i.e. all dimensions are handled equally, by application of 3D convolutions and 3D max pooling layers. A schematic of the architecture can be seen in Figure 6.2. The model was trained for classification of YouTube sports video clips given in the Sport-1M dataset [172] featuring 1.1 million videos of 487 sports categories. Weights of the trained model are publicly available [184].

The model was originally pretrained on colored RGB video clips of spatial size 112×112 featuring 16 time frames, i.e. input size was given by $3 \times 16 \times 112 \times 112$. To account for the smaller size in the temporal dimension, the first max pooling layer featured a kernel size and stride of $1 \times 2 \times 2$, to not reduce the temporal dimension too early in the architecture. The two other max pooling layers had kernels and strides of size $2 \times 2 \times 2$. For all convolutional layers, kernel size was given by $3 \times 3 \times 3$ and stride by $1 \times 1 \times 1$, with zero padding of $1 \times 1 \times 1$ applied for preservation of spatiotemporal dimensions. ReLU was used as an activation function for all convolutional layers.

All dense layers were removed from the model and replaced with randomly initialized ones of appropriate size. The basic structure of the weight transfer can be seen in Figure 6.3. Weights of all convolutional layers were kept fix during training. All hidden dense layers were followed by a dropout layer, ReLU was

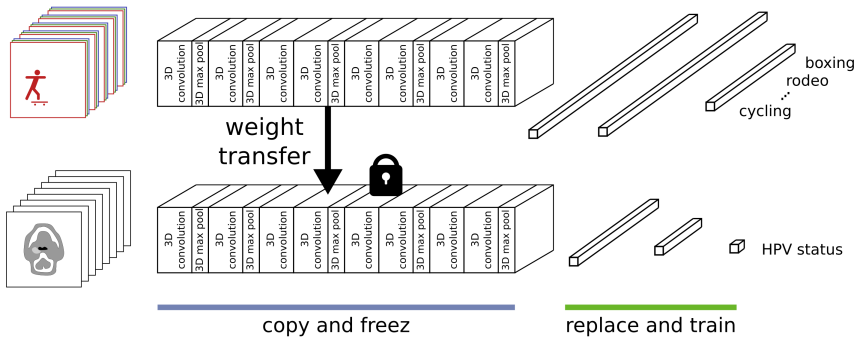


Figure 6.3: Transfer learning approach. Weights of the 3D convolutional part of the C3D architecture, shown in the upper part, are transferred to the HPV classification network and kept fix during training. Dense layers in the classification head are replaced with randomly initialized ones of appropriate size. Utilization of 3D layers in the pretraining step allows for full exploitation of 3D information in the downstream task of CT imaging based HPV detection.

again used as an activation function. The output layer consisted of a single neuron followed by Sigmoid activation for binary classification (c.f. Figure 3.5).

Input size of the downstream architecture was kept the same as for the pretrained model. Patches of size $112 \times 112 \times 48$ were cropped from the CT images and then rearranged to a shape of $3 \times 16 \times 112 \times 112$, such that the input requirements of the pretrained model were fulfilled. To do so, three consecutive CT slices were fed to the color channels.

Augmentation techniques applied during training involved, flipping patches on the coronal and sagittal plane, rotation by a multiple of 90° and shifting of the orientation point, utilized for cropping, by a value between 0 and 7 pixels in both directions of the transverse plane.

The weighted binary crossentropy loss was minimized using the Adam optimizer. Hyperparameter optimization was performed based on the validation loss. Manual variations involved adoption of: number and size of dense hidden layers, dropout rate, batch size and learning rate. The best performing model featured two hidden dense layers of size 1024 and 64, with a dropout rate of 0.35 and 0.25, respectively. Learning rate was given by 10^{-4} and batch size by 16. All models were trained for 200 epochs.

For comparison, a 3D CNN architecture was trained from scratch and the transfer learning approach relying on a ImageNet pretrained model was tested. Augmentation techniques and optimization strategies stayed the same as for the C3D pretrained model, only the size and number of trainable layers were reduced to prevent model overfitting.

The 3D CNN trained from scratch followed the same basic structure as the C3D model, i.e. 3D convolutional layers and max pooling were utilized

Figure 6.4: CNN model, adopted form Tran et al. [95]. Convolutional layers featured kernels of size $3 \times 3 \times 3$ and a stride and zero padding of $1 \times 1 \times 1$. *Pool1* utilized a kernel and stride of $1 \times 2 \times 2$, all other pooling layers had kernels and strides of $2 \times 2 \times 2$. Dropout rate was given by 0.25 for all layers. Feature map sizes and number of neurons for dense layers are denoted in the boxes. All hidden layers used ReLU as an activation function. The architecture ended in one output neuron with a Sigmoid activation.

for feature extraction and dense layers employed for classification, Figure 6.4. However, dropout was already applied after the last pooling layer to further reduce models' capability for overfitting. The best performing model featured convolutional layers of size 16, 32, 32, 64, 128, and 128, and two fully connected layers of size 256 and 128. Dropout rate was given by 0.25 for all layers. Again, ReLU was used as an activation function for all hidden layers. Also, input patches were cropped from the CT images in the same way as for the C3D based model. However, no rearrangement into color channels was performed, i.e. model input featured a size of $48 \times 112 \times 112$ voxels.

For the ImageNet pretrained architecture, the VGG16 model of Simonyan and Zisserman [94], depicted in Figure 3.12, was chosen. All dense layers were removed from the model and replaced with randomly initialized ones of appropriate size. Weights of convolutional layers were kept fix during training, as for the C3D based model. Patches were again cropped in the same way as before. However, ImageNet pretraining requires 2D input from three color channels. Therefore, patches were split into 16 slices of size $3 \times 112 \times 112$, i.e. three consecutive slices were input to the color channels of the model. Slices were individually processed during training and validation. During testing, an overall prediction score was constructed taking the mean of those 16 slice level predictions. The best performing model employed two dense layers of size 512 and 64, followed by a dropout layer of rate 0.5 and 0.25 and a ReLU activation layer. Data augmentation was performed as before.

Due to the limited size of the test and validation set, all three models were trained ten times with the same hyperparameter settings. In this way, a rough estimate of performance error could be introduced, enabling better comparability.

6.2.2 Self-supervised learning

Transfer learning assumes patterns learned in the pretrain task to be of value in the downstream task. Even though Fujima et al. [171] have proven the principle power of natural imaging data based pretraining for a CT based HPV classification, both domains are not that similar. Therefore, the question if pretraining on a more similar task leads to improvements in performance arises.

Self-supervised learning provides an opportunity for in-domain pretraining.

Ability of the masked autoencoder (MAE) architecture of He et al. [20], for a CT based HPV classification was studied in this thesis. The model was proven to outperform other self-supervised training approaches in the natural imaging domain. MAE relies on the transformer models introduced in Section 3.3.10, namely the vision transformer (ViT) model developed by Dosovitskiy et al. [110]. As depicted in Figure 3.18, a fraction of the input patches [3] is removed/masked. Self-supervised task is given by restoration of the uncorrupted image, utilizing mean squared error (MSE) as a loss function in pixel space.

He et al. [20] found that masking of a very high portion of patches works best for the approach. For their final model they removed 75% of patches in the input. Transformer architectures feature lots of parameters, as attention layers compute relations between all tokens in the input. This leads to a large demand for GPU memory during training. To handle this, ViT models embed whole patches and not single pixels, but for 3D data even insertion of patches leads to an extensive memory demand. As MAE removes a large portion of its input patches and only processes parts of them, the architecture depicts a perfect self-supervised transformer model candidate to be applied onto 3D medical imaging data.

The vision transformer approach [110] introduced three different variants of ViT models, featuring 12, 24, and 32 layers (layer in this context refers to the gray block depicted on the right hand side of Figure 3.20). Those sizes were also tested in the encoder of MAE [20]. Due to the memory demand of 3D data, only the *base* model comprising 12 layers was studied in this thesis. The developed approach will be called *masked autoencoder for medical imaging* (MAEMI) in the following. Furthermore, the dimension of the latent space was reduced from 512 to 384 and the depth of the decoder from 8 to 4 layers. Code for the publicly available MAE model [185] was modified to handle medical imaging data of 3D grayscale format. During construction of the MAEMI model, Feichtenhofer et al. [186] modified MAE in a similar way to be trained on video recognition data, like Kinetics-400 [187], but did not publish their code. However, they observe a even larger masking ratio as high as 90% to work best, associating it with a higher information redundancy in video data. Following those findings, a masking ratio of 85% was probed for pretraining of the MAEMI model, assuming the information redundancy in CT imaging to be less distinct than in video data.

Despite all the measures taken to reduce models' memory demand, processing of input sizes of $112 \times 112 \times 48$ voxels, used in the C3D transfer learning approach, was not possible with the NVIDIA M60 GPU available for training. Therefore, a smaller volume of $112 \times 112 \times 16$ voxels was studied.

Availability of patients HPV status was essential for inclusion of cases in the transfer learning dataset. However, goal of self-supervised approaches is inclusion of unlabeled data, this allowed for involvement of additional cases without a tested HPV status available. Moreover, it is likely that general knowledge

[3] Notation of *patches* for the ViT model [110] collides with the notation of patches used before, revering to the clipping of CT images to smaller size, which will therefore be denoted as CT-patches in this Section.

about the head and neck region will improve models' performance when trained to distinguish between HPV positive and HPV negative oropharyngeal cases. Therefore, also the need for cases to be of oropharyngeal subtype and existence of tumor segmentations were abandoned. This led to inclusion of more cases from the four cohorts used before, but also two additional TCIA cohorts, namely *QIN-HeadNeck* [188, 189] and *ACRIN-HNSCC-FDG-PET/CT (ACRIN 6685)* [190, 191], could be employed for pretraining. In total, this resulted in a dataset of 2067 CT images of head and neck cancer patients, with a tested HPV status available for 906 of them.

In the transfer learning approach, an external cohort was used for final performance testing, featuring 80 cases. Due to the small dataset size of this cohort, models were retrained 10 times for rough estimation of error margins. Considering the computational workload of the MAEMI model, each model was only trained once in the self-supervised setting. However, size of the test set was expanded and bootstrap resampling, featuring 10,000 samples, utilized for evaluation of error margin, asserted by 5^{th} and 95^{th} percentiles. All cases were merged and then split into a pretrain, train, validation and test set. The pretrain set was used for pretraining on the image reconstruction task, while train, validation and test sets were utilized in the downstream task of HPV classification. Therefore, for construction of train, validation and test sets, cases with available HPV status were split twice using a stratified split, preserving the ratios of HPV cases, gender and medical center, resulting in 509 train cases, 170 validation cases and 227 test cases [4]. For construction of the pretrain set, cases without an available HPV status were added to the train set, resulting in 1670 cases.

All images were again resampled to a uniform voxel size of $1.0\,\text{mm} \times 1.0\,\text{mm} \times 1.0\,\text{mm}$ and intensity values cropped at -250 HU and 250 HU to then be linearly rescaled to range from 0 to 1, as for the transfer learning studies.

Pretraining

Again, rough identification of the region of interest was applied to crop images to smaller CT-patches. However, segmentation of tumor regions was not available for all cases. Therefore, the dimensional extend of the human head found by Vasavada et al. [192], depicted in Figure 6.5, has been used for orientation. In the study, a mean head height of 190.5 mm has been found, while the values for neck length and head depth were given by 107.5 mm and 194.5 mm. For identification of the gross head region, threshold based segmentation of CT intensity values larger than -800 HU was performed. The top CT slice was assumed to coincide with patients skullcap. While the center of the oropharynx was assumed to roughly be at the same height as the lower mark utilized for measurement of head height, c.f. Figure 6.5. In lateral and anterior-posterior direction the center of the head was assumed to very roughly coincide with the

[4] Relaxed restrictions in pattern recognition, employed to identify tumor segmentation files during data mining, resulted in an increased dataset, in comparison to that depicted in Table 6.1.

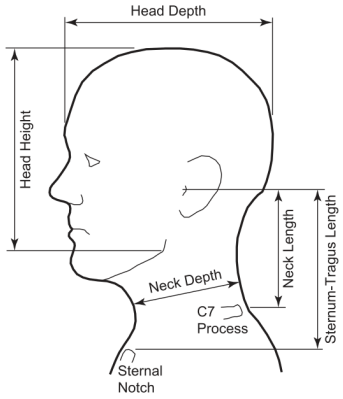


Figure 6.5: Dimensional extend of the human head, taken from Vasavada et al. [192]. A mean head height of 190.5 mm, neck length of 107.5 mm and head depth of 194.5 mm has been found [192]. Those measures have been used to roughly identify the desired region of interest, i.e. the oropharynx, and crop CT images to smaller size.

center of the oropharynx. Therefore, orientation points lying 190 mm from the top slice, 100 mm from the anterior segmentation boundary, and in the center of the lateral dimension were identified. Patches of size 120 mm × 120 mm × 52 mm were then cropped using this orientation point. Clearly, this procedure will fail for some of the cases and only contain part of the oropharynx region or even completely different parts of the head. However, for the majority of cases valuable information, essential for the downstream task, is assumed to be involved.

During pretraining, learning rate of the MAEMI model was set to 0.001, a batch size of 4 has been chosen and a patch size of $8 \times 8 \times 4$ has been employed. Notably, selection of batch and patch size were restricted by memory limitations of the computing resources available. Weights of the encoder were initialized with the publicly available ImageNet pretrained weights of the MAE approach [185]. Notably, weight initialization with ImageNet pretrained weights for a 3D model is possible due to the embedding space utilized in transformer models. The model was trained for 400 epochs. As for the approach of He et al. [20], no validation performance was captured during pretraining. Reasoning behind this is the assumption that for such a high fraction of random masking overfitting is unlikely to set in. Hence, no estimation of the generalization gap is needed.

CT-patches were randomly cropped to an input size of $16 \times 112 \times 112$ for data augmentation. To reduce the computational workload, only flip and rotation augmentations were used during training.

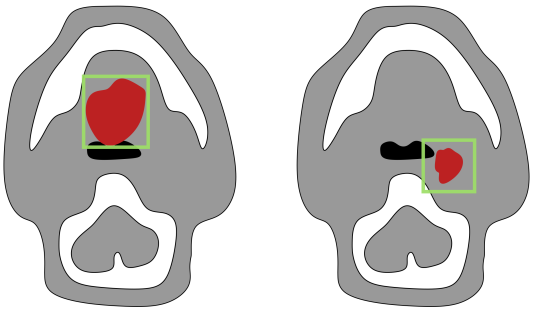


Figure 6.6: Selection of the region used to crop CT images to smaller size. Larger ROIs were cropped using the smallest bounding box surrounding the area (left), while ROIs of smaller extension were center cropped using a bounding box of 120 mm × 120 mm × 24 mm (right). Such that a minimal CT-patch size could be secured.

HPV classification

As for the transfer learning approach, tumor segmentations have been utilized to crop CT images to smaller size. However, cropping has not been performed in a uniform way, but CT-patches, taking the tumor extent into account, were generated. A size of 120 mm × 120 mm × 24 mm has been chosen for small tumors, while for larger tumors the smallest bounding box surrounding the ROI was used for cropping, a schematic can be seen in Figure 6.6.

The decoder part of the pretrained model was replaced with a dense layer, ending in a binary output score. The encoder was initialized with the weights obtained in the pretraining step. Initial layers were kept fix during training, while weights of later layers were optimized. Data augmentations were performed in the same way as for the pretraining task. Weighted binary crossentropy has been chosen as a loss to be minimized using the Adam optimizer.

Manual hyperparameter optimization was performed based on the validation loss. Variations involved: changes in the number of initial layers kept fix during training, learning rate and dropout rate. The best final model featured freezing of the first ten transformer layers, a learning rate of 0.0001 and a dropout rate of 0.1 in transformer layers. All other hyperparameters followed that of the *base* model of He et al. [20].

Due to the changes in data preparation and input size, the C3D transfer learning approach was also retrained for comparison. The modified input dimension was handled in such a way that the same slice of the 112 × 112 × 16 CT-patches was fed to all the color channels of the model, resulting in the desired 3 × 16 × 112 × 112 input dimension. Also, hyperparameter optimization was repeated due to the changes in the datasets. However, the best performing model again featured the same structure as for the dataset utilized before.

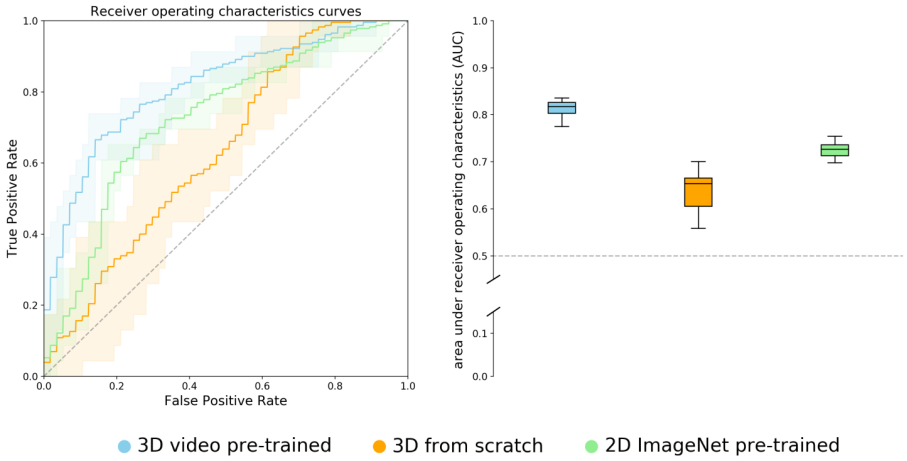


Figure 6.7: Test set results for the C3D based model, 3D CNN architecture and the VGG16 based network obtained on the *HN1* cohort, taken from Lang et al. [183].

6.3 Results

On the external data split of the transfer learning approach, the C3D based model achieved an AUC score of 0.949 (0.903 - 0.978) and 0.734 (0.686 - 0.770) for training and validation. The 3D CNN model trained with randomly initialized weights accomplished an AUC score of 0.827 (0.664 - 0.918) and 0.713 (0.670 - 0.744), respectively. Performance of the VGG16 based model was given by 0.776 (0.764 - 0.788) and 0.621 (0.583 - 0.644). Test set results can be seen in Figure 6.7. AUC scores were given by 0.814 (0.775 - 0.836), 0.638 (0.558 - 0.701) and 0.726 (0.698 - 0.754) for the C3D based model, 3D CNN architecture and the VGG16 based network. ROC curves generated from the mean predictions of all ten training runs were significantly different with a p-value of 0.027 between the C3D based and the 3D CNN model. P-values between the C3D and VGG16 approaches and between the 3D CNN and VGG16 frameworks were given by 0.110 and 0.435.

Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), for a threshold based binary classification at a value of 0.5 in the output layer are shown in Table 6.2. Furthermore, F_1 score, given by the harmonic mean of precision and recall, and balanced accuracy, reflecting the mean value of sensitivity and specificity, are listed.

Example reconstructions from the validation set, learned by the MAEMI model, can be seen in Figure 6.8.

On the mixed dataset, the MAEMI approach achieved train and validation

	3D video pre-trained	3D from scratch	2D ImageNet pre-trained
AUC	0.814	0.638	0.726
sensitivity	0.752	0.665	0.843
specificity	0.721	0.491	0.398
PPV	0.533	0.346	0.367
NPV	0.880	0.789	0.868
F_1 score	0.618	0.452	0.508
balanced accuracy	0.737	0.578	0.621

Table 6.2: Test results for the three different models probed in the transfer learning approach. Scores represent the mean of the values obtained by training each model ten times. Binary classification metrics were calculated using a threshold based discrimination at a value of 0.5 in the output layer.

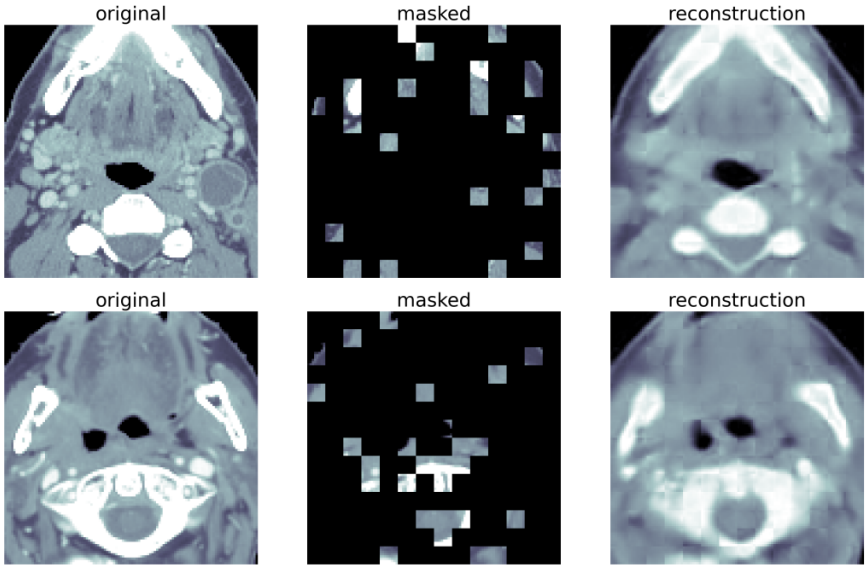


Figure 6.8: Example reconstructions learned from the self-supervised MAEMI approach. The left column shows the unaltered image, while the masked model input is shown in the middle column. Reconstructions learned from the model can be seen on the right.

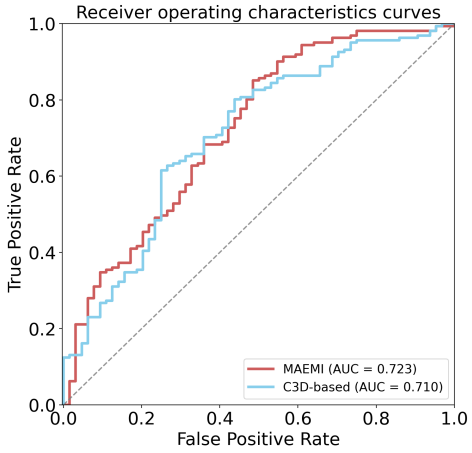


Figure 6.9: ROC curves for MAEMI and the C3D based model on the mixed datasets. AUC is given by 0.723 for the MAEMI approach and 0.710 for C3D based transfer learning. P-value for a difference of both ROC curves is given by 0.720.

AUC scores of 0.856 (0.827 - 0.885) and 0.836 (0.774 - 0.893). The C3D based model was able to reach a training AUC score of 0.875 (0.845 - 0.902), while validation performance was given by 0.791 (0.722 - 0.855). On the test set, MAEMI accomplished an AUC score of 0.723 (0.657 - 0.787) and the C3D based model of 0.710 (0.644 - 0.773), ROC plots are shown in Figure 6.9. P-value for a difference of both ROC curves was given by 0.720. Binary classification metrics, again obtained by a threshold based discrimination at a value of 0.5 in the output layer, can be seen in Table 6.3.

6.4 Discussion

In the transfer learning studies, the C3D based model performed best with a improved AUC score in comparison to the 3D CNN trained from scratch and the ImageNet pretrained VGG16 model. Difference between the ROC curve of the C3D based model and the 3D CNN was significant. Between both transfer learning approaches, 3D and 2D, ROC curves were only different with a p-value of 0.110. However, the C3D based model performed better, reaching a mean AUC score of 0.81 in comparison to a mean AUC score of 0.73 for the VGG16 based model. This improved performance is likely to be caused by the three dimensional structure, allowing for full exploitation of information, in comparison to the VGG16 based model. Significant difference has to be investigated in further studies, featuring larger test sets. Moreover, the approach should be evaluated on other problem settings in the area of 3D medical imaging.

	MAEMI	C3D based
accuracy	0.667	0.653
balanced accuracy	0.654	0.650
F_1 score	0.746	0.731
NPV	0.440	0.427
PPV	0.821	0.822
sensitivity	0.683	0.658
specificity	0.625	0.641

Table 6.3: Test set results on the mixed datasets. Binary classification metrics were calculated using a threshold based discrimination at a value of 0.5 in the output layer.

General benefit of natural imaging based pretraining for a downstream task in the medical imaging domain has been challenged by Raghu et al. [100]. Who were able to associate improvements in the downstream task with an application of overparameterized models. However, the smallest data regime employed in their study still featured 5000 2D images for training, while the dataset utilized here contained 675 training cases of 3D images. Involvement of 3D data hinders construction of lightweight CNN models with a low number of trainable parameters. Therefore, a general overparametrization of models is likely, transfer learning helps counteracting this overparametrization. Moreover, due to the training set being a magnitude smaller than data utilized by Raghu et al., benefit of real transferred knowledge can be inferred but must be tested in further studies.

Hendrycks et al. [193] were able to show that utilization of pretrained weights leads to improvements in terms of robustness. Radiomics features are known to be affected by application of different scanners [194–196], and convolutional neural networks are sensitive to such domain shifts [197, 198]. Therefore, application of transfer learning improves models' ability to generalize well on external medical imaging data. For the dataset in the transfer learning study, external testing has been applied, featuring different scanner types and imaging protocols. Hence, robustness gained from transfer learning is likely to have caused the superior performance of the C3D model.

On the mixed dataset, the C3D based model achieved a test set AUC score of 0.723 (0.657 - 0.787) which was lower than the 0.814 (0.775 - 0.836) reached for the external data split. In principle one would expect a converse behavior, i.e. better test performance on the mixed dataset. However, both test sets only involved 80 and 227 cases. For such low numbers, datasets do only partially reflect the underlying distribution, which forbids direct comparison between models. This is not only the case during testing. In general, deep learning models trained on different datasets simply behave dissimilar. Hence, comparisons between models trained and/or tested on unequal datasets is not possible, at least in the small data domain.

Saint-Esteven et al. [115] utilized a 2D model pretrained on ImageNet and finetuned it for HPV classification based on CT imaging. They cropped 2D images from the axial, sagittal and coronal planes, to feed those to the expected 3 color channels, calling it a 2.5D approach. The model was trained on the *OPC* dataset and an internal cohort. Testing was performed on *HNSCC* and *HN1*. They reported an AUC score of 0.83 on the *HN1* cohort, and compared it with the 0.814 achieved by the model developed for this thesis. Which is not directly possible, due to utilization of different datasets employed during training.

Radiomics models trained on CT based HPV classification typically reach AUC scores ranging between 0.70 and 0.80 [169, 170]. A fundamental limitation of the approach is introduced by the dependence on predefined volumes utilized for feature extraction. Typically, features from the gross tumor volume (GTV) are extracted by hand-crafted filters. However, oropharynx cancer cases caused by an infection with HPV are known to be associated with areas outside of the tumor volume, i.e. cystic lymph nodes [199, 200]. Therefore, models considering features from inside the tumor area only, discard essential information, valuable for prediction.

MAEMI reached an AUC score of 0.723, but performance was not significantly different from that of the C3D based model. However, different aspects forbid direct comparison between both procedures. First and foremost, both approaches feature different architectures and dataset sizes. If one would like to compare self-supervised pretraining and transfer learning both frameworks should feature the same structure, i.e. both should utilize either convolutional or transformer layers. Furthermore, both models were pretrained on different dataset sizes. C3D was trained on 1.1 million sports clips, while the data set utilized for pretraining of MAEMI included 1670 cases. Influence of dataset size on the pretraining method has to be investigated in future studies. One could imagine superiority of the self-supervised approach for equal dataset sizes. However, video data is available in large numbers, while generation of a head and neck dataset including 1.1 million CT images not feasible. Therefore, different dataset sizes have to be probed to foster understanding of which method should be preferred based on the available data. Last but not least, optimization of hyperparameters and selection of model architecture for MAEMI was limited by the computing resources available during training. Larger computing resources are needed for determination of the right parameters.

However, it was proven that modification to the MAE approach of He et al., that allow for three-dimensional medical imaging based classification tasks, are possible. Modern transformer layers have been utilized to process CT imaging, which has been done by very few studies in a 3D manner up till now. The approach is valued promising to be applied in future settings. However, further investigations have to compare video data based pretraining of the masked autoencoder model, following the approach of Feichtenhofer et al. [186], with the same-domain pretraining strategy developed in this thesis. Also, in-domain pretraining should be investigated, i.e. pretraining on CT imaging data from another entity. One example would be given by self-supervised pretraining on lung CT data, available in larger numbers than head and neck CT, to then be

finetuned on HPV detection in oropharynx cancer.

A problem generally associated with the processing of head and neck CT imaging data is given by the vulnerability against artifacts. Leijenaar et al. [201] analyzed a subset of the *OPC* cohort and found that 49% of patient's CT images were affected by visible artifacts. In a subsequent study, it was proven that artifacts have significant influence on machine learning models by training of a radiomics approach that achieved AUC scores ranging between 0.70 and 0.80, depending on the distribution of cases containing artifacts in the train and validation sets [202]. Assumption of the same influence on deep learning model is reasonable, but also has to be tested in further studies.

6.5 Conclusion

Pretraining convolutional neural networks on video clip data allows for processing of three-dimensional CT imaging data in the downstream task. For CT based HPV classification, the approach results in better performance than networks trained from scratch or relying on 2D input. Deep learning features a inherent superiority for CT based detection of HPV infections in oropharynx cancer patients in comparison to radiomics models. However, for fair comparison both approaches have to be trained on the same datasets. Self-supervised learning allows for in-domain and same-domain pretraining. The masked autoencoder approach of He et al. [20], utilizing modern transformer layers, can be adopted to be used in combination with 3D medical imaging data. Studies investigating the power of both approaches in more detail have to be performed on larger datasets. Influence of different scanners and imaging protocols as well as image quality, has to be investigated for a clinical applicability in the area of HPV detection.

Chapter 7

Head and neck cancer progression free survival

HPV depicts a surrogate marker for the endpoint of overall survival (OS). Deep learning features the power for direct prediction of survival endpoints. But, as pointed out in Section 4.4, handling of time dependent data requires special architectures. Ability of the approach of Gensheimer and Narasimhan [22] in conjunction with the C3D based transfer learning approach developed in Chapter 6 was studied. The model was again tested in the competitive setting of a public challenge, namely the MICCAI 2021 HEad and neCK TumOR (HECKTOR) segmentation and outcome prediction challenge [29]. Task of the challenge was prediction of progression free survival in head and neck cancer patients.

Part of the studies presented in this Chapter were published in the proceedings of the challenge: "Deep learning based GTV delineation and progression free survival risk score prediction for head and neck cancer patients", Lang et al. [203]. Code of the survival model was published online [1].

7.1 Introduction

Head and neck tumors feature a large diversity in subtypes and treatment options, as pointed out in Chapter 6. Therefore, chances for a complete cure are still relatively low, with 40% of patients experiencing locoregional failure during the first two years after therapy [204]. On one hand, this led Troost et al. [205] to identify tumor subvolumes with high proliferative activity eligible for dose escalation in order to enhanced tumor control. On the other, head an neck cancer patients, undergoing radiation therapy, can experience significant side effects ranging from skin changes to secondary cancers [206], which led different studies to seek for dose de-escalation [207]. Gillison et al. [163] showed that patients with high risk HPV positive head and neck tumors have a 59%

[1] https://github.com/LangDaniel/HECKTOR2021

reduction in risk of death from cancer when compared to HPV negative HNSCC patients. Based on that, the ECOG 1308 trial [208] studied radiation dose reduction for HPV positive head and neck cancer patients. The study found stable tumor control and survival rates for patients identified to qualify for a treatment with the de-intensified dose. However, HPV was not utilized solely as a marker but only patients with a favorable response to induction chemotherapy were stratified to the low dose regime. Studies were performed on a relatively small cohort and findings of the trial were also discussed controversially [209]. The De-ESCALaTE and RTOG 1016 trials studied dose de-escalation in chemo-radiotherapy by replacing cisplatin with cetuximab, two chemotherapeutic agents, for HPV positive head and neck cancer patients. Both trials found no benefit in terms of toxicity reduction but significant decline in tumor control [210, 211]. Therefore, a marker, available prior to treatment, that allows for identification of head and neck cancer patients eligible for dose de-escalation is still to be found.

In general, patient survival, i.e. overall and progression free survival, depicts a essential endpoint in clinical oncology [21]. Architectures able to model such endpoints have the power to identify novel tumor markers. Haider et al. [212] were able to improve clinical patient stratification in terms of overall and progression free survival, utilizing a PET/CT based radiomics model. Therefore, the capability of deep learning architectures for modeling of time dependent data in head and neck cancer patients was studied here.

Task of the MICCAI 2021 HEad and neCK TumOR (HECKTOR) segmentation and outcome prediction challenge [29] was PET/CT imaging based prediction of progression free survival (PFS). In this thesis the ability of the transfer learning approach, relying on the C3D pretrained model as a backbone, in combination with the discrete-time survival loss of Gensheimer and Narasimhan [22] has been tested.

Following the findings of Section 6, the video clip pretrained convolutional part of C3D has been utilized as a feature extractor, independently applied on the CT and the PET image. Resulting feature vectors were stacked and processed by dense layers, ending in a output layer, reflecting discrete time intervals. Similar to the approach in Chapter 5, a U-Net like architecture has been trained for identification of the region of interest, i.e. the gross tumor volume, in order to crop images to smaller size.

7.2 Material and methods

The training dataset of HECKTOR involved 224 cases coming from five centers. Data from four of those centers was also involved in the *HN PET CT* cohort utilized in Chapter 6. The test dataset involved 101 cases coming from two centers. Data stemming from one of those centers was only present in the test cohort, while data from the other center was involved in both, training and testing, cohorts. All cases involved a PET/CT image, with PET image intensity values given in units of standardized uptake value (SUV) and CT

image intensity values in Hounsfield units. For each case, a bounding box of size $144\,\text{mm} \times 144\,\text{mm} \times 144\,\text{mm}$, locating the oropharynx region, was provided. Apart from imaging data, clinical information was also available including:

- center,
- age,
- gender,
- TNM stage,
- TNM edition,
- tobacco consumption,
- alcohol consumption,
- performance status,
- HPV status,
- treatment,

treatment indicated involvement of radiotherapy only or application of chemo-radiotherapy and *TNM edition* distinguished between different versions of the TNM standard. Furthermore, the training set involved ground truth segmentation maps of the GTV and patient survival data, given by time and event status.

Imaging data and segmentation maps were resampled to a voxel size of $1.0\,\text{mm} \times 1.0\,\text{mm} \times 1.0\,\text{mm}$. CT images were again clipped at -250 and 250 HU and linearly rescaled to range from 0 to 255, as in Chapter 6. PET images were clipped at 0 and 100 SUV and also linearly rescaled to range from 0 to 255.

7.2.1 GTV segmentation

As for the models presented before, region of interest delineations have been used to crop images to smaller size. GTV segmentations were only available in the training cohort. Therefore, a segmentation model was trained for generation of labels in the test set. In a first step, the very rough bounding boxes, provided in the dataset, were used to crop PET and CT images to smaller size. Hence, input to the segmentation model was given by $2 \times 144 \times 144 \times 144$ voxels, featuring a CT and a corresponding PET image, fed to the color channels of the network.

For segmentation of the GTV area, again a U-Net like architecture was utilized. However, instead of using a 2D model, like in Section 5, a 3D architecture has been modified for reduction of network parameters. Such that the framework could be fitted on the NVIDIA M60 GPU available for training. In contrary to the max pooling layers employed for spatial dimension reduction in the original U-Net model of Ronneberger et al. [13], depicted in Figure 3.15, a convolutional layer with a stride larger than 1 has been used. Therefore, main building blocks of the model were given by a convolutional layer with a stride of 1, followed by a convolutional layer with a stride larger than 1. Instead of two convolutional layers of stride 1 followed by a max pooling layer. Ronneberger et al. [13] utilized skip connections to forward the last generated feature maps in each stage of the encoder to the decoder path. However, for the model developed in this thesis, the first generated feature maps, obtained before feature map extension, were forwarded. In this way, size of copied feature maps was

reduced by a factor of 1/2. Additionally, extent of the final output block has been reduced and chosen to consist of one convolutional layer only. A sketch of the network can be seen in Figure 7.1.

All convolutional hidden layers were followed by a dropout and a ReLU activation layer. In the final output layer sigmoid activation has been used. The Dice loss of eq. (3.16) has been chosen as a objective function to be minimized by the Adam optimizer. Optimization of network hyperparameters was again performed manually, guided by the validation loss. Each model was trained for 150 epochs.

The encoder of the final model consisted of five basic building blocks. The first two of those blocks featured convolutional layers with kernels of size $5 \times 5 \times 5$ and spatial dimensionality reduction layers with kernels of size $3 \times 3 \times 3$. For the three remaining blocks kernel sizes were given by $3 \times 3 \times 3$ and layers reducing spatial dimensionality featured a stride of $2 \times 2 \times 2$. In the decoder, the same basic structure in a mirrored fashion has been used, with transpose convolutions utilized for recovery of spatial dimensionality. Dropout rate was given by 0.25, learning rate by 10^{-4} and batch size by 4.

During training data augmentation was applied to prevent the model from overfitting. Images were flipped on the coronal and sagittal plane, rotated by a multiple of 90°, Gaussian noise with zero mean and a standard deviation of two was added, and voxel values were randomly shifted by a global offset between -2 and 2. Furthermore, elastic deformation, featuring a grid sampled from a normal distribution with standard deviation randomly chosen between zero and one and a voxel size of $3 \times 3 \times 3$, has been applied. The `elasticdeform` python package [213] has been utilized to do so.

During inference, scalar intensity values in the segmentation maps have been thresholded for formation of a binary segmentation mask using a value of 0.50. The largest connected area has been identified using the `label` filter given in the `ndimages.measurements` class of the `scipy` python package [214]. Only this area has been kept and further processed by the `ndimage.gray_dilation` filter with a kernel size of $2 \times 2 \times 1$, for smoothing of edges.

7.2.2 Prediction of progression free survival

Preprocessing in the progression free survival model involved cropping of PET and CT images. As pointed out in Section 5, removal of image parts, that do not contain any information essential for the task of interest, can facilitate learning, which is especially valuable in the small data regime of medical imaging. Therefore, the center point of the GTV has been used to crop patches from PET/CT images. For the train and validation set, segmentation maps provided were used, while for the test set predictions obtained from the U-Net model were used. Cropping of patches was preformed in the same manner as for the approach depicted in Figure 6.6, sustaining a minimal size of $100\,\text{mm} \times 100\,\text{mm} \times 100\,\text{mm}$. Voxel size resampling and intensity value rescaling was performed in the same way as for the segmentation model.

As pointed out in Section 4.4, proper modeling of survival data requires

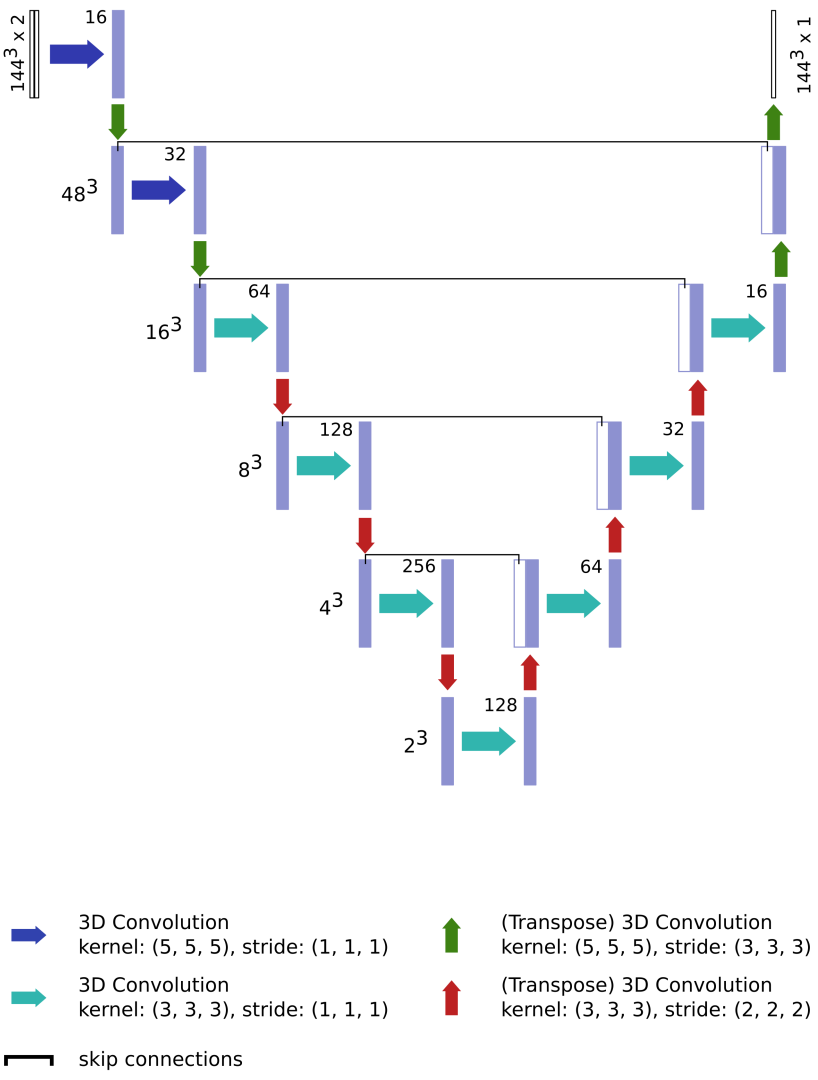


Figure 7.1: Modified U-Net model, adopted from Lang et al. [203]. PET/CT images were used as an input. For downsampling, convolutional layers with a stride > 1 were utilized, instead of max pooling layers employed by Ronneberger et al. [13]. Forward skipping was performed on the layers obtained after spatial dimension reduction and before expansion of feature map size. This reduced the size of copied features by a factor of 1/2.

architectures able to handle time dependent data. The discrete-time survival model of Gensheimer and Narasimhan [22] provides an approach well suited to be applied in conjunction with deep neural networks that also allows for utilization of batches during training. Therefore, the loss depicted in eq. (4.6) has been utilized for modeling of the progression free survival data.

The convolutional part of the C3D model, already utilized in Chapter 6, was again used as a feature extractor. Here, next to a CT image a corresponding PET image was given. Hence, a way to fuse features extracted from both modalities had to be found. In Chapter 5, the element-wise maximum was calculated to fuse features stemming from both kidney regions of the same CT image. However, for the case at hand input modalities featured different acquisition methods, namely CT and PET, that do possibly contain complementary information. Taking the element-wise maximum would destroy such complementary information. Therefore, both feature vectors were stacked to be then further processed by dense, randomly initialized layers, such that no information was lost during merging. Clinical features were again fused with dense features before construction of the final output. The architecture ended in a layer reflecting a given number of time intervals, c.f. Figure 4.3. Number and size of those time intervals had to be fixed during hyperparameter optimization. A schematic of the model can be seen in Figure 7.2.

Patches were cropped to a uniform input size of $96 \times 96 \times 96$ voxels. This was performed randomly during training and by center cropping during validation and testing. Other data augmentation techniques involved again flipping on the coronal and sagittal plane, rotations by a multiple of $90°$, and the same elastic deformations as for the segmentation model. Gaussian noise sampled with a variance between zero and one and zero mean was added.

The Adam optimizer with a learning rate of 5×10^{-5} was used to minimize the negative log likelihood loss. Batch size was choose to be 16. Manual hyperparameter tuning was applied for determination of the best performing architecture. Model performance was measured based on the c-index score of the validation set. All models were trained for 75 epochs. Hyperparameters modified included: size of dense layers, inclusion point of clinical features, dropout rate, and number and size of time intervals in the output layer.

The best performing model featured dense layers of size 512 and 256 ending in a output layer reflecting 15 time intervals, with the first 10 intervals depicting a time span of half a year and the remaining 5 intervals representing a time span of one year, leading to a maximum survival time of 10 years.

Output of the architecture was given by an array containing survival probabilities for given time intervals. Commonly, the c-index metric, introduced in Section 4.4, measuring the order of all datapoints, is utilized for performance evaluation in survival models, However, no inherent order is given for the output of the model. In their paper Gensheimer and Narasimhan [22] fixed this by computation of the c-index at a certain timepoint, e.g. at one year survival. However, selection of a single time point for calculation of the c-index disregards performance at all other timepoints.

Therefore, in this thesis, the expected survival time per patient has been

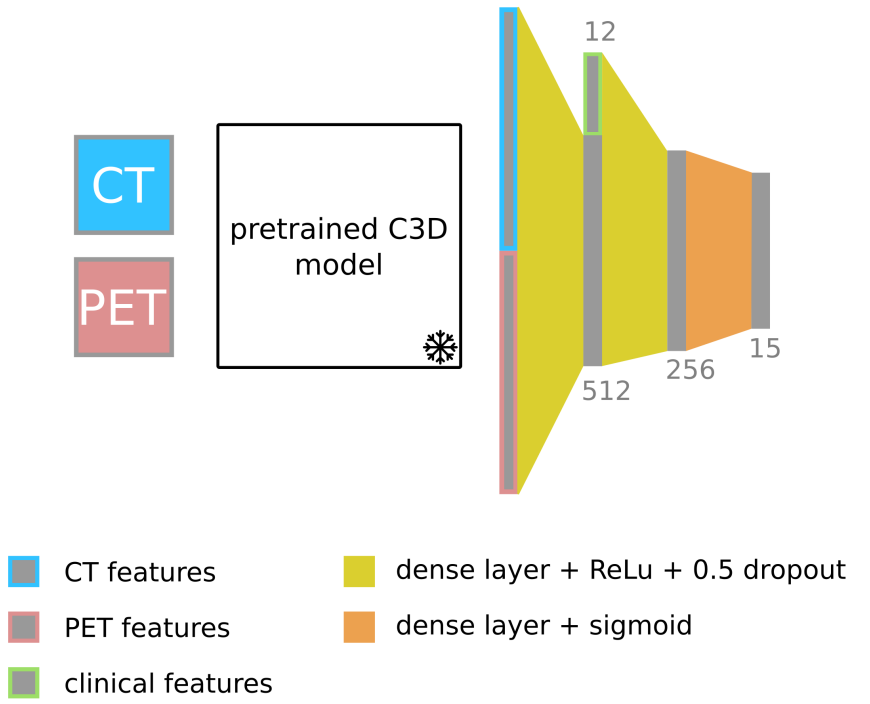


Figure 7.2: Progression free survival model, adopted from Lang et al. [203]. The convolutional part of the pretrained C3D model was used as a feature extractor on the CT and the PET image. Resulting feature vectors were stacked to be then further processed by dense layers. Clinical features were merged with dense features before classification. Hidden dense layers employed ReLU activation and were followed by a dropout layer. The final output layer featured Sigmoid activation as the model was trained with the discrete-time loss of Gensheimer and Narasimhan [22].

calculated via

$$E(T) = \sum_{k=1}^{n} t^k \; p^k = \sum_{k=1}^{n} t^k \; h^k \; \prod_{l=0}^{k-1} (1 - h^l), \tag{7.1}$$

with t^k the timepoints of the n intervals chosen during model construction, p^k the probability of the patient to receive the event, i.e. tumor progression, in the time interval from t^{k-1} to t^k but not before, and h^l the hazard probability for the interval ranging form t^{l-1} to t^l, cf. Figure 4.3. Min-max scaling was then applied for construction of a normalized survival score

$$\hat{t} = \frac{T - T^{\min}}{T^{\max} - T^{\min}}, \tag{7.2}$$

ranging between 0 and 1, with $T^{\min}$ and $T^{\max}$ the minimum and maximum predicted survival time for the cohort at hand, i.e. the test set. Finally, this survival score was inverted for construction of a normalized risk score using

$$r = 1 - \hat{t}. \tag{7.3}$$

In this way, a risk score, taking into account the complete duration of modeled survival time, was generated. This risk score could then be evaluated using the c-index metric.

Two input modalities were available for model development, clinical and imaging data. For evaluation of input from clinical variables, a Cox proportional hazards model has been trained using the python `lifelines` library [215].

To foster reproducibility, code of the developed C3D based survival model has been published online [2].

In total, 44 teams registered for the segmentation task of the HECKTOR2021 challenge, while 30 teams took part in the survival prediction task. Inclusion in the final ranking required publication of developed methods, 22 teams fulfilled this requirement for the segmentation task and 17 teams for the survival prediction task. For evaluation of segmentation results, the 95^{th} percentile of the Hausdorff Distance (HD) has been used, i.e. replacing the maximum operation in eq. (3.17) with the 95^{th} percentile. For construction of the final score, the median Hausdorff distance of all test set cases was taken. [29] The c-index was utilized as a performance measure in the survival task.

7.3 Results

The adopted U-Net architecture achieved a Dice similarity score of 0.733 and 0.762 on the train and validation set. Test set performance was given by a Dice score of 0.705. From the 22 teams in the final ranking the segmentation model was placed 16^{th}. Performance of the top 5 models can be seen in Table 7.1, the developed model participated under the team name *DMLang*. Example predictions for three training set cases can be seen in Figure 7.3.

[2] https://github.com/LangDaniel/HECKTOR2021

Rank	Team	Dice	HD95 [mm]
1	Pengy [216]	0.779	3.088
2	SJTU EIEE 2-426Lab [217]	0.773	3.088
3	HiLab [218]	0.774	3.088
4	BCIOQurit [219]	0.771	3.088
5	Aarhus Oslo [220]	0.779	3.155
	⋮		
16	**DMLang**	0.705	4.027

Table 7.1: Segmentation test set results of the HECKTOR2021 challenge.

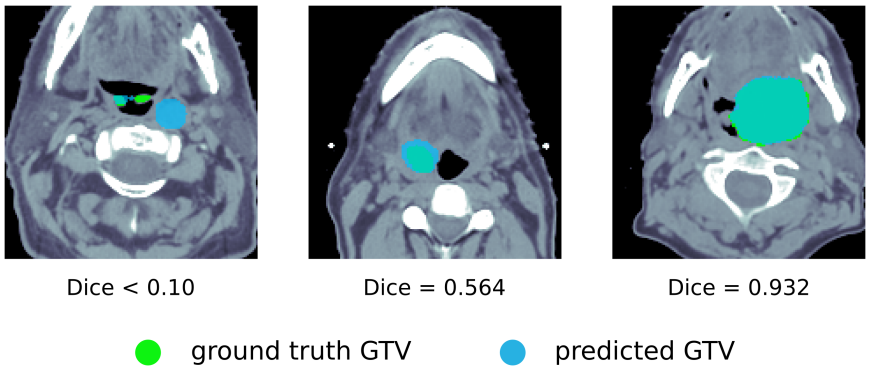


Figure 7.3: Slices of predicted and ground truth segmentation maps for examples from the training set. Dice scores are given in reference to the complete GTV, i.e. not only in reference to the slices depicted.

Rank	Team	c-index
1	BioMedIA [221]	0.720
2	Fuller MDA [222]	0.694
3	Qurit Tecvico [223]	0.683
4	BMIT USYD [224]	0.671
5	**DMLang**	0.668
	⋮	

Table 7.2: Progression free survival prediction results for the HECKTOR2021 challenge.

For the survival model, c-index performance was given by 0.899 (0.865 - 0.931) and 0.833 (0.728 - 0.930) on the train and validation set, error margins were computed using bootstrap resampling of 10,000 samples and evaluation of 5% and 95% percentiles. On the test set a score of 0.668, has been achieved. The model was ranked 5^{th} of the 17 participating teams. Top 5 results are shown in Table 7.2.

The Cox proportional hazards model, relying on clinical variables, achieved a c-index of 0.725 (0.655 - 0.792) and 0.739 (0.572 - 0.881) for the train and validation set, respectively. Error margins were again computed using bootstrap resampling. Performance on the test set could not be evaluated, as ground truth scores were not publicly available.

7.4 Discussion

The basic U-Net structure of Ronneberger et al. [13] was modified for development of a lightweight 3D segmentation model. The model achieved a Dice score of 0.705 and a Hausdorff distance of 4.027 on the test dataset. Participation in the HECKTOR2021 challenge revealed an inferiority in terms of segmentation performance, with the model only being ranked 16^{th}. Notably, a U-Net like architecture was used by 21 of the 22 teams that participated in the segmentation task. All of the top five teams used an ensemble of U-Nets for training, with four of them applying the nnU-Net of Isensee et al. [98].

However, aim of the segmentation approach presented in this thesis was development of a lightweight model, that could be fitted on the NVIDIA M60 GPU available for training, but also achieved sufficient performance, that allows identification of the rough ROI. The Dice score of 0.705 is valued sufficient for rough identification of the GTV. This claim can also be visually backed by the example predictions depicted in Figure 7.3. Hence, the segmentation model is valued as sufficient.

Ground truth segmentation, performed by medical experts, suffers from interobserver variability. Gudi et al. [225] identified a Dice similarity score of 0.69 between three experienced radiation oncologists, for PET/CT based head and neck cancer GTV segmentation. This variability in ground truth labels limits

the performance of segmentation algorithms trained on the data, and raises the question of comparability between models.

The developed survival architecture was able to achieve a c-index performance of 0.668 on the test set, and was ranked 5^{th} of the 17 teams that participated in the survival prediction task.

The winning architecture of the survival task, constructed by Saeed et al. [221], was able to achieve a c-index of 0.720. They fused CT and PET images to be then processed by 3D convolutional and max pooling layers, the resulting feature vector was merged with clinical variables, similar to the architecture developed for this thesis. For modeling of survival data a Multi-Task Logistic Regression (MTLR) model [226] was applied. For final prediction, risk probabilities obtained in this way were joint with probabilities obtained by a Cox proportional hazards model that was trained on the clinical variables.

The approach placed second, developed by Naser et al. [222], reshaped clinical variables into a matrix form, such that they could be fused with PET/CT imaging data and processed by convolutional layers. Clinical, PET and CT input was then fed to different channels of a 3D DenseNet121 CNN model [227]. They also utilized the log-likelihood loss of Gensheimer and Narasimhan [22] to train their model.

However, absence of error margins depicts a major limitation of the challenge. For the training set, error margins were roughly given by ± 0.03 for the CNN model and by ± 0.06 for the Cox model developed within this thesis. Differences between c-index performance of neighboring approaches in the final ranking are smaller than those scales. Therefore, no inherent superiority or inferiority can directly be derived from the ranking. The developed model is considered to feature the principle ability for prediction of progression free survival in head and neck cancer patients, but for comparison with other model architectures larger patient cohorts and inclusion of error margins are needed.

Integration of survival models in clinical decision making does not only require further improvements in terms of performance but also robustness. Generalization error, measured on the validation set, suffered a large drop when probed on the external test set, with c-index performance declining from 0.833 to 0.668. Influence of external testing was already discussed in previous sections. In the test set, 53 of the 101 cases came from an external center not involved in training. Hence, performance loss on the test set is likely to be caused by usage of this external test data. Furthermore, vulnerability of head and neck CT imaging to artifacts was already discussed in Section 6. Such artifacts are also likely to impair model performance. Future studies should investigate their impact on the performance of survival models. Moreover, pathological mechanisms can lead to abnormal uptake of FDG in PET images [228]. Such FDG irregularities can also cause a loss in performance, but further studies of their effect on survival models are needed.

The Cox model, trained on clinical variables, achieved a train and validation c-index performance of 0.725 and 0.739, performance on the test set could not be evaluated. Successful training of the model proves the validity of information contained in the clinical data, for a prediction of progression free survival.

Influence of imaging data is yet to be determined. The different nature of the validation set utilization in the Cox model, employing it for unbiased testing, and in the deep learning model, applying it for generalization error determination, forbids direct comparison of both measures. However, the fact that the biased c-index performance of the deep learning model was significantly higher than the unbiased Cox model performance suggest valuable influence of imaging data.

7.5 Conclusion

A lightweight segmentation model for identification of the rough region of interest in head and neck cancer patients has been developed. The ability of deep learning models for prediction of progression free survival based on PET/CT imaging and clinical data has been proven. The log-likelihood loss of Gensheimer and Narasimhan [22] was combined with the transfer learning model developed in Chapter 6. A tumor progression risk score was introduced. For determination of model robustness and influence of imaging and clinical data, larger patient cohorts are needed. Fair comparison of different architectures, developed within a competitive setting, demands involvement of error margins in model ranking.

Chapter 8

Summary and discussion

Different deep learning architectures, able to process three-dimensional medical imaging data, have been developed.

In Chapter 5, a convolutional neural network was trained on pre-surgical CT images and clinical data for classification of a pathological risk score, utilized for decision making in follow-up treatment of renal masses. Risk score prediction based on pre-surgical imaging data allows for non-invasive assessment prior to therapy and therefore enables adoption of treatment, such that patients can be prevented from overdosage. The model was developed on data from the publicly accessible KNIGHT challenge [28], which allowed for comparison with other approaches. First, a U-Net like segmentation model was trained for identification of the region of interest. Then, two patches, one for each kidney, were cropped from the CTs to be input into a Siamese network. An element-wise fusion layer, merging resulting features vectors of both patches, was established for final classification. For the clinically relevant discrimination between cases requiring adjuvant therapy and cases without the need for adjuvant therapy, the model achieved an AUC score of 0.814 on the test set.

In Chapter 6, different approaches for identification of an infection with the human papilloma virus (HPV) in head and neck cancer patients based on CT imaging data were developed. HPV testing is essential for selection of the right treatment regime in oropharynx cancer patients. HPV positive tumors are more radiosensitive, and different dose de-escalation studies seek reduction of therapy induced side effects [208, 229]. HPV status prediction based on CT imaging would allow for fast, non-invasive testing and comes with no extra costs when performed on routinely acquired data. The public TCIA archive was mined for appropriate cases, creating a dataset that allows other researchers to reproduce the findings of this thesis. Techniques, well established in the natural imaging domain, were modified to be used in combination with medical imaging data.

First, a transfer learning approach, that was pretrained on sports video clips, was introduced. Video data features a 3D structure with two spatial and one temporal dimensions. This allowed for utilization of a three-dimensional architecture during pretraining and therefore exploitation of full dimensional

information in the CT images of the downstream task. For comparison, ability of a 3D convolutional neural network and a 2D transfer learning approach, relying on ImageNet for pretraining, were studied. The developed 3D transfer learning model was able to outperform the two other networks, reaching an AUC score of 0.814 on the test set.

In a second step, a self-supervised training strategy was developed. The masked autoencoder approach of He et al. [20] was altered to be able to handle medical imaging data. The architecture utilizes modern transformer layers instead of convolutional layers. Self-supervised learning allows for in-domain pretraining. Possible advances of the approach in comparison to the transfer learning architecture studied before were investigated. The head and neck dataset was modified and augmented by cases without a tested HPV status available. The network was able to achieve an AUC score of 0.723 on the test set. The 3D transfer learning approach was retrained on the modified dataset and reached an AUC score of 0.710. Hence, performance of both architectures was comparable, but optimization of the self-supervised transformer model was limited by the computational resources available during training. Model development should be reperformed on larger architectures. Future studies have to investigate the individual influence of transformer layers and utilization of the self-supervised training strategy. However, the architecture of He et al. [20] was successfully modified and modern transformer layers were utilized for an application on three-dimensional medical imaging data. Both, the transfer and the self-supervised model, have to be tested on larger dataset for evaluation of robustness.

In Chapter 7, a model was trained for prediction of progression free survival in head and neck cancer patients. An architecture able to evaluate patient survival based on medical imaging data may outperform existing risk estimations. Patients could be divided into high and low risk cases based on their predicted survival time and possibly administered to different treatment regimens, i.e. preventing low risk cases from overdosage and identifying high risk cases for which dose escalation is required. However, survival architectures require special handling of time dependent data and training strategies in deep learning impede adoption of classical approaches like the Cox proportional hazards model. For development of a convolutional neural network, capable of handling time dependent data, the strategy of Gensheimer and Narasimhan [22] has been adopted and fused with the 3D transfer learning approach developed in Chapter 6. Ability of the method for prediction of progression free survival based on PET/CT imaging and clinical data in head and neck cancer patients has been tested in the public HECKTOR challenge [29]. A 3D segmentation model was trained for identification of the gross tumor volume and achieved a Dice similarity score of 0.705 on the test set. Patches were cropped from this region to be input into the CNN based survival model. The output layer reflected 15 time intervals spanning a total survival time of 10 years. A risk score based on patients' expected survival times was constructed. On the test set the model was able to achieve a concordance index of 0.668. The HECKTOR training set included 224 cases, future studies have to be conducted on larger patient cohorts.

A limitation of studies caused by a low number of cases was concluded for all problem settings presented here. Involvement of a limited number of patients and institutions produces two problems in deep learning based medical imaging studies. First, utilization of different imaging scanners introduces domain shifts in the data. Intensity values in MRI and PET do not display quantitative values and can vary strongly, but also Hounsfield units in CTs can change substantially between different scanners and imaging protocols [230, 231]. Deep learning models are known to be sensitive to such domain shifts [197, 198]. Therefore, architectures trained and tested on a very limited number of examples result in models that are likely to fail on external cohorts not seen during training [25]. Accordingly, diverse and large training cohorts, involving a variety of scanners and imaging protocols, are needed to improve the robustness of models. The second limitation that arises from a low number of examples is given by the dependence of deep learning models on large datasets for successful training. Datasets outside of medicine are constantly growing, and while natural language processing (NLP) models like GTP-3 are trained on billions of dataset points [107], medical imaging tasks only involve a few hundred or thousands of cases. Sun et al. [232] found that the power of deep learning based vision models increases logarithmically with the amount of training data. Therefore, for general improvement of medical imaging based deep learning models, larger patient cohorts are needed.

However, curation of large medical cohorts is limited by factors like costs and privacy protection regulations. Even though, some of those problems could be handled, for example by construction of architectures able to process sensitive data in a privacy preserving way (see e.g. Kaissis et al. [233]), construction of large public medical imaging datasets, taking every possible oncological endpoint into account, is not feasible. But, in the same study mentioned above, Sun et al. [232] also found that performance of many vision tasks can be improved by training of better baseline representation learning approaches. Transfer and self-supervised learning models developed within this thesis present such representation learning approaches. Self-supervised learning can be performed without the need for any labels, which are possibly affected by privacy preserving issues or costs for generation. Therefore, a future step for the domain could be given by construction of larger unlabeled public datasets that can be utilized for training of representation learning models in a self-supervised way. Those pretrained models can than be adopted to specific downstream tasks, featuring a lower amount of data. The MAEMI model developed in Section 6.2.2 depicts a valuable approach to be tested for such an application.

All studies developed within this thesis were performed on open datasets or in the setting of public challenges, fostering reproducible research. Comparison of approaches, by development in the setting of public challenges, has become the standard validation method in the biomedical imaging domain. However, Maier-Hein et al. [26] also reported certain problems associated with the approach. Most notably, an instability of rankings against small changes in evaluation metrics. Missing error bars in final rankings was also criticized in Chapter 7. Therefore, one should not rely too closely on the exact ranking or score

achieved, especially in the small data domain of medical imaging, but rather value competitions as a rough estimate of model performance. Also, the plurality of approaches, developed by different researchers, facing the same problem setting, gives valuable insight.

Right now artificial intelligence based assessment of radiological imaging data is still in its infancy. Even though first CE and United States Food and Drug Administration (FDA) approved algorithms are available, widespread application is not achieved yet [234, 235]. One can separate applications of algorithms into two categories: human interpretable and non-interpretable. Algorithms trained on tasks usually performed by medical experts, like GTV segmentation or prediction of TNM stage, can be reviewed and tested by humans, they are human interpretable. Therefore, networks can also be applied in a supportive manner, which eases applicability and also acceptability in users. In contrast, models aiming for replacement of biological tests, like HPV status prediction, or for introduction of completely new clinical scores, like generation of a survival risk score, cannot be interpreted by human readers. Predictions in non-interpretable problem settings have to be of very high quality, as no intervention by users is possible and models cannot be applied in a supportive way, but have to be fully trusted, which could impede acceptability in users. However, also interpretations of radiologists and biological tests have to be trusted to a certain degree. Mazurowski [236] argues that there exists a fundamental gap between expectations placed on the performance of AI algorithms and human readers. He claims that human cognitive processes are also imperfect, biased and inconsistent and states that an exact interpretation of how decisions by radiologist are actually made can also not be given. The arguments formulated by Mazurowski, about the expectations on AI algorithms in terms of radiological data, also hold in comparison with other testing methods. For assessment of patients' HPV status immunohistochemistry (IHC) staining of p16 is visual interpreted by a pathologist. IHC staining also comes with different limitations like inter-observer [237] or inter-laboratory variability [238]. Therefore, application of deep learning models in the medical domain does not introduce uncertainties in an area that is otherwise fully interpretable. The same pitfalls are also encountered by other methods. Deep learning has to be tested in the same rigorous manner that other methods have to fulfill, i.e. by testing its value in large clinical trials, but models should not be treated with overly, arbitrary skepticism just because of a general mistrust against computer generated predictions. Risk assessment guidelines are developed for all kinds of medical devices and information systems. Such guidelines also have to be established for deep learning based algorithms.

However, also techniques that allow for better interpretability of deep learning models' decision making have to be developed [239]. Such explainable artificial intelligence (XAI) is needed to improve trust in model predictions but also for advancement in traceability of tasks that are currently non-interpretable. Saliency maps are often used for identification of image regions essential for decision making. However, those techniques are black boxes themselves, that can result in misleading interpretations [240, 241]. The approach developed by Chen

et al. [242] is better suited to be applied for an enhancement of interpretability. The approach called "This looks like that" learns class prototypes that can be reviewed and analyzed by human readers, which has the ability to foster trust in decision making of algorithms.

Deep learning models are influenced by the data they are trained on, such that imbalanced datasets can lead to biased models. The influence of different image scanners and scan protocols has already been discussed above, but also differences in population are causing bias. Gender or race imbalanced datasets influence performance of deep learning models [243, 244]. This problem is not newly introduced by deep learning itself, but already present in medicine as a whole [245]. However, guidelines on how to uniformly report demographic variables for trained models have to be established, such that applicability of algorithms in new environments can be estimated. Also, studies investigating possible invariance of models under certain demographic variable shifts are needed.

Adversarial attacks depict a real risk factor for the application of deep learning models. Subtle perturbations, not visible to the human eye, are applied to the input data in order to cause the model to predict incorrect outputs [246, 247]. Such vulnerabilities pose a major risk, especially on models utilized for decisions on life threatening actions. Therefore, strong defense mechanisms have to be developed to close those security gaps. First approaches focusing on the medical domain were developed, e.g. Li and Zhu [248], but further studies are needed. Moreover, not only algorithms can be misguided but also medical imaging data can be altered in such a way that disease patterns are artificially introduced or removed from data. Mirsky et al. [249] have proven that PACSs can be infiltrated to inject or remove lung cancer from CT images with such accuracy that expert radiologists are misguided. Approaches able to detect such manipulated images have to be established.

Aim of this thesis was development of deep learning techniques in the field of oncology, with a special focus on radiation therapy. Non-invasive imaging based risk score prediction would be of real value for an application of stereotactic body radiation therapy in renal mass patients. Head and neck cancer therapy is an essential part of radiation oncology and testing for an infection with HPV comprises a important information source for treatment decision, which may be improved by survival based risk scores. However, there are several other aspects of deep learning in radiation oncology, that were not studied here. First and foremost, dose plans are a fundamental part of radiation therapy. Targeting high doses onto the tumor while at the same time sparing surrounding healthy tissue represents the major challenge of treatment planning. Overdosage of surrounding tissue and organs at risk can lead to severe side effects, affecting patients' quality of life [250]. Therefore, a model able to predict such side effects from pre-treatment data would be highly valuable. Calculated dose plans could be tested for possible implications and possibly adapted for prevention. Moreover, dose plans include essential information about possible therapy success. Hence, inclusion of dose data into algorithms trained on therapy related endpoints, like the progression free survival model of Chapter 7, would provide

essential information to the models. At the moment, there are only a few deep learning studies that involve dose data [251], which is most likely due to the fact that development of such models is limited by the availability of appropriate datasets. However, current studies like the REQUITE trial [252] are aiming for incorporation of dose information. Therefore, inclusion of dose data into deep learning models in radiation therapy depicts one major next step. Apart from classification problems that could be approached with the help of deep learning, there are also other tasks given, specifically in the area of adaptive radiation therapy (ART). ART aims for incorporation of any changes in patient's body during treatment, online-ART even attempts immediate adoption. This requires fast algorithms for image reconstruction and registration. Deep learning models have proven to be capable of those tasks, see Wang et al. [253] and De Vos et al. [254], and could be utilized for a speed up during online-ART. The basic ability of deep learning models for a prediction of RT dose has been proven by different studies, e.g. [255, 256], but further development for clinical applicability is needed. However, successfully trained models could be applied for adjustment of dose in ART, as well as for a speed up of Monte Carlo based dose calculation algorithms [257].

Chapter 9

Conclusion

Deep learning has great power to be used for medical imaging based detection of oncological properties. However, there are several issues that have to be solved and obstacles that have to be tackled for a successful clinical application. Models have to be designed under consideration of the features specific to the data. Absence of large datasets depicts one of the main problems of the research area, but also tumor expansion has to be taken into account for selection of the right input volumes, and temporal endpoints require models able to handle time-to-event data.

Architectures developed within this thesis involve methods that facilitate training on a small number of datapoints. For the renal mass risk score classification model this was achieved by utilization of a Siamese model and development of a element-wise merging layer. The engineered architecture features a lower number of trainable parameters as a conventional convolutional neural network, trained on the complete input volume, and is therefore not as prone to overfit on the data. Siamese models are usually utilized for similarity computation between objects [150]. However, in this thesis the approach was modified to allow for lightweight input from two regions of interest. In Chapter 5 the architecture was used to process input coming from both kidneys of renal mass patients' CT images, but the developed network can also be used to handle input from the same anatomy but different imaging modalities, e.g. different MRI sequences.

All other models employed transfer and self-supervised training strategies. Both methods are well studied in the general research field of deep learning, i.e. for an application on natural images. For utilization in the medical domain, networks have been redesigned in this thesis.

Transfer learning is used to facilitate training on sparse data by injection of prior knowledge into the model. Often, studies pretrain their models on large 2D image dataset, like the ImageNet [115, 258]. However, this approach forbids input in the downstream task to be three-dimensional. Within this thesis a transfer learning model pretrained on video data, featuring two spatial and one temporal dimensions, was developed. The model allows three-dimensional pro-

cessing of medical imaging data in the downstream task. The studies conducted in section 6.2.1 were able to verify superior ability of the approach in comparison to 3D convolutional neural networks trained from scratch and 2D transfer learning approaches pretrained on ImageNet. Large video datasets are publicly available, which makes pretraining easy. The 3D transfer learning method developed in this thesis can be utilized to foster training on small datasets for all problem settings related to three-dimensional medical imaging data.

Self-supervised learning has even greater potential than transfer learning. The method enables in-domain pretraining and can be applied on larger public datasets for development of representation learning approaches, which can then be finetuned on specific problem settings with a lower amount of data available. Such feature extraction models, able to transform input data into a lower dimensional space, are of great value in the medical domain, as generation of large datasets, allowing models to be trained from scratch, is not feasible for all possible (sub-)diseases. In section 6.2.2 the masked autoencoder (MAE) model of He et al. [20] has been adopted for an utilization on medical imaging data. Code of the MAE model was published open access and has been modified to be able to handle three-dimensional data within this thesis. The architectures employs transformers, a powerful deep learning model, developed only recently. Transformer layers are heavily utilized in natural language processing models but not much used in the medical imaging domain yet. Section 6.2.2 showed that the approach features the ability to perform better than the transfer learning approach developed in section 6.2.1, but studies on larger datasets and computing resources are needed. However, a modern self-supervised learning approach featuring transformer layers has been developed within this thesis. The model has the power to be trained as a general feature extractor on input like CT or MR images, facilitating training on all types of sparse medical datasets.

Time dependent endpoints, like overall survival, are frequently encountered in oncology. Modeling of such endpoints requires architectures to be able to handle time-to-event data. Most often, studies only predict survival at a given time point, see e.g. Hosny et al. [259]. In this thesis, temporal dependent data was handled by utilization of a discrete-time survival model, allowing for incorporation of censored cases and detailed risk estimation. The model was merged with the transfer learning approach developed in section 6.2.1. Furthermore, a novel risk score, allowing for utilization of common survival metrics, was developed in Chapter 7. Significance of the risk score could be proven for prediction of tumor progression in head and neck cancer patients. The developed approach has the power to identify low and high-risk cases, which should receive dose (de-)escalation, and therefore allows for improvements in therapy outcome and a reduction of side effects.

All studies performed within this thesis were conducted on open access data or in the setting of publicly held challenges. The code for the video data based transfer learning approach and the progression free survival model has been made publicly available, making the studies reproducible.

Therefore, within this thesis customized three-dimensional approaches able to be trained on the small datasets encountered in medical imaging were de-

veloped. Furthermore, the ability of deep learning architectures for modeling of novel endpoints has been proven. As public datasets have been employed and code was published online, the models can be used by other researchers to build upon them. For advances in clinical applicability, studies on larger patient cohorts are needed.

Bibliography

[1] Hyuna Sung et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: a cancer journal for clinicians* 71.3 (2021), pp. 209–249.

[2] L.M. Franks and N.M. Teich. *Introduction to the Cellular and Molecular Biology of Cancer*. Oxford science publications. Oxford University Press, 1997. ISBN: 9780198548546.

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

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

[5] Athanassios Argiris et al. "Head and neck cancer". In: *The Lancet* 371.9625 (2008), pp. 1695–1709.

[6] Francesca Pezzuto et al. "Update on head and neck cancer: current knowledge on epidemiology, risk factors, molecular features and novel therapies". In: *Oncology* 89.3 (2015), pp. 125–136.

[7] Changxing Liu et al. "The molecular mechanisms of increased radiosensitivity of HPV-positive oropharyngeal squamous cell carcinoma (OPSCC): an extensive review". In: *Journal of Otolaryngology-Head & Neck Surgery* 47.1 (2018), pp. 1–8.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).

[9] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[10] Yann LeCun. *The MNIST database of handwritten digits*. URL: http://yann.lecun.com/exdb/mnist/ (visited on July 2, 2022).

[11] Alex Krizhevsky, Geoffrey Hinton, et al. *Learning multiple layers of features from tiny images*. URL: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf (visited on July 2, 2022).

[12] Ahmed Hosny et al. "Artificial intelligence in radiology". In: *Nature Reviews Cancer* 18.8 (2018), pp. 500–510.

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

[14] Margarita Kirienko et al. "Convolutional neural networks promising in lung cancer T-parameter assessment on baseline FDG-PET/CT". In: *Contrast Media & Molecular Imaging* 2018 (2018).

[15] Jeong Hoon Lee, Eun Ju Ha, and Ju Han Kim. "Application of deep learning to the diagnosis of cervical lymph node metastasis from thyroid cancer with CT". In: *European radiology* 29.10 (2019), pp. 5452–5457.

[16] Tafadzwa L Chaunzwa et al. "Deep learning classification of lung cancer histology using CT images". In: *Scientific reports* 11.1 (2021), pp. 1–12.

[17] Nima Tajbakhsh et al. "Convolutional neural networks for medical image analysis: Full training or fine tuning?" In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1299–1312.

[18] Hoo-Chang Shin et al. "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning". In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1285–1298.

[19] Sebastian Starke et al. "2D and 3D convolutional neural networks for outcome modelling of locally advanced head and neck squamous cell carcinoma". In: *Scientific reports* 10.1 (2020), pp. 1–13.

[20] Kaiming He et al. "Masked autoencoders are scalable vision learners". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 16000–16009.

[21] F Fiteni et al. "Endpoints in cancer clinical trials". In: *Journal of visceral surgery* 151.1 (2014), pp. 17–22.

[22] Michael F Gensheimer and Balasubramanian Narasimhan. "A scalable discrete-time survival model for neural networks". In: *PeerJ* 7 (2019), e6257.

[23] *Regulating the internet giants: the world's most valuable resource is no longer oil, but data*. 2017. URL: https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data (visited on August 26, 2022).

[24] Roger D Peng. "Reproducible research in computational science". In: *Science* 334.6060 (2011), pp. 1226–1227.

[25] Devran Ugurlu et al. "The Impact of Domain Shift on Left and Right Ventricle Segmentation in Short Axis Cardiac MR Images". In: *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer. 2021, pp. 57–65.

[26] Lena Maier-Hein et al. “Why rankings of biomedical image analysis competitions should be interpreted with care”. In: *Nature communications* 9.1 (2018), pp. 1–13.

[27] Kenneth Clark et al. “The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository”. In: *Journal of digital imaging* 26.6 (2013), pp. 1045–1057.

[28] *KNIGHT Challenge*. URL: https://research.ibm.com/haifa/Workshops/KNIGHT/ (visited on August 1, 2022).

[29] Vincent Andrearczyk et al. “Overview of the HECKTOR challenge at MICCAI 2021: automatic head and neck tumor segmentation and outcome prediction in PET/CT images”. In: *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer. 2021, pp. 1–37.

[30] Peter Hoskin, Thankamma Ajithkumar, and Vicky Goh. *Imaging for Clinical Oncology*. Oxford University Press, 2021.

[31] Michael Wannenmacher, Frederik Wenz, and Jürgen Debus. *Strahlentherapie*. Springer-Verlag, 2013.

[32] Wolfgang Schlegel, Christian P Karger, and Oliver Jäkel. *Medizinische Physik: Grundlagen–Bildgebung–Therapie–Technik*. Springer-Verlag, 2018.

[33] Daniela Schulz-Ertner, Oliver Jäkel, and Wolfgang Schlegel. “Radiation therapy with charged particles”. In: *Seminars in radiation oncology*. Vol. 16. 4. Elsevier. 2006, pp. 249–259.

[34] M Kara Bucci, Alison Bevan, and Mack Roach III. “Advances in radiation therapy: conventional to 3D, to IMRT, to 4D, and beyond”. In: *CA: a cancer journal for clinicians* 55.2 (2005), pp. 117–134.

[35] Faiz M Khan, John P Gibbons, and Paul W Sperduto, eds. *Khan’s Treatment Planning in Radiation Oncology*. 4th ed. Lippincott Williams and Wilkins, 2016.

[36] Jalil ur Rehman et al. “Intensity modulated radiation therapy: A review of current practice and future outlooks”. In: *Journal of Radiation Research and Applied Sciences* 11.4 (2018), pp. 361–367.

[37] Byungchul Cho. “Intensity-modulated radiation therapy: a review with a physics perspective”. In: *Radiation oncology journal* 36.1 (2018), p. 1.

[38] Simon S Lo et al. “Stereotactic body radiation therapy: a novel treatment modality”. In: *Nature reviews Clinical oncology* 7.1 (2010), pp. 44–54.

[39] Di Yan et al. “Adaptive radiation therapy”. In: *Physics in Medicine & Biology* 42.1 (1997), p. 123.

[40] Stephanie Lim-Reinders et al. “Online adaptive radiation therapy”. In: *International Journal of Radiation Oncology* Biology* Physics* 99.4 (2017), pp. 994–1003.

[41] Janet Husband, Rodney H Reznek, and Janet E Husband. *Imaging in Oncology*. CRC Press, 2009.

[42] James D Brierley, Mary K Gospodarowicz, and Christian Wittekind. *TNM classification of malignant tumours*. John Wiley & Sons, 2017.

[43] World Health Organization et al. *WHO handbook for reporting results of cancer treatment*. World Health Organization, 1979.

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

[45] Thierry Berghmans et al. "Primary tumor standardized uptake value (SUVmax) measured on fluorodeoxyglucose positron emission tomography (FDG-PET) is of prognostic value for survival in non-small cell lung cancer (NSCLC): a systematic review and meta-analysis (MA) by the European Lung Cancer Working Party for the IASLC Lung Cancer Staging Project". In: *Journal of Thoracic Oncology* 3.1 (2008), pp. 6–12.

[46] Seung Hyup Hyun et al. "Volume-based assessment by 18F-FDG PET/CT predicts survival in patients with stage III non-small-cell lung cancer". In: *European journal of nuclear medicine and molecular imaging* 41.1 (2014), pp. 50–58.

[47] Peng Xie et al. "18F-FDG PET or PET-CT to evaluate prognosis for head and neck cancer: a meta-analysis". In: *Journal of cancer research and clinical oncology* 137.7 (2011), pp. 1085–1093.

[48] Kunio Doi. "Diagnostic imaging over the last 50 years: research and development in medical imaging science and technology". In: *Physics in Medicine & Biology* 51.13 (2006), R5.

[49] Robert M Nishikawa et al. "Initial experience with a prototype clinical intelligent mammography workstation for computer-aided diagnosis". In: *Medical Imaging 1995: Image Processing*. Vol. 2434. International Society for Optics and Photonics. 1995, pp. 65–71.

[50] Heang-Ping Chan et al. "Improvement in radiologists' detection of clustered microcalcifications on mammograms". In: *Arbor* 1001 (1990), pp. 48109–0326.

[51] Samuel G Armato et al. "Lung cancer: performance of automated lung nodule detection applied to cancers missed in a CT screening program". In: *Radiology* 225.3 (2002), pp. 685–692.

[52] Philippe Lambin et al. "Radiomics: extracting more information from medical images using advanced feature analysis". In: *European journal of cancer* 48.4 (2012), pp. 441–446.

[53] Virendra Kumar et al. "Radiomics: the process and the challenges". In: *Magnetic resonance imaging* 30.9 (2012), pp. 1234–1248.

[54] David R Cox. “Regression models and life-tables”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), pp. 187–202.

[55] Hugo JWL Aerts et al. “Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach”. In: *Nature communications* 5.1 (2014), pp. 1–9.

[56] Marta Bogowicz et al. “Computed tomography radiomics predicts HPV status and local tumor control after definitive radiochemotherapy in head and neck squamous cell carcinoma”. In: *International Journal of Radiation Oncology* Biology* Physics* 99.4 (2017), pp. 921–928.

[57] Daniel Pinto dos Santos, Matthias Dietzel, and Bettina Baessler. *A decade of radiomics research: are images really data or just patterns in the noise?* 2021.

[58] Alex Zwanenburg et al. “The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping”. In: *Radiology* 295.2 (2020), pp. 328–338.

[59] Joost JM Van Griethuysen et al. “Computational radiomics system to decode the radiographic phenotype”. In: *Cancer research* 77.21 (2017), e104–e107.

[60] Jeremy J Erasmus et al. “Interobserver and intraobserver variability in measurement of non–small-cell carcinoma lung lesions: implications for assessment of tumor response”. In: *Journal of clinical oncology* 21.13 (2003), pp. 2574–2582.

[61] Matea Pavic et al. “Influence of inter-observer delineation variability on radiomics stability in different tumor sites”. In: *Acta Oncologica* 57.8 (2018), pp. 1070–1074.

[62] Jeffrey Wong et al. “Effects of interobserver and interdisciplinary segmentation variabilities on CT-based radiomics for pancreatic cancer”. In: *Scientific reports* 11.1 (2021), pp. 1–12.

[63] Maria Luisa Belli et al. “Quantifying the robustness of [18F] FDG-PET/CT radiomic features with respect to tumor delineation in head and neck and pancreatic cancer patients”. In: *Physica Medica* 49 (2018), pp. 105–111.

[64] Binsheng Zhao et al. “Reproducibility of radiomics for deciphering tumor phenotype with imaging”. In: *Scientific reports* 6.1 (2016), pp. 1–7.

[65] RWY Granzier et al. “MRI-based radiomics in breast cancer: Feature robustness with respect to inter-observer segmentation variability”. In: *Scientific reports* 10.1 (2020), pp. 1–11.

[66] Christoph Haarburger et al. “Radiomics feature reproducibility under inter-rater variability in segmentations of CT images”. In: *Scientific reports* 10.1 (2020), pp. 1–10.

[67] Borros Arneth. “Tumor microenvironment”. In: *Medicina* 56.1 (2019), p. 15.

[68] Fabian Spill et al. “Impact of the physical microenvironment on tumor progression and metastasis”. In: *Current opinion in biotechnology* 40 (2016), pp. 41–48.

[69] Marie-Caline Z Abadjian, W Barry Edwards, and Carolyn J Anderson. “Imaging the tumor microenvironment”. In: *Tumor Immune Microenvironment in Cancer Progression and Cancer Therapy* (2017), pp. 229–257.

[70] Warren S McCulloch and Walter Pitts. “A logical calculus of the ideas immanent in nervous activity”. In: *The bulletin of mathematical biophysics* 5.4 (1943), pp. 115–133.

[71] Frank Rosenblatt. “The perceptron: a probabilistic model for information storage and organization in the brain.” In: *Psychological review* 65.6 (1958), p. 386.

[72] Stephanie Dick. “Artificial Intelligence”. In: *Harvard Data Science Review* 1 (2019). URL: https://doi.org/10.1162/99608f92.92fe150c.

[73] L Floridi. “The 4th Revolution. How the Infosphere is Reshaping Human Reality L. Floridi”. In: *Oxford University Press*. 2014, p. 248.

[74] Nithin Buduma, Nikhil Buduma, and Joe Papa. *Fundamentals of deep learning*. ” O’Reilly Media, Inc.”, 2022.

[75] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. ” O’Reilly Media, Inc.”, 2019.

[76] Michael A Nielsen. *Neural networks and deep learning*. Vol. 25. Determination press San Francisco, CA, USA, 2015.

[77] Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. URL: probml.ai.

[78] Andreas Maier et al. “A gentle introduction to deep learning in medical image processing”. In: *Zeitschrift für Medizinische Physik* 29.2 (2019), pp. 86–101.

[79] Xavier Glorot and Yoshua Bengio. “Understanding the difficulty of training deep feedforward neural networks”. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 249–256.

[80] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. *Learning internal representations by error propagation*. Tech. rep. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[81] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. “Learning representations by back-propagating errors”. In: *nature* 323.6088 (1986), pp. 533–536.

[82] Augustin Cauchy et al. “Méthode générale pour la résolution des systemes d’équations simultanées”. In: *Comp. Rend. Sci. Paris* 25.1847 (1847), pp. 536–538.

[83] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv:1412.6980* (2014).

[84] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[85] Paula Branco, Luís Torgo, and Rita P Ribeiro. "A survey of predictive modeling on imbalanced domains". In: *ACM Computing Surveys (CSUR)* 49.2 (2016), pp. 1–50.

[86] Abdul Ghaaliq Lalkhen and Anthony McCluskey. "Clinical tests: sensitivity and specificity". In: *Continuing education in anaesthesia critical care & pain* 8.6 (2008), pp. 221–223.

[87] James A Hanley and Barbara J McNeil. "The meaning and use of the area under a receiver operating characteristic (ROC) curve." In: *Radiology* 143.1 (1982), pp. 29–36.

[88] Francois Chollet. *Deep learning with Python*. Simon and Schuster, 2021.

[89] Geoffrey E Hinton et al. "Improving neural networks by preventing co-adaptation of feature detectors". In: *arXiv:1207.0580* (2012).

[90] Connor Shorten and Taghi M Khoshgoftaar. "A survey on image data augmentation for deep learning". In: *Journal of big data* 6.1 (2019), pp. 1–48.

[91] David H Hubel and Torsten N Wiesel. "Receptive fields of single neurones in the cat's striate cortex". In: *The Journal of physiology* 148.3 (1959), p. 574.

[92] Damian Podareanu et al. "Best Practice Guide - Deep Learning". In: *Partnership for Advanced Computing in Europe (PRACE), Tech. Rep* (2019).

[93] Vincent Dumoulin and Francesco Visin. "A guide to convolution arithmetic for deep learning". In: *arXiv:1603.07285* (2016).

[94] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv:1409.1556* (2014).

[95] Du Tran et al. "Learning spatiotemporal features with 3d convolutional networks". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 4489–4497.

[96] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. "Highway networks". In: *arXiv:1505.00387* (2015).

[97] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[98] Fabian Isensee et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature methods* 18.2 (2021), pp. 203–211.

[99] Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. "Comparing images using the Hausdorff distance". In: *IEEE Transactions on pattern analysis and machine intelligence* 15.9 (1993), pp. 850–863.

[100] Maithra Raghu et al. "Transfusion: Understanding transfer learning for medical imaging". In: *Advances in neural information processing systems* 32 (2019).

[101] Pascal Vincent et al. "Extracting and composing robust features with denoising autoencoders". In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 1096–1103.

[102] Deepak Pathak et al. "Context encoders: Feature learning by inpainting". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2536–2544.

[103] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. "Unsupervised representation learning by predicting image rotations". In: *arXiv:1803.07728* (2018).

[104] Carl Doersch, Abhinav Gupta, and Alexei A Efros. "Unsupervised visual representation learning by context prediction". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1422–1430.

[105] Mehdi Noroozi and Paolo Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles". In: *European conference on computer vision*. Springer. 2016, pp. 69–84.

[106] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv:1810.04805* (2018).

[107] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[108] Aston Zhang et al. "Dive into deep learning". In: *arXiv:2106.11342* (2021).

[109] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

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

[111] Patrice Y Simard, David Steinkraus, John C Platt, et al. "Best practices for convolutional neural networks applied to visual document analysis." In: *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.* Vol. 3. 2003, pp. 958–963.

[112] Phillip Chlap et al. "A review of medical image data augmentation techniques for deep learning applications". In: *Journal of Medical Imaging and Radiation Oncology* 65.5 (2021), pp. 545–563.

[113] Mohammad Reza Hosseinzadeh Taher et al. "A systematic benchmarking analysis of transfer learning for medical image analysis". In: *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*. Springer, 2021, pp. 3–13.

[114] Yang Wen et al. "Rethinking pre-training on medical imaging". In: *Journal of Visual Communication and Image Representation* 78 (2021), p. 103145.

[115] Agustina La Greca Saint-Esteven et al. "A 2.5 D convolutional neural network for HPV prediction in advanced oropharyngeal cancer". In: *Computers in biology and medicine* (2022), p. 105215.

[116] Kaiming He et al. "Momentum contrast for unsupervised visual representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9729–9738.

[117] Ting Chen et al. "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607.

[118] Liang Chen et al. "Self-supervised learning for medical image analysis using image context restoration". In: *Medical image analysis* 58 (2019), p. 101539.

[119] Shekoofeh Azizi et al. "Big self-supervised models advance medical image classification". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 3478–3488.

[120] John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Vol. 1230. Springer, 2003.

[121] Edward L Kaplan and Paul Meier. "Nonparametric estimation from incomplete observations". In: *Journal of the American statistical association* 53.282 (1958), pp. 457–481.

[122] Frank E Harrell et al. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Vol. 3. Springer, 2015.

[123] Jared L Katzman et al. "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network". In: *BMC medical research methodology* 18.1 (2018), pp. 1–12.

[124] Travers Ching, Xun Zhu, and Lana X Garmire. "Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data". In: *PLoS computational biology* 14.4 (2018), e1006076.

[125] Harald Steck et al. "On ranking in survival analysis: Bounds on the concordance index". In: *Advances in neural information processing systems* 20 (2007).

[126] In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2022. URL: https://ieeexplore.ieee.org/xpl/conhome/9761376/proceeding.

[127] Daniel M Lang et al. “Risk Score Classification of Renal Masses on CT Imaging Data Using a Convolutional Neural Network”. In: *2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*. IEEE. 2022, pp. 1–4.

[128] Brian D Ballard and Nilmarie Guzman. “Renal Mass”. In: *StatPearls*. StatPearls Publishing, 2022.

[129] Phillip M Pierorazio et al. “Management of renal masses and localized renal cancer: systematic review and meta-analysis”. In: *The Journal of urology* 196.4 (2016), pp. 989–999.

[130] Steven C Campbell et al. “Renal mass and localized renal cancer: evaluation, management, and follow-up: AUA guideline: part I”. In: *The Journal of urology* 206.2 (2021), pp. 199–208.

[131] Scott P Campbell et al. “Stereotactic ablative radiotherapy for the treatment of clinically localized renal cell carcinoma”. In: *Journal of oncology* 2015 (2015).

[132] Carlos Nicolau et al. “Imaging characterization of renal masses”. In: *Medicina* 57.1 (2021), p. 51.

[133] Alejandro Sanchez, Adam S Feldman, and A Ari Hakimi. “Current management of small renal masses, including patient selection, renal tumor biopsy, active surveillance, and thermal ablation”. In: *Journal of Clinical Oncology* 36.36 (2018), p. 3591.

[134] Robert Abouassaly, Brian R Lane, and Andrew C Novick. “Active surveillance of renal masses in elderly patients”. In: *The Journal of urology* 180.2 (2008), pp. 505–509.

[135] Morton A Bosniak. “The current radiological approach to renal cysts.” In: *Radiology* 158.1 (1986), pp. 1–10.

[136] Ole Graumann, Susanne Sloth Osther, and Palle Jörn Sloth Osther. “Characterization of complex renal cysts: a critical evaluation of the Bosniak classification”. In: *Scandinavian journal of urology and nephrology* 45.2 (2011), pp. 84–90.

[137] Ivo G Schoots et al. “Bosniak classification for complex renal cysts reevaluated: a systematic review”. In: *The Journal of urology* 198.1 (2017), pp. 12–21.

[138] Stuart G Silverman et al. “Bosniak classification of cystic renal masses, version 2019: an update proposal and needs assessment”. In: *Radiology* 292.2 (2019), p. 475.

[139] Vincent B Ho and Peter L Choyke. “MR evaluation of solid renal masses”. In: *Magnetic Resonance Imaging Clinics* 12.3 (2004), pp. 413–427.

[140] Alexander Kutikov et al. “Incidence of benign pathologic findings at partial nephrectomy for solitary renal mass presumed to be renal cell carcinoma on preoperative imaging”. In: *Urology* 68.4 (2006), pp. 737–740.

[141] Alexander Kutikov and Robert G Uzzo. "The RENAL nephrometry score: a comprehensive standardized system for quantitating renal tumor size, location and depth". In: *The Journal of urology* 182.3 (2009), pp. 844–853.

[142] Vincenzo Ficarra et al. "Preoperative aspects and dimensions used for an anatomical (PADUA) classification of renal tumours in patients who are candidates for nephron-sparing surgery". In: *European urology* 56.5 (2009), pp. 786–793.

[143] Matthew N Simmons et al. "Kidney tumor location measurement using the C index method". In: *The Journal of urology* 183.5 (2010), pp. 1708–1713.

[144] H Gilbert Welch and William C Black. "Overdiagnosis in cancer". In: *Journal of the National Cancer Institute* 102.9 (2010), pp. 605–613.

[145] Nicholas Heller et al. "The kits19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes". In: *arXiv:1904.00445* (2019).

[146] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[147] Fabian Isensee and Klaus H Maier-Hein. "An attempt at beating the 3D U-Net". In: *arXiv:1908.02182* (2019).

[148] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[149] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[150] Davide Chicco. "Siamese neural networks: An overview". In: *Artificial Neural Networks* (2021), pp. 73–94.

[151] Suman Chaudhary, Wanting Yang, and Yan Qiang. "Deep Learning-Based Methods for Directing the Management of Renal Cancer Using CT Scan and Clinical Information". In: *2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*. IEEE. 2022, pp. 1–4.

[152] S Varsha et al. "Multi-Modal Information Fusion for Classification of Kidney Abnormalities". In: *2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*. IEEE. 2022, pp. 1–4.

[153] Sercan Ö Arik and Tomas Pfister. "Tabnet: Attentive interpretable tabular learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 8. 2021, pp. 6679–6687.

[154] Ethel-Michele De Villiers et al. "Classification of papillomaviruses". In: *Virology* 324.1 (2004), pp. 17–27.

[155] Cary A Moody and Laimonis A Laimins. "Human papillomavirus oncoproteins: pathways to transformation". In: *Nature Reviews Cancer* 10.8 (2010), pp. 550–560.

[156] Harald zur Hausen. "Condylomata acuminata and human genital cancer". In: *Cancer research* 36.2 pt 2 (1976), pp. 794–794.

[157] Lutz Gissmann et al. "Presence of human papillomavirus in genital tumors". In: *Journal of Investigative Dermatology* 83.1 (1984), S26–S28.

[158] *The Nobel Prize in Physiology or Medicine 2008*. URL: https://www.nobelprize.org/prizes/medicine/2008/summary (visited on August 16, 2022).

[159] Mark Schiffman et al. "Human papillomavirus and cervical cancer". In: *The lancet* 370.9590 (2007), pp. 890–907.

[160] Jon Mork et al. "Human papillomavirus infection as a risk factor for squamous-cell carcinoma of the head and neck". In: *New England Journal of Medicine* 344.15 (2001), pp. 1125–1131.

[161] Sara I Pai and William H Westra. "Molecular pathology of head and neck cancer: implications for diagnosis, prognosis, and treatment". In: *Annual review of pathology* 4 (2009), p. 49.

[162] James S Lewis Jr et al. "Human papillomavirus testing in head and neck carcinomas: guideline from the College of American Pathologists". In: *Archives of pathology & laboratory medicine* 142.5 (2018), pp. 559–597.

[163] Maura L Gillison et al. "Evidence for a causal association between human papillomavirus and a subset of head and neck cancers". In: *Journal of the National Cancer Institute* 92.9 (2000), pp. 709–720.

[164] Anil K Chaturvedi et al. "Human papillomavirus and rising oropharyngeal cancer incidence in the United States". In: *Journal of clinical oncology* 29.32 (2011), p. 4294.

[165] Martyn Plummer et al. "Global burden of cancers attributable to infections in 2012: a synthetic analysis". In: *The Lancet Global Health* 4.9 (2016), e609–e616.

[166] Carole Fakhry et al. "Improved survival of patients with human papillomavirus–positive head and neck squamous cell carcinoma in a prospective clinical trial". In: *Journal of the National Cancer Institute* 100.4 (2008), pp. 261–269.

[167] Andrew G Schache et al. "Evaluation of human papilloma virus diagnostic testing in oropharyngeal squamous cell carcinoma: sensitivity, specificity, and prognostic discrimination". In: *Clinical Cancer Research* 17.19 (2011), pp. 6262–6271.

[168] Richard C Jordan et al. "Validation of methods for oropharyngeal cancer HPV status determination in United States cooperative group trials". In: *The American journal of surgical pathology* 36.7 (2012), p. 945.

[169] Chao Huang et al. "Development and validation of radiomic signatures of head and neck squamous cell carcinoma molecular features and subtypes". In: *EBioMedicine* 45 (2019), pp. 70–80.

[170] Marta Bogowicz et al. “Privacy-preserving distributed learning of radiomics to predict overall survival and HPV status in head and neck cancer”. In: *Scientific reports* 10.1 (2020), pp. 1–10.

[171] Noriyuki Fujima et al. “Prediction of the human papillomavirus status in patients with oropharyngeal squamous cell carcinoma by FDG-PET imaging dataset using deep learning analysis: a hypothesis-generating study”. In: *European Journal of Radiology* 126 (2020), p. 108936.

[172] Andrej Karpathy et al. “Large-scale video classification with convolutional neural networks”. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2014, pp. 1725–1732.

[173] Sarfaraz Hussein et al. “Risk stratification of lung nodules using 3D CNN-based multi-task learning”. In: *International conference on information processing in medical imaging*. Springer. 2017, pp. 249–260.

[174] Lei Zhou et al. “Self pre-training with masked autoencoders for medical image analysis”. In: *arXiv:2203.05573* (2022).

[175] Daniel M Lang et al. “Deep learning based hpv status prediction for oropharyngeal cancer patients”. In: *Cancers* 13.4 (2021), p. 786.

[176] Jennifer Yin Yee Kwan et al. “Radiomic biomarkers to refine risk models for distant metastasis in HPV-related oropharyngeal carcinoma”. In: *International Journal of Radiation Oncology* Biology* Physics* 102.4 (2018), pp. 1107–1116.

[177] JYY Kwan et al. *Data from Radiomic Biomarkers to Refine Risk Models for Distant Metastasis in Oropharyngeal Carcinoma*. The Cancer Imaging Archive. 2019. URL: https://doi.org/10.7937/tcia.2019.8dho2gls.

[178] Aaron J Grossberg et al. “Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy”. In: *Scientific data* 5 (2018), p. 180173.

[179] H Elhalawani et al. *Radiomics outcome prediction in Oropharyngeal cancer [Dataset]*. The Cancer Imaging Archive. 20178. URL: https://doi.org/10.7937/tcia.2020.2vx6-fy46.

[180] Martin Vallieres et al. “Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer”. In: *Scientific reports* 7.1 (2017), pp. 1–14.

[181] Vallières Martin et al. *Data from Head-Neck-PET-CT*. The Cancer Imaging Archive. 2017. URL: https://doi.org/10.7937/K9/TCIA.2017.8oje5q00.

[182] L Wee and A Dekker. *Data from Head-Neck-Radiomics-HN1*. The Cancer Imaging Archive. 2019. URL: https://doi.org/10.7937/tcia.2019.8kap372n.

[183] Daniel M Lang et al. “Deep Learning Based HPV Status Prediction for Oropharyngeal Cancer Patients”. In: *arXiv:2011.08555* (2020).

[184] Du Tran et al. *C3D: Generic Features for Video Analysis*. URL: http://vlg.cs.dartmouth.edu/c3d/ (visited on August 17, 2022).

[185] Kaiming He and Xinleid Chen. *Masked Autoencoders: A PyTorch Implementation*. URL: https://github.com/facebookresearch/mae (visited on August 19, 2022).

[186] Christoph Feichtenhofer et al. "Masked Autoencoders As Spatiotemporal Learners". In: *arXiv:2205.09113* (2022).

[187] Will Kay et al. "The kinetics human action video dataset". In: *arXiv:1705.06950* (2017).

[188] Andriy Fedorov et al. "DICOM for quantitative imaging biomarker development: a standards based approach to sharing clinical data and structured PET/CT analysis results in head and neck cancer research". In: *PeerJ* 4 (2016), e2057.

[189] RR Beichel et al. *Data From QIN-HEADNECK*. The Cancer Imaging Archive. 2015. URL: https://doi.org/10.7937/K9/TCIA.2015.K0F5CGLI.

[190] Val J Lowe et al. "Multicenter trial of [18F] fluorodeoxyglucose positron emission tomography/computed tomography staging of head and neck cancer and negative predictive value and surgical impact in the N0 neck: results from ACRIN 6685". In: *Journal of Clinical Oncology* 37.20 (2019), p. 1704.

[191] P Kinahan et al. *Data from the ACRIN 6685 Trial HNSCC-FDG-PET/CT [Data set]*. TCIA. 2019. URL: https://doi.org/10.7937/K9/TCIA.2016.JQEJZZNG.

[192] Anita N Vasavada, Jonathan Danaraj, and Gunter P Siegmund. "Head and neck anthropometry, vertebral geometry and neck strength in height-matched men and women". In: *Journal of biomechanics* 41.1 (2008), pp. 114–121.

[193] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. "Using pre-training can improve model robustness and uncertainty". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2712–2721.

[194] Lin Lu et al. "Assessing agreement between radiomic features computed for multiple CT imaging settings". In: *PloS one* 11.12 (2016), e0166550.

[195] Binsheng Zhao et al. "Exploring variability in CT characterization of tumors: a preliminary phantom study". In: *Translational oncology* 7.1 (2014), pp. 88–93.

[196] Ruben THM Larue et al. "Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study". In: *Acta oncologica* 56.11 (2017), pp. 1544–1553.

[197] Dan Hendrycks and Thomas Dietterich. “Benchmarking neural network robustness to common corruptions and perturbations”. In: *arXiv:1903.12261* (2019).

[198] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. “Adversarial examples in the physical world”. In: *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[199] David Goldenberg et al. “Cystic lymph node metastasis in patients with head and neck cancer: an HPV-associated phenomenon”. In: *Head & Neck: Journal for the Sciences and Specialties of the Head and Neck* 30.7 (2008), pp. 898–903.

[200] Shahnaz Begum et al. “Detection of human papillomavirus in cervical lymph nodes: a highly effective strategy for localizing site of tumor origin”. In: *Clinical Cancer Research* 9.17 (2003), pp. 6469–6475.

[201] Ralph TH Leijenaar et al. “External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma”. In: *Acta oncologica* 54.9 (2015), pp. 1423–1429.

[202] Ralph TH Leijenaar et al. “Development and validation of a radiomic signature to predict HPV (p16) status from standard CT imaging: a multicenter study”. In: *The British journal of radiology* 91.1086 (2018), p. 20170498.

[203] Daniel M Lang et al. “Deep learning based GTV delineation and progression free survival risk score prediction for head and neck cancer patients”. In: *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer. 2021, pp. 150–159.

[204] Enrique Chajon et al. “Salivary gland-sparing other than parotid-sparing in definitive head-and-neck intensity-modulated radiotherapy does not seem to jeopardize local control”. In: *Radiation oncology* 8.1 (2013), pp. 1–9.

[205] Esther GC Troost et al. “18F-FLT PET/CT for early response monitoring and dose escalation in oropharyngeal tumors”. In: *Journal of Nuclear Medicine* 51.6 (2010), pp. 866–874.

[206] Itzhak Brook. “Late side effects of radiation treatment for head and neck cancer”. In: *Radiation Oncology Journal* 38.2 (2020), p. 84.

[207] Ari J Rosenberg and Everett E Vokes. “Optimizing treatment de-escalation in head and neck cancer: current and future perspectives”. In: *The Oncologist* 26.1 (2021), pp. 40–48.

[208] Shanthi Marur et al. “E1308: phase II trial of induction chemotherapy followed by reduced-dose radiation and weekly cetuximab in patients with HPV-associated resectable squamous cell carcinoma of the oropharynx—ECOG-ACRIN Cancer Research Group”. In: *Journal of Clinical Oncology* 35.5 (2017), p. 490.

[209] Adam S Garden and Pierre Blanchard. "ECOG-ACRIN 1308: Commentary on a Negative Phase II Trial." In: *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 35.17 (2017), pp. 1969–1970.

[210] Hisham Mehanna et al. "Radiotherapy plus cisplatin or cetuximab in low-risk human papillomavirus-positive oropharyngeal cancer (De-ESCALaTE HPV): an open-label randomised controlled phase 3 trial". In: *The Lancet* 393.10166 (2019), pp. 51–60.

[211] Maura L Gillison et al. "Radiotherapy plus cetuximab or cisplatin in human papillomavirus-positive oropharyngeal cancer (NRG Oncology RTOG 1016): a randomised, multicentre, non-inferiority trial". In: *The Lancet* 393.10166 (2019), pp. 40–50.

[212] Stefan P Haider et al. "Potential added value of PET/CT radiomics for survival prognostication beyond AJCC 8th edition staging in oropharyngeal squamous cell carcinoma". In: *Cancers* 12.7 (2020), p. 1778.

[213] Gijs van Tulder. "elasticdeform: Elastic deformations for N-dimensional images (v0.4.9)". In: (2021). URL: https://doi.org/10.5281/zenodo.4569691.

[214] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272.

[215] Cameron Davidson-Pilon. "lifelines: survival analysis in Python". In: *Journal of Open Source Software* 4.40 (2019), p. 1317. URL: https://doi.org/10.21105/joss.01317.

[216] Juanying Xie and Ying Peng. "The head and neck tumor segmentation based on 3D U-Net". In: *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer. 2021, pp. 92–98.

[217] Chengyang An, Huai Chen, and Lisheng Wang. "A coarse-to-fine framework for head and neck tumor segmentation in CT and PET images". In: *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer. 2021, pp. 50–57.

[218] Jiangshan Lu et al. "Priori and posteriori attention for generalizing head and neck tumors segmentation". In: *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer. 2021, pp. 134–140.

[219] Fereshteh Yousefirizi and Arman Rahmim. "GAN-based bi-modal segmentation using mumford-shah loss: Application to head and neck tumors in PET-CT images". In: *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer. 2020, pp. 99–108.

[220] Jintao Ren et al. "PET Normalizations to Improve Deep Learning Auto-Segmentation of Head and Neck Tumors in 3D PET/CT". In: *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer. 2021, pp. 83–91.

[221] Numan Saeed et al. “An ensemble approach for patient prognosis of head and neck tumor using multimodal data”. In: *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer. 2021, pp. 278–286.

[222] Mohamed A Naser et al. “Progression Free Survival Prediction for Head and Neck Cancer Using Deep Learning Based on Clinical and PET/CT Imaging Data”. In: *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer. 2021, pp. 287–299.

[223] Mohammad R Salmanpour et al. “Advanced automatic segmentation of tumors and survival prediction in head and neck cancer”. In: *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer. 2021, pp. 202–210.

[224] Mingyuan Meng et al. “Multi-task deep learning for joint tumor segmentation and outcome prediction in head and neck cancer”. In: *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer. 2021, pp. 160–167.

[225] Shivakumar Gudi et al. “Interobserver variability in the delineation of gross tumour volume and specified organs-at-risk during IMRT for head and neck cancers and the impact of FDG-PET/CT on such variability at the primary site”. In: *Journal of medical imaging and radiation sciences* 48.2 (2017), pp. 184–192.

[226] Chun-Nam Yu et al. “Learning patient-specific cancer survival distributions as a sequence of dependent regressors”. In: *Advances in neural information processing systems* 24 (2011).

[227] Forrest Iandola et al. “Densenet: Implementing efficient convnet descriptor pyramids”. In: *arXiv:1404.1869* (2014).

[228] Bela S Purohit et al. “FDG-PET/CT pitfalls in oncological head and neck imaging”. In: *Insights into imaging* 5.5 (2014), pp. 585–602.

[229] BS Chera et al. “Initial results from a phase 2 prospective trial of deintensified chemoradiation therapy for low-risk HPV-associated oropharyngeal squamous cell carcinoma”. In: *International Journal of Radiation Oncology, Biology, Physics* 100.5 (2018), pp. 1309–1310.

[230] B Zurl et al. “Hounsfield units variations”. In: *Strahlentherapie und Onkologie* 190 (2014), pp. 88–93.

[231] Erlend Peter Skaug Sande et al. “Interphantom and interscanner variations for Hounsfield units—establishment of reference values for HU in a commercial QA phantom”. In: *Physics in Medicine & Biology* 55.17 (2010), p. 5123.

[232] Chen Sun et al. “Revisiting unreasonable effectiveness of data in deep learning era”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 843–852.

[233] Georgios Kaissis et al. “End-to-end privacy preserving deep learning on multi-institutional medical imaging”. In: *Nature Machine Intelligence* 3.6 (2021), pp. 473–484.

[234] Kicky G van Leeuwen et al. “Artificial intelligence in radiology: 100 commercially available products and their scientific evidence”. In: *European radiology* 31.6 (2021), pp. 3797–3804.

[235] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó. “The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database”. In: *NPJ digital medicine* 3.1 (2020), pp. 1–8.

[236] Maciej A Mazurowski. “Do we expect more from radiology AI than from radiologists?” In: *Radiology: Artificial Intelligence* 3.4 (2021).

[237] Wei Chang Colin Tan et al. “Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy”. In: *Cancer Communications* 40.4 (2020), pp. 135–153.

[238] Cornelia M Focke et al. “Interlaboratory variability of Ki67 staining in breast cancer”. In: *European Journal of Cancer* 84 (2017), pp. 219–227.

[239] Bas HM van der Velden et al. “Explainable artificial intelligence (XAI) in deep learning-based medical image analysis”. In: *Medical Image Analysis* (2022), p. 102470.

[240] Julius Adebayo et al. “Sanity checks for saliency maps”. In: *Advances in neural information processing systems* 31 (2018).

[241] Cynthia Rudin. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215.

[242] Chaofan Chen et al. “This looks like that: deep learning for interpretable image recognition”. In: *Advances in neural information processing systems* 32 (2019).

[243] Esther Puyol-Antón et al. “Fairness in cardiac magnetic resonance imaging: assessing sex and racial bias in deep learning-based segmentation”. In: *Frontiers in cardiovascular medicine* (2022), p. 664.

[244] Agostina J Larrazabal et al. “Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis”. In: *Proceedings of the National Academy of Sciences* 117.23 (2020), pp. 12592–12594.

[245] Christopher J Pannucci and Edwin G Wilkins. “Identifying and avoiding bias in research”. In: *Plastic and reconstructive surgery* 126.2 (2010), p. 619.

[246] Samuel G Finlayson et al. “Adversarial attacks on medical machine learning”. In: *Science* 363.6433 (2019), pp. 1287–1289.

[247] Naveed Akhtar and Ajmal Mian. “Threat of adversarial attacks on deep learning in computer vision: A survey”. In: *IEEE Access* 6 (2018), pp. 14410–14430.

[248] Xin Li and Dongxiao Zhu. “Robust detection of adversarial attacks on medical images”. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 1154–1158.

[249] Yisroel Mirsky et al. “CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning”. In: *28th USENIX Security Symposium (USENIX Security 19)*. 2019, pp. 461–478.

[250] Lara Barazzuol, Rob P Coppes, and Peter van Luijk. “Prevention and treatment of radiotherapy-induced side effects”. In: *Molecular oncology* 14.7 (2020), pp. 1538–1554.

[251] AL Appelt et al. “Deep Learning for Radiotherapy Outcome Prediction Using Dose Data–A Review”. In: *Clinical Oncology* 34.2 (2022), e87–e96.

[252] Petra Seibold et al. “REQUITE: a prospective multicentre cohort study of patients undergoing radiotherapy for breast, lung or prostate cancer”. In: *Radiotherapy and Oncology* 138 (2019), pp. 59–67.

[253] Ge Wang, Jong Chul Ye, and Bruno De Man. “Deep learning for tomographic image reconstruction”. In: *Nature Machine Intelligence* 2.12 (2020), pp. 737–748.

[254] Bob D De Vos et al. “A deep learning framework for unsupervised affine and deformable image registration”. In: *Medical image analysis* 52 (2019), pp. 128–143.

[255] C Kontaxis et al. “DeepDose: towards a fast dose calculation engine for radiation therapy using deep learning”. In: *Physics in Medicine & Biology* 65.7 (2020), p. 075013.

[256] Dan Nguyen et al. “A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning”. In: *Scientific reports* 9.1 (2019), pp. 1–10.

[257] Ti Bai et al. “Deep dose plugin: towards real-time Monte Carlo dose calculation through a deep learning-based denoising algorithm”. In: *Machine Learning: Science and Technology* 2.2 (2021), p. 025033.

[258] Raul Victor Medeiros Da Nóbrega et al. “Lung nodule classification via deep transfer learning in CT lung images”. In: *2018 IEEE 31st international symposium on computer-based medical systems (CBMS)*. IEEE. 2018, pp. 244–249.

[259] Ahmed Hosny et al. “Deep learning for lung cancer prognostication: a retrospective multi-cohort radiomics study”. In: *PLoS medicine* 15.11 (2018), e1002711.

Acronyms

AI artificial intelligence

ART adaptive radiation therapy

AUA American Urological Association

AUC area under the receiver operating characteristics curve

CNN convolutional neural network

CT computed tomography

DL deep learning

EBRT external beam radiation therapy

FDA United States Food and Drug Administration

GAN generative adversarial network

GPU graphics processing unit

GTV gross tumor volume

HNSCC head and neck squamous cell carcinoma

HPV human papilloma virus

HU Hounsfield units

IEEE Institute of Electrical and Electronics Engineers

IHC immunohistochemistry

IMRT intensity modulated radiotherapy

ISBI International Symposium on Biomedical Imaging

MDI mean decrease in impurity

MLP multilayer perceptron

MRI magnetic resonance imaging

MSE mean squared error

NLP natural language processing

NPV negative predictive value

OPC oropharynx cancer

OS overall survival

PACS picture archiving and communication system

PCA principle component analysis

PET positron emission tomography

PFS progression free survival

PPV positive predictive value

RCC renal cell carcinoma

ReLU rectified linear unit

ROC receiver operating characteristics

ROI region of interest

RT radiation therapy

SBRT stereotactic body radiation therapy

SPECT single photon emission computed tomography

TCIA The Cancer Imaging Archive

TME tumor microenvironment

VMAT volumetric modulated arc therapy

XAI explainable artificial intelligence

Acknowledgments

This thesis has only been possible due to the help of several people, all of which I am very grateful to. First of all, I would like to thank Jan Wilkens for always providing support, for sharing his scientific experience, for giving valuable feedback to all of our papers, and for supervising this thesis. I am particularly thankful to Stefan Bartzsch. He gave me the opportunity to conduct the research for this thesis in his working team. Fruitful discussions with him gave me inspiration to investigate new approaches that paved the way for this thesis. I gained a lot from his knowledge in scientific writing and in research as a whole and his comments improved this thesis but also all papers we wrote together. Furthermore, I am thankful that he gave me the freedom to work on my own and for the trust he put in my work.

I would like to thank Carsten Marr for being part of my thesis committee. He gave me valuable advice on how to proceed with my research, and feedback on the transfer learning study presented in this thesis. Jan Peeken provided me with a lot of insights into medicine and supported me with his knowledge in medical machine learning, for which I am grateful.

Thanks to all the members of the group of Stefan Bartzsch and Thomas Schmid at the Institute of Radiation Medicine at Helmholtz Munich as well as to the members of Jan Wilken's group at the Klinikum rechts der Isar. I valued working with all those people and enjoyed the welcoming atmosphere. Especially, I would like to thank Johanna Winter for being a perfect office colleague, for answering all my questions about official regulations concerning this thesis, and for providing valuable comments to different parts of this work.

Part of the research for this thesis has been conducted at Raja Giryes' lab at Tel Aviv University. I would like to thank Raja for hosting this stay and providing valuable advice on the transformer based self-supervised approach. Furthermore, I would like to thank the Helmholtz Information & Data Science Academy for financing the research performed in Tel Aviv.

Standing on the shoulders of giants, I would like to thank all the software developers that have chosen to dedicate their free time on the open source tools I used during my research.

Last but not least, I would like to thank all my friends and my family for supporting me during this time, without them this thesis would not have been possible.