Aus der Klinik für Radiologie
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

Prediction of lymph node infiltration by prostate cancer using deep learning on CT imaging

zur Erlangung des akademischen Grades
Doctor medicinae (Dr. med.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Alexander Hartenstein

aus San Diego, CA, USA

Datum der Promotion:

3. Dezember 2021

Table of Contents

Abstract ***1***

Abstrakt (Deutsch) ***3***

Synopsis ***5***

1 Introduction **5**

1.1 Computer diagnosis and computer aided diagnosis 5

1.2 Deep Learning 6

1.3 Neural network explainability techniques 7

1.4 Prostate Cancer 8

1.5 PSMA PET/CT 9

2 Materials and Methods **11**

2.1 Patient cohort and imaging studies 11

2.2 Dataset generation 11

2.3 Input to the neural network 12

2.4 Developing neural network architecture 12

2.5 Training neural networks 13

2.6 Non CNN classifiers 14

2.7 Development of zero-footprint radiological viewer 14

2.8 Feature visualization implementation 16

3 Results **16**

3.1 Performance of CT only status and location balanced CNNs 16

3.2 Performance of CT and mask status and location balanced CNNs, xMask 17

3.3 Performance of non-CNN classifiers 17

3.4 Performance of radiologist readers 18

3.5 Heatmaps as explanation tool 18

3.6 Feature visualization as explanation tool 20

4 Discussion **21**

4.1 Detecting and overcoming bias 21

4.2 Use of explainability tools in the clinical setting 22

4.3 Limitations of the presented study 23

5 Conclusion 24

References 25

Eidesstattliche Versicherung 31

Journal Summary List 33

Publication 35

Curriculum Vitae 45

Author Publication List 47

Acknowledgements 49

Abstract

Background

Computer aided diagnostic tools have been developed for many decades but are only widely used in very specific diagnostic areas. New algorithmic tools, specifically deep learning, have achieved high performance and may find their way into broader clinical practice in the near future. However, the high complexity of these algorithmic tools renders them effectively 'black boxes', meaning that users are unable to understand how they are able to make decisions. This 'black box' nature of deep learning severely inhibits their introduction into high risk fields such as medicine.

Objective

In this dissertation, deep learning models were used to test the feasibility of using deep learning to aid in the diagnosis of lymphatic infiltration by prostate cancer (PCa). In order to detect the presence of PCa metastasis into the lymphatic system, ^{68}Ga-PSMA-PET/CT is increasingly being performed. However, due to limitations of cost and availability, it is unlikely that ^{68}Ga-PSMA-PET/CT will be useful for large segments of the population. For this reason, computed tomography (CT) has remained the most important modality for PCa staging, despite low sensitivity and specificity being reported. The goal of this work was to train deep learning models to distinguish normal from PCa-infiltrated lymph nodes based on conventional CT scan.

Methods

From 549 patients where ^{68}Ga-PSMA-PET/CT was performed, a dataset of 2616 segmented lymph nodes was used. A label of positive or negative for infiltration was generated for each lymph node on the basis of the PET reference standard. Five convolutional neural networks (CNNs), a type of deep learning model, were trained. In order to assess radiologist performance, a zero-footprint web based radiological viewer was developed. Using this viewer, the performance two radiologist reader was assessed.

Results

The CNNs performed with an Area-Under-the-Curve between 0.95 and 0.86, compared to an average AUC of 0.81 for the experience radiologists. Of note is that CNNs were able to use anatomical surroundings to increase performance, effectively learning probabilities of infiltration by anatomical location. Two neural network explainability methods were employed to attempt understanding how CNNs achieve high classification performance. One of these methods, namely saliency map generation, provided valuable information, showing that one CNN used anatomical surroundings to increase performance. The other, known as feature visualization, did not provide useful information.

Conclusion

From this study, we find that CNNs have the potential to form the basis of a CT-based biomarker for lymph node metastasis in PCa. Additionally, segmentation masks are not required to achieve high classification performance.

Abstrakt (Deutsch)

Hintergrund

Computergestützte diagnostische Methoden sind bereits seit mehreren Jahrzehnten in der Entwicklung, finden aber bisher nur in sehr begrenzten Gebieten Anwendung. Neue algorithmische Methoden der letzten zehn Jahre, speziell “Deep-Learning-Modelle”, zeigen eine außerordentliche Leistungsfähigkeit, und könnten daher in der Zukunft Eingang in eine weitreichende klinische Praxis finden. Einschränkend muss jedoch bemerkt werden, dass diese neuen Methoden aufgrund ihrer hohen Komplexität essentiell “Black Boxes” darstellen; in anderen Worten, es ist zur Zeit für den Benutzter nicht nachvollziehbar, wie ein Deep-Learning-Modell zu bestimmten Entscheidungen gelangt. Dieser Umstand limitiert die Anwendung von Deep-Learning-Modellen in risikobehafteten medizinischen Gebieten.

Zielsetzung und Problematik

In der vorliegenden Dissertation wurden Deep-Learning-Modelle daraufhin getestet, ob sie zur radiologischen Diagnose von Lymphknoteninfiltration durch Prostatakarzinome (PCa) tauglich sind. Die lymphatische Ausdehnung eines Prostatakarzinoms ist ein wesentlicher Faktor bei der Auswahl therapeutischer Maßnahmen. Zum Nachweis lymphatischer PCa Metastasen wird in zunehmendem Maße ^{68}Ga-PSMA-PET/CT angewandt. Angesichts der Beschränkungen hinsichtlich Kosten und Verfügbarkeit ist jedoch zweifelhaft ob ^{68}Ga-PSMA-PET/CT für weite Teile der Bevölkerung eingesetzt werden kann. Aus diesem Grund verbleibt die Computertomographie (CT), trotz geringer Sensitivität und Spezifität, die wichtigste Methode zur Stadienbestimmung des PCa. Zielsetzung der vorliegenden Dissertation war es, Deep-Learning-Modelle unter Benutzung herkömmlicher Computertomographie in die Lage zu versetzen, normale von PCa-infiltrierten Lymphknoten zu unterscheiden.

Methodik

Es wurden ein Datenatz von 2616 Lymphknoten aus ^{68}Ga-PSMA-PET/CT Aufnahmen von 549 PCa Patienten verwendet. Auf der Basis des PET Referenzstandards wurde jedem dieser Lymphknoten die Beurteilung positiv oder negativ für Lymphknotenbefall zugeordnet. Fünf konvolutionelle Netzwerke (CNNs; eine spezielle Art von Deep-Learning Modellen) mit identischer Architektur wurden getestet. Der Unterschied der CNNs bestand in der Verwendung verschiedener Trainingsdaten, was es ermöglichte, die Leistungsfähigkeit der CNNs mit der Art der Eingabedaten, speziell der An- oder Abwesenheit einer Segmentierungsmaske, zu korrelieren. Zur Bestimmung der diagnostischen Treffsicherheit menschlicher Experten im Vergleich zu den CNNs wurde ein zero-footprint webbasierter radiologischer Viewer entwickelt.

Ergebnisse

Die CNNs erzielten eine Fläche unter der Kurve (AUC) zwischen 0.95 und 0.86, im Vergleich zu einem Durchschnittswert von 0.81, der von den Radiologen erreicht wurde. Interessanterweise waren CNNs in der Lage, den anatomischen Kontext zur Optimierung ihrer Leistung zu nutzen, wobei sie die Wahrscheinlichkeit des Lymphknotenbefalls in Relation zur anatomischen Lage der Lymphknoten erlernten.

Zwei ‚Explainability Methoden' wurden hinzugezogen, um die hohe Klassifizierungsleistung der CNNs zu analysieren. Eine dieser Methoden, die Erstellung von „saliency maps", ergab aussagekräftige Resultate, die darauf hinwiesen, dass das CNN die anatomische Umgebung der Lymphknoten hinzuzog, um die Unterscheidung zwischen "metastatisch-befallen" und "normal" zu treffen. Demgegenüber erbrachte die andere Methode, Merkmalsvisualisierung ("feature visualization"), keine nützlichen Erkenntnisse.

Schlussfolgerung

Unsere Studie ergibt, dass CNNs das Potential aufweisen, unter Verwendung von CT-Daten eine Beurteilung von Lymphknoten im Hinblick auf Metastasen vornehmen zu können. Des Weiteren zeigen unsere Resultate, dass Segmentierungsmasken nicht erforderlich sind, um eine hohe diagnostische Treffsicherheit der CNNs zu gewährleisten.

Synopsis

1 Introduction

1.1 Computer diagnosis and computer aided diagnosis

For several decades, medical imaging has steadily increased in importance in the early detection, diagnosis and treatment of disease[1]. It is therefore not surprising that medical imaging is performed increasingly[2]. For a large majority of these cases, expert radiologists are required in order to view and diagnose pathologies in images that are produced. However, it can be argued that artificially intelligent computer systems have been aiding radiologists for several decades. For example, the automatic exposure device developed in the 1980s[3] aided the radiologist in determining how best to take a radiograph. This device did not provide diagnostic advice, but it did remove an element of decision making from the radiologist[4]. Indeed, modern medical imaging devices undertake many thousands of automated decisions in order to transform physical input into high quality images.

Despite the fact that image generation has become a highly automated, computer aided procedure, the critical step of diagnosis has remained the purview of highly trained medical professionals, with some notable exceptions. Since the 1960's the concept of completely replacing human readers with computers has been explored in the field of automated computer diagnosis (CD), while the concept of human experts using computer output in conjunction with imaging, called computer aided diagnosis (CAD) became popularized in the 1980's[5]. The difference between CD and CAD lies in how the output of the computer is used; in CAD it is used a form of second opinion, while in CAD it is a final output. CD devices are essentially not utilized in present day practices, if not simply because CD systems to not outperform humans enough to justify their deployment, but also due to more mundane questions such as reimbursement. CAD devices, on the other hand, have been used in many instances, are since the early 2000s have become a part of routine clinical work for the detection of breast cancer on mammograms in the United States[5]. Indeed, studies have reported a 164% increase in the detection of small (less than 1 cm) invasive cancers with the use of CAD[6]. CAD systems have also been deployed for the detection of lung nodules and vertebral fractures on radiographs, intracranial aneurysms on MRA, detection of colorectal polyps, and diabetic retinopathy, and others[7–10]. The extent to which these tools are utilized and beneficial is still under rigorous investigation.

CD and CAD systems have traditionally been composed of statistical techniques and simple machine learning techniques such as principle component analysis, support vector machines and k-nearest neighbors[1112–14]. However, increasing interest and resources are being poured into a subset of machine learning techniques known as deep learning, as these methods have been shown to have increased performance. Even more interestingly, deep learning methods to a large extent remove the need of domain experts in the process of developing CAD tools, as the process of feature engineering, where the important aspects required for classification, is automated. In this dissertation, deep learning was used to created CAD-like tools. However, as will be discussed in following sections, the use of deep learning is inhibited by the lack of human interpretable explanations for automated decision processes.

1.2 Deep Learning

Deep Learning refers to the usage of a subset of machine learning algorithms that are based on cascading layers of small units known as 'neurons'[15,16]. Based very loosely on concepts from neurobiology, deep learning models, known as neural networks, are actually highly non-linear mathematical models with millions of parameters that, with the implementation of optimization algorithms, are able to learn patterns in data with remarkable accuracy[17].

The foundations for deep learning were set in the mid-20th century with the description of artificial neuron-like units by McCulloch and Pitts[18], followed by developments such as the creation of a single layer neural network and learning rule known as the perceptron by Rosenblatt[19]. Criticisms levied at the nascent field and the lack of computational power led to a loss of interest in deep learning, though backpropagation, a critical method for learning representations in the 'hidden' layers of multi-layer networks, was explored in the 80's[20]. It was only in the early 2010's that deep learning again became of interest, based upon the trifecta of deep learning models, low-cost, highly parallel computational power provided by graphics processing units (GPUs), and the existence of large labelled datasets, with curation aided by the internet. In 2011[21] and most famously in 2012 with AlexNet, neural networks outperformed all available image recognition algorithms to date by a large margin, having an error 10.8% points lower than the next runner up[22]. These achievements spurred explosive interest in the field of deep learning.

With performance often rivaling or besting that of humans in image recognition tasks, it is of great interest to implement deep learning models as decision support systems in fields such as radiology and pathology, which are to a large extent based on pattern recognition in images. As the volume of imaging performed increases, deep learning has the potential to aid the radiologist in routine tasks, for example performing a background 'triage' that flags images with severe problems to be viewed immediately by a medical expert, or automatically performing time consuming tasks such as segmentation and volume measurements of regions of interest, which over time and repeat imaging increase in value. There are also tantalizing possibilities that deep learning methods, by virtue of accessing and processing massive datasets, are able to distinguish between patterns that humans are not yet capable of distinguishing, and thus aid in areas humans cannot reliably perform, or provide a better understanding of pathology[23–25].

In this dissertation, a type of deep learning model known as a convolutional neural network (CNN) was developed and tested. These CNNs, in a process known as supervised learning, were shown many CT images of lymph nodes for which the infiltration status by PCa was known, on the basis of ^{68}Ga-PSMA PET/CT scans. Following this training process, the CNNs were shown CT images of lymph nodes it had never seen before and were asked to predict the infiltration status. The resulting performance on this test set was then compared to human readers of the same CT images. Our results showed that CNNs were able to perform remarkably well. Curiously, we did not have an explanation for how they performed as well as they did.

A serious factor inhibiting the deployment of deep learning techniques in the medical field is a lack of understanding of how decisions are reached. Because the internal mechanism of output generation (i.e. prediction) is not readily comprehensible to humans, neural networks are referred to as black box models. The black box nature of

deep learning is due to 1) the high number of parameters and 2) it's non-linear and hierarchical architecture[26]. Neural networks regularly are composed of many millions to billions of parameters, with each single neuron receiving input from many thousands of other neurons, all working together to reach a prediction, after which a non-linear activation function is applied and the output shunted to the next neuron; humans already find it difficult to understand the interplay of a linear system with two variables.

It is important to note that machine learning techniques that are traditionally considered 'interpretable' in comparison to neural networks are more often than not equally unintelligible to human interpretation. For example, a linear model or decision tree are often considered interpretable, with a serious of thresholds and simple rules to follow. However, a linear model can quickly become uninterpretable as the number of parameters and dependent variables increases in size; often linear models, too, have many thousands of parameters.

In high risk fields such as medicine, where clinical practitioners are held accountable for treatment decisions and failure to act properly can have consequences on patient health and life, it is imperative that users have, at a minimum, great confidence in a tool they use as a decision aid, if not a full and extensive understanding of how decisions are reached. The field of explainability and interpretability has been growing rapidly to fill exactly this need.

1.3 Neural network explainability techniques

Explainability techniques are categorized as global or local, with global explanations providing an explanation for the entire representation that a model has learned, and local explanations explaining a single decision on a case-by-case basis.

In the medical field, it is clear that some kind of local explanation is required in order for a clinician to have confidence in a decision. For example, if a clinician begins treatment for a tension pneumothorax based on the output of a deep learning tool presented with a chest radiograph, the deep learning tool should be able to provide evidence for why a pneumothorax is present for this particular chest radiograph. Additionally, a global explanation, which in this case would reveal what the neural network understands under the concept of 'pneumothorax' generally (and not for this particular patient), would be of significant interest to gain confidence in the system.

Current local explainability techniques include the generation for saliency maps, more generally known as heatmaps. These are well suited for imaging datasets, as the saliency map is a visual tool. As the name suggests, saliency maps depict areas of 'salience' or importance in the input space. In the radiological sense, this means that areas of an input radiographic image are highlighted if they are somehow important to decision making. For the example of a chest radiograph of a patient with pneumothorax, the hope is that a saliency map would highlight a visible visceral pleural edge or collapsed lung, as well as the lack of lung markings peripheral to this line. Unfortunately, there are many methods to generate saliency maps, and for each the 'salience' that is depicted is not exactly clear. For example, sensitivity decomposition methods highlight regions of the image that, if changed, are most likely to change the output decision[27], while layerwise relevance techniques hope to depict regions that positively led to the output decision[28].

Current global explainability techniques include representative dataset sampling and feature visualization. As previously mentioned, global techniques hope to provide an explanation for what concepts or internal representations a model has learned and do not provide explanations for any particular instance of input. In representative dataset sampling, inputs that result in high activation of a particular class are gathered the define that class. For example, all images that a neural network classifies with high output probability as 'pneumothrax' are gathered, and thus hint at the internal representation of pneumothorax.

Feature visualization is another technique developed in the hope of revealing internal model representations, and consists of using the same numerical optimization techniques used for model training to change, not model parameters, but inputs[29–31]. Model parameters are frozen, and then inputs are adjusted in an iterative process until they result in maximal output at some selected point of the model. For models that are trained using images, this results in images that maximally stimulate some selected neuron in a neural network. This has led to the discovery of 'feature detector' neurons within neural networks, that are maximally stimulated by some specific feature. For example, it has been found that there is a 'dog fuzzy ear detector' neuron within InceptionNet that 'fires' or is active whenever a dog ear is present, but does not respond to cat ears or anything else[29].

Using feature visualization, a hierarchical organization of concepts has been revealed within many neural networks. Low level neurons (neurons close to the input) act as feature detectors to simple features such as vertical or horizontal lines at varying frequencies. Neurons in intermediate layers have been found to use composite features of lower levels, for example responding maximally to corners (the joining of a horizontal and vertical lines) and in later layers joining groups of lines and edges into shapes. Even later layers reveal abstract concepts, such as human faces or animals. While these insights have been interesting, how such features will be useful for providing explanations of model decision making is not clear. Later in this dissertation, an example of feature visualization applied to a medical neural network is shown.

Despite significant shortcomings, current explainability techniques have proven useful and do provide an increased level of confidence in models generated. In the publication in this dissertation, saliency maps were used in order to a detect an error in the deep learning tool developed, which could then be rectified with further effort.

1.4 Prostate Cancer

In this dissertation, a tool to aid in the detection of the lymphatic spread of prostate cancer was developed. Prostate Cancer (PCa) is the most common malignant cancer in men, and is the second most common cause of cancer related death among men[32]. In areas with regular access to medical care, a larger proportion of men are diagnosed at younger ages and with tumors confined to the prostate, with some pointing to the advent of prostate-specific antigen (PSA) screening as a cause[33].

Upon diagnosis of PCa, an initial evaluation based on digital rectal examination (DRE), pretreatment PSA level, and the Gleason score/grade group in the initial biopsy, as well as how many and to what extent biopsy cores contain cancerous cells, is performed[34]. Based upon these factors, patients are stratified into risk categories, upon which further staging procedures and treatment are contingent. Common risk classification tools include the D'Amico classification[35], or variants thereof, such as that

of the European Association of Urology[34] . In these classifications, patients are classified as very low, low, intermediate, high or very high risk.

Whether further imaging studies or treatment is performed depends on which risk category is defined. Patients with very-low risk disease, with increased serum PSA (<10 ng/mL) but no abnormality on DRE or imaging, and biopsy histologic grade group 1 (Gleason score ≤ 6) are recommended to follow active surveillance. In active surveillance regimens, no invasive treatment is initiated, and the progression of the disease is followed. Beginning with low-risk disease, in which there is additional presence of an abnormality limited to one lobe of the prostate, patients have the choice as to whether active surveillance should be pursued, as well as the option to undergo a definitive treatment such as radical prostatectomy or radiation therapy (RT). Definitive treatment is recommended for intermediate risk patients, where tumor is restricted to the prostate but involves more than one half of one lobe or is bilateral, as well as high and very-high risk PCa, in which serum PSA ≥20 ng/mL or a histology shows a Gleason grade group of 4 or 5.

Imaging studies, including a radionuclide bone scan and computed tomography (CT) of the abdomen and pelvis are used in patients with intermediate, high or very-high risk PCa in order to determine if the primary tumor is confined to the prostate (T), if regional nodes have been infiltrated (N) and whether distant metastases are present (M)[36]. Patients with clinical evidence of lymph node involvement or disseminated metastases are not officially categorized as intermediate, high or very-high risk, regardless of extent or grade of tumor within the prostate. Presence of regional lymph node infiltration (N1) are automatically defined as prognostic stage group IVA of the Union for International Cancer Control (UICC 8th edition)[37], while those with distant metastasis to nonregional lymph nodes, bone, or other sites (M1) are stage IVB. Young men with minimal regional lymphatic spread are recommended to undergo a combination strategy with radical prostatectomy and postoperative ADT or RT, with ADT continued for 18 to 24 months after RT. Meanwhile, if distant metastasis is present and definitive locoregional therapy is not an option, therapy is centered on ADT.

As the presence of extra prostatic infiltration, either in the regional lymph nodes, or distant nonregional lymph nodes or bones, has a large impact of tumor prognostic stage group and treatments undergone, it is important that it can be reliably detected. Current imaging studies recommended, as mentioned, are the radionuclide bone scan (^{99m}Tc-bone scintigraphy) and CT imaging. Unfortunately, the sensitivity and specificity of CT imaging has been found to be lacking, at 42% and 82% respectively[38]. This is largely attributable to the fact that limited morphological criteria are used to assess lymph node infiltration status[39], with a threshold set at 8-10mm often used. This usage of a size criteria remains, even though it has been reported that 80% of infiltrated lymph nodes are smaller than 8 mm[40].

1.5 PSMA PET/CT

Efforts to improve detection of extra prostatic infiltration by PCa are areas of active interest. As discussed previously, CT imaging is the recommended modality for intermediate, high, and very-high risk patients[34]. The gold standard to detect nodal infiltration is the diagnostic pelvic lymph-node dissection (PLND); however, such invasive procedures are not suggested as a staging procedure. Beginning in the 90's, there has been an interest in preempt PLND or augment CT imaging with a more targeted imaging approach, fueled by the knowledge of a type II membrane protein

highly specific to prostate tissue. This membrane protein, known as Prostate Specific Membrane Antigen (PSMA) or glutamate carboxypeptidase II (BCPII), is produced by the prostatic epithelium and has been found to be highly restricted in its expression[41]. Not only has the expression of PSMA been found to be specific to benign and malignant prostate epithelial cells, but expression of PSMA has also been found to have a high positive correlation with grade of adenocarcinoma; it is has been shown that there is a 100-1000 fold increase in expression on the membrane of PCa cells compared to prostate cells[41,42,43]. Thus, PSMA expression is positively correlated with aggressive disease, metastasis and disease recurrence. These two qualities, 1) specificity to prostate cells and 2) positive correlation with grade, make PSMA an ideal marker for imaging of PCa and the search for metastases in compartments outside of the prostate.

Many radiolabeled small molecules with high affinity to the extracellular domain of PSMA have been introduced in the last decade[44,45], and the ^{68}Ga- and 18F-labeled PSMA-targeted ligands have been widely used in clinical practice. A number of studies have validated the use of these ligands, and have found a high sensitivity and specificity of PSMA-targeted PET in imaging PCa progression in men[46,47]. In a systematic review and meta-analysis of five studies, a sensitivity and specificity of 80% and 97% respectively was found for predicting lymph node infiltration (LNI)[46]. PSMA PET/CT has been deployed in the detection of PCa in the prostate, other soft tissues, as well as bone. Importantly, even lymph nodes under 10mm in size that have been infiltrated can be detected, and a study has reported a 60% detection rate for nodes 2-5 mm in size[45,48]. PSMA-targeted PET has been shown to be superior to ^{99m}Tc-bone scintigraphy and anatomic imaging and has been shown to identify more skeletal lesions, as well as bone marrow seeding and osteolytic metastases; the sensitivity and specificity for ^{68}Ga-PSMA PET/CT was 96.2% and 99.1% compared to 73.1% and 84.1% for ^{99m}Tc-bone scintigraphy[49].

Despite the many benefits, a number of factors inhibit the widespread use of PSMA PET/CT in the clinical setting. An important consideration in the usage and deployment of PSMA PET/CT are matters of cost and logistics, as a cyclotron facility able to deliver radionuclides within short distance is required. This fact alone will inhibit the usage of PSMA PET/CT imaging in most parts of the world. Although data is not well collected, the WHO reports the number of PET scanners per 100,000 people as 0.05 for members of the EU after May 2004 (EU13)[50], and data outside of the EU is difficult to find. A study which directly contacted nuclear medicine providers and asked if PET scanners were available for usage found that only 22.4% of high-income countries and 10.9% of low and middle income countries responded affirmatively[51]. Meanwhile, the WHO reports that there are 1.6 CT scanners available per 100,000 in the EU13[52].

Due to the lack of PSMA PET/CT availability in large parts of the world for the foreseeable future, it would be beneficial to improve the more widely available imaging modalities currently used in PCa staging. The purpose of the publication included in this dissertation is to test if deep learning methods, based on information gained from PSMA PET/CT, are able to increase the sensitivity and specificity of using CT imaging alone.

2 Materials and Methods

This dissertation utilized a large dataset of CT images of lymph nodes, labelled with infiltration status based on a ^{68}Ga-PSMA PET/CT reference standard. Using this labelled dataset, a number of convolutional neural networks (CNNs) and other classifiers were developed, trained and tested. Following this, two radiologist readers were asked to assess the infiltration status of the test set. Classification performance of a final set of five CNNs, two random forests and two radiologist readers was then compared. Two explainability techniques were tested.

2.1 Patient cohort and imaging studies

Patients were included in the dataset if ^{68}Ga-PSMA PET/CT examination with contrast-enhanced CT examination in parallel had been performed. Histological verification of the presence of prostate cancer that warranted further staging examinations was also required. 738 patients had been initially screened for adherence to inclusion criteria, and 549 patients (of age 68.7 ± 7.54 [45-87] years, PSA 20.9 ± 94.6 [0-1423] ng/ml) were finally included in the dataset. From each patient, a number of lymph nodes had been identified and semi-automatically segmented using CT imaging. A categorical label, of either positive or negative for the presence of tumor infiltration, had been assigned to each lymph node using ^{68}Ga-PSMA PET/CT scans as a reference standard, in consensus of two radiologists experienced in hybrid imaging, correlated with SUV_{max}. An additional categorical label for anatomical position had been assigned to each lymph node, manually assigning each lymph node as located in the (ascending) retroperitoneal, perirectal, cervical, mediastinal, axillary, iliacal (including obturator fossa), supra or infraclavicular, or inguinal region.

2.2 Dataset generation

Using the labelled dataset, a hold-out test set and two training datasets were then created using the pool of all labelled lymph node images created. It was clear that some kind of class balancing was necessary, as there were more lymph nodes without tumor infiltration than with tumor infiltration. Without any steps taken to rectify this inbalance, the neural network simply learns to classify most images as the larger class, as the larger class is more frequent. There are three main ways in which a class inbalance is normally accommodated for during training[16]. The first method is known as undersampling, in which the smaller class is used as the size limiting factor, and an equal amount of the other classes is used as the smaller class. This results in essentially ‘wasted’ datapoints, as the datapoints from the larger class are discarded and not used for training. Another method of class balancing is called oversampling. In this case, datapoints from the smaller class are duplicated until the model is presented with an equal number of all classes during training. Finally, the method of class weighting does not discard datapoints or duplicate the smaller class. Datapoints are presented to the model in the frequency at which they naturally appear, but the loss function used for training is adjusted by a weighting factor based on the frequency of each class. This means that the error from the smaller class has a larger impact on parameter adjustments than the error for the larger class; each example from the smaller class is ‘more important’ for training.

Class balancing by undersampling was performed in this study. First, a hold-out test set containing 130 lymph node images with exactly equal numbers of positive/negative

images was set aside. Then the first of the training datasets, known as the 'status balanced' training dataset, contained 366 positive nodes and 366 randomly selected negative nodes. Due to variations in frequency at which lymph nodes had been selected during dataset generation, lymph nodes from the inguinal, iliacal and retroperitoneal region were overrepresented in the status balanced training set and test set at 32, 23, and 19% respectively. The second training dataset, known as the 'location balanced' training dataset, was generated by taking matching positive and negative nodes equally within each anatomical category; thus, there were always an equal number of positive and negative nodes at each anatomical location for the location balanced training dataset.

2.3 Input to the neural network

Initially, three dimensional volumes were used as input to the networks, with isotropic resolution of 48x48x48. However, the large image size created using 3D volumes (with 110,592 voxels!) meant that very few images could be used in each batch during training, which had a negative impact on model training and final performance. Thus, in order to decrease model size and increase batch sizes, two dimensional images as input. Only a single central slice (along the cranial-caudal axis) was taken from each lymph node and used as input.

In order to create input images, a volume centered at the lymph node of interest was extracted and resampled to an isotropic resolution of 1 x 1 x 1 mm^3. During training only, these images were the augmented in an online manner, and for both training and testing, a single central slice in the axial plane was provided to the model as input. The resulting two-dimensional image of 48 x 48 voxels alone was used to train two models described in this dissertation (which were not described in the included publication). The segmentation of the lymph node was additionally augmented in parallel to the CT image. For two further neural networks, the CT image of the lymph node and an identically sized segmentation mask were used to train the neural network. The segmentation mask was included as an additional channel, meaning that at every voxel position there was the CT intensity value as well as the mask value of 'belonging to the lymph node' or 'belonging to the background'. Finally, for the xMask model, the CT image was multiplied with the segmentation mask. Thus the input was a single image of size 48 x 48, where all pixels not within the central lymph node had the intensity value set to 0, while all intensity values within the lymph node in question remained unchanged. For all models, the output was a binary prediction of whether the single lymph node displayed was positive or negative for tumor or not. The process of input image generation described here is shown in Fig. 2.

2.4 Developing neural network architecture

A number of neural networks architectures were experimented with in order to find a well performing network. The hold-out test set was *not* used in this process. Instead, k-fold validation was performed with k=4, iteratively setting one fourth of the training dataset aside and performing training on the remaining three fourths. Following this training, the model performance is determined using the unseen quarter. The average

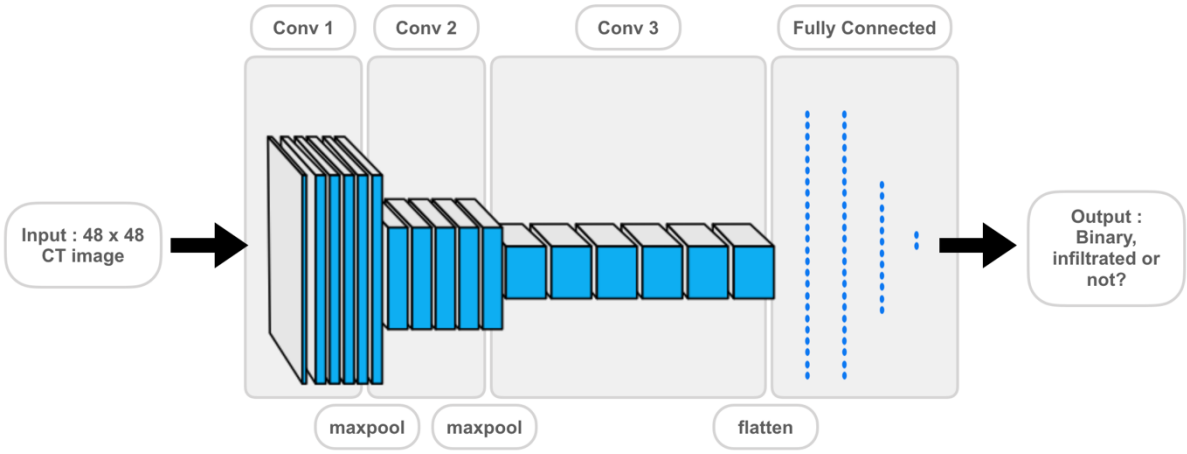


Figure 1 ***Final neural network architecture****. The three neural networks developed shared the architecture shown here and differed only in the input data used for training (either by balancing or masking). The architecture developed was inspired by successes in image classification by convolutional networks such as VGG16. Input CT images of size 48x48 are fed to a series of convolutional layers, which have maxpooling operations interspersed between them. Maxpool operations act to low the resolution of the image, thus increasing the field of view of each neuron in later layers and decreasing the model size. After the convolutional layers, the 2D image (with an additional depth channel) is flattened (i.e. made 1D) and fed to a series of fully connected layers. Finally, the output of the network results in a binary classification that reports a (pseudo) probability as to whether the input image contains tumor infiltrated lymph node or not.*

of model performance on the four validation sets was then used to compare neural network architectures.

Hyperparameters of the network architecture that were adjusted were the number of maxpool operations (from 0-3), the depth of kernels at each convolutional layer (either 8, 16, 32, 64 or 128), the number of convolutional kernels (from 2-16), the size of convolutional kernels (either 5x5 or 3x3), as well as the number and size of fully connected layers (from 2-5 layers with various sizes). This step of model architecture generation was performed using unmasked CT images and segmentations from the status balanced dataset. The final neural network architecture had 16 convolutional layers interpolated with three maxpool operations, three final fully connected layers. A depiction of the final network architecture used is shown in Fig. 1.

2.5 Training neural networks

Using the network architecture developed in the previous section, a total of five neural networks were trained (three of which were included in the final publication). The neural networks differed only by training procedure, more specifically, by which dataset was used (status or location balanced) and if or how the segmentation mask was provided. The segmentation mask was either not provided at all, was provided as an additional channel of information (such as an color image is determined by a red, a green and a blue channel), or was multiplied by the input image, thus creating a single image that was zero at all values except for inside of the segmentation (refer to Fig. 2 for an example of masked input data).

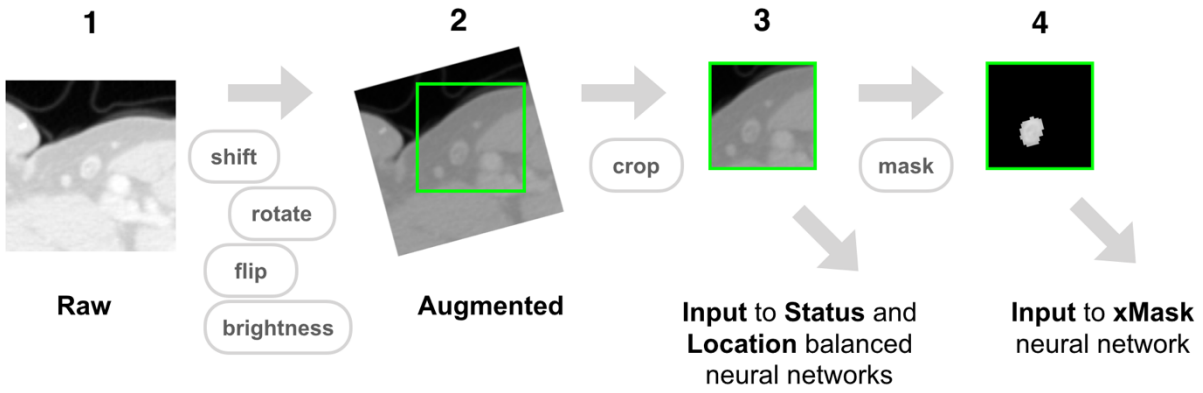


Figure 2 ***Generating inputs for neural networks.*** *A series of steps were taking before feeding images to the neural network that greatly impacted network learning. Images were initially resampled to isotropic resolution. 1) A 'raw input' image patch of size 64x64 was extracted from the isotropic image volume, centered at the lymph node in question. 2) Random image augmentation was performed during the training process only to artificially increase the dataset size so as to prevent overfitting. 3) The image was cropped to a size of 48x48, which represented the final input to the status and location balanced neural networks. 4) For the xMask neural network, the image was multiplied by the segmentation mask.*

A total of three networks were trained using the status balanced dataset (CT-only, CT+Mask, and xMask) while two networks were trained using the location balanced dataset. Heatmaps were created with the PatternAttribution method[53] using the implementation provided by the Innvestigate (v. 1.0.2) package[54].

2.6 Non CNN classifiers

It was important to see how simpler methods performed as classification tools, as neural networks, with many millions of parameters, require often unnecessary complication. The first classifier created used a size threshold of 1cm to classify lymph node infiltration. All nodes larger than 1cm were considered positive for infiltration and all below negative. As discussed in the introduction, size is currently the most relevant criteria used by radiologists to classify nodal infiltration. In order to take into account location information as well as size information, we developed decision trees and random forests. Both classifiers predicted nodal infiltration status taking only two variables, namely nodal anatomical location and volume in mm^3 into account, with the 9-category anatomical label encoded as a one-hot vector. Finally, an ensemble model was created. This model took the averaged outputs of three CNNs, namely the CT+mask status balanced, location balanced and xMask networks. The hope in this case was that some models may be better at particular nodes than others, and the combined output would be more accurate in those instances.

2.7 Development of zero-footprint radiological viewer

Knowledge of the performance of a neural network is not useful in a clinical setting unless it is known how human radiologist readers perform on the same input images. In order to perform the comparison, we required a way to present test images to radiologist readers and query their responses. A less than ideal scenario to assess radiologist performance involves providing radiologist readers either a software to download or a dedicated computer fixed to a physical location. Through this software or terminal, a directory of image files would be provided in conjunction with information as to where to locate the exact regions of interest in question. For example, we

considered providing an excel spreadsheet that defined the index of the slice at which the lymph node in question could be found. We would hope that the reader scrolls to the correct lymph node in question and then input an assessment into the excel spreadsheet.

We believed such a system of physical terminal and excel data entry to perform the reader study to be untenable. I therefore created a web-based zero-footprint viewer that ran on a server at the Charité using a python backend (utilizing Flask v. 1.0.1) and a JavaScript frontend. Readers were asked to log in to a personal account and were then presented with a randomized list of lymph nodes in sequential order (See Fig. 4a). Clicking on the 'begin study' would then enter the 'viewer' , which presented a single lymph node in three axes (transversal, sagittal, and coronal), as shown in Fig. 4b. Full radiological viewer functionality of scrubbing, panning, zooming, and window width and centering controls were provided using the CornerstoneJS tool[55]. Results were sent directly to the central server, thus allowing for fast and accurate gathering of radiologist performance with the minimum possible sharing of patient data, as image datasets did not have to stored on local computers and were only ever presented as need.

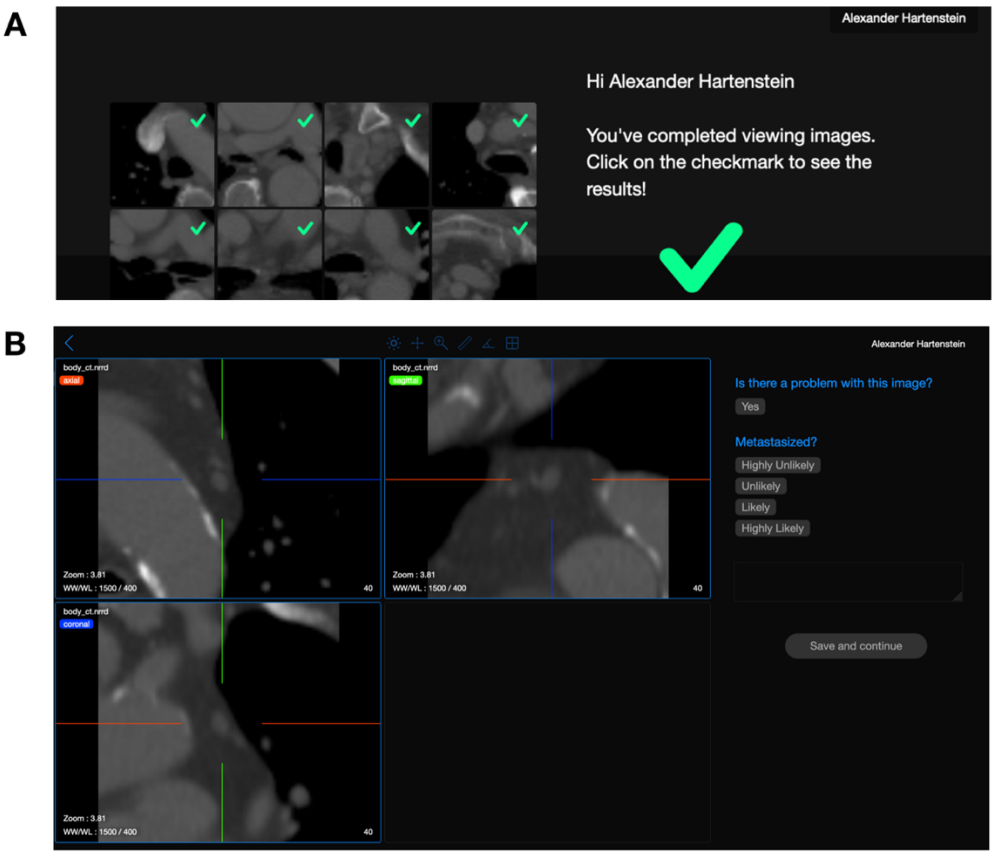


Figure 3. ***Zero-Footprint Viewer created to assess radiologist performance****. A) Upon logging into a personal account from any web browser, users are presented with a randomized list of images that they will have to assess. Clicking on the begin button leads to a radiological viewer. B) The radiological viewer shows a single region of interest, in this case a lymph node, in all three axes. The region of interest that should be assessed is automatically centered in the field of view. Readers are asked to respond to questions displayed in the right panel i.e. whether or not they believe prostate cancer has infiltrated the node in question. Responses are collected and analyzed on the central server.*

Using the zero-footprint viewer, two readers were asked to assess the hold-out test CT images (n=130) for lymph node infiltration. Each radiologist had five or more years of experience with urogenital imaging. Expert readers had access to larger images than the CNNs, as well as 3D images, at 80 x 80 x 80 mm^3, compared to the 48 x 48 mm^2 provided to the neural networks. Images were cropped so that the current lymph node was centered with a 1 x 1 x 1 mm^3 resolution. Readers reported if the tumor was infiltrated on a four value scale from very likely to very unlikely. The segmentation and ^{68}Ga-PSMA PET/CT images were not provided to the readers.

The calculated area under the receiver operating characteristic curve (AUC) for the CNNs was compared to the performance of the non-CNN classifiers and expert radiologists. As output of the CNNs is a continuous value, while the reporting of the random forest classifier and radiologist reader are discrete values, thresholds at which to compare CNNs were set by maximizing Youden's index (sensitivity+specificity-1), and all outputs below were counted as negative, and all above positive for infiltration. Classification reports including accuracy, sensitivity, specificity, PPV and NPV were calculated using these binary predictions. The four categories provided for study readers were simplified to a binary prediction of likely/unlikely.

2.8 Feature visualization implementation

Feature visualization, like saliency map generation, is a technique used to attempt to explain machine learning understanding and provide interpretations for decision making. The technique was implemented using python and tensorflow. All model weights were held fixed. Each convolutional layer was sequentially selected as the output to be optimized. Input images were initialized with gaussian randomized values between 0 and 1, and were adjusted over 200 iterations using stochastic gradient descent to maximize the selected neuron's output. Laplacian pyramid methods were used a method of frequency penalization, enforcing regularizers that reduced the amount of high frequency noise in images, as suggested by Mordvintsev[31].

3 Results

3.1 Performance of CT only status and location balanced CNNs

The status balanced CNN trained on only the CT images that received no segmentation mask performed with an AUC of 0.91. The location balanced CNN that received only CT images performed with an AUC of 0.86. Curiously, the location balanced CNN without the segmentation mask performed better than the location balanced CNN with segmentation mask, (AUC of 0.88 versus 0.86), contrasting the improved performance with the mask of the status balanced CNNs. Refer to Fig. 4.4

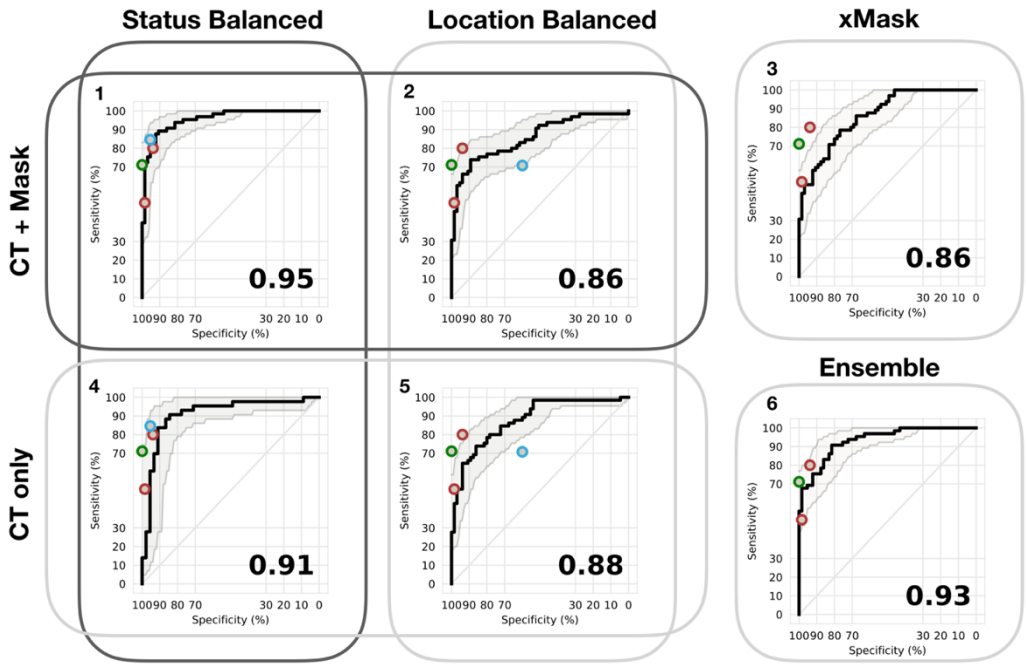


Figure 4 ***Classification performance.*** *Receiver Operator Characteristic (ROC) curves for 5 neural networks and trained to classify presence of tumor infiltration into lymph nodes. Neural networks differed by the type of input data used for training (either only CT images, CT images and a segmentation mask as a separate channel, or CT images multiplied by the segmentation mask) and the balancing of data, either by status alone or by status within each location category. The ROC of an ensemble model that averages the output of the models labeled with an asterisk is also shown. Performance of two radiologist readers is displayed as red dots on all plots. Green dots show performance using a size cut-off of 1cm. Performance of the two random forests trained on status and location balanced data is shown on the corresponding plot as a blue dot. The area under the ROC curve, AUC, is shown in the bottom right. A value of 1 is perfect performance, while a value of 0.5 is random classification. 1-3 are based on data found in Hartenstein et al [35] and are included for direct comparison of the two networks trained without segmentation masks in 4,5 and the ensemble model shown in 6.*

and 4.5 for the ROC curves of the two unmasked, CT only status and location balanced CNNs. The performance of the two radiologist readers, size classifier, and random forest classifiers are also displayed as points.

3.2 Performance of CT and mask status and location balanced CNNs, xMask

As discussed at length in the publication included in this dissertation, the masked status balanced CNN performed the best, with calculated AUC of 0.955 [56]. The xMask and location balanced CNN with mask had very similar performance with AUCs calculated as 0.863 and 0.858[56], respectively. Refer to Fig. 4.1 and 4.2 for a reiteration of the ROC curves of the CT+mask status and location balanced CNNs. The status balanced neural networks were better able to identify inguinal lymph nodes as (true) negative, and retroperitoneal lymph nodes as (true) positive, as can be seen in Fig. 5.

3.3 Performance of non-CNN classifiers

Using size as a classifier resulted in a calculated AUC of 0.85, with an accuracy, sensitivity and specificity of 81%, 71% and 98%, respectively. The size classifier performance is shown in all panels of Fig.4 as a green point.

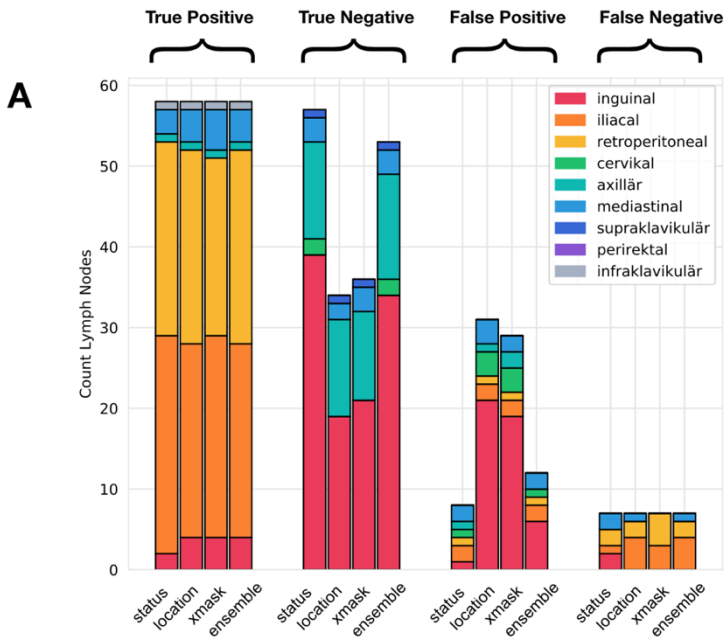
As discussed in the included publication, the status balanced random forest classifier had an AUC, accuracy, sensitivity, and specificity of .90, 90%, 84%, 95%, while the location balanced random forest had an AUC, sensitivity, specificity and accuracy of 0.654, 70%, 60% and 65%[56]. The status balanced RF is shown as a blue point in Fig. 4.1 and 4.4, while the location balanced RF is shown as a blue point in Fig. 4.2 and 4.5. The ensemble classifier did not perform better than the CT+mask status balanced CNN, but did perform better than the location balanced CNNs, with an AUC of 0.933 (95% CI from 0.894 - 0.972). The ROC of the ensemble model is shown in Fig. 4.6.

3.4 Performance of radiologist readers

As discussed at length in the included publication, the radiologist readers had AUC, accuracy, sensitivity, and specificity of 0.81, 81%, 65%, and 96% respectively when averaging their results[56]. One radiologist received a calculated AUC of 0.75, while the other had a calculated AUC of 0.87. The differences between error rate for CNNs and expert readers were not found to be statistically significant.

3.5 Heatmaps as explanation tool

Heatmaps produced for the location balanced and xMask CNN, as well as a majority of the status balanced CNN, showed high levels of attention to the lymph node in



B

	True Positive	True Negative	False Positive	False Negative
status	58	57	8	7
location	58	34	31	7
xmask	58	36	29	7
ensemble	58	53	12	7

Figure 5 ***Comparison of neural network classification performance by region****. In order to view the bias of the best performing neural network (the status balanced neural network), it is necessary to view performance within each anatomical location category. For purposes of comparison, the sensitivity of all neural networks was set at 90%. A) a stacked bar chart displaying number of lymph nodes in each location category that were correctly or incorrectly classified, each bar representing a neural network. Notice that the status balanced network is able to correctly classify inguinal lymph nodes (red) as true negative as compared to the location and xMask networks. B) The confusion matrix displaying the counts shown in A (height of each bar) for all networks and ensemble.*

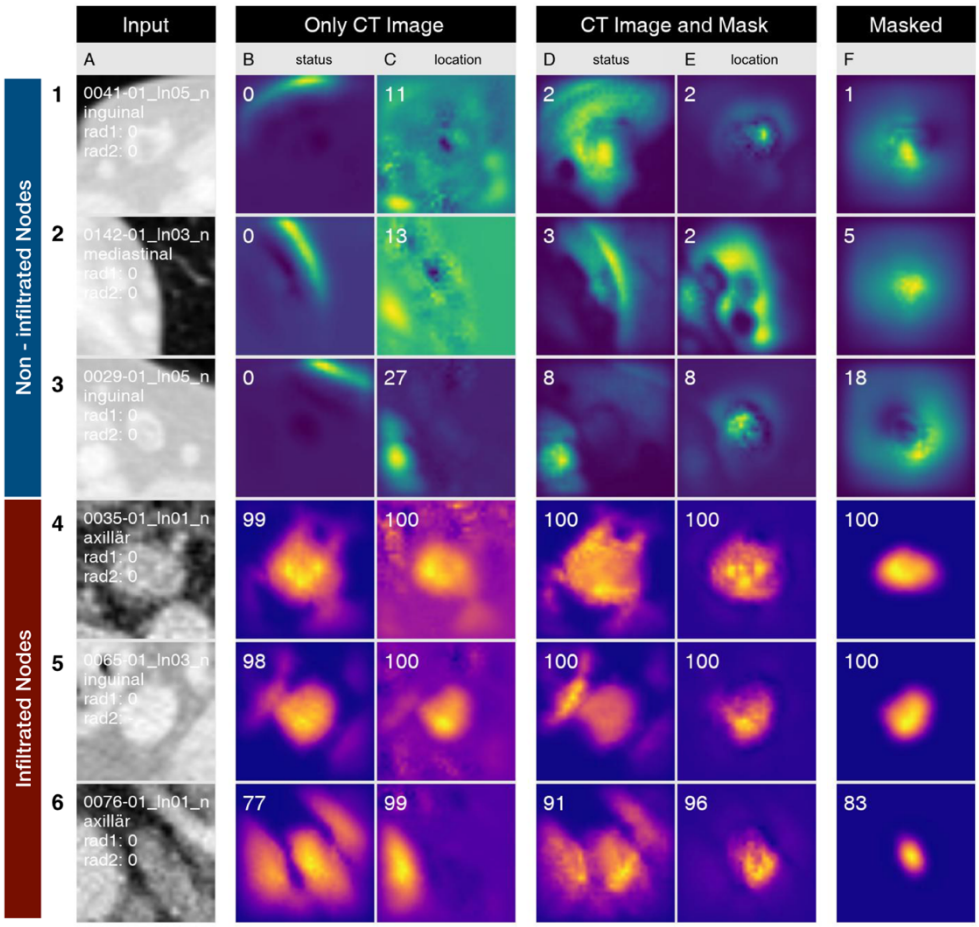


Figure 6 ***Heatmaps generated using the explainability technique of LRP****. Each row represents a distinct lymph node, with 3 non-infiltrated lymph nodes and 3 infiltrated lymph nodes in rows 1-3 and 4-6, respectively. Column A shows the input CT image for each row, along with location category for the respective node and the assessment of the two radiologist readers. All nodes displayed were assessed as 'unlikely' to contain infiltration (a score of zero) by both readers, thus representing three true positives and three false negatives for rows 1-3 and 4-6, respectively. In contrast to radiologist readers, all nodes were correctly classified by all five neural networks. Each column B-F displays heatmaps generated to explain network output for 5 distinct, separately trained networks. The CNNs in B and C received the CT image only, those in D and E received the CT image and the segmentation mask in a separated channel, while the CNN in F received a single image created by multiplying the CT image and the segmentation. The NNs in B and D received a set of 732 lymph node images balanced by lymph node infiltration only (LNI) for training (Naively Balanced, NB), while C received 555 training images balanced by location in addition to LNI (Location Balanced, LB). The value in the upper left of each image is output of the NN for the column lymph node, a pseudo probability that the lymph node in that row is positive for tumor infiltration. Stars to the right signify true output predictions, ie true positives in columns 1-3 and true negative in columns 3-6. Within heatmaps, light colors represent areas that contribute to output prediction, while dark regions do not contribute little to output prediction.*

question; further information other than 'the lymph node was relevant to output classification' could not be extrapolated from these images. Some heatmaps for xMask images had rings of demarcated relevance around the lymph node in question. For heatmaps produced for the status balanced CNN using inguinal lymph nodes as input, the air/skin border was often well demarcated, meaning that the air/skin border (a feature of 'inguinality') contributed heavily to final output classification. The demarcated air/skin border is most obvious in the images from the CT only status balanced CNN in Fig. 6 B 1-3, though it is also present in the best performing CT+mask status balanced CNN, seen in Fig. 6 D 1-2. Interestingly, there is no air/skin border seen in

CT+mask status balanced CNN in Fig. 6 D3; the presence of the mask seems to have forced attention to a non-central region of interest. The location balancing, seen in b

3.6 Feature visualization as explanation tool

The method of feature visualization produced images that suggest that neurons of lower levels resemble simple edge detectors or simple pattern detectors. Optimized images created to maximize output of low level neurons, such as Fig. 7.1 and 7.2, show repeated linear patterns in a uniform orientation; this means that the neurons in 7.1 are 'vertical line' detectors, and those in 7.2 are 'horizontal line' detectors. Meanwhile, optimized images created for neurons at higher levels are either combinations of simple features or more 'abstract concepts'; in optimized images for high level neurons (such as Fig.7.4), a circular region of high intensity in the center of the image is apparent, resembling a centered lymph node. Little information about what

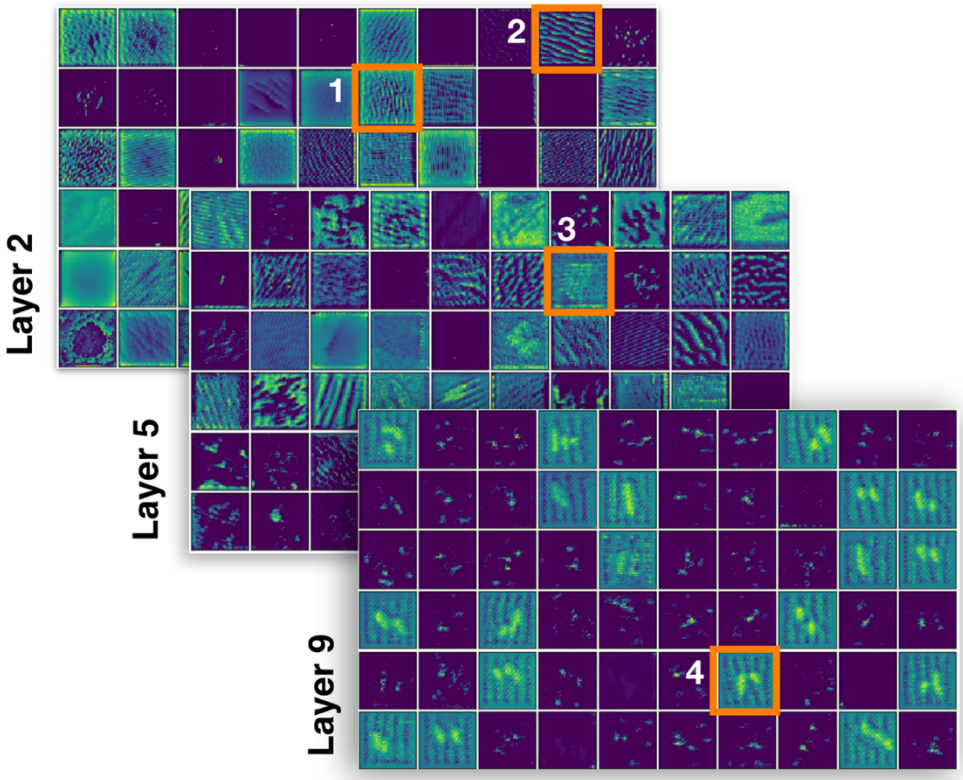


Figure 7 ***Output of feature visualization explainability technique****. The LRP method of heatmap generation used to help identify bias of our best performing neural network represents one field of explainability research. Feature visualization represents another attempt to explain neural network performance. In feature visualization, input images are generated through iterative optimization techniques to maximally activate single neurons. The hope is to find feature detector neurons that respond to obvious (human interpretable) discrete features, in the hope of gaining a global understanding of model function. Here, optimized images are shown for a set of neurons at low, middle, and high level layers (layers 2,5, and 9 respectively, with 'low' meaning near the input layer and high near the output). While the typical insights previously described for feature visualization are present, with high and low frequency pattern detectors at low levels and more complicated shape detectors at high levels, no further insights to model performance could be derived for our purposes.*

the CNN ‘understands’ about positive or negative lymph nodes is revealed by feature visualization.

4 Discussion

In the study included in this dissertation, we explored how class balancing and the presence or absence of the segmentation mask affects the performance of CNNs as imaging-based biomarkers to predict the metastatic infiltration of lymph nodes by PCa on contrast-enhanced CT images. In addition, we explored how explainability tools could help build confidence in CNNs deployed in the clinic, finding one method to be helpful and another to provide no additional information. All CNNs developed had comparable performance to two experienced human readers. CNNs also performed better than random forest classifiers that took only size and anatomical location into consideration. In addition, in our detection of bias using saliency maps, we showed that the use of explainability tools is a critical component of medical applications of deep learning-based systems.

This is, to our knowledge, the first attempt to use convolutional neural networks to find metastasis into lymph nodes by PCa using CT imaging. Current staging procedures, initiated at classification of a patient as intermediate, high or very-high risk based of DRE, serum PSA and histopathology on biopsy, include ^{99m}Tc-bone scintigraphy and CT imaging. This is despite the fact that sensitivity and specificity of CT imaging for LNI is reported as 42% and 82% respectively[38,57,58]. Currently, size is the most relevant diagnostic criteria for classification of LNI; nodes greater than 10mm are deemed as suspicious while those below are classified as benign[40]. Additional criteria used are not quantifiable and highly dependent on reader experience. Indeed, our two radiologist readers, with averaged sensitivity/specificity of 65% and 96%, performed better than reported. The higher performance of our expert readers may be attributable to their inclusion of these non-quantifiable characteristics gained during their greater than 5 years of experience; it is likely that not every CT staging procedure is read by highly experienced uroradiologists. This points to a possible benefit of a deep learning based decision support system, even one that does not greatly outperform highly specialized experts.

In addition to comparing our CNNs to radiologist readers, we compared them to random forest classifiers that took only size and anatomical location into consideration. Our CNN classifier also performed better than these random forests.

4.1 Detecting and overcoming bias

A key contribution of this work is the use of explainability tools to reveal the large impact that class balancing has on the usage of deep learning systems, and indeed all machine learning approaches. Our highest performing CNN, the status balanced CNN, was able to perform with an AUC of 0.95. However, using explainability tools, we found a strange pattern of attention to anatomical structures outside of the lymph node in question, for example the air/skin border in images of inguinal lymph nodes and structures such as the aorta. As we had also collected coarse data of anatomical location, with each lymph node given a single label of nine anatomical categories. Therefore, we were able to see that our dataset had a large bias we had been unaware of; a large majority (97%) of the inguinal lymph nodes were negative for LNI while a large majority of retroperitoneal lymph nodes were positive. Using explainability tools,

we were able to detect that our status balanced CNN had learned not to detect LNI in the lymph node, but to learn anatomical position. Indeed, the random forest that takes only size and anatomical position into account also performs well.

In order to overcome this undesirable behavior, we developed two further neural networks that would not have access to anatomical location. The first was the xMask CNN, which had all regions outside of the lymph node set to zero. The second was the location balanced CNN, which received a smaller dataset where there was exactly the same amount of positive and negative lymph nodes in each location category, and thus frequencies of infiltration in each location could not be learned. These two networks did not perform as well as the status balanced CNN, but we could be confident that one bias in our dataset was not represented in the models, and thus the results would generalize better to external datasets.

4.2 Use of explainability tools in the clinical setting

In this dissertation, two types of explainability methods were employed, namely saliency maps, and feature visualization. As previously discussed, the purpose of explainability tools is to increase confidence in machine learning models, so that they can eventually be deployed as diagnostic aids or even diagnostic tools. At present, it has been shown that machine learning models, such as neural networks, are able to perform similarly well to highly trained humans under highly controlled settings with relatively small, usually single institution datasets. Even under these controlled settings, the high numbers of parameters (into the hundreds of millions) and non-linearities mean that it is not clear how the networks achieve high classification performance.

The two explainability methods used here represent two avenues of explainability research, namely global and local explanations. These two avenues of research attempt to answer different questions, either ‘how does the *model as a whole understand* this concept?’ or ‘why did the model make *this specific decision*?’. Saliency maps are a type of local explainability technique, and thus hope to provide an explanation on a case-by-case basis; a doctor would provide a single image to the network, and then receive an explanation for why the model made that specific decision for output classification. Meanwhile, feature visualization is a method of global explainability, and hopes to reveal how abstract concepts are structured within the network. For example, with feature visualization it has been shown that there are ‘dog ear detector’ and ‘cat ear detector’ neurons, which activate only when shown that specific features; higher level neurons then use these concepts compositionally to compose representations such as ‘dog’ and ‘cat’[29].

Our use of feature visualization was underwhelming, as it seems that feature visualization itself require interpretation and may only be useful for well understood and distinct entities, such as ‘dog’ and ‘cat’; the concepts of ‘infiltrated’ and ‘non-infiltrated’ lymph node are unfortunately very similar, as both appear as circular regions of high intensity on a low intensity background. Whether or not some of the circles are bigger or represent concepts which could be used compositionally to understand infiltration is not clear.

We found the use saliency maps to be of more use to this study, as with this local explainability technique, we were able to underscore the bias that was present within our models. Using the saliency maps, we were forced to acknowledge that our network

was learning a solution to our problem that would not generalize well to real world scenarios. This usage is one of the most obvious cases in which explainability tools are useful, namely in debugging model errors. However, the saliency maps were not able to provide any additional information as to what features the model took into consideration.

4.3 Limitations of the presented study

There are several limitations to the study presented in this dissertation. We used ^{68}Ga-PSMA PET/CT as the reference standard, and used this as a proxy for presence or absence of infiltration of lymph nodes by PCa. The reported sensitivity and specificity of ^{68}Ga-PSMA PET/CT is 80% and 97%, respectively, and thus higher than that for CT only, reported at 42% and 82%, respectively. Thus the usage was justified if we wished to achieve the performance of PET/CT using CT only. However, it must be taken into consideration that some number of infiltrated nodes were not included in the training dataset due to the sensitivity of 80%. The use of this reference standard is justified due to the high specificity of our procedure; it is unlikely, with a 97% specificity, that non-infiltrated nodes were incorrectly labeled as positive. The use of ^{68}Ga-PSMA PET/CT allowed us to include patients who had not undergone PLND, as well as nodes that were in distant locations, which are not regularly biopsied. Another limitation is the large number of inguinal lymph nodes that were included in our data set. As these lymph nodes are large, they were often included, despite the fact that they were in the large majority of cases negative. This led to a class inbalance in our dataset, which required extensive balancing and use of explainability tools to overcome. We also used a binary value for presence of tumor infiltration, despite the fact that the intensity in PET scans is a continuous value. This kept a human in the process to determine what was positive and what negative. A direction of future work could be to directly predict SUVmax on ^{68}Ga-PSMA PET using only CT imaging.

In this dissertation, the CNNs created only classified images of lymph nodes as positive or negative. Finding the lymph nodes of interest on a full body CT scan was not in the scope of this project. A future direction of research is to build a tool that analyzes a full body CT scan and identifies lymph nodes positive for infiltration. Alternatively, further research could implement a workflow in which a radiologist clicks on a lymph node of interest and receives a prediction score of lymph node infiltration status in real time. Of course, usage of such a tool requires extensive research into human-machine interaction, exploring how presentation of classifier predictions affects radiologist performance.

Another limitation of this study is the lack of explainability tools which exist to provide explanations for neural networks. While we were able to use tools that exist to allow some ‘de-bugging’ of our network, in which we were able to identify a failure of the network to perform as desired, and then rectify this issue with the usage of further class balancing, further explanations are lacking. The heatmaps generated for the location balanced CNN appear visually to be similar. The heatmaps show that the CNN focuses on the lymph node itself to determine output classification, with the lymph node itself highlighted as important to output classification. However, further information is not provided, and heatmaps generated for positive and negative lymph nodes appear identical. What features the neural networks are determining are relevant to infiltration status are not revealed, and if those features are similar to the features that a radiologist would use is not clear. If size is a relevant criterion is not clear. Current

explainability tools in the field of artificial intelligence are lacking in explanatory power, and thus we cannot provide additional information of how our well-performing classifiers function.

Another limitation of this study is the use of data from only a single institute. It has been found that deep learning tools that work very well in one context or institute do not perform well when applied to dataset from different scanners or machines. A benefit of using CT imaging is that intensity values are standardized, as opposed to MRI imaging where intensity values vary even within a single patient. As the CNN classifiers trained in this network are trained with CT images, one can imagine that they will perform well. However, this has not been tested. Due to patient privacy regulations, publicly accessible datasets or sharing of datasets is not feasible. A future direction of research is to attempt federated learning, for exampling sharing the CNNs developed in this dissertation with other institutes, and having further training at these other institutes, with patient data never moving between institutes. This would overcome the single-institute overfitting that may occur and result in better real-world performance.

5 Conclusion

In this dissertation, we investigated how the presence or absence of the segmentation maps as input to neural networks affected the performance of classification of tumor infiltration of lymph nodes by PCa on CT imaging. In addition, we compared performance of our CNNs to two experienced uroradiologists. We used explainability tools in order to explore how our neural network was able to perform. One of the explainability tools, namely saliency maps, was able to show that anatomical context has a large impact on performance of CNNs and thus should be carefully considered when building imaging-based biomarkers. The other explainability tool, feature visualization, did not provide any further information to improve confidence in model decision making.

References

1. Brody, H. Medical imaging. *Nature* **502**, S81–S81 (2013).

2. Smith-Bindman, R., Kwan, M. L., Marlow, E. C., Theis, M. K., Bolch, W., Cheng, S. Y., Bowles, E. J. A., Duncan, J. R., Greenlee, R. T., Kushi, L. H., Pole, J. D., Rahm, A. K., Stout, N. K., Weinmann, S. & Miglioretti, D. L. Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016. *JAMA* **322**, 843–856 (2019).

3. Sterling, S. Automatic exposure control: a primer. *Radiol Technol* **59**, 421–427 (1988).

4. Hardy, M. & Harvey, H. Artificial intelligence in diagnostic imaging: impact on the radiography profession. *BJR* **93**, 20190840 (2019).

5. Doi, K. Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential. *Comput Med Imaging Graph* **31**, 198–211 (2007).

6. Cupples, T. E., Cunningham, J. E. & Reynolds, J. C. Impact of Computer-Aided Detection in a Regional Screening Mammography Program. *American Journal of Roentgenology* **185**, 944–950 (2005).

7. Hagiwara, Y., Koh, J. E. W., Tan, J. H., Bhandary, S. V., Laude, A., Ciaccio, E. J., Tong, L. & Acharya, U. R. Computer-aided diagnosis of glaucoma using fundus images: A review. *Computer Methods and Programs in Biomedicine* **165**, 1–12 (2018).

8. Ahmad, O. F., Soares, A. S., Mazomenos, E., Brandao, P., Vega, R., Seward, E., Stoyanov, D., Chand, M. & Lovat, L. B. Artificial intelligence and computer-aided diagnosis in colonoscopy: current evidence and future directions. *The Lancet. Gastroenterology & Hepatology* **4**, 71–80 (2019).

9. Fujita, H., Uchiyama, Y., Nakagawa, T., Fukuoka, D., Hatanaka, Y., Hara, T., Lee, G. N., Hayashi, Y., Ikedo, Y., Gao, X. & Zhou, X. Computer-aided diagnosis: the emerging of three CAD systems induced by Japanese health care needs. *Computer Methods and Programs in Biomedicine* **92**, 238–248 (2008).

10. Kasai, S., Li, F., Shiraishi, J. & Doi, K. Usefulness of computer-aided diagnosis schemes for vertebral fractures and lung nodules on chest radiographs. *AJR. American journal of roentgenology* **191**, 260–265 (2008).

11. Chen, S., Suzuki, K. & MacMahon, H. Development and evaluation of a computer-aided diagnostic scheme for lung nodule detection in chest radiographs by means of two-stage nodule enhancement with support vector classification. *Med Phys* **38**, 1844–1858 (2011).

12. Giger, M. L., Chan, H.-P. & Boone, J. Anniversary paper: History and status of CAD and quantitative image analysis: the role of Medical Physics and AAPM. *Medical Physics* **35**, 5799–5820 (2008).

13. Shiraishi, J., Li, Q., Appelbaum, D. & Doi, K. Computer-aided diagnosis and artificial intelligence in clinical imaging. *Seminars in Nuclear Medicine* **41**, 449–462 (2011).

14. Nie, Y., Li, Q., Li, F., Pu, Y., Appelbaum, D. & Doi, K. Integrating PET and CT information to improve diagnostic accuracy for lung nodules: A semiautomatic computer-aided method. *Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine* **47**, 1075–1080 (2006).

15. Bishop, C. *Pattern Recognition and Machine Learning*. (Springer-Verlag, 2006).

16. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. (Springer-Verlag, 2009). doi:10.1007/978-0-387-84858-7.

17. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

18. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* **5**, 115–133 (1943).

19. Rosenblatt, F. The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain. *Psychological Review* 65–386 (1958).

20. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).

21. Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M. & Schmidhuber, J. *Flexible, High Performance Convolutional Neural Networks for Image Classification*. (2011).

22. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1* 1097–1105 (Curran Associates Inc., 2012).

23. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* **25**, 44–56 (2019).

24. Jha, S. & Topol, E. J. Adapting to Artificial Intelligence: Radiologists and Pathologists as Information Specialists. *JAMA* **316**, 2353–2354 (2016).

25. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. *Nature Reviews. Cancer* **18**, 500–510 (2018).

26. Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M. & Kagal, L. Explaining Explanations: An Overview of Interpretability of Machine Learning. *arXiv:1806.00069 [cs, stat]* (2019).

27. Montavon, G., Binder, A., Lapuschkin, S., Samek, W. & Müller, K.-R. Layer-Wise Relevance Propagation: An Overview. in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (eds. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. & Müller, K.-R.) 193–209 (Springer International Publishing, 2019). doi:10.1007/978-3-030-28954-6_10.

28. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R. & Samek, W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE* **10**, e0130140 (2015).

29. Olah, C., Mordvintsev, A. & Schubert, L. Feature Visualization. *Distill* **2**, e7 (2017).

30. Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K. & Mordvintsev, A. The Building Blocks of Interpretability. *Distill* **3**, e10 (2018).

31. Mordvintsev, A. & Olah, C. Inceptionism: Going Deeper into Neural Networks. *Google AI Blog* http://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.

32. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2018. *CA Cancer J Clin* **68**, 7–30 (2018).

33. Barry, M. J. & Nelson, J. B. Patients Present with More Advanced Prostate Cancer since the USPSTF Screening Recommendations. *J. Urol.* **194**, 1534–1536 (2015).

34. Mottet, N., Bellmunt, J., Bolla, M., Briers, E., Cumberbatch, M. G., De Santis, M., Fossati, N., Gross, T., Henry, A. M., Joniau, S., Lam, T. B., Mason, M. D., Matveev, V. B., Moldovan, P. C., van den Bergh, R. C. N., Van den Broeck, T., van der Poel, H. G., van der Kwast, T. H., Rouvière, O., Schoots, I. G., Wiegel, T. & Cornford, P. EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. *European Urology* **71**, 618–629 (2017).

35. D'Amico, A. V., Whittington, R., Malkowicz, S. B., Schultz, D., Blank, K., Broderick, G. A., Tomaszewski, J. E., Renshaw, A. A., Kaplan, I., Beard, C. J. & Wein, A. Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. *JAMA* **280**, 969–974 (1998).

36. Carroll, P. R., Parsons, J. K., Andriole, G., Bahnson, R. R., Castle, E. P., Catalona, W. J., Dahl, D. M., Davis, J. W., Epstein, J. I., Etzioni, R. B., Farrington, T., Hemstreet, G. P., Kawachi, M. H., Kim, S., Lange, P. H., Loughlin, K. R., Lowrance, W., Maroni, P., Mohler, J., Morgan, T. M., Moses, K. A., Nadler, R. B., Poch, M., Scales, C., Shaneyfelt, T. M., Smaldone, M. C., Sonn, G., Sprenkle, P., Vickers, A. J., Wake, R., Shead, D. A. & Freedman-Cass, D. A. NCCN Guidelines Insights: Prostate Cancer Early Detection, Version 2.2016. *J Natl Compr Canc Netw* **14**, 509–519 (2016).

37. Paner, G. P., Stadler, W. M., Hansel, D. E., Montironi, R., Lin, D. W. & Amin, M. B. Updates in the Eighth Edition of the Tumor-Node-Metastasis Staging Classification for Urologic Cancers. *Eur. Urol.* **73**, 560–569 (2018).

38. Hövels, A. M., Heesakkers, R. a. M., Adang, E. M., Jager, G. J., Strum, S., Hoogeveen, Y. L., Severens, J. L. & Barentsz, J. O. The diagnostic accuracy of CT and MRI in the staging of pelvic lymph nodes in patients with prostate cancer: a meta-analysis. *Clin Radiol* **63**, 387–395 (2008).

39. Maurer, T., Gschwend, J. E., Rauscher, I., Souvatzoglou, M., Haller, B., Weirich, G., Wester, H.-J., Heck, M., Kübler, H., Beer, A. J., Schwaiger, M. & Eiber, M. Diagnostic Efficacy of (68)Gallium-PSMA Positron Emission Tomography Compared to Conventional Imaging for Lymph Node Staging of 130 Consecutive Patients with Intermediate to High Risk Prostate Cancer. *J. Urol.* **195**, 1436–1443 (2016).

40. Heesakkers, R. A., Hövels, A. M., Jager, G. J., van den Bosch, H. C., Witjes, J. A., Raat, H. P., Severens, J. L., Adang, E. M., van der Kaa, C. H., Fütterer, J. J. & Barentsz, J. MRI with a lymph-node-specific contrast agent as an alternative to CT scan and lymph-node dissection in patients with prostate cancer: a prospective multicohort study. *The Lancet Oncology* **9**, 850–856 (2008).

41. Silver, D. A., Pellicer, I., Fair, W. R., Heston, W. D. & Cordon-Cardo, C. Prostate-specific membrane antigen expression in normal and malignant human tissues. *Clin. Cancer Res.* **3**, 81–85 (1997).

42. Bostwick, D. G., Pacelli, A., Blute, M., Roche, P. & Murphy, G. P. Prostate specific membrane antigen expression in prostatic intraepithelial neoplasia and adenocarcinoma. *Cancer* **82**, 2256–2261 (1998).

43. Perner, S., Hofer, M. D., Kim, R., Shah, R. B., Li, H., Möller, P., Hautmann, R. E., Gschwend, J. E., Kuefer, R. & Rubin, M. A. Prostate-specific membrane antigen expression as a predictor of prostate cancer progression. *Hum. Pathol.* **38**, 696–701 (2007).

44. Maurer, T., Eiber, M., Schwaiger, M. & Gschwend, J. E. Current use of PSMA–PET in prostate cancer management. *Nat Rev Urol* **13**, 226–235 (2016).

45. Hofman, M. S., Hicks, R. J., Maurer, T. & Eiber, M. Prostate-specific Membrane Antigen PET: Clinical Utility in Prostate Cancer, Normal Patterns, Pearls, and Pitfalls. *Radiographics* **38**, 200–217 (2018).

46. Perera, M., Papa, N., Christidis, D., Wetherell, D., Hofman, M. S., Murphy, D. G., Bolton, D. & Lawrentschuk, N. Sensitivity, Specificity, and Predictors of Positive 68Ga-Prostate-specific Membrane Antigen Positron Emission Tomography in Advanced Prostate Cancer: A Systematic Review and Meta-analysis. *Eur. Urol.* **70**, 926–937 (2016).

47. Eiber, M., Maurer, T., Souvatzoglou, M., Beer, A. J., Ruffani, A., Haller, B., Graner, F.-P., Kübler, H., Haberkorn, U., Eisenhut, M., Wester, H.-J., Gschwend, J. E. & Schwaiger, M. Evaluation of Hybrid ^{68}Ga-PSMA Ligand PET/CT in 248 Patients with Biochemical Recurrence After Radical Prostatectomy. *J. Nucl. Med.* **56**, 668–674 (2015).

48. Leeuwen, P. J. van, Emmett, L., Ho, B., Delprado, W., Ting, F., Nguyen, Q. & Stricker, P. D. Prospective evaluation of 68Gallium-prostate-specific membrane antigen positron emission tomography/computed tomography for preoperative lymph node staging in prostate cancer. *BJU International* **119**, 209–215 (2017).

49. Lengana, T., Lawal, I. O., Boshomane, T. G., Popoola, G. O., Mokoala, K. M. G., Moshokoa, E., Maes, A., Mokgoro, N. P., Van de Wiele, C., Vorster, M. & Sathekge, M. M. 68Ga-PSMA PET/CT Replacing Bone Scan in the Initial Staging of Skeletal Metastasis in Prostate Cancer: A Fait Accompli? *Clin Genitourin Cancer* **16**, 392–401 (2018).

50. WHO European health information at your fingertips. https://gateway.euro.who.int/en/indicators/hlthres_181-positron-emission-tomography-scanners-per-100-000/visualizations/#id=28397&tab=table.

51. Vitola, J. V., Dondi, M., Prado, P., Shaw, L. & Paez, D. Worldwide Availability and Utilization of PET/CT from IAEA Survey. *Annals of Nuclear Cardiology* **5**, 44–46 (2019).

52. WHO European health information at your fingertips. https://gateway.euro.who.int/en/indicators/hlthres_37-computed-tomography-scanners-per-100-000/visualizations/#id=27697&tab=table.

53. Kindermans, P.-J., Schütt, K. T., Alber, M., Müller, K.-R., Erhan, D., Kim, B. & Dähne, S. Learning how to explain neural networks: PatternNet and PatternAttribution. *arXiv:1705.05598 [cs, stat]* (2017).

54. Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., Samek, W., Müller, K.-R., Dähne, S. & Kindermans, P.-J. iNNvestigate neural networks! *arXiv:1808.04260 [cs, stat]* (2018).

55. *cornerstonejs/cornerstone*. (cornerstone.js, 2020).

56. Hartenstein, A., Lübbe, F., Baur, A. D. J., Rudolph, M. M., Furth, C., Brenner, W., Amthauer, H., Hamm, B., Makowski, M. & Penzkofer, T. Prostate Cancer Nodal Staging:

Using Deep Learning to Predict 68Ga-PSMA-Positivity from CT Imaging Alone. *Sci Rep* **10**, 3398 (2020).

57. Engeler, C. E., Wasserman, N. F. & Zhang, G. Preoperative assessment of prostatic carcinoma by computerized tomography: Weaknessnes and new perspectives. *Urology* **40**, 346–350 (1992).

58. Flanigan, R. C., McKay, T. C., Olson, M., Shankey, T. V., Pyle, J. & Waters, W. B. Limited efficacy of preoperative computed tomographic scanning for the evaluation of lymph node metastasis in patients before radical prostatectomy. *Urology* **48**, 428–432 (1996).

Eidesstattliche Versicherung

„Ich, Alexander Hartenstein, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema: „Prediction of lymph node infiltration by prostate cancer using deep learning on CT imaging“ selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.

Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren/innen beruhen, sind als solche in korrekter Zitierung kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen) werden von mir verantwortet.

Ich versichere ferner, dass ich die in Zusammenarbeit mit anderen Personen generierten Daten, Datenauswertungen und Schlussfolgerungen korrekt gekennzeichnet und meinen eigenen Beitrag sowie die Beiträge anderer Personen korrekt kenntlich gemacht habe (siehe Anteilserklärung). Texte oder Textteile, die gemeinsam mit anderen erstellt oder verwendet wurden, habe ich korrekt kenntlich gemacht.

Meine Anteile an etwaigen Publikationen zu dieser Dissertation entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit dem/der Erstbetreuer/in, angegeben sind. Für sämtliche im Rahmen der Dissertation entstandenen Publikationen wurden die Richtlinien des ICMJE (International Committee of Medical Journal Editors; www.icmje.og) zur Autorenschaft eingehalten. Ich erkläre ferner, dass ich mich zur Einhaltung der Satzung der Charité – Universitätsmedizin Berlin zur Sicherung Guter Wissenschaftlicher Praxis verpflichte.

Weiterhin versichere ich, dass ich diese Dissertation weder in gleicher noch in ähnlicher Form bereits an einer anderen Fakultät eingereicht habe.

Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§§156, 161 des Strafgesetzbuches) sind mir bekannt und bewusst.“

Datum Unterschrift

Anteilserklärung an den erfolgten Publikationen

Publikation 1: Hartenstein A, Lübbe F, Baur ADJ, Rudolph M, Furth C, Brenner W, Amthauer H, Hamm B, Makowski M, Penzkofer T, **Prostate Cancer Nodal Staging: Using Deep Learning to Predict ^{68}Ga-PSMA-Positivity from CT Imaging Alone,** Scientific Reports, Feb 25, 2020

Beitrag im Einzelnen:

Alexander Hartenstein ist maßgeblicher Autor folgender Textbestandteile : Introduction, Materials and Methods, Results, Discussion, Conclusion. Alexander Hartenstein fertigte Abbildung 1-7 und Tabelle 1 an. Alexander Hartenstein hat die der Arbeit zugrundeliegenden neuronalen Netze und den zwei Random Forests-Classifier entwickelt, trainiert und getestet. Alexander Hartenstein hat die in der Arbeit verwendeten explainable AI Methoden angewendet (Heatmap-Generierung).

_____

Unterschrift, Datum und Stempel des/der erstbetreuenden Hochschullehrers/in

_____

Unterschrift des Doktoranden/der Doktorandin

Journal Data Filtered By: **Selected JCR Year: 2017** Selected Editions: SCIE,SSCI Selected Categories: **"MULTIDISCIPLINARY SCIENCES"** Selected Category Scheme: WoS

Gesamtanzahl: 64 Journale

Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfactor Score
1	NATURE	710,766	41.577	1.355810
2	SCIENCE	645,132	41.058	1.127160
3	Nature Communications	178,348	12.353	0.926560
4	Science Advances	10,194	11.511	0.057080
5	PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA	637,268	9.504	1.108220
6	National Science Review	952	9.408	0.004340
7	GigaScience	1,694	7.267	0.011030
8	Scientific Data	1,567	5.305	0.008550
9	Journal of Advanced Research	1,843	4.327	0.003820
10	Annals of the New York Academy of Sciences	46,160	4.277	0.033270
11	Science Bulletin	1,952	4.136	0.005900
12	Scientific Reports	192,841	4.122	0.718960
13	Journal of the Royal Society Interface	11,357	3.355	0.030960
14	Research Synthesis Methods	1,374	3.218	0.006030
15	PLoS One	582,877	2.766	1.862350
16	PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A-MATHEMATICAL PHYSICAL AND ENGINEERING SCIENCES	17,807	2.746	0.028220
17	Royal Society Open Science	2,145	2.504	0.009260
18	PROCEEDINGS OF THE ROYAL SOCIETY A-MATHEMATICAL PHYSICAL AND ENGINEERING SCIENCES	17,157	2.410	0.018270
19	PeerJ	7,377	2.118	0.031600
20	NPJ Microgravity	94	2.000	0.000350
21	SCIENCE AND ENGINEERING ETHICS	1,496	1.859	0.002520
22	COMPLEXITY	1,369	1.829	0.002380
23	Science of Nature	324	1.789	0.001260

Publication

SCIENTIFIC REPORTS
nature research

OPEN

Prostate Cancer Nodal Staging: Using Deep Learning to Predict ^{68}Ga-PSMA-Positivity from CT Imaging Alone

A. Hartenstein[1], F. Lübbe[1], A. D. J. Baur[1], M. M. Rudolph[1], C. Furth[2], W. Brenner[2], H. Amthauer[2], B. Hamm[1], M. Makowski[1,4,5] & T. Penzkofer[1,3,5*]

Lymphatic spread determines treatment decisions in prostate cancer (PCa) patients. 68Ga-PSMA-PET/CT can be performed, although cost remains high and availability is limited. Therefore, computed tomography (CT) continues to be the most used modality for PCa staging. We assessed if convolutional neural networks (CNNs) can be trained to determine 68Ga-PSMA-PET/CT-lymph node status from CT alone. In 549 patients with 68Ga-PSMA PET/CT imaging, 2616 lymph nodes were segmented. Using PET as a reference standard, three CNNs were trained. Training sets balanced for infiltration status, lymph node location and additionally, masked images, were used for training. CNNs were evaluated using a separate test set and performance was compared to radiologists' assessments and random forest classifiers. Heatmaps maps were used to identify the performance determining image regions. The CNNs performed with an Area-Under-the-Curve of 0.95 (status balanced) and 0.86 (location balanced, masked), compared to an AUC of 0.81 of experienced radiologists. Interestingly, CNNs used anatomical surroundings to increase their performance, "learning" the infiltration probabilities of anatomical locations. In conclusion, CNNs have the potential to build a well performing CT-based biomarker for lymph node metastases in PCa, with different types of class balancing strongly affecting CNN performance.

Prostate cancer (PCa) is the most common malignant cancer in men worldwide, and the second most common cause of cancer related death in men[1]. Patients with intermediate or high-risk PCa undergo regular staging examinations in order to determine if the tumor has spread beyond the prostate. As treatment success is highly dependent on the presence of systemic spread[2,3], staging procedures with high sensitivity and specificity are necessary.

Standard of care imaging for PCa staging typically includes contrast-enhanced computed tomography (CT) and Technetium-99m-methylene diphosphonate bone scans[4,5]. Despite the continued recommendation of CT in staging, it has been shown that predicting lymph node infiltration (LNI) with CT scans is not very reliable[6,7], with one study reporting a sensitivity and specificity of only 42% and 82%[8]. This low performance is most likely due to the limited morphological criteria used to define a lymph node as positive for infiltration, with size being the most relevant[9]. A threshold of 8–10 mm is often used despite the fact that 80% of lymph node metastases are less than 8 mm in the short axis[10]. Further criteria, such as status of hilum fat, nodal shape, and enhancement characteristics are used to aid diagnosis, but it remains difficult to exclude LNI in large benign hyperplastic nodes or detect it in small nodes below the size threshold[11].

In 2012 imaging agents binding to Prostate Specific Membrane Antigen (PSMA) were introduced, leading to the development of PSMA PET/CT[8]. PSMA, an integral membrane glycoprotein expressed 100–1000 fold on membranes of PCa cells compared to prostate cells, has been shown to correlate with aggressive disease, disease

[1]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Department of Radiology, Augustenburger Platz 1, 13353, Berlin, Germany. [2]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Department of Nuclear Medicine, Charitéplatz 1, 13353, Berlin, Germany. [3]Berlin Institute of Health (BIH), Anna-Louisa-Karsch-Str. 2, 10178, Berlin, Germany. [4]Institute for Diagnostic and Interventional Radiology, Klinikum rechts der Isar der Technischen Universität München, Ismaninger Straße 22, D-81675, München, Germany. [5]These authors contributed equally: M. Makowski and T. Penzkofer. *email: tobias.penzkofer@charite.de

recurrence, and metastasis[12–14], and radio-tracer targeting of PSMA in conjunction with CT has been shown in a systematic review and meta-analysis of 5 studies to predict LNI with a sensitivity and specificity of 80% and 97% respectively[15]. PSMA PET/CT has been used to detect PCa in the prostate, soft tissue, and bone, and has been shown to detect LNI in nodes even under 10 mm in size, with one study reporting a 60% detection rate for nodes between 2–5 mm[16,17].

Even though PSMA PET/CT has proven to be very valuable in PCa staging, it remains of limited availability and hybrid imaging such as PET/CT is associated with high costs. The goal of this study was to evaluate – using 68Ga-PSMA PET/CT as a reference standard – if it is possible to elucidate the status of lymph nodes based on contrast-enhanced CT images alone using deep learning in the form of convolutional neural networks (CNNs).

Materials and Methods

Imaging datasets. Inclusion criteria for this retrospective study was the availability of a 68Ga-PSMA PET/CT examination with parallel contrast-enhanced CT examination performed between September 2013 and April 2017. All patients had histopathologically verified prostate cancer that warranted staging examinations. Exclusion criteria were non-contrast or low-dose only CT examination, insufficient image quality, and follow-up studies (only the first 68Ga-PSMA PET/CT of each patient was included). Of 738 patients, 549 patients (68.7 ± 7.54 [45–87] years, PSA 20.9 ± 94.6 [0–1423] ng/ml) fulfilled our inclusion criteria. The study was approved by the Charité Ethics Committee, and due to the retrospective design, the need for informed written consent was waived by the same review board, in accordance with institutional guidelines and regulations. The study was performed in accordance with the Declaration of Helsinki.

All patients had received 68Ga-PSMA PET/CT examinations for clinical purposes during the course of treatment. A standard 68Ge/68Ga generator (Eckert and Ziegler Radiopharma GmbH, Berlin, Germany) was used for 68Ga production, and PSMA- HBED-CC (ABX GmbH, Radeberg, Germany) labelling with 68Ga was performed according to the previously described method[18]. All PET/CT images were acquired using a Gemini Astonish TF 16 PET/CT scanner (Phillips Medical Systems, Best, The Netherlands) after intravenous injection of 68Ga-PSMA-HBED-CC[19] using 3-D acquisition mode for all PET scans.

Semi-automated manual three dimensional segmentation of lymph nodes was performed using the MITK software suite (MITK v. 2016.3.0, DKFZ, Heidelberg, Germany)[20]. Using the PSMA PET image as ground truth, a label of positive or negative for tumor infiltration was generated for each lymph node in consensus of two radiologists experienced in hybrid imaging, correlated with SUVmax. Figure 1 shows an example of a 68Ga-PSMA PET/CT full body scan and two selected lymph nodes, one positive and one negative for infiltration. In addition to the tumor infiltration label, the position of each lymph node in the body was manually assigned a categorical variable from a set of 9 possible categories (inguinal, iliacal (including obturator fossa), perirectal, (ascending) retroperitoneal, axillary, mediastinal, supra or infraclavicular, and cervical).

Patient collective and dataset generation. A final set of 549 patients fulfilled the inclusion criteria. An average of 4.72 ± 0.77 (SD) lymph nodes were segmented and labelled in each patient resulting in a total of 2,616 labelled lymph nodes, with 431 of these labelled as positive for infiltration. Figure 2 shows how these images were used to generate test and training datasets, and is explained as follows. A set of 130 lymph nodes was set aside for testing all CNNs and experts. This test set was created by taking 15% of the available positive nodes (65 nodes) and matching with 65 randomly selected negative nodes to create a 50:50 class balanced set. The remaining 366 positive nodes were matched with 366 randomly selected negative nodes to create a 50:50 class balanced set referred to as the 'status balanced' training set, with a total of 732 lymph nodes. The majority of lymph nodes in the status balanced and test dataset were in the inguinal region (32%), followed by the iliacal region (23%), and retroperitoneal region (19%). Figure 3a shows anatomical distribution by training set. To investigate effects of anatomical localization on classification results, the same 366 positive nodes used to create the status balanced set were sorted by anatomical category and matched to randomly selected negative nodes from within the same anatomical category, thus creating a 50:50 class balanced set with 548 lymph nodes, referred to as the 'location balanced' training set.

Neural network training. Images were resampled to an isotropic resolution of $1 \times 1 \times 1\,mm^3$. A volume of $80 \times 80 \times 80\,mm^3$ was cropped around the lymph node, centered at the center point of the manual lymph node segmentation. Image augmentation was performed online during model training, while only non-augmented images were provided to the model during validation and testing. A total of four random augmentations were performed: brightness was augmented by a factor between 0.5 and 1.5, after which images were rotated between ± 180 degrees, translated by a maximum of 5 voxels in the x, y and/or z axis, and finally flipped across the sagittal or axial plane or both. In order to ensure that no 'black borders' (i.e. areas with no image data due to rotation and shifting during augmentation) would be fed to the model, images were again cropped to a final volume of $48 \times 48 \times 48\,mm^3$. Finally, a single axial central slice was provided to the model as input.

Networks received two-dimensional images of 48×48 voxels and output a binary prediction whether or not the single lymph node displayed contained tumor or not. A final network architecture with 16 convolutional layers and three densely connected layers, inspired by the success of similar architectures by the Visual Geometry Group (VGGNet)[21], was selected using k-fold validation with $k = 10$. Figure 4 shows the architecture used by all CNNs. CNNs were not pre-trained. Batch normalization was performed after every layer, with rectified linear units (ReLU) used as the activation function. The output of the convolutional layers was fed to a fully connected feed forward network with 3 hidden layers. Adam optimization was used to update network weights[22], with parameters for alpha, beta1, beta2 and epsilon set at 0.0001, 0.9, 0.999 and 1e-08.

Three separate CNNs were trained. All models shared identical network architecture and were distinguished by the dataset used to train them: the status balanced model received status balanced CT images and

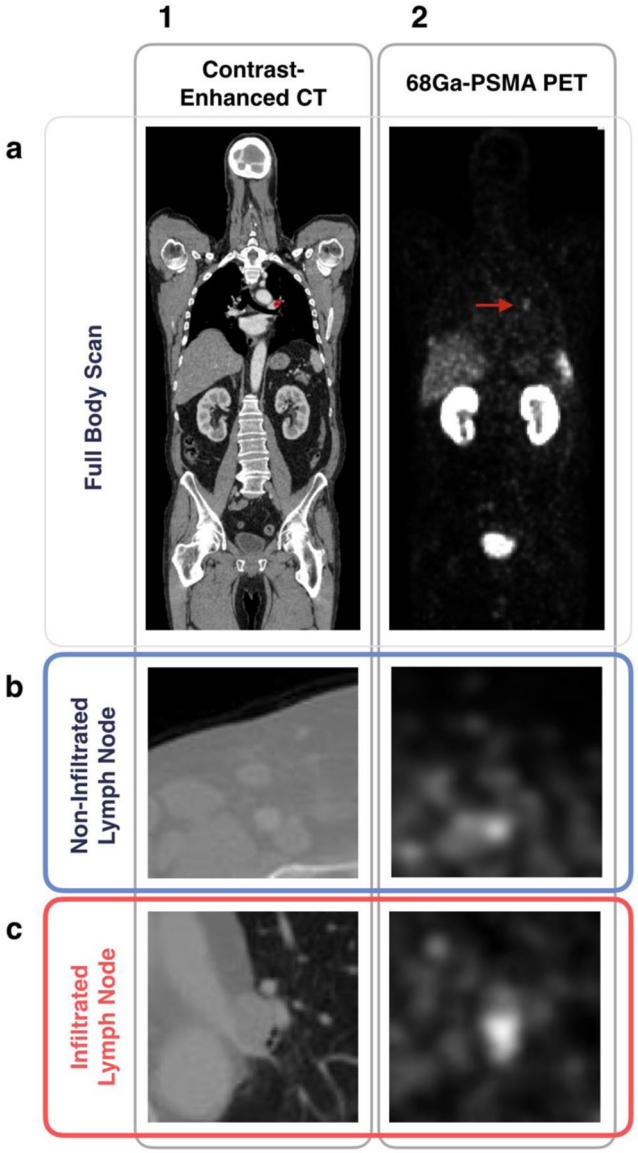


Figure 1. Generation of Labelled Dataset. (**a**) Imaging of a single patient with (1) a contrast-enhanced CT scan and (2) a 68Ga-PSMA PET scan. An average of 4.72 ± 0.77 lymph nodes were selected and semi-automatically segmented for each patient. A single lymph node positive for infiltration by PCa can be seen in the mediastinal region outlined in red in the CT image in (a1), and demarcated by a red arrow in PET scan in (a2). Using the 68Ga-PSMA PET/CT as our reference standard, a label for infiltration status by prostate cancer (either positive or negative) was assigned on a per lymph node basis. (**b**) An example of a negative 68Ga-PSMA PET/CT image pair in which the centered lymph node does not exceed background. (**c**) An example of a positive image pair.

segmentations, the location balanced model received location balanced CT images, and the xMask model received status balanced CT images multiplied by their corresponding segmentation mask. All models were implemented in Keras and Tensorflow (v. 1.10.1) and run on a Nvidia TITAN Xp graphics card (NVIDIA Titan Xp, Rev A1, Santa Clara, CA, United States). Heatmaps were generated using the Innvestigate (v. 1.0.2) package[23] using the PatternAttribution method[24].

Random forests. In order to validate neural network performance, random forests were generated to predict nodal infiltration status taking only nodal volume in mm^3 and nodal anatomical location into account. Two random forests were trained for each of the training sets used (status balanced, location balanced). Anatomical location was encoded as a one hot vector. Random forests were implemented using the sklearn python package[25] with maximum depth set at 5 to prevent overfitting to the training data.

Study readers. Two radiologists, with at least 5 years of experience in urogenital imaging, were presented with all test CT images ($n = 130$). Radiologists were presented an $80 \times 80 \times 80\,mm^3$ volume centered on the lymph node in question at $1 \times 1 \times 1\,mm^3$ resolution, and were asked to categorize the likelihood of lymph node

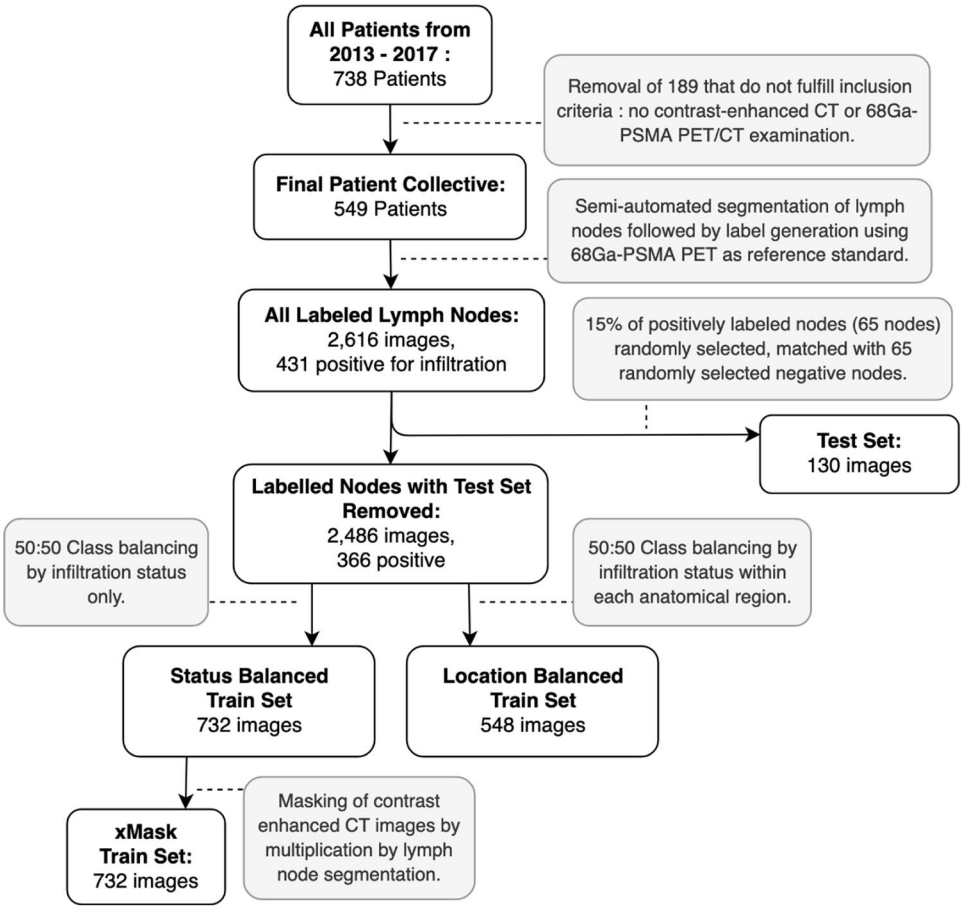


Figure 2. Dataset Generation Flowchart. Diagram describing generation of train and test datasets. Three train datasets shown, (status balanced, location balanced and xMask) were used to train three distinct neural networks. All neural networks and experts were tested and compared using a separate test set of 130 images, which was withheld from the neural networks during training. 50:50 class balancing was performed by taking all available infiltrated lymph nodes and randomly selecting an equally sized set of non-infiltrated lymph nodes, either from all available non-infiltrated nodes or from nodes within the same location category.

infiltration by tumor from the following four categories: very unlikely, unlikely, likely, and very likely. Neither the segmentation, 68Ga-PSMA PET/CT images, nor label were provided.

Statistical analysis. Model performance was evaluated for each CNN on the independent test set ($n = 130$) using the area under curve (AUC) of the receiver operating characteristic (ROC) curve. AUCs and confidence intervals were calculated using the pROC package in R[26], with confidence intervals computed using the bootstrap method with 10,000 stratified replicates. To allow for model comparison, the optimal threshold at which to consider CNN output as positive was set by maximizing Youden's index (sensitivity + specificity − 1), from which binary predictions were generated. Accuracy, sensitivity, specificity, PPV and NPV were calculated using the binary predictions. For study readers, the four categories were simplified to a dichotomous prediction of likely/unlikely. AUC for each radiologist is equivalent to the average of specificity and sensitivity[27]. McNemar's test was applied to all pairs of CNNs and experts. Results were considered statistically significant at a reduced $P < 0.005$ level to correct for multiple comparison. All variables are given as mean along with standard deviation and range where applicable.

Results

Evaluation of CNN classifiers and experts. The best performing Neural Network was trained using the status balanced training set, with an AUC of 0.955 (95% CI from 0.923–0.987). The CNNs trained with datasets where implicit frequency data was stripped using 50:50 class balancing by location category (the location balanced training set) or masking by the segmentation masks (xMask) performed comparably well, with an AUC of 0.858 (95% CI from 0.793–0.922) and 0.863 (95% CI from 0.804–0.923), respectively. Setting the sensitivity at 90% for all CNN models, the specificities of status balanced, location balanced, and xMask models was 88%, 52%, and 55%, respectively. Figure 5a shows ROC curves of all CNNs. Figure 5b shows histograms of CNN classification performance. Table 1 presents classification performance.

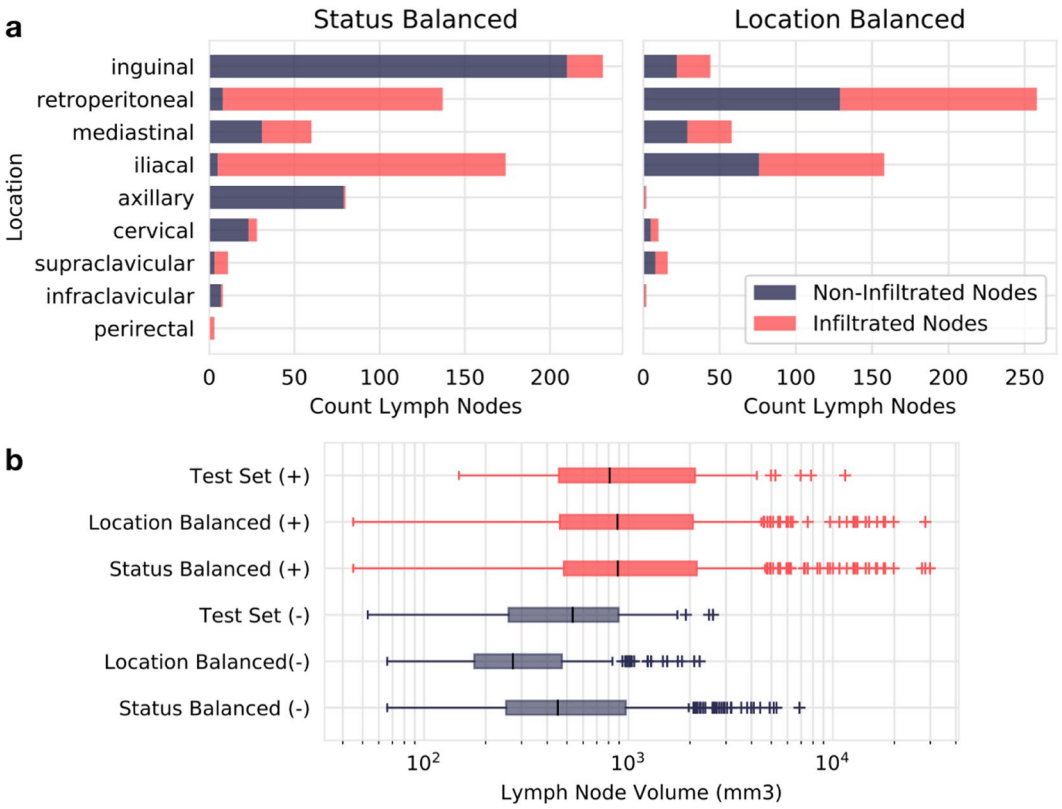


Figure 3. Dataset Regional and Volume Distributions. (**a**) The final distribution of lymph node images by location and infiltration status for the two training sets, referred to as 'status balanced' with 732 images and 'location balanced' with 548 images. (**b**) Boxplots depicting volume distribution for the location and status balanced training sets and test set grouped by infiltration status. Due to considerable overlap of the two distributions, size or volume is not a powerful indicator of infiltration.

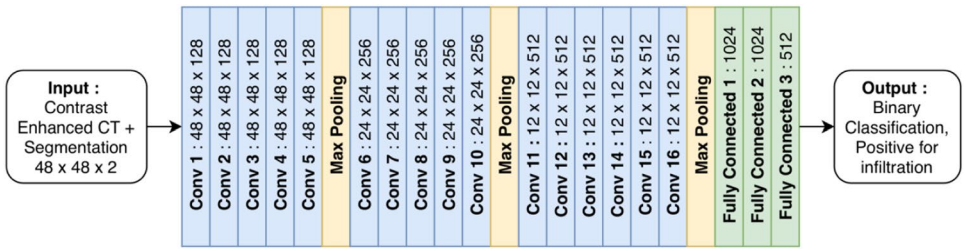


Figure 4. Convolutional neural network architecture. All three CNNs developed shared a common architecture and differed by the data used for training. CNNs received 2D contrast-enhanced CT images and segmentation masks as input, with input images augmented randomly during training. All convolutional layers used a kernel size of 3×3. A rectified linear unit (ReLU) activation function followed by batch normalization was performed at every layer. Adam optimization was used to update network weights, with parameters for alpha, beta1, beta2 and epsilon set at 0.0001, 0.9, 0.999 and 1e-08. Training was continued for 50 epochs.

The experienced uroradiologists achieved an average AUC, sensitivity, specificity and accuracy of 0.81, 65%, 96% and 81% respectively. The first radiologist performed with a calculated AUC of 0.86, while the second radiologist achieved a calculated AUC of 0.75. All differences in error rate between CNNs and expert readers was not statistically significant using McNemar's test and p set at a reduced 0.005.

The random forest trained with the status balanced training set achieved an AUC, sensitivity, specificity and accuracy of 0.900, 84%, 95% and 90% respectively on the test set. The random forest trained with the location balanced set performed significantly worse with an AUC, sensitivity, specificity and accuracy of 0.654, 70%, 60% and 65% respectively on the test set.

Use of heatmaps to explain differences in performance. Using heatmaps, we sought to elucidate how deep learning achieves a high classification performance. Examples of heatmaps are shown in Figs. 6 and 7. It appears that the CNNs are able to learn features within the lymph node and more surprisingly, outside the

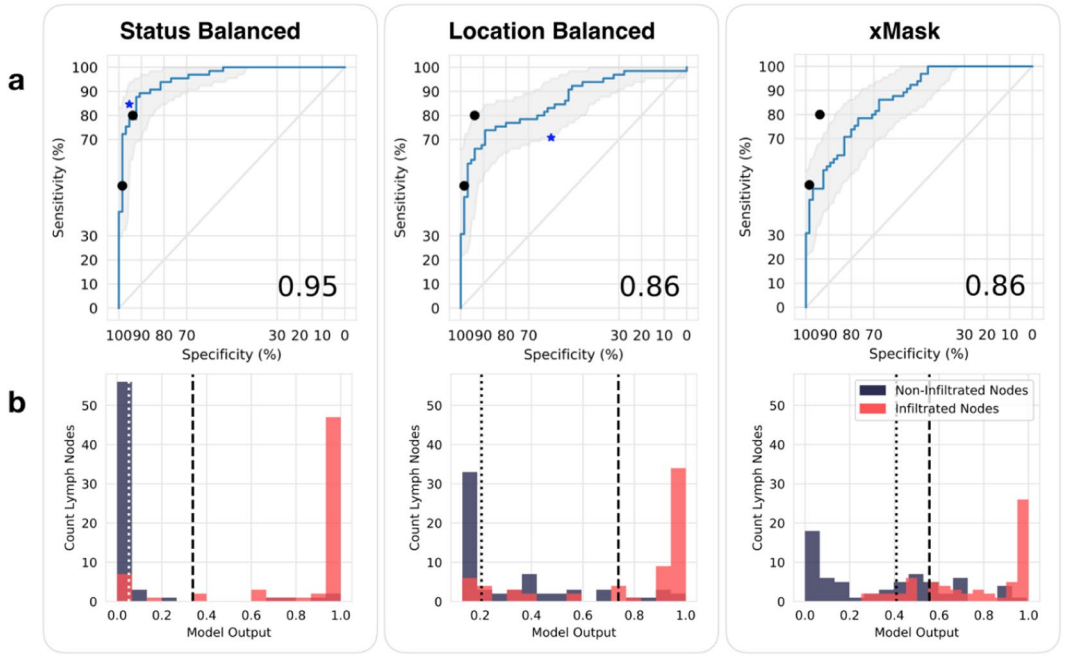


Figure 5. Classification performance. (**a**) Shown are the ROC curves for the three trained CNNs on the separate test set (n = 130) with 95% confidence interval of the sensitivity at given specificities in shaded gray. Displayed in the lower right hand corner is the corresponding AUC. Classification by individual radiologists on the same test set are displayed as black dots. Blue stars show random forest performance on the separate test set using the corresponding training dataset (status or location balanced). (**b**) Histograms of CNN model classification performance on the test set. The threshold that maximizes Youden's index is shown as a dashed line. The threshold which corresponds to a 90% sensitivity is shown as a dotted line. Infiltrated nodes (red bars) to the right of the given threshold are 'true positive', while those to the left are 'false negative': non-infiltrated nodes (blue) to the left are true negative, to the right are false positive.

Classifier	AUC	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1 Score
CNN: Status Balanced	0.95	89	86	92	91	86	88
CNN: Location Balanced	0.86	80	72	89	87	76	78
CNN: xMask	0.86	76	76	76	76	76	76
RF: Status Balanced	0.90	90	84	95	94	86	89
RF: Location Balanced	0.65	65	70	60	63	67	67
Expert 1	0.86	86	80	93	92	82	85
Expert 2	0.75	74	50	98	97	66	66

Table 1. Classification performance. Classification results are displayed in percentages. The optimal threshold for the three CNNs was selected by maximizing Youden's Index. RF: Random Forest.

boundaries of the lymph node (such as the aorta or air/skin borders), that correlate with lymph node infiltration status. It is critical to note that our best performing model, trained on status balanced data, appears to rely on features outside of the lymph node in question. This can be most clearly seen on images of inguinal or mediastinal lymph nodes, where areas of skin/air border (often found in the inguinal region) or lung/mediastinum border contribute heavily to final classification output, and the lymph node centered in the image is not highlighted. Heatmaps from the same CNN show that the lymph node itself is more important in true positive considerations, suggesting that 'inguinality', i.e. features of the inguinal region are important considerations in a negative infiltration status. Heatmaps generated can also be diffuse, with CNN attention displayed in many regions of the image but not particularly focused on the lymph node or surrounding region.

Discussion

In this study we trained and tested three CNNs that predict metastatic infiltration of lymph nodes by PCa using contrast-enhanced CT images and assessed their performance versus that of experienced human readers. The CNNs performed at the same level of two expert radiologists.

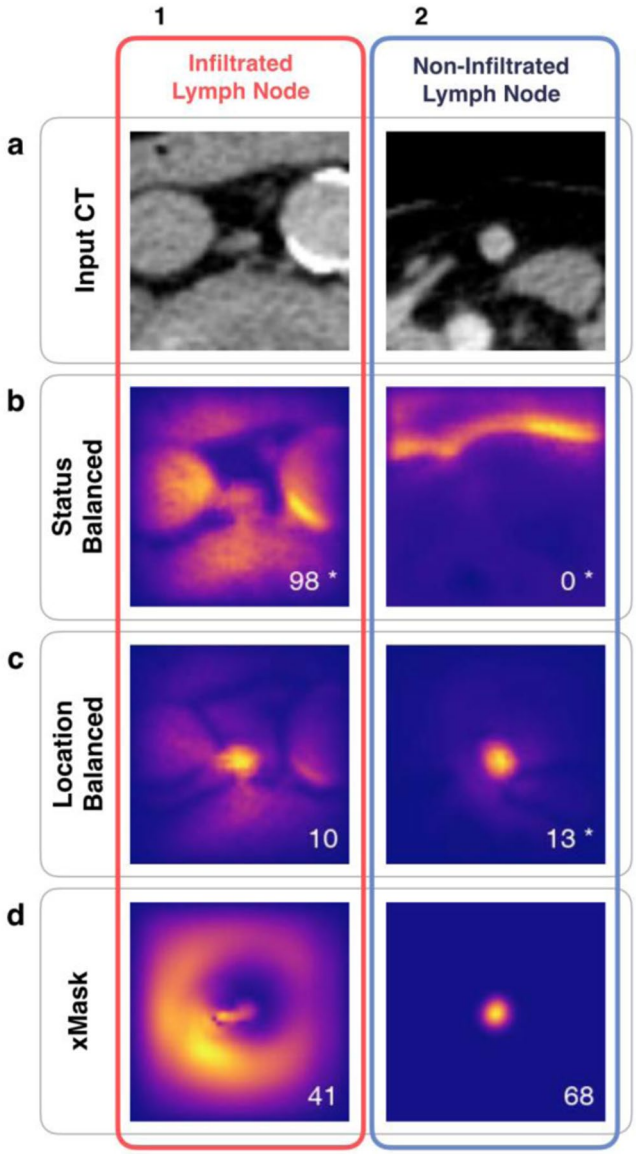


Figure 6. Heatmaps display neural network attention. (**a**) Contrast-enhanced CT images for two lymph nodes that were used as input to generate all heatmaps displayed, with (1) a retroperitoneal lymph node positive for infiltration by PCa, and (2) an inguinal lymph node negative for infiltration. In (**b–d**) heatmaps for the lymph nodes shown in (**a**), produced by three CNNs trained with status balanced training data, location balanced training data, or masked input data, respectively. CNN output, a pseudo probability score that the lymph node was classified as positive for tumor infiltration, is shown in the bottom right of each heatmap in (**b–d**). Stars signify true output predictions (either true positives for the lymph node in column 1 or true negative for column (2), with thresholds set by optimizing Youden's index for each CNN, set at 34, 73 and 54 for b,c, and d respectively. Within heatmaps, light colors represent areas that contribute to output prediction, while dark regions contribute little to output prediction. CNNs often highlight regions within the lymph node that expert radiologists recognize as important for infiltration status, such as nodal center density and contrast enhancement. In true positive images it appears that high central density is the most relevant parameter in designating a 'positive' label. We postulate that the 'halo' surrounding the lymph node in the xMask CNN (d1), depicts the CNN attention to size. Heatmaps produced by the CNN trained with status balanced data highlight anatomical regions which aid in classification of lymph nodes, often demarcating the air-skin border seen in images of inguinal lymph nodes, as in b2.

Current attempts at detecting lymph node metastases in PCa by radiological reading have been shown to be suboptimal, with a sensitivity and specificity shown in one study to be 42% and 82%, respectively[6–8]. Size is often the most relevant diagnostic criteria, with nodes greater than 10 mm deemed as suspicious and all below as benign[9,10]. Other criteria are difficult to quantify and are highly dependent on reader experience. Thus, the use of quantitative or algorithmic methods to detect LNI is desired. By radiomic analysis, in which a host of quantitative features are extracted from images and analyzed for statistical correlations, it has been suggested

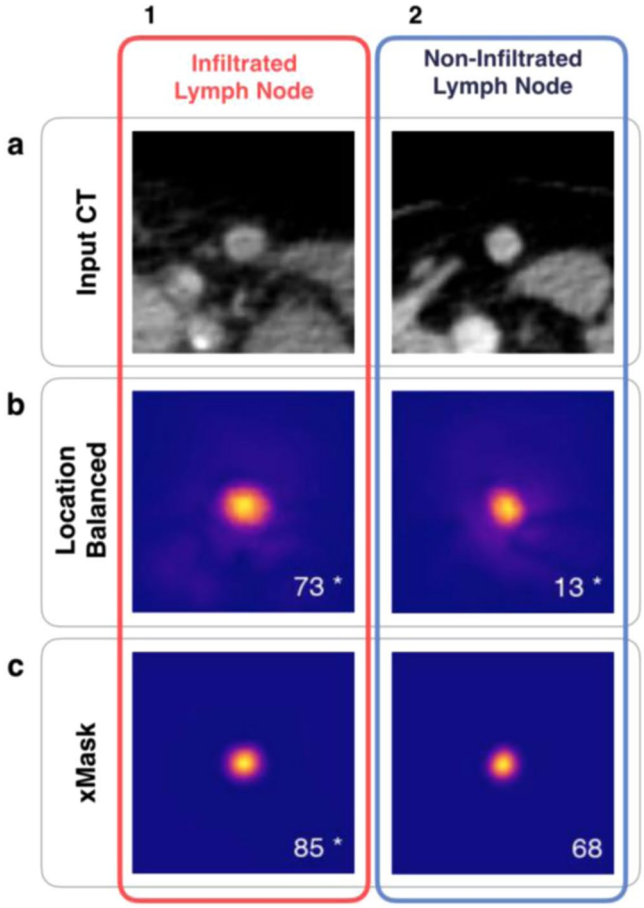


Figure 7. Limitations of heatmaps as tool to explain black box predictions. (**a**) Contrast-enhanced CT images for two inguinal lymph nodes that were used as input to generate all heatmaps displayed, with (1) a lymph node positive for infiltration by PCa, and (2) a lymph node negative for infiltration. In (**b,c**) heatmaps produced by two CNNs trained with location balanced training data, or masked input data, respectively. Beyond verifying that the lymph node is important for classification, heatmaps provide little additional information as to why classification output was either true positive (b1,c1), true negative (b2), or false positive (c2).

that a 7.5–Hounsfield CT density threshold could act as a surrogate parameter to differentiate LNI from benign processes[28], with 89% of non-infiltrated LNs below this threshold and 92% infiltrated LN above, though this study used many different cancer types and a mix of PET tracers as a standard of reference. Deep learning, in which optimization algorithms are used to train neural network models in classification tasks, have shown mixed success in detecting LNI. It has been previously found that CNNs are able to predict SUVmax in a PET scan using CT images of lymph nodes with a moderate accuracy, with an AUC of 0.85[29]. A number of studies predicting mediastinal LNI by lung cancer and breast cancer have been performed, with one study finding an AUC of 0.76[30] and another study classifying LNI in axillary lymph nodes by breast cancer achieving an AUC of 0.84[31]. CNNs were also found to classify head and neck tumor extranodal extension with an AUC of 0.91 using 3D CT images[32]. It has also been shown feasible to identify tumor infiltrated lymph nodes in MRI using deep learning[33]. To our knowledge, no study using deep learning to identify metastases of PCa into the lymphatic system by CT has been performed so far.

Generation of heatmaps is an attempt to explain how deep learning models reach classification decisions on a per-image basis, and represents a growing field of research known as 'explainability'. Each heatmap can be interpreted as displaying CNN attention; regions of an input image that influenced the classification decision are demarcated. From the heatmaps produced in our study, it becomes clear that identical CNN architectures learn different methods to solve the same problem, depending on which data is used for training. It appears that the CNN trained with status balanced data learned not only to recognize features of the lymph node in question, but also to recognize anatomical features surrounding the lymph node. Using these anatomical features, it appears that the status balanced CNN implicitly learned frequency of infiltration in different anatomical regions and used these frequencies or probabilities to improve output prediction. For example, in the status balanced dataset, 91% of inguinal lymph nodes were negative (see Fig. 3a). Thus, labeling all inguinal lymph nodes as negative is highly rewarded during the training process, and recognizing 'inguinality' aided in achieving high classification accuracy. Indeed, the air/skin border found in inguinal lymph nodes was often well demarcated in heatmaps, as seen in Fig. 6. However, it is unclear to what extent such anatomical features influenced classification; the CNN trained

with status balanced data did classify some inguinal lymph nodes as positive, and some retroperitoneal lymph nodes (of which 94% were positive in the status balanced dataset) as negative, as shown in Fig. 5b. The fact that learning anatomical features within the image (as proxy for anatomical location) greatly improves classification performance in the status balanced dataset is underscored by the high performance of the random forest trained on the this dataset; using nodal volume and location alone, high classification performance was achieved (AUC 0.90). Thus, our best performing neural network is most likely essentially useless on external datasets not sharing the anatomical bias found in the status balanced dataset.

We created two additional CNNs to eliminate anatomical clues within images in an attempt to force neural network attention to the lymph node. First, we created a new training dataset created by balancing positive and negative lymph nodes within each location category. By doing so we eliminated the possibility of learning infiltration frequency at each anatomical location. While it is clear from generated heatmaps that the CNN trained with this location balanced set did focus more on the lymph node and not on anatomical features, it was not able to achieve the same classification performance as the status trained CNN. However, a random forest receiving nodal volume and location information trained on this location balanced dataset performed poorly, considerably worse than the CNN (AUC 0.677 vs 0.858). This leads us to believe that the neural network is indeed focusing on features within the lymph node to perform classification. Secondly, a new CNN was provided images created by multiplying the CT image by the manually generated segmentation (xMask), thus setting all values outside of the lymph node to zero. This removed all contextual information, such as location in the body or presence of neighboring structures. The resulting performance was similar to the location balanced CNN. Interestingly, heatmaps created by the xMask CNN often showed a diffuse halo like pattern of attention outside of the lymph borders, which we postulate may be the CNNs attention to size. We cannot definitively state that any of the CNNs developed are able to determine nodal size due to intrinsic limitations of heatmaps as an explainability tool and the black box nature of neural networks, which often created very similar looking heatmaps (see Fig. 7). Regardless, size alone is a poor predictor of infiltration, as can be intuited by the considerable overlap of volume distributions for lymph nodes positive and negative for infiltration (see Fig. 3b) and shown quantitatively by the poor performance of the random forests trained with location balanced data.

There are a number of limitations to our study. It is important to note that the usage of PSMA PET/CT is an imperfect method of label generation. In comparison to the gold standard for detecting LN metastases, namely histopathological analysis after extended pelvic lymph node dissection (PLND)[34], PSMA PET/CT was found to have a sensitivity of 80% and specificity of 97% in a systematic review and meta-analysis[15,16,35,36]. Due to the high specificity, it is unlikely that our models were trained with large numbers of false positive lymph nodes. In addition, we relied on manual detection of segmentation of lymph nodes, and we do not perform lymph node detection. The tendency to select easily definable and large lymph nodes for analysis led to a large amount of inguinal lymph nodes being included in our dataset, a limitation we sought to overcome by various means of class balancing.

The obvious attention to anatomical features demonstrated by our best performing CNN raises a number of issues in the implementation of deep learning in the medical field. Deep learning models are able to learn frequencies and summary statistics, known as biases, within datasets, which can lead to high classification performance based upon undesirable features. This problem is distinct from overfitting to the training dataset, and instead points to the need for a more rigorous explainability of deep learning models. Our results represent a moderate success in the use of saliency maps (heatmaps), as through this instance-based analysis of CNN attention, we were able to determine that our best performing model was using anatomical features of the lymph node environment in addition to features within the lymph node. We were able to compensate for anatomical variations in infiltration frequency because we had collected coarse data on anatomical location. However, not only does class balancing at ever higher levels of abstraction encroach on the notion of 'automated feature generation', it is not feasible in the medical field due to lack of knowledge of what constitutes a relevant category. The lack of explainability methods for deep learning models is also a limitation. Our use of heatmaps, known as an attribution method, of which there are several, is problematic not just because of inconsistencies in implementation and performance[24], but the underpinning assumption that individual pixels in an input image should be the primary unit of relevance for classification.

Current deep learning systems can perform remarkably well and will most likely continue to improve with larger datasets and access to more contextual information, such as blood serum values and genomic data. Our results show that CNNs are capable of classifying lymphatic infiltration by PCa on contrast-enhanced CT scans alone as compared to the 68Ga-PSMA PET/CT reference standard. Anatomical context influences the performance of CNNs and should be carefully considered when building such imaging based biomarkers.

Received: 29 July 2019; Accepted: 11 February 2020;
Published online: 25 February 2020

References

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2018. *CA Cancer J. Clin.* **68**, 7–30, https://doi.org/10.3322/caac.21442 (2018).
2. Oderda, M., Joniau, S., Spahn, M. & Gontero, P. Debulking surgery in the setting of very high-risk prostate cancer scenarios. *BJU Int.* **110**, E192–E198, https://doi.org/10.1111/j.1464-410X.2012.10942.x (2012).
3. Luchini, C. *et al.* Extranodal extension of lymph node metastasis influences recurrence in prostate cancer: a systematic review and meta-analysis. *Sci. Rep.* **7**, 2374, https://doi.org/10.1038/s41598-017-02577-4 (2017).
4. Carroll, P. R. *et al.* NCCN Guidelines Insights: Prostate Cancer Early Detection, Version 2.2016. *J. Natl Compr. Canc Netw.* **14**, 509–519 (2016).
5. Mottet, N. *et al.* EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. *Eur. Urol.* **71**, 618–629, https://doi.org/10.1016/j.eururo.2016.08.003 (2017).

6. Engeler, C. E., Wasserman, N. F. & Zhang, G. Preoperative assessment of prostatic carcinoma by computerized tomography: Weaknessnes and new perspectives. *Urol.* **40**, 346–350, https://doi.org/10.1016/0090-4295(92)90386-B (1992).
7. Flanigan, R. C. *et al.* Limited efficacy of preoperative computed tomographic scanning for the evaluation of lymph node metastasis in patients before radical prostatectomy. *Urol.* **48**, 428–432, https://doi.org/10.1016/S0090-4295(96)00161-6 (1996).
8. Hövels, A. M. *et al.* The diagnostic accuracy of CT and MRI in the staging of pelvic lymph nodes in patients with prostate cancer: a meta-analysis. *Clin. Radiol.* **63**, 387–395, https://doi.org/10.1016/j.crad.2007.05.022 (2008).
9. Maurer, T. *et al.* Diagnostic Efficacy of (68)Gallium-PSMA Positron Emission Tomography Compared to Conventional Imaging for Lymph Node Staging of 130 Consecutive Patients with Intermediate to High Risk Prostate Cancer. *J. Urol.* **195**, 1436–1443, https://doi.org/10.1016/j.juro.2015.12.025 (2016).
10. Heesakkers, R. A. M. *et al.* MRI with a lymph-node-specific contrast agent as an alternative to CT scan and lymph-node dissection in patients with prostate cancer: a prospective multicohort study. *Lancet Oncol.* **9**, 850–856, https://doi.org/10.1016/S1470-2045(08)70203-1 (2008).
11. Gillessen, S. *et al.* Management of Patients with Advanced Prostate Cancer: The Report of the Advanced Prostate Cancer Consensus Conference APCCC 2017. *Eur. Urol.* **73**, 178–211, https://doi.org/10.1016/j.eururo.2017.06.002 (2018).
12. Silver, D. A., Pellicer, I., Fair, W. R., Heston, W. D. & Cordon-Cardo, C. Prostate-specific membrane antigen expression in normal and malignant human tissues. *Clin. Cancer Res.* **3**, 81–85 (1997).
13. Bostwick, D. G., Pacelli, A., Blute, M., Roche, P. & Murphy, G. P. Prostate specific membrane antigen expression in prostatic intraepithelial neoplasia and adenocarcinoma. *Cancer* **82**, 2256–2261, 10.1002/(SICI)1097-0142(19980601)82:11<2256::AID-CNCR22>3.0.CO;2-S (1998).
14. Perner, S. *et al.* Prostate-specific membrane antigen expression as a predictor of prostate cancer progression. *Hum. Pathol.* **38**, 696–701, https://doi.org/10.1016/j.humpath.2006.11.012 (2007).
15. Perera, M. *et al.* Gallium-68 Prostate-specific Membrane Antigen Positron Emission Tomography in Advanced Prostate Cancer—Updated Diagnostic Utility, Sensitivity, Specificity, and Distribution of Prostate-specific Membrane Antigen-avid Lesions: A Systematic Review and Meta-analysis. *Eur. Urol.* https://doi.org/10.1016/j.eururo.2019.01.049 (2019).
16. Leeuwen, P. J. V. *et al.* Prospective evaluation of 68Gallium-prostate-specific membrane antigen positron emission tomography/computed tomography for preoperative lymph node staging in prostate cancer. *BJU Int.* **119**, 209–215, https://doi.org/10.1111/bju.13540 (2017).
17. Hofman, M. S., Hicks, R. J., Maurer, T. & Eiber, M. Prostate-specific Membrane Antigen PET: Clinical Utility in Prostate Cancer, Normal Patterns, Pearls, and Pitfalls. *Radiographics* **38**, 200–217, https://doi.org/10.1148/rg.2018170108 (2018).
18. Afshar-Oromieh, A. *et al.* PET imaging with a [68Ga]gallium-labelled PSMA ligand for the diagnosis of prostate cancer: biodistribution in humans and first evaluation of tumour lesions. *Eur. J. Nucl. Med. Mol. Imaging* **40**, 486–495, https://doi.org/10.1007/s00259-012-2298-2 (2013).
19. Surti, S. *et al.* Performance of Philips Gemini TF PET/CT scanner with special consideration for its time-of-flight imaging capabilities. *J. Nucl. Med.* **48**, 471–480 (2007).
20. Wolf, I. *et al.* The medical imaging interaction toolkit. *Med. Image Anal.* **9**, 594–604, https://doi.org/10.1016/j.media.2005.04.005 (2005).
21. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]* (2014).
22. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]* (2014).
23. Alber, M. *et al.* iNNvestigate neural networks! *arXiv:1808.04260 [cs, stat]* (2018).
24. Kindermans, P.-J. *et al.* Learning how to explain neural networks: PatternNet and PatternAttribution. *arXiv:1705.05598 [cs, stat]* (2017).
25. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
26. pROC: Display and Analyze ROC Curves v. 1.13.0 (2018).
27. Haenssle, H. A. *et al.* Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* **29**, 1836–1842, https://doi.org/10.1093/annonc/mdy166 (2018).
28. Giesel, F. L. *et al.* Correlation Between SUVmax and CT Radiomic Analysis Using Lymph Node Density in PET/CT-Based Lymph Node Staging. *J. Nucl. Med.* **58**, 282–287, https://doi.org/10.2967/jnumed.116.179648 (2017).
29. Shaish, H. *et al.* Prediction of Lymph Node Maximum Standardized Uptake Value in Patients With Cancer Using a 3D Convolutional Neural Network: A Proof-of-Concept Study. *American Journal of Roentgenology*, 1–7, https://doi.org/10.2214/AJR.18.20094 (2018).
30. Beig, N. *et al.* Perinodular and Intranodular Radiomic Features on Lung CT Images Distinguish Adenocarcinomas from Granulomas. *Radiology*, 180910, https://doi.org/10.1148/radiol.2018180910 (2018).
31. Ha, R. *et al.* Axillary Lymph Node Evaluation Utilizing Convolutional Neural Networks Using MRI Dataset. *J Digit Imaging*, https://doi.org/10.1007/s10278-018-0086-7 (2018).
32. Kann, B. H. *et al.* Pretreatment Identification of Head and Neck Cancer Nodal Metastasis and Extranodal Extension Using Deep Learning Neural Networks. *Sci. Rep.* **8**, 14036, https://doi.org/10.1038/s41598-018-32441-y (2018).
33. Lu, Y. *et al.* Identification of Metastatic Lymph Nodes in MR Imaging with Faster Region-Based Convolutional Neural Networks. *Cancer Res.* **78**, 5135–5143, https://doi.org/10.1158/0008-5472.CAN-18-0494 (2018).
34. EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part II: Treatment of Relapsing, Metastatic, and Castration-Resistant Prostate Cancer. *Eur. Urol.* **71**, 630–642, https://doi.org/10.1016/j.eururo.2016.08.002 (2017).
35. Hijazi, S. *et al.* Pelvic lymph node dissection for nodal oligometastatic prostate cancer detected by 68Ga-PSMA-positron emission tomography/computerized tomography. *Prostate* **75**, 1934–1940, https://doi.org/10.1002/pros.23091 (2015).
36. Jilg, C. A. *et al.* Diagnostic Accuracy of Ga-68-HBED-CC-PSMA-Ligand-PET/CT before Salvage Lymph Node Dissection for Recurrent Prostate Cancer. *Theranostics* **7**, 1770, https://doi.org/10.7150/thno.18421 (2017).

Acknowledgements

We acknowledge support from the German Research Foundation (DFG) and the Open Access Publication Funds of Charité – Universitätsmedizin Berlin.

Author contributions

All authors contributed to the study design. A.H., F.L., A.B., M.M.R., C.F., H.A., W.B. collected and curated the data. A.H. and T.P. wrote the manuscript. All authors critically revised and reviewed the manuscript.

Competing interests

A.H., F.L., C.F., M.M., W.B. and M.M.R. declare no competing interest with respect to the relation to the work described. A.B. received fee as a speaker for Bayer, and Bender Gruppe, both outside of the current work. H.A. reports grants from Sirtex Medical, Bayer and GE Healthcare; lecture and/or travel fees from Sirtex Medical, GE Healthcare, Novartis, Eisai and Terumo. All outside the submitted work. B.H. declares the following competing interests: Grant money from the following companies or nonprofit organizations to the Dept of Radiology: Abbott, AbbVie, Ablative Solutions, Accovion, Achaogen Inc., Actelion Pharmaceuticals, ADIR,

Aesculap, AGO, AIF Arbeitsgemeinschaft industrieller Forschungsvereinigungen, AIO: Arbeitsgemeinschaft Internistische Onkologie, Alexion Pharmaceuticals, Amgen, AO Foundation, Arena Pharmaceuticals, ARMO Biosciences, Inc., art photonics GmbG Berlin, ASR Advanced sleep research, Astellas, AstraZeneca, BARD, Bayer Healthcare, Bayer Schering Pharma, Bayer Vital, BBraun (Sponsoring a workshop), Berlin-Brandenburger Centrum für Regenerative Therapien (BCRT), Berliner Krebsgesellschaft, Biotronik, Bioven, BMBF, Boehring Ingelheimer, Boston Biomedical Inc., BRACCO Group, Brainsgate, Bristol-Myers Squibb, Cascadian Therapeutics, Inc., Celgene, CELLACT Pharma, Celldex, Therapeutics, CeloNova BioSciences, Charité research organisatin GmbH, Chiltern, CLOVIS ONCOLOGY, INC., Covance, CUBIST, CureVac AG, Tübingen, Curis, Daiichi, DC Devices, Inc. USA, Delcath Systems, Dermira Inc. Deutsche Krebshilfe, Deutsche Rheuma Liga, DFG, DSM Nutritional Products AG, Dt. Stiftung für Herzforschung, Dynavax, Eisai Ltd., European Knowledge Centre, Mosquito Way, Hatfield, Eli Lilly and Company Ltd. EORTC, Epizyme, INC., Essex Pharma, EU Programmes, Euroscreen S.A., Fibrex Medical Inc., Foxused Ultrasound Surgery Foundation, Fraunhofer Gesellschaft, Galena Biopharma, Galmed Research and Development Ltd., Ganymed, GE, Genentech Inc., GETNE (Grupo Espanol de Tumores Neuroendocrinos), Gilead Sciences, Inc, Glaxo Smith Kline, Glycotope GmbH, Berlin, Goethe Uni Frankfurt, Guerbet, Guidant Europe NV, Halozyme, Hewlett Packard GmbH, Holaira Inc. ICON (CRO), Idera Pharmaceuticals, Inc., Ignyta, Inc. Immunomedics Inc., Immunocore, Incyte, INC Research, Innate Pharma, InSightec Ltd., Inspiremd, inVentiv Health Clinical UK Ltd., Inventivhealth, IOMEDICO, IONIS, IPSEN Pharma, IQVIA, ISA Therapeutics, Isis Pharmaceuticals Inc., ITM Solucin GmbH, Jansen, Kantar Health Gmbh (CRO), Kartos Therapeutics, Inc., Karyopharm Therapeutics, Inc., Kendle/MorphoSys Ag, Kite Pharma, Kli Fo Berlin Mitte, La Roche, Land Berlin, Lilly GmbH, Lion Biotechnology, Lombard Medical, Loxo Oncology, Inc., LSK BioPartners; USA, Lundbeck GmbH, Lux Biosciences, LYSARC, MacroGenics, MagForce, MedImmune Inc., Medpace, Medpace Germany GmbH (CRO), MedPass (CRO), Medronic, Merck, Merromack Pharmaceuticals Inc., MeVis Medical Solutions AG, Millennium Pharmaceuticals Inc., Mologen, Monika Kutzner Stiftung, MSD Sharp, NeoVacs SA, Newlink Genentics Corporation, Nexus Oncology, NIH, Novartis, novocure, Nuvisan, Ockham oncology, OHIRC Kanada, Orion Corporation Orion Pharma, Parexel CRO Service, Perceptive, Pfizer GmbH, Pharma Mar, Pharmaceutical Research Associates GmbH (PRA), Pharmacyclics Inc., Philipps, PIQUR Therapeutics Ltd., Pluristem, PneumRX.Inc, Portola Pharmaceuticals, PPD (CRO), PRAint, Premier-research, Provectus Biopharmaceuticals, Inc., psi-cro, Pulmonx International Sàrl, Quintiles GmbH, Regeneron Pharmaceuticals, Inc., Respicardia, Roche, Samsung, Sanofi, sanofis-aventis S.A., Schumacher GmbH (Sponsoring a workshop), Seattle Genetics, Servier (CRO), SGS Life Science Services (CRO), Shore Human Genetic Therapies, Siemens, Silena Therapeutics, Spectranetics GmbH, Spectrum Pharmaceuticals, St. Jude Medical, Stiftung Wolfgang Schulze, Symphogen, Taiho Oncology, Inc., Taiho Pharmaceutical Co., TauRx Therapeutics Ltd., Terumo Medical Corporation, Tesaro, tetec-ag, TEVA, Theorem, Theradex, Threshold Pharmaceuticals Inc., TNS Healthcare GmbH, Toshiba, UCB Pharma, Uni München, VDI/VDE, Vertex Pharmaceuticals Incorporated, winicker-norimed, Wyeth Pharma, Xcovery Holding Company, Zukunftsfond Berlin (TSB). TP receives grant support from the Berlin Institute of Health within the Clinician Scientist Programme. TP declares no additional conflict of interest with respect to the relation to the work described. Outside of the current work there are institutional relationship with the following entities (no personal payments to TP): research support from Siemens Healthcare and Philips Healthcare, clinical trials with AGO, Aprea AB, Astellas Pharma Global Inc., AstraZeneca, Celgene, Genmab A/S, Incyte Coporation, Lion Biotechnologies, Inc., Millennium Pharmaceuticals, Inc., Morphotec Inc., MSD, Tesaro Inc., and Roche.

Additional information

Correspondence and requests for materials should be addressed to T.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

Author Publication List

Hartenstein, A., F. Lübbe, A. D. J. Baur, M. M. Rudolph, C. Furth, W. Brenner, H. Amthauer, B. Hamm, M. Makowski, and T. Penzkofer. "Prostate Cancer Nodal Staging: Using Deep Learning to Predict 68ga-Psma-Positivity from Ct Imaging Alone." Scientific Reports 10, no. 1 (2020/02/25/ 2020): 3398. https://dx.doi.org/10.1038/s41598-020-60311-z.

Impact factor Scientific Reports : **4.525**

Rudolph, M. M., Baur, A. D. J., Cash, H., Haas, M., Mahjoub, S., Hartenstein, A., Hamm, C. A., Beetz, N. L., Konietschke, F., Hamm, B., Asbach, P. & Penzkofer, T. Diagnostic performance of PI-RADS version 2.1 compared to version 2.0 for detection of peripheral and transition zone prostate cancer. Scientific Reports 10, 15982 (2020).

Impact factor Scientific Reports : 4.525

Mahjoub, S., Baur, A. D. J., Lenk, J., Lee, C. H., Hartenstein, A., Rudolph, M. M., Cash, H., Hamm, B., Asbach, P., Haas, M. & Penzkofer, T. Optimizing size thresholds for detection of clinically significant prostate cancer on MRI: Peripheral zone cancers are smaller and more predictable than transition zone tumors. European Journal of Radiology 129, 109071 (2020).

Impact factor European Radiology : 4.014

Rudolph, M. M., Baur, A. D. J., Haas, M., Cash, H., Miller, K., Mahjoub, S., Hartenstein, A., Kaufmann, D., Rotzinger, R., Lee, C. H., Asbach, P., Hamm, B. & Penzkofer, T. Validation of the PI-RADS language: predictive values of PI-RADS lexicon descriptors for detection of prostate cancer. Eur Radiol (2020) doi:10.1007/s00330-020-06773-1.

Impact factor European Radiology : 4.014

Kunde, Felix, Alexander Hartenstein, and Petra Sauer. Spatio-Temporal Traffic Flow Forecasting on a City-Wide Sensor Network. Edited by Igor Ivan, Jiří Horák, and Tomáš Inspektor: Springer International Publishing, 2018.

I would like to express my deepest appreciation to my group leader, PD Dr. med. Tobias Penzkofer, who provided invaluable guidance and mentorship during my doctoral work. I would like to extend my sincere thanks to PD Dr. med. Patrick Asbach for mentorship during this thesis, as well as sincerest gratitude Prof. Dr. med. Bernd Hamm for guidance. My completion of this dissertation would not have been possible without the hard work of Falk Lübbe, as well as that of my colleagues and collaborators.