Computerized Medical Imaging and Graphics 107 (2023) 102241

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Investigation and benchmarking of U-Nets on prostate segmentation tasks

Shrajan Bhandary [a,*], Dejan Kuhn [b,c,d], Zahra Babaiee [a], Tobias Fechter [b,c,d], Matthias Benndorf [e], Constantinos Zamboglou [c,d,f,g], Anca-Ligia Grosu [c,d,f], Radu Grosu [a,h]

[a] Cyber-Physical Systems Division, Institute of Computer Engineering, Faculty of Informatics, Technische Universität Wien, Vienna, 1040, Austria
[b] Division of Medical Physics, Department of Radiation Oncology, Medical Center University of Freiburg, Freiburg, 79106, Germany
[c] Faculty of Medicine, University of Freiburg, Freiburg, 79106, Germany
[d] German Cancer Consortium (DKTK), Partner Site Freiburg, Freiburg, 79106, Germany
[e] Department of Diagnostic and Interventional Radiology, Medical Center University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, 79106, Germany
[f] Department of Radiation Oncology, Medical Center University of Freiburg, Freiburg, 79106, Germany
[g] German Oncology Center, European University, Limassol, 4108, Cyprus
[h] Department of Computer Science, State University of New York at Stony Brook, NY, 11794, USA

ARTICLE INFO

Keywords:
Medical imaging
Automatic prostate segmentation
U-net variations
Comparison framework

ABSTRACT

In healthcare, a growing number of physicians and support staff are striving to facilitate personalized radiotherapy regimens for patients with prostate cancer. This is because individual patient biology is unique, and employing a single approach for all is inefficient. A crucial step for customizing radiotherapy planning and gaining fundamental information about the disease, is the identification and delineation of targeted structures. However, accurate biomedical image segmentation is time-consuming, requires considerable experience and is prone to observer variability. In the past decade, the use of deep learning models has significantly increased in the field of medical image segmentation. At present, a vast number of anatomical structures can be demarcated on a clinician's level with deep learning models. These models would not only unload work, but they can offer unbiased characterization of the disease. The main architectures used in segmentation are the U-Net and its variants, that exhibit outstanding performances. However, reproducing results or directly comparing methods is often limited by closed source of data and the large heterogeneity among medical images. With this in mind, our intention is to provide a reliable source for assessing deep learning models. As an example, we chose the challenging task of delineating the prostate gland in multi-modal images. First, this paper provides a comprehensive review of current state-of-the-art convolutional neural networks for 3D prostate segmentation. Second, utilizing public and in-house CT and MR datasets of varying properties, we created a framework for an objective comparison of automatic prostate segmentation algorithms. The framework was used for rigorous evaluations of the models, highlighting their strengths and weaknesses.

1. Introduction

Prostate cancer (PCa) is one of the most common non-cutaneous malignancies diagnosed in men in Europe, and there is an urgent need to reduce the rate of mortality (Marhold et al., 2022). Radiotherapy (RT) is one type of curative treatment option that has advanced tremendously, ranging from implementation of advanced diagnostics in staging and treatment planning, up to precise delivery of ablative RT doses in fewer fractions (Spohn et al., 2021; Zamboglou et al., 2021). In RT, medical images play a key role in the complete process, from diagnosis to treatment planning, including patient care after the required clinical procedures. The high volume of applied doses demands precise and accurate identification of tumours and surrounding tissues (Spohn et al., 2021). Therefore, prior to executing PCa clinical tasks, it is essential to locate and segment the prostate gland from medical images (Thompson et al., 2016). However, the correct identification and segmentation of the anatomical structures is a time-consuming approach, that demands proficiency in healthcare (Steenbergen et al., 2015). Furthermore, exactness in delineation of the region of interest (ROI) during manual segmentation is still hampered by inter-observer variability (Steenbergen et al., 2015).

The rise of deep learning (DL) in recent years revolutionized the field of medical image segmentation (Litjens et al., 2017b; Singh et al., 2020; Isensee et al., 2021; Santoro et al., 2022). DL algorithms, in particular convolutional neural networks (CNNs) have shown outstanding

* Corresponding author.
E-mail address: shrajan.bhandary@tuwien.ac.at (S. Bhandary).

https://doi.org/10.1016/j.compmedimag.2023.102241
Received 30 November 2022; Received in revised form 3 May 2023; Accepted 3 May 2023
Available online 12 May 2023

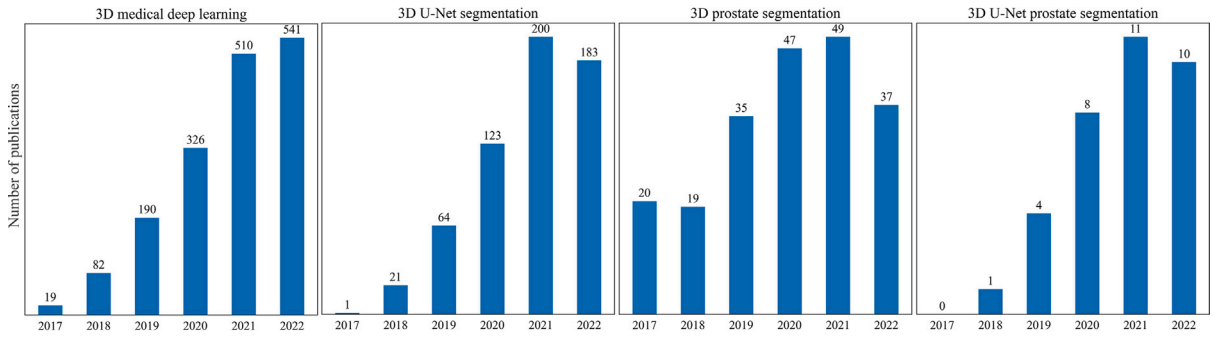

Fig. 1. Year-wise number of publications while searching for *3D medical deep learning*, *3D U-Net segmentation*, *3D prostate segmentation*, and *3D U-Net prostate segmentation* in the PubMed registry.

segmentation performances for almost every anatomical site (Litjens et al., 2017b; Isensee et al., 2021). In certain facets of computer vision, CNNs have even surpassed the classification accuracies of human observers (Hekler et al., 2019; Pham et al., 2021). Over the years, numerous DL works have performed 3D medical image segmentations, including the prostate (Gillespie et al., 2020), as evident in Fig. 1. A major factor that promoted the research on prostate segmentation is the organization of challenges, such as, the Prostate MR Image Segmentation 2012 challenge (PROMISE-12) (Litjens et al., 2014b), the NCI-ISBI 2013 Challenge (Bloch et al., 2015), the SPIE-AAPM-NCI Prostate MR Classification Challenge (PROSTATEx) (Litjens et al., 2014a, 2017a), the medical segmentation decathlon (MSD-prostate) (Simpson et al., 2019) and the Prostate Imaging: Cancer AI challenge (PI-CAI) (Saha et al., 2022). These platforms provided separate datasets, and a leaderboard for comparing performances of networks.

Among the top performing methods of the challenges, the majority use a 3D U-Net (Çiçek et al., 2016; Litjens et al., 2017b; Ghavami et al., 2019; Singh et al., 2020; Isensee et al., 2021) based architecture. However, the comparison and the subsequent selection of a single algorithm based on the leaderboards is limited due to various factors. First, the leaderboards show only the results for one dataset, and many teams participate in one challenge. A ranking across different datasets would reflect how an algorithm can deal with heterogeneous data, but this does not exit. Second, the challenges prior to the year 2018 did not require the publication of source code for participation. This makes it almost impossible to reproduce results from the leaderboards, as incomplete documentations are recorded in the corresponding manuscripts. Third, the challenges organized so far have mainly focused on magnetic resonance imaging (MRI) data, however, the treatment of prostate cancer might involve information attained from multi-modal images, such as, computed tomography (CT) or transrectal ultrasound (TRUS). Therefore, it is virtually infeasible to objectively measure which algorithm is ideal for the task of prostate segmentation. A common platform for the evaluation across multiple modalities would be of importance.

Ghavami et al. (2019) tackled the problem by evaluating six DL algorithms (out of which four were U-Net variants) on a common dataset. In their results, they show that the original U-Net is outperformed by its variants. However, the work by Ghavami et al. (2019) has its limits. The networks were trained on a dataset comprising T2 weighted MR images from 232 patients, and the evaluation was done by utilizing a hold-out test set. The MRI scans were obtained from three different trials that shared the same imaging protocols, and the prostate boundaries were annotated by the same clinical group. Additionally, since the publication of Ghavami et al. (2019), medical image processing has undergone an enormous change by the introduction of state-of-the-art (SOTA) frameworks such as nnU-Net (Isensee et al., 2021). Moreover, recent algorithms such as Attention U-Net (inspired by the success of attention gates in natural language processing) (Oktay et al., 2018), SegResnet (uses residual connections within each encoder–decoder block) (Myronenko, 2019), and U-Net++ (using deep supervision to extract important features using densely connected nested decoder subnetworks) (Zhou et al., 2019), have achieved tremendous performances in medical image segmentation tasks.

Santoro et al. (2022) conducted a survey of 921 research articles and analysed the recent applications of artificial intelligence (AI) in clinical RT. The authors considered the phase of the RT workflow based on the AI approaches to improve patient care coordination and optimization, image registration and segmentation, synthetic image generation, treatment planning, and outcome prediction. Most of the segmentation algorithms reviewed in Santoro et al. (2022) used the U-Net architecture (Ronneberger et al., 2015; Çiçek et al., 2016), and the networks were trained to delineate different anatomical areas, such as the brain, head, neck, thorax, and female or male pelvis (including the prostate glands and tumours). These methods were also applied in the clinical RT setting across different image modalities to pinpoint the exact locations of lesions and organs at risk throughout several parts of the human body (Choi et al., 2020; Nemoto et al., 2020; Jeong et al., 2021; Zhong et al., 2021). The U-Net architectures showcased promising results with improved accuracy, efficiency, and robustness. During our investigation, we found multiple DL-based CNN algorithms that were deployed either as commercial or open-source research tools for automatic ROI segmentation tasks (Estienne et al., 2020; Wong et al., 2020; Consortium, 2020).

Despite the noteworthy contributions, several professionals have raised concerns about the involvement of AI in healthcare, including the need for harmonization while overcoming barriers (Santoro et al., 2022). There are several major difficulties that block the AI models like U-Nets from real clinical use. One of the major difficulties is the need for more standardization in imaging and contouring protocols across different clinics and institutions (Haga et al., 2019). This can lead to variations in the data quality used to train and validate the AI models, making it difficult to generalize the results to other clinical settings (Syed et al., 2020). Another obstacle is the need for high-quality data to train and validate the models, which can be time-consuming and expensive (Litjens et al., 2017b; Singh et al., 2020). Furthermore, the black-box nature of these models makes it challenging for clinicians to trust their predictions and understand their reasoning (Vayena et al., 2018). The uncertainty in quantification makes it challenging to know when the model may fail or produce unreliable results (McBee et al., 2018; Korreman et al., 2021), which raises one of the main ethical concerns: *"if AI fails to deliver the correct output, who will take responsibility for the mistake?"* It is paramount that the biomedical imaging community addresses these shortcomings before AI models can be adopted in clinical practice. Therefore, there is a need for rigorous testing and validation of the AI models, to ensure their safety and effectiveness before they can be implemented in the clinical workflow (Goldenberg et al., 2019; Isensee et al., 2021; Punn and Agarwal, 2022; Saha et al., 2022; Santoro et al., 2022).

A successful transfer of AI models to clinical routine could enable new treatment techniques that have not been feasible until now due

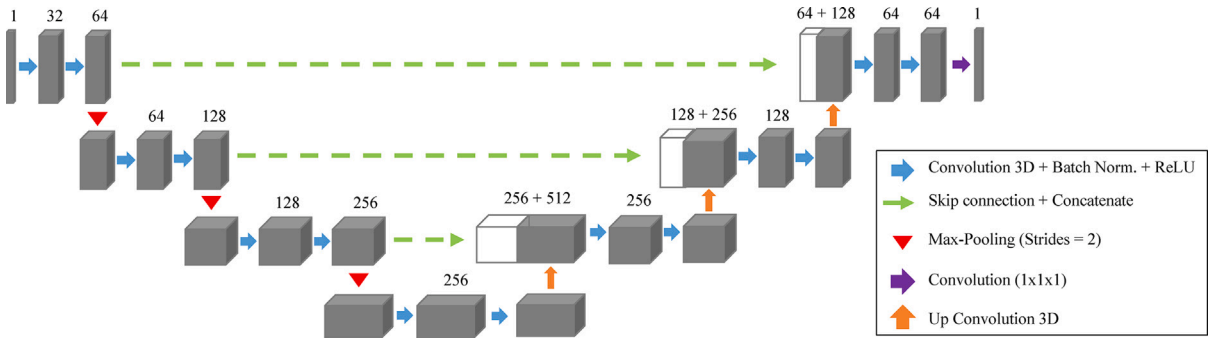

Fig. 2. 3D architecture of the U-Net re-created from Çiçek et al. (2016).

to complex processing tasks. One example might be online adaptive radiotherapy (ART) using hybrid linear accelerators (linac) (Hall et al., 2021). ART is a novel technique where the CT or MR volumes are obtained as part of the treatment delivery process. Simultaneous beam treatment with imaging enables the adaptation of the irradiated volume to account for changes in the physical properties of the organ and tumour (Hall et al., 2021). Kawula et al. (2023) trained a 3D U-Net to segment bladder, rectum, and clinical target volume for prostate cancer patients treated with 0.35 T MR-linacs. The experiments were conducted using a total of 332 in-house images, and demonstrated that CNNs can enhance automatic segmentation for MR-guided RT (Kawula et al., 2023). A possible scenario in the future could be the usage of DL techniques to tailor ART parameters in real time as required, while simultaneously minimizing the risk of side effects (Hall et al., 2021; Kawula et al., 2023).

In this work, we evaluated and benchmarked five CNNs: U-Net (Çiçek et al., 2016), V-Net (Milletari et al., 2016), Attention U-Net (Oktay et al., 2018), SegResNet (Myronenko, 2019) and U-Net++ (Zhou et al., 2019), on the task of segmenting the prostate from 3D medical images across different settings. To elaborate, initially, we scrutinized the recent developments in DL-based biomedical image segmentation techniques, and based on this review, we chose the best-suited algorithms for further evaluation. In the next part, we created a framework for investigating the performance of the algorithms, contingent on network architecture and the influence of factors such as dataset size, imaging protocols and image modality. The studies were carried out on two private and two public datasets (the medical segmentation decathlon-prostate set (Simpson et al., 2019) and PROMISE-12 dataset (Litjens et al., 2014b)) with varying training sample sizes and two modalities; CT and MRI. We also conducted the same experiments using the nnU-Net framework to establish a baseline. Additionally, we performed a statistical analysis to facilitate a fair comparison, and to determine networks with pronounced improvement in segmentation accuracies. Finally, we discussed in detail the merits and limitations of the U-Net variants on their ability to segment the prostate.

The proposed comparison framework[1] is open-source, and although our work focused on prostate related segmentation tasks, it can be easily extended for other ROIs as well. We believe that this would eventually lead to trustworthy deep learning methods, that are reliable and acceptable to doctors and physicians (Goldenberg et al., 2019; Singh et al., 2020; Isensee et al., 2021; Hall et al., 2021; Santoro et al., 2022).

2. A brief history of U-Nets

Since this paper explicitly deals with the U-Net architecture, it is crucial to understand its origin, structure, and why it is important in the context of 3D image segmentation. When AlexNet (Krizhevsky et al., 2012) achieved a top-5 error of 15.3% in the ImageNet-2012 classification challenge (Deng et al., 2009), the field of deep learning and subsequently, convolutional neural networks (CNNs) became the de-facto routine to solve computer vision tasks. This influential 2D architecture spurred numerous articles that employed CNNs and graphical processor units (GPUs) to accelerate DL (Singh et al., 2020). Since then, CNNs have been deployed in a variety of computer vision applications in real-world scenarios such as robotics, self-driving cars, and healthcare (Goldenberg et al., 2019; Singh et al., 2020; Isensee et al., 2021). Three years later, 2D U-Net (Ronneberger et al., 2015) was created to segment neuronal structures in electron microscopic stacks, and was the highest ranking algorithm in the competition.

The architecture of the 2D U-Net has two symmetric pathways; a contracting path to capture the essence of an image, and an expanding path that enables precise localization of the required target. Although the contracting path is made up of the same structure as that of the AlexNet, the U-Net improved by inserting skip connections between both the paths. If we closely inspect all the work since the success of the CNNs, the 2D U-Net was instrumental in reshaping the field of medical image analysis, by being a superior and objective tool. A year later, in 2016 a 3D version of the U-Net was created (Çiçek et al., 2016); Fig. 2 illustrates the architecture of the 3D U-Net. The 3D U-Net was taught to learn dense volumetric segmentation by replacing all 2D operations in the network with their 3D counterparts. This opened up a completely new avenue in the application of DL as most medical images such as MRI, CT, and positron emission tomography (PET) are 3-dimensional in nature.

Fig. 1 showcases a gradual increase in the number of publications of different 3D-medical image based techniques since the introduction of the 3D U-Net. Currently, U-Net is one of the most popular biomedical segmentation architectures that has approximately, 28 000 citations (both 2D and 3D combined). It is still one of the top ranking models in several grand challenges (Litjens et al., 2017b; Isensee et al., 2021; Singh et al., 2020), surpassing new DL approaches such as, recurrent neural networks, generative adversarial networks, and transformers (Crimi and Bakas, 2022). Since its introduction, a huge number of models, with the U-Net as their architectural backbone (template), have been deployed. Many of these variations propose extensions and advances, and have achieved SOTA performances: Attention U-Net (Oktay et al., 2018), U-Net++ (Zhou et al., 2019), SegResNet (Myronenko, 2019) and nnUNet (Isensee et al., 2021).

3. Related work

3.1. Prostate segmentation without any U-Nets

As mentioned earlier, Fig. 1 gives a clear picture that a lot of research has been done to improve the task of automatic prostate segmentation. A significant portion of it was carried out with network architectures other than the U-Net. Here we give an outline of a few of those methods, focusing on the representative ones.

[1] Source code at https://github.com/Shrajan/Prostate-Segmentation.

A CNN and training strategy based on statistical shape models was developed for prostate segmentation from MRIs in Karimi et al. (2018). This method consisted of a dataset with T2-weighted axial images in which the prostate was manually labelled by a radiologist. To overcome the insufficiency of training data, the subtle modes of variation in prostate shapes were learned in a process known as statistical shape modelling. Furthermore, synthetic images were generated to expand the training set, provided these examples were plausible representations of real data, such as translated, rotated, mirrored, or deformed shapes. This method achieved an overlap of 88% between the ground-truth and the prediction; the result was calculated using a metric called Dice-Sørensen Coefficient (DSC = 0.88). The proposed method put forward an innovative data augmentation technique by taking advantage of the prostate shape and out performed the V-Net (Milletari et al., 2016). That being said, the CNN model used fully-connected layers after the encoder blocks to predict the final output. This strategy creates a bottleneck, and reconstruction of the target region to original image size is not always accurate.

A semi-automatic prostate segmentation method from CT images was introduced by Shahedi et al. (2018). Their approach utilized local texture classification and statistical shape modelling, and attained a mean DSC of 0.88. Although the proposed work surpassed the other procedures, overall time required for segmenting a 3D image was considerably more than U-Net based networks. As mentioned by Shahedi et al. (2018), another limitation was that the semi-automatic prostate segmentation method was tested on a small size of the non-brachytherapy dataset (10 images).

Lei et al. (2020) put forward a CNN architecture that first created synthetic MRIs (sMRI) from CT images using a cyclic generative adversarial network. Then, the sMRIs were given as input to a deep attention fully convolution network (DAFCN) to segment the prostate contours. The authors reported DSC, Hausdorff-Distance (HD) and mean surface distance of 0.92 ± 0.09, 4.38 ± 4.66, and 0.62 ± 0.89 mm, respectively, on the test set. In the work by Comelli et al. (2021), a segmentation network, based on a popular natural image classification architecture called Efficient Net (ENet), was created to predict the prostate gland from MR scans. The ENet achieved a mean DSC of 0.908 with cross-validation on 85 samples. However, the performances of the ENet and U-Net (Çiçek et al., 2016) were statistically identical.

The source-code of all the research work described in this subsection is not publicly available, which makes it impossible to replicate the results. Moreover, most of the experiments described in the literature were carried out using private datasets with tuned parameters, such as the learning rate. The performances of these technologies could be improved by utilizing robust data processing and augmentation techniques. Finally, to conduct a fair and rigorous evaluation, all the models should be trained using the same hyperparameters and settings.

3.2. Prostate segmentation with SOTA U-Net variants

As discussed in Section 1, the 3D U-Net architecture is one of the top performing models in most prostate segmentation challenges. Similarly, the U-Net variants that we selected were also used in some of these challenges, and have made it to the top of the leaderboards. To the best of our knowledge, except for the V-Net, the remaining U-Net variants that we selected have not been deployed to prostate segmentation tasks. Moreover, the V-Net architecture has been employed numerous times on both private and public datasets.

Lei et al. (2019) trained a V-Net with reliable contour refinement on TRUS images of 44 patients, and achieved a prostate volume DSC and HD of 0.92 ± 0.03 and 3.94 ± 1.55 mm, respectively. Lee et al. (2020) used an ellipsoid formula to compare the predictions of their V-Net with the ground-truth labels. The dataset contained 330 image samples, and the mean DSC for the entire prostate was 0.87, whereas, the mean DSC for the transition zone was 0.77. A recent method by Jin et al. (2021) extended the V-Net with bi-cubic interpolation technique to train and test 106 clinical prostate MR volumes, and attained a mean DSC of 0.97, and a mean HD of 0.93 mm.

We were unable to find published articles on the applications of Attention U-Net or SegResNet or U-Net++ to segment the prostate zones. However, there are a few publications available involving these models that show promising results for segmentation of prostate tumours in multi-modal MRI scans (Machireddy et al., 2020; Saha et al., 2021). Nonetheless, the aforementioned works provide sufficient evidence that even now, CNN based U-Nets are still the top contenders for segmenting ROIs from 3D medical images.

3.3. Surveys of the past

Several surveys and reviews pertaining to 3D medical segmentation and their applications have been conducted in the past few years. Since not all of them are relevant to our objectives of comparing 3D U-Nets, we have narrowed down to a few works that closely resemble the practices we envisioned for this study.

As mentioned in Section 1, Ghavami et al. (2019) surveyed six CNNs, and assessed their performances using an in-house 3D MRI dataset. The results show that there exists a statistically significant difference in the performances of these models.

Gillespie et al. (2020) did a general review of deep learning in prostate segmentation that enumerated and compared the performances of diverse algorithms, including the SOTA. The experiments in the manuscript also evaluated a 2D U-Net on four publicly available datasets. Moreover, the results outlined the performance of the U-Net on the testing data obtained from other datasets. Although, this work does a good job of eliciting the strengths and weakness of the U-Net, the comparison appears to be incomplete, since only one model and one modality were used.

The creators of the popular nnU-Net (Isensee et al., 2021) framework did a thorough research on the diverse biomedical image segmentation tasks across 53 datasets. The robust nnU-Net segmentation algorithm scored the top places in almost all the challenges. The prostate segmentation tasks were trained on the two public MRI datasets: MSD-prostate (Simpson et al., 2019) and PROMISE-12 (Litjens et al., 2014b). The authors recommend using the original U-Net architecture, as it generalizes well across all the tasks. However, when considering the results of Ghavami et al. (2019), it would be unwise to use the generic U-Net for a specific task (such as prostate segmentation) without comparing it against other network variations across different modalities. One minor drawback of the nnU-Net as compared to other frameworks is longer training times that depends on several factors, such as the model configuration (2D, 3D or cascaded), and large number of epochs (the default value is 1000).

Singh et al. (2020) reviewed the important research ideas in the field of 3D medical imaging analysis using 3D CNNs (and its variants) in different vision application areas such as classification, segmentation, detection, and localization. The article discusses the current challenges associated with the use of 3D CNNs in the medical imaging domain, and the possible future trends in the field. Punn and Agarwal (2022) presented a thorough analysis of the U-Net variants for different medical imaging modalities such as MRI, X-ray, CT, ultrasound, PET, etc. Unfortunately, both manuscripts conducted pure surveys, and no experiments were performed to substantiate the results.

3.4. Public databases

In this work, we evaluated the U-Net variants on prostate segmentation tasks using two public MRI datasets: MSD-prostate (Simpson et al., 2019) and PROMISE-12 (Litjens et al., 2014b). These databases were obtained from biomedical imaging grand-challenges, particularly the Medical Segmentation Decathlon (Simpson et al., 2019), and the MICCAI Grand Challenge: Prostate MR Image Segmentation 2012 (Litjens et al., 2014b), respectively. Over the years, both the challenges have

attracted numerous participants, who have submitted a wide range of computational works and approaches to automatically segment distinct ROIs from an assortment of imaging modalities. Some examples of computational methods include U-Net based models, deep adversarial networks, multi-stage models and hybrid approaches that combined CNNs with other machine learning algorithms, such as support vector machines (SVMs) or random forests (Litjens et al., 2014b, 2017b; Singh et al., 2020; Isensee et al., 2021; Hatamizadeh et al., 2022a).

MSD-prostate: The aim of the medical segmentation decathlon is to provide a comprehensive benchmarking platform for general purpose algorithmic validation and testing that covers a large span of challenges (Simpson et al., 2019). This is achieved through the open sourcing of large medical imaging datasets on several highly different tasks, and by standardizing the analysis and validation process. For this survey, we only used the sub-set of medical images from the prostate segmentation task. Presently, the MSD leaderboard reports the ranks of 54 unique participants, however, it is not possible to ascertain the names of all the associated works or authors. Nonetheless, two out of the top-five are based on the U-Net architecture. Unfortunately, we could not find any research articles corresponding to the works occupying the fifth and the fourth positions (mean positions = 12.8 and 12.5, respectively) on the MSD leaderboard. When the nnU-Net (Isensee et al., 2021) was first published, it was the top ranking algorithm, but it is now in the third spot (mean position = 12.3). The nnU-Net framework is a flexible segmentation method that automatically adapts to a given dataset using 2D, 3D and cascaded U-Nets, and a robust data processing pipeline.

Inspired by the success of transformers for natural language processing, Hatamizadeh et al. (2022b) proposed a new architecture called UNEt TRansformers (UNETR). With a U-Net backbone (but without pure convolutional layers for feature extraction), the UNETR used a pure vision transformer as its encoder. With the further development and success of shifted windows (Swin) transformers (Liu et al., 2021), an advanced version of the UNETR called Swin UNETR was developed (Hatamizadeh et al., 2022a). The Swin UNETR achieved exceptional results on multiple segmentation tasks, and gained the second place with a mean position of 10.1. Liu et al. (2023) created a universal model for organ segmentation and tumour detection using embeddings learned from contrastive language–image pretraining. The model was developed with 3410 CT scans collated from 14 datasets, and attained the first rank on the public leaderboard of the MSD (mean position = 12.3). However, for tasks of the MSD challenge that have MR images (for example the prostate segmentation task), Liu et al. (2023) submitted the output predictions obtained directly using the nnU-Net. As the universal model and the Swin UNETR are transformer based, we did not investigate them further in our work.

PROMISE-12: The Prostate MR Image Segmentation (PROMISE12) challenge was set up to evaluate interactive and automatic prostate segmentation algorithms on the basis of performance and robustness. The challenge dataset consists of transversal T2-weighted MR images from multiple centres, with differences in scanner manufacturers, field strengths and protocols. The leaderboard has over 350 submissions (including duplicate entries), and all the top-five algorithms use some form of CNN-based architectures. With a score of 89.5858, the fifth position is occupied by a boundary-weighted domain adaptive neural network (BOWDA- Net) (Zhu et al., 2019). BOWDA-Net was trained to focus more on the boundaries during segmentation using a boundary-weighted segmentation. The nnU-Net framework used an ensemble approach consisting of 2D and 3D U-Nets to win the fourth rank (score = 89.6507) in the leaderboard (Isensee et al., 2021).

In third place with a total score of 90.3441, the Hybrid Discriminative Network (HD-Net) (Jia et al., 2019) consists of a shared encoder, a segmentation decoder and a boundary decoder. It also incorporates cascaded pyramid convolutional blocks and residual refinement blocks along with attention blocks to extract contextual information. Qin (2019) first trained a ResNet101 (He et al., 2016) on a private MRI dataset, and then used transfer learning to achieve the second spot with 90.7993 points. The method proposed by Qin (2019) designed a multi-level edge attention module to overcome the difficulties of ambiguous boundary in prostate MRI segmentation tasks. The highest ranked network, called as a multi-scale synergic discriminative network (MSD-Net) (Jia et al., 2022), reached an overall score of 91.9072. The MSD-Net is a successor of the HD-Net with near identical components and internal workings. Unfortunately, since the MSD-Net (Jia et al., 2022), Qin (2019), HD-Net (Jia et al., 2019) and BOWDA-Net (Zhu et al., 2019) are not U-Net based architectures, we did not consider them for our evaluation.

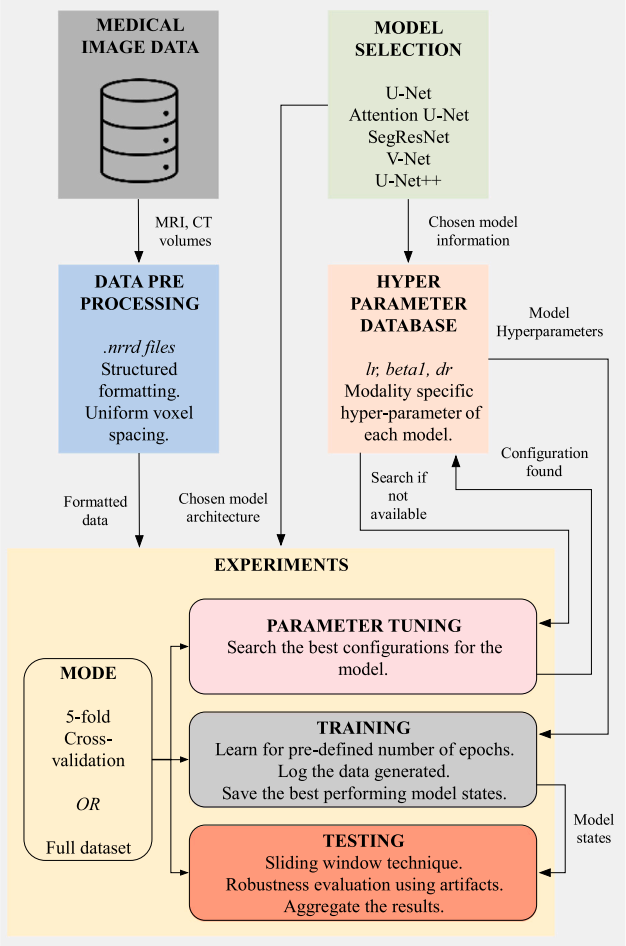

Fig. 3. Structure of the framework.

4. Materials and methods

To ensure an objective comparison between the selected networks, we created a framework as illustrated in Fig. 3. The framework serves as a complete segmentation pipeline, where only the network architecture varies. Each experiment begins with the selection of a model and a dataset by the user. Since we want to achieve the finest possible results, it is paramount to utilize the best parameters and settings. To handle this, our framework provides an all-inclusive data-processing pipeline (including formatting and data augmentations), automatic hyperparameter selection based on modality, training-and-test procedures, and a set of carefully selected evaluation metrics. Each part is described in detail in this section.

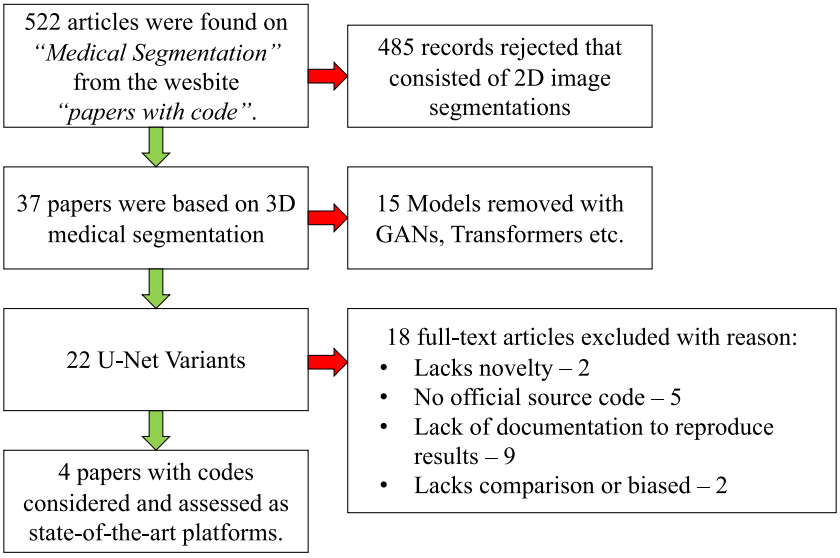

Fig. 4. Selection criteria for assessing SOTA models.

Table 1
Original information of the datasets.

Dataset name	Voxel spacing (in mm^3) [Min/Max]	Volume shape [Min/Max]
Internal CT	$1.099 \times 1.099 \times 3.0$	$[353/636] \times [353/636] \times [80/642]$
Internal MRI	$[0.260/1.786] \times [0.260/1.786] \times [2.0/6.6]$	$[112/768] \times [112/768] \times [22/46]$
MSD-prostate	$[0.600/0.750] \times [0.600/0.750] \times [3.0/4.0] \times 1$	$[256/384] \times [256/384] \times [11/24] \times 2$
PROMISE-12	$[0.273/0.750] \times [0.273/0.750] \times [2.2/4.0]$	$[256/512] \times [256/512] \times [15/54]$

Table 2
Configurations of dataset used in the experiments after processing and conversion[a].

Dataset name	#Samples (Train and Test)	Voxel spacing (in mm^3)	Patch shape	Batch size
Internal CT	Train = 135, Test = 49	$1.099 \times 1.099 \times 3.0$	$128 \times 128 \times 64$	2
Internal MRI	Train = 209, Test = 0	$0.781 \times 0.781 \times 3.0$	$176 \times 176 \times 16$	4
MSD-prostate	Train = 32, Test = 0	$0.625 \times 0.625 \times 3.6$	$228 \times 228 \times 16$	2
PROMISE-12	Train = 50, Test = 0	$0.613 \times 0.613 \times 3.6$	$224 \times 224 \times 16$	2

[a]In case of the MSD-prostate and PROMISE-12, the original datasets do not provide labels for the test set. Furthermore, the Internal MRI dataset does not have a test set. In these instances, we calculated the metrics of the validation samples.

4.1. Selection of U-Net variations

The literature review showed clearly that, the U-Net and its variants achieve SOTA performance in many applications. With these insights, we determined the criteria for the model selection process.

Fig. 4 depicts the strategy we used to select the models based on the following factors:

- The official source code of CNN's architecture in PyTorch (Paszke et al., 2019) was publicly available at https://paperswithcode.com (Stojnic et al., 2021).
- The algorithm has an auto-encoder structure that resembles the 3D U-Net, or is a variant/expansion of the 3D U-Net network.
- The model's experimental (training and testing) results should be majorly based on 3D medical image segmentation tasks. A higher preference was given to data concerning the prostate and PCa.
- The research introduced a novel multi-view and/or multiscale network that aims to enhance its receptive fields and reduce bottlenecks.
- Any alterations of network architecture should be limited to the encoder–decoder blocks of the U-Net variations.

We subjected a substantial number of papers through the selection criteria, and after an exhaustive inspection, four U-Net variants were nominated (Model Selection block in Fig. 3). Their unique features are delineated in the remaining part of this subsection.

The V-Net architecture replaced pooling layers of the U-Net with down (strided) and up convolutions, so that there is a smaller memory footprint during training (Milletari et al., 2016). It also used skip connections within every encoder block to learn residual functions, which further helped to expedite network convergence. In the original publication, the V-Net was trained on 50 MRI volumes (with their respective ground-truth annotations) of the PROMISE-12 challenge.

The authors of the attention U-Net (Oktay et al., 2018) found that by employing the attention gate (AG) model in the U-Net, the network automatically learned to emphasize on target structures of varying shapes and sizes. Initially, this model was trained and evaluated using two separate abdominal 3D CT scans with a focus on segmenting pancreatic boundaries. The attention U-Net was able to perform better than the 3D U-Net by highlighting the important features, and therefore, effectively suppressing the irrelevant regions in the images.

U-Net++ (Zhou et al., 2019) is a variation of the U-Net that introduced additional convolutional layers on the skip pathways, intended to bridge the semantic gap between encoder and decoder feature maps. In addition, it also incorporated dense skip connections on the skip pathways to improve gradient flow, and it also offered deep supervision capable of pruning the model.

SegResNet (Myronenko, 2019) has a segmentation architecture identical to the V-Net, with skip connections and strided convolutions. In addition, due to a limited training dataset size, a variational auto-encoder (VAE) branch was added to reconstruct the input image itself,

Table 3
Hyperparameter ranges used in the configuration search. MD stands for model dependent, that is, the dropout rate used in the original implementation.

Parameter	Lower bound	Upper bound	Default value
Starting lr	10^{-6}	0.1	10^{-4}
β_1 of Adam optimizer	0.1	0.9	0.9
Dropout rate	0.0	0.7	MD
No. of initial channels	16	48	32
Weight decay	10^{-6}	0.0	10^{-5}

Table 4
Robustness of U-Net architectures on four separate prostate datasets using DC and HD95 metrics.

Dataset name	Model name	DC (↑)	HD95 (↓)
Internal CT Train = 135 Test = 49	U-Net	0.832 ± 0.053	3.434 ± 1.099
	Attention U-Net	0.835 ± 0.048	3.488 ± 1.014
	SegResNet	0.832 ± 0.045	3.315 ± 0.671
	V-Net	0.814 ± 0.053	3.774 ± 1.456
	U-Net++	0.834 ± 0.052	3.322 ± 0.959
	nnU-Net	**0.838 ± 0.050**	**3.245 ± 0.896**
Internal MRI Train = 209 Test = 0	U-Net	0.843 ± 0.165	3.176 ± 10.439
	Attention U-Net	0.848 ± 0.150	2.897 ± 9.852
	SegResNet	0.845 ± 0.153	2.971 ± 9.726
	V-Net	0.840 ± 0.164	3.442 ± 14.388
	U-Net++	0.848 ± 0.152	2.911 ± 9.705
	nnU-Net	**0.850 ± 0.149**	**2.825 ± 9.686**
MSD-prostate Train = 32 Test = 0	U-Net	0.826 ± 0.100	14.180 ± 5.349
	Attention U-Net	0.873 ± 0.077	**1.958 ± 1.333**
	SegResNet	0.868 ± 0.085	2.414 ± 1.845
	V-Net	0.820 ± 0.110	13.872 ± 8.515
	U-Net++	0.871 ± 0.109	1.970 ± 1.375
	nnU-Net	**0.876 ± 0.117**	15.078 ± 75.203
PROMISE-12 Train = 50 Test = 0	U-Net	0.865 ± 0.141	2.861 ± 0.830
	Attention U-Net	0.900 ± 0.091	**1.436 ± 0.740**
	SegResNet	0.893 ± 0.085	1.573 ± 0.909
	V-Net	0.865 ± 0.142	2.844 ± 0.921
	U-Net++	0.878 ± 0.100	2.809 ± 0.825
	nnU-Net	**0.910 ± 0.121**	9.735 ± 7.566

to regularize the shared decoder and impose additional constraints on its layers. SegResNet ranked second in the validation stage of the Brain Tumour Segmentation Challenge 2021 (Baid et al., 2021), and in doing so outperformed a transformer-based network.

The original 3D U-Net and the four variations: Attention U-Net, SegResNet, V-Net and U-Net++, were used in our framework experiments across all the datasets.

4.2. Datasets

In this work, four prostate datasets were used: the internal CT and MRI datasets, the MSD-prostate set (Simpson et al., 2019), and the PROMISE-12 (Litjens et al., 2014b) samples. All four datasets have a different sample size of 135, 209, 32, and, 50 respectively. It was imperative that each dataset, irrespective of the modality, was pre-processed appropriately. To ensure this, we have provided detailed information on the four datasets, and how they were individually curated in the following. Table 1 provides information about the original voxel spacings, and the range of shapes for each dataset, prior to processing and reformatting.

Internal CT dataset. For this study, the internal CT dataset comprised of scans that were collated from three different centres. Each centre: Medical Center – University of Freiburg (C1), Hannover Medical School, Hannover, Germany (C2), and Medical School of Nanjing University, Affiliated Drum Tower Hospital, China (C3), contributed 114, 21, and 49 samples, respectively. Two experts from C1 delineated prostate volumes for all patients in consensus. Complete information about the data collection and contouring process is described in Kostyszyn et al. (2020). Datasets of the C1 and C2 were combined as the training set, and the scans from C3 were utilized as an independent test set.

Internal MRI dataset. The internal MRI dataset was collected by the clinical collaborators from C1. This dataset consists of 209 volumes of multimodal (T2, ADC) MRIs, and similar to the MSD-prostate volumes, we only opted for the T2-weighted scans for the experiments. Manual contouring of the prostate was done by the physicians of the group, and the detailed description of the private MRI dataset can be found in the work done by Gunashekar et al. (2022).

MSD-prostate dataset. The MSD-prostate (Simpson et al., 2019) dataset consists of 48 (training = 32, testing = 16) multimodal (T2, ADC) 3D MRI samples. The volumes were obtained from a single source, however, they have non-uniform voxel-spacings. We selected the MSD-prostate dataset for our comparisons as it has an abundance of inter-subject variability. Since the ADC modality is often taken into account for tumour characterization (detection, classification, or segmentation), we only considered the T2-weighted modality of the MSD-prostate dataset for our experiments. This is because T2-weighted images provide most of the important anatomical information required for the task. Additionally, the annotated labels of the original MSD-prostate dataset are separated into two regions: the central prostate gland, and the peripheral zone. Therefore, we combined the two areas into a single ROI, so that the networks could learn whole prostate segmentations.

PROMISE-12 dataset. The public, PROMISE-12 (Litjens et al., 2014b) dataset provides 80 (training = 50, testing = 30) volumes of transversal T2-weighted MRIs with prostate contours. This dataset is a compilation of scans acquired from multiple centres and vendors with different acquisition protocols. Due to this, there is a difference in the voxel-spacings and the slice thickness of the volumes. Though, unlike the MSD-prostate dataset, the PROMISE-12 dataset has only one modality, and the foreground annotations encompass the complete prostate glands.

4.3. Implementation

All the necessary source code for the framework was implemented in PyTorch (Paszke et al., 2019), to objectively compare and evaluate different U-Net variants. The design principles were inspired from the nnU-Net (Isensee et al., 2021) framework, and the scripts were constructed with a few modules from the Medical Open Network for Artificial Intelligence (MONAI) (Consortium, 2020) framework. The network architectures were altered to ensure that the feature maps had a minimum of four voxels in each dimension. We did not further alter the original implementations of the SOTA networks, including the convolution and normalization layers (Ioffe and Szegedy, 2015), activation functions, and dropout layers and probabilities.

Other than the internal CT dataset, the volumes in the other three datasets have varying voxel spacings, as shown in Table 1. CNNs disregard information about heterogeneous spacings while operating on the voxel grids. To handle this heterogeneity in the datasets, all the images and their corresponding ground-truth labels were resampled to the median target voxel spacings of their respective datasets, as listed in Table 2. We employed trilinear and nearest-neighbour interpolations to resample the images and annotated labels, respectively. Due to the large image sizes in most biomedical cases, it is not possible to fit the complete volume into a GPU. For that reason, 3D patches of the foreground and background, as listed in Table 2, were randomly sampled with equal probabilities. This was done to ascertain that the trainings would be smooth, and the networks would converge faster.

In the CT dataset, a few of the patients had metal objects inside their bodies, such as markers, for instance. To avoid high Hounsfield Unit values (bright spots) and a badly distributed normalized space, each patch of a CT scan was clipped to its 0.5 and 99.5 percentiles, and then standardized with the z-score normalization. The resultant patch (X') was computed using Eq. (1).

$$X'(i) = \frac{X(i)-\mu}{\sigma}, \qquad x_i \in X = x_0, \ldots, x_n \tag{1}$$

Table 5
Friedmann test performed on predictions of the U-Net variants to determine the disparities using DC and HD95 metrics.

Dataset name	DC *p-value*	HD95 *p-value*
Internal CT	0.0001	0.0000
Internal MRI	0.0000	0.0000
MSD-prostate	0.0000	0.0000
PROMISE-12	0.0000	0.0000

Table 6
Statistical comparison of U-Net variants on the internal datasets. Wilcoxon signed-rank tests were performed to evaluate statistical differences in the CNNs' segmentation performances. A *p-value* lower than 0.005 indicates that the medians between the results of two models differ significantly. To check which model performed superiorly, see Table 4.

Dataset name	U-Net variant	DC *p-value*	HD95 *p-value*
	U-Net *vs.* Attention U-Net	0.4647	0.8320
	U-Net *vs.* SegResNet	0.0742	0.2393
	U-Net *vs.* V-Net	0.0205	0.0997
	U-Net *vs.* U-Net++	0.9643	0.2764
	U-Net *vs.* nnU-Net	0.0717	0.3402
	Attention U-Net *vs.* SegResNet	0.0742	0.8209
Internal CT	Attention U-Net *vs.* V-Net	0.0941	0.1051
Train = 135	Attention U-Net *vs.* U-Net++	0.0977	0.3277
Test = 49	Attention U-Net *vs.* nnU-Net	0.0175	0.0193
	SegResNet *vs.* V-Net	0.0973	0.0725
	SegResNet *vs.* U-Net++	0.2346	0.7804
	SegResNet *vs.* nnU-Net	0.0159	0.0205
	V-Net *vs.* U-Net++	0.0828	0.0383
	V-Net *vs.* nnU-Net	0.0078	0.0069
	U-Net++ *vs.* nnU-Net	0.0120	0.0366
	U-Net *vs.* Attention U-Net	0.1207	0.2520
	U-Net *vs.* SegResNet	0.3131	0.3044
	U-Net *vs.* V-Net	0.5457	0.6381
	U-Net *vs.* U-Net++	0.0972	0.3026
	U-Net *vs.* nnU-Net	0.0534	0.2597
	Attention U-Net *vs.* SegResNet	0.3004	0.4159
Internal MRI	Attention U-Net *vs.* V-Net	0.4950	0.3400
Train = 209	Attention U-Net *vs.* U-Net++	0.1972	0.7701
Test = 0	Attention U-Net *vs.* nnU-Net	0.2148	0.5492
	SegResNet *vs.* V-Net	0.3160	0.4473
	SegResNet *vs.* U-Net++	0.2582	0.3860
	SegResNet *vs.* nnU-Net	0.4168	0.4391
	V-Net *vs.* U-Net++	0.5433	0.5889
	V-Net *vs.* nnU-Net	0.6947	0.3911
	U-Net++ *vs.* nnU-Net	0.2407	0.6016

where $X(i)$ is a voxel value at position i, μ the mean and σ the standard deviation of patch X. Although data clipping was not done for any of the MRI datasets, z-score normalization was performed after each patch crop. During training, additional samples were synthesized; the data pipeline carried out augmentations such as, elastic deformations, gamma correction, random Gaussian noise, blurring, rotations, and scaling.

Most of the 3D medical segmentation networks, including the U-Net and its variants discussed so far, make use of one of these loss functions: Binary Cross Entropy loss (BCE), or Dice Loss (dice) (Sudre et al., 2017) or their combination (BCE-dice) (Milletari et al., 2016). For a predicted segmentation P and a ground-truth G, the loss functions are defined in Eqs. (2), (3) and (4).

$$L_{\text{BCE}} = -\frac{1}{N} \cdot \sum_{i=1}^{N} g_i \cdot log(p_i) + (1 - g_i) \cdot log(1 - p_i) \tag{2}$$

$$L_{\text{dice}} = 1 - \frac{\sum_{i=1}^{N} p_i g_i + \epsilon}{\sum_{i=1}^{N} p_i + g_i + \epsilon} + \frac{\sum_{i=1}^{N} (1-p_i)(1-g_i) + \epsilon}{\sum_{i=1}^{N} 2 - p_i - g_i + \epsilon} \tag{3}$$

$$L_{\text{BCE-dice}} = L_{\text{BCE}} + L_{\text{dice}} \tag{4}$$

where N is the batch size, p_i and g_i are the predictions and ground-truth labels for a given batch respectively, and $\epsilon = 10^{-6}$ (to avoid divide-by-zero error). We used BCE-dice loss to evaluate the models, and then computed the gradients.

Table 7
Quantitative comparison of U-Net variants on the public datasets (an extension of Table 6). The U-Net variant that has statistically significant values among all the networks is marked in bold font. If a column does not have a value in bold, it means there exists no statistical difference.

Dataset name	U-Net variant	DC *p-value*	HD95 *p-value*
	U-Net *vs.* Attention U-Net	0.0000	**0.0000**
	U-Net *vs.* SegResNet	0.0000	0.0001
	U-Net *vs.* V-Net	0.7791	0.3272
	U-Net *vs.* U-Net++	0.0000	0.0000
	U-Net *vs.* nnU-Net	0.0000	0.0000
	Attention U-Net *vs.* SegResNet	0.0143	**0.0002**
MSD-prostate	Attention U-Net *vs.* V-Net	0.0000	**0.0005**
Train = 32	Attention U-Net *vs.* U-Net++	0.7593	**0.0009**
Test = 0	Attention U-Net *vs.* nnU-Net	0.7248	**0.0009**
	SegResNet *vs.* V-Net	0.0006	0.0089
	SegResNet *vs.* U-Net++	0.0002	0.0710
	SegResNet *vs.* nnU-Net	0.0004	0.0082
	V-Net *vs.* U-Net++	0.0000	0.0000
	V-Net *vs.* nnU-Net	0.0000	0.0000
	U-Net++ *vs.* nnU-Net	0.7648	0.0005
	U-Net *vs.* Attention U-Net	0.0000	**0.0000**
	U-Net *vs.* SegResNet	0.0000	0.0000
	U-Net *vs.* V-Net	0.0247	0.6115
	U-Net *vs.* U-Net++	0.0004	0.0002
	U-Net *vs.* nnU-Net	**0.0000**	0.0000
	Attention U-Net *vs.* SegResNet	0.1876	**0.0000**
PROMISE-12	Attention U-Net *vs.* V-Net	0.0000	**0.0000**
Train = 50	Attention U-Net *vs.* U-Net++	0.0000	**0.0000**
Test = 0	Attention U-Net *vs.* nnU-Net	**0.0001**	**0.0000**
	SegResNet *vs.* V-Net	0.0000	0.0000
	SegResNet *vs.* U-Net++	0.0000	0.0026
	SegResNet *vs.* nnU-Net	**0.0001**	0.0003
	V-Net *vs.* U-Net++	0.0096	0.0116
	V-Net *vs.* nnU-Net	**0.0000**	0.0036
	U-Net++ *vs.* nnU-Net	**0.0000**	0.0019

Throughout the training phases, we used the Adam optimization algorithm with an initial learning rate of 1e−04, that decayed using a polynomial scheduler (Myronenko, 2019) for 500 epochs. We used L2 regularization on the convolution kernel parameters, with a weight of 1e−05. The mini-batch sizes for all the datasets are listed in Table 2. In the inference stages, we employed a sliding window approach with an overlap of 25% to predict the prostate contours.

Several of the works discussed in Section 2 use DSC and HD metrics, however, these metrics can be biased and limited (Maier-Hein et al., 2018; Reinke et al., 2021). We selected the surface Dice similarity coefficient (DC) (Nikolov et al., 2018), and the 95th percentile of Hausdorff Distance (HD95) as metrics to measure the performances of the U-Net variants. The metrics DC and HD95 are more stable to small outliers, and are the current standard in the biomedical segmentation community (Maier-Hein et al., 2022). The open-source implementation of the DC metric can be found at DeepMind's GitHub repository.[2] The HD95 metric is defined in Eq. (5).

$$HD95 = 95\% \left[\max_{p \in P} \left(\min_{g \in G} d(p, g) \right) + \max_{g \in G} \left(\min_{P \in P} d(g, p) \right) \right] \tag{5}$$

where $d(g, p)$ is the Euclidean distance between ground-truth tensor (g) and predicted output (p) of the network.

4.4. Hyperparameter optimization

The hyperparameters settings can considerably influence the performance of a CNN. These parameters are not limited to the architectural choices (e.g., the number of layers, initial weights) but are also applicable to learning rates, the optimizer's momentum factors and regularization techniques. Within the framework, we provide the option

[2] Source code of surface Dice similarity coefficient: https://github.com/deepmind/surface-distance.

Table 8
Average inference time (in s) of each model required for predicting a label.

Model name	Internal CT dataset	MSD-prostate	PROMISE-12	Internal MRI dataset
U-Net	218.89	1.94	6.08	4.80
Attention U-Net	177.16	1.51	4.14	3.75
SegResNet	151.61	1.24	4.02	2.36
V-Net	180.75	1.87	5.17	4.38
U-Net++	279.04	3.05	10.32	6.47
nnU-Net	250.39	1.53	4.47	3.01

of tuning the hyperparameters, such as the initial learning rate (lr), Adam's momentum parameter (β_1), the dropout rate (dr), the number of initial channels in the convolutional layers, and weight decay. The state space of these hyperparameters is listed in Table 3. This was implemented using the Python package called HpBandSter (Falkner et al., 2018). When needed, a total of 22 runs will be executed for each model, out of which 18 unique configurations are sampled for the final deliberation of the parameters.

5. Experiments and results

To analyse the capabilities of the networks, we examined the models for all available volumes using a 5-fold cross-validation (CV). For the nnU-Net framework, we followed the steps detailed in Isensee et al. (2021). We ran our experiments on a NVIDIA Titan RTXs with 24 GB memory. The CV method, for a single network on one dataset, was completed in four and six days with our framework and the nnU-Net framework, respectively. In our study, we did not optimize the hyperparameters of the networks, however, for reference, each search per model can be completed in approximately $5-7$ days.

Since multiple groups based on different models are compared, we used the Friedmann-test to check whether the repeated measurements of the same volumes have the same distribution. When the appraisals were inconsistent, then a statistical analysis was done with the Wilcoxon signed-rank test to identify the difference in performance among the networks. The Wilcoxon signed-rank test, which checks for the null hypothesis that the median group difference is zero, was chosen due to the abnormal distribution and heteroscedasticity of the data. We performed a total of 120 statistical tests, between 120 pairs of models. As the number of tests increased, so did the likelihood of a type I error, that is, a considerable number of false-negative differences could be present. This was accounted by adjusting the significance threshold from 0.05 to 0.005 using the Bonferroni correction. Consequently, for our experiments, the confidence alpha was set to 0.5% for both Friedmann and Wilcoxon signed-rank tests.

5.1. Segmentation accuracy and robustness

In Table 4, we have summarized the robustness of the U-Net and other SOTA networks with the help of the DC and HD95 metrics. Fig. 5 displays the predicted labels of each model for different datasets. Throughout our experiments, we found out that the for datasets with larger training sample size (greater than 100), all the U-Net variants produce nearly identical results. As evident in Table 4, for both the internal datasets, the segmentation accuracies of the nnU-Nets are equivalent to that of the Attention U-Nets, SegResNets, and U-Net++.

For datasets with a smaller sample size (fewer than 100), such as the MSD-prostate and PROMISE-12, we noticed a clear difference in the segmentation accuracies of the U-Net variants. The nnU-Net outperformed the other algorithms when the DC metric was considered. On the contrary, the Attention U-Net appears to be the superior network, when assessed from the perspective of the HD95 metric. It can be noted that the generic U-Net and the V-Net architectures do not achieve good results for small sized datasets.

5.2. Statistical analysis

The results of the Friedmann tests for each dataset using DC and HD95 metrics are listed in Table 5. With a confidence level of 0.5%, it can be inferred that there exists a difference in the performance of the models across all the datasets. Subsequent statistical evaluation of the different U-Net variants using the Wilcoxon test on the DC and HD95 values is shown in Tables 6 and 7. The details of the statistical evaluation for each dataset is described below:

Internal CT: In Table 6, for the DC metric the p-values range from 0.0078 to 0.9643, and for the Hausdorff distance (95th percentile) the p-values range from 0.0069 to 0.8320. It is evident that there is no single U-Net variant with a superior performance, however, it is worth noting that the p-values for the nnU-Net when compared to the other networks is slightly better.

Internal MRI: Similar to the p-values for the internal CT dataset, the models do not exhibit significantly different behaviours for the in-house MRI volumes. Interestingly, the performance of the highest ranked nnU-Net model in Table 4 is similar to the networks in the lower position, namely the V-Net and U-Net++. Overall, the range of p-values is 0.0534 to 0.6947 and 0.2520 to 0.7701 for DC and HD95 metrics, respectively.

MSD-prostate: In Table 7, the detailed pairwise multiple comparison results listed for DC and HD-95 metrics are contradictory. For the DC metric, although the p-values range from 0.0000 to 0.7791, the performances of almost all U-Net variants (except the U-Net and V-Net with lower accuracies) are similar. This is denoted by high p-values for Attention U-Net, V-Net, U-Net++ and nnU-Net. On the contrary, the p-values (range: 0.0000 to 0.3272) for the HD95 metric convey the distinguished performance of the Attention U-Net (highlighted in bold). As before, the U-Net and V-Net produce identical predictions, and the p-value of the SegResNet is greater than 0.005 against U-Net++ and nnU-Net, respectively.

PROMISE-12: The results of the Wilcoxon signed-rank test enumerated in Table 7 resemble the segmentation accuracies of the U-Net variants recorded in Table 4. When considering the DC metric, the nnU-Net considerably outperforms the other networks, as indicated by the low p-values. The other five U-Net based models showcase homogeneous outcomes. Furthermore, it is also apparent the Attention U-Net surpasses the rest of the architectures for the HD95 metric. Unsurprisingly, the higher HD95 value for the nnU-Net diminished the overall standing for the PROMISE-12 dataset.

To summarize, there is a substantial resemblance in the predicted images among all the networks for the internal datasets. Whereas, in the case of the public datasets, the nnU-Net and Attention U-Net secured the top rankings, the SegResNet and U-Net++ occupied the middle spots, and the U-Net and V-net were the worst performing models.

5.3. Computation time

As outlined in Table 8, predicting using SegResNet, Attention U-Net, nnU-Net, and V-Net takes the least amount of time. U-Net takes slightly more time to produce the labels due to extra kernels in the auto-encoder structure. Also, the U-Net++ takes approximately 125% of time as needed for the U-Net as it contains multiple convolutional layers in the skip connections. Even though the nnU-Net utilized deep supervision during training, it was disabled during inference, essentially, speeding up the process. The required time to segment is large for the internal CT dataset, due to the large shapes of the images, and because each volume is subjected to 5-folds using the sliding window approach.

6. Discussion

CNNs have played a vital role in medical diagnosis and clinical advancements. Despite the efficiency of the U-Net and its variants for prostate gland segmentation, they have flaws and associated challenges. One such major challenge is the expense of data acquisition

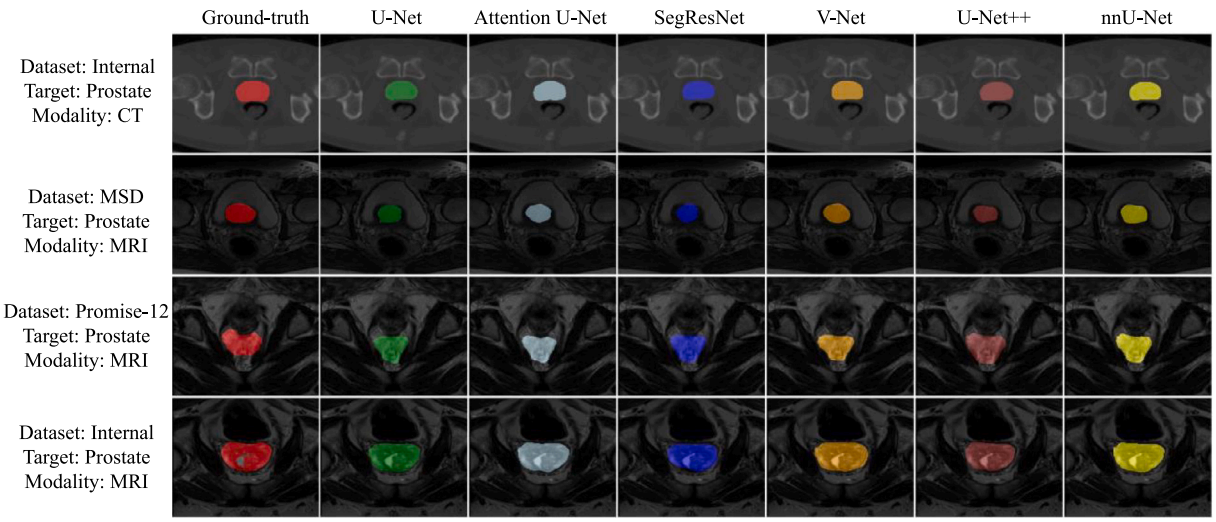

Fig. 5. A qualitative comparison of the accuracy of 3D U-Nets, Attention U-Nets, SegResNets, V-Nets, U-Net++ and nnU-Nets on the prostate-gland segmentation task.

and delineation, which ultimately leads to the limited availability of the data. Furthermore, in most experiments, the patient images belong to a particular study group, and due to ethical and privacy concerns, the authors (owners) are unable to make them public. Another shortcoming is that, although large amounts of research get published, the documentation needs to be improved to reproduce the performances. Additionally, nearly all the openly available prostate annotated datasets have MR modalities. Therefore, authors often limit their experiments to in-house datasets. Consequently, these reasons make it almost impossible to judge the performances of multiple and make a fair comparison objectively.

Another major obstacle to applying the outstanding anatomical segmentation results of U-Nets in clinical routine (for example, prostate contours before radiotherapy) is the complexity of medical software. Generally, radiology and radiation oncology departments streamline workflows for efficient reporting and image post-processing (Dias-Santagata et al., 2022). Furthermore, we believe the local PACS (picture archiving and communication system) and reporting platforms should seamlessly integrate AI technologies for broader applications. But, at the moment, there is a gap in AI training in the curriculum of medical staff (Santomartino and Yi, 2022). If a comprehensive integration into the clinical workflows is not provided, clinical departments would need additional staff trained in imaging informatics to apply research-based algorithms. We theorize that the acceptance of AI algorithms will increase with the release of software that requires fewer manual steps to obtain results in a clinical image viewer.

We tried to address all the aforementioned issues by introducing a comprehensive evaluation framework covering the complete workflow, beginning with data pre-processing, data augmentation, training, inference, and culminating with post-processing. Based on an extensive review of the existing literature, we selected five SOTA architectures for further evaluation. The properties of our chosen datasets comprised of different clinical settings with multi-modal data, varying training samples, and homogeneous and heterogeneous datasets. A crucial aspect of this study was the choice of evaluation metrics and statistical comparison methods. We selected two of the foremost metrics to validate the efficiency of the segmentation models, yet they provide contradicting assessments in certain situations. For instance, when we examined the DC and HD95 scores of the volumes from the private datasets, we noticed that both the metrics endorse each other. However, this consistency degrades when there are fewer samples in the datasets, and then the nnU-Net and Attention U-Net compete with one another for the highest rank. Therefore, we performed statistical analysis to figure out whether there were noteworthy differences in performances. With this, we were able to achieve conclusive proof and determine the best networks.

Amongst the SOTA, the Attention U-Net and the nnU-Net are the most versatile segmentation tools exhibiting generalizability, as they top most of the tables, and consume less time to predict the labels. It is worth noting that the models exhibit dissimilar performances depending on the dataset properties. In the case of the internal datasets and irrespective of the modality, all the models have indistinguishable accuracies. With that being said, the DC metric standard deviations for the MRI scans are greater than 0.1, and one reason is that a sizeable number of MR volumes from this set have heterogeneous voxel spacings. The large variances are also present in certain models for the public datasets, including the nnU-Net, U-Net, U-Net++ and V-Net.

Additionally, we acknowledge that strengths and contributions of other frameworks, such as MONAI (Consortium, 2020) and NiftyNet (Gibson et al., 2018). We obtained the script for SegResNet from the MONAI repository (contributed by the original authors of the model), and integrated it in our framework. Although MONAI and NiftyNet frameworks have a comprehensive list of networks, loss functions and other APIs, our framework is built with an emphasis on prostate gland segmentation from 3D images. Our work aims to contribute in the direction of existing frameworks, so that we could collaborate and collectively improve medical deep learning algorithms. Therefore, we recommend not to directly deploy any of the SOTA architectures and frameworks for a particular task, including the U-Net, Attention U-Net and nnU-Net. Rather, thoroughly investigate the frameworks with unique networks, analyse the performances of each combination for a given dataset, and then pick the most suitable solution.

We thoroughly investigated the individual SOTA CNNs to the best of their ability. However, we did not explore the combination of the selected U-Net variants for clinical use. One possible option would be to create a new architecture using the best features of the existing ones, perhaps the attention-gate of the Attention U-Net (Oktay et al., 2018) or the deep supervision technique of the U-Net++ (Zhou et al., 2019). Another course of action would be to create a cascade of two or three different (multi-stage) networks placed in conjunction to one another. Several works have already implemented one or all of these alternatives to exploit the spatial contextual information across multiple modalities and extract semantically consistent features of the prostate gland (Isensee et al., 2021; Jia et al., 2019, 2022). For instance, depending on the properties of a given dataset, the model configuration of the nnU-Net framework can be set to 2D, 3D-low resolution, 3D-full resolution or 3D-cascade (two-stage: low resolution plus full resolution) (Isensee et al., 2021). Furthermore, the nnU-Net

has a modified version of the deep supervision technique to tune the hidden layers of the decoder blocks. As stated in Section 2, both the HD-Net (Jia et al., 2019) and MSD-Net (Jia et al., 2022) have architectures with two decoders that share the same encoder. In addition, the HD-Net and MSD-Net use pyramid convolutional blocks and residual refinement blocks in a cascaded fashion, as well as channel attention blocks (loosely inspired by the V-Net Milletari et al., 2016, SegResNet Myronenko, 2019 and Attention U-Net Oktay et al., 2018). But, we could not find the original source-code of the MSD-Net and HD-Net, and therefore were not able to reproduce the results. Regardless, the impressive results of the nnU-Net (Isensee et al., 2021, 2022) in the MSD challenge (Simpson et al., 2019) and the multi-modality abdominal multi-organ segmentation (AMOS) challenge (Ji et al., 2022) have demonstrated that DL algorithms can now achieve state-of-the-art performance without human intervention for hyperparameter optimization. These successes are beneficial as the images for both the MSD and AMOS challenges were acquired across multiple institutions and modalities during real-world clinical applications. Antonelli et al. (2022) monitored the generalizability of the top ranked methods in the MSD challenge for 2 years (since 2018), and surmised that precise semantic segmentation networks can now be fully automated. The free availability of medical image segmentation methods permits clinical researchers with intermediate computer skills to use the software with minimal or lack of AI specific knowledge.

As mentioned in Section 1, we analysed a few commercial products for automatic medical images and extracting crucial information (Estienne et al., 2020; Wong et al., 2020; Consortium, 2020). The accomplishments of the commercially available software tools have proved that the AI driven systems can alleviate the workloads of the healthcare workers (Wong et al., 2021; D'Aviero et al., 2022; Radici et al., 2022; Caba et al., 2023). The Division of Medical Physics at Freiburg is currently evaluating two different AI-systems (Estienne et al., 2020; Wong et al., 2020) in a clinical setting. The results convey that the products can accurately contour different organs, for example, the bladder, heart, and lungs. However, certain ROIs need minute modifications to become clinically adaptable. This might stem from the fact that every clinic has its own distinctive procedures in the imaging and contouring process, that render the application of a general model difficult. These findings are in line with the results of our comparison framework, which showed that the best performing models depend on the used dataset. It is our opinion that a full automation of the workflow is not yet credible. But with a thoroughly conducted network selection process or an additional fine-tuning step of vendor models, segmentation performance could be further increased.

Most of our experiments with the four datasets for the prostate screening application focused on diagnostic MR and CT sequences. This, however, may not be applicable for real-world clinical phases. Often low dose CT or fast MR sequences are used. For example, when positioning the patient and optimizing the treatment plan in ART. Such techniques usually show increased noise and artefacts in the scanned images. One strategy to address this weakness was proposed by Gong et al. (2022), who showed a generative adversarial network to restore and improve the image quality of the low-dose CT volumes. The authors integrated a CT-linac platform with a DL algorithm to reconstruct images in order to meet the clinical feasibility requirements of ART. Transfer learning can usher significant improvement in the delineation of ROI and PCa, as evident in the study conducted by Kawula et al. (2023) with images collated from two MR-linacs. The use of ART for aggressive tumours remains an essential component of organ-sparing and multimodality cancer treatment (Skup, 2010; Hall et al., 2021). By implementing automatic segmentation techniques in the context of an online adaptive workflow could help by shortening the re-contouring time and reducing inter-observer variability. Nevertheless, considerable prospective evaluations are vital to substantiate that online ART with automatic DL approaches improve clinical outcomes for patients with a variety of malignancies (Hall et al., 2021).

A potential robotic strategy consisting of a versatile automatic segmentation algorithm in conjunction with ART could be used to optimize treatment planning and delivery. With the planned approach, the DL network would provide feedback from imaging by localizing and capturing anatomical changes of the ROIs, so that the RT treatment parameters could be customized for each patient (Hall et al., 2021; Kawula et al., 2023). There is a wide spectrum of hypothesized advantages, from shorter workflows (since manual contouring is time-consuming) (Kawula et al., 2023), to small toxicity improvements, and hopefully, meaningful gains in overall survival (Hall et al., 2021).

During the course of this research survey and evaluation, we discovered a difference in perspectives that computer scientists and clinicians have regarding the implementation of AI adjuncts for various manual tasks in clinical settings. For computer scientists, the focus may have to shift from the pure development of algorithms to ongoing performance monitoring, which would require collaboration with clinical researchers and epidemiologists. On the other hand, the algorithms could provide clinicians with an excellent opportunity to reduce the time spent on manual segmentations and cumbersome detection tasks, such as detecting small prostate metastases in a CT scan. Clinicians can use this extra time to focus on more patient-centric work, such as preparing multi-disciplinary meetings like in-hospital tumour conferences for optimal patient management, increasing patient–doctor interaction, and refocusing on diagnostic and therapeutic procedures in radiology departments.

In the future, we plan to further carry out experiments on the models using samples from PROSTATEx (Litjens et al., 2014a), NCI-ISBI (Bloch et al., 2015), and PI-CAI (Saha et al., 2022) challenges, and observe how the models behave on intra-prostatic tumour segmentation tasks. We also hope to test these models on other modalities, such as PET and multi-parametric MRI. With the required and appropriate modifications, we aim to utilize these automatic segmentation algorithms in various clinical applications, such as the diagnosis, prognosis, and treatment of prostate cancer with an online adaptive workflow.

7. Conclusion

This project explored the capabilities of five different variations of U-Nets (including the original version), and applied them to segment the prostate gland from MR and CT images. We conducted our experiments on four prostate-contour-annotated datasets: internal CT, MSD-prostate, PROMISE-12, and internal MRI. The networks were chosen based on criterias such as, source code availability and reproducibility, and whether they were already applied in the biomedical segmentation scenarios. This ease of integration opens a wide spectrum of application for U-Net, with endless possibilities of novel architecture designs. Considering the implementation strategies, most authors applied an end-to-end training-from-scratch tactic with minimal pre-processing, i.e. resizing and normalization. We also performed statistical analysis to compare the significance of the predictions of each model. Our results show that on volumes that were obtained from small sized source, the networks produce statistically different results. However, this difference in performance diminishes with a sufficiently large number of samples, irrespective of their modality. We investigated and benchmarked that the U-Nets can predict unbiased features of the prostate biology from medical images. This combined with their impressive pace of operation could further help in the detection of tumours, and hopefully, alleviate the arduous procedure of radiation therapy.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

This research was funded in part by the Austrian Science Fund (FWF): I 4718, and the Federal Ministry of Education and Research (BMBF) Germany, under the frame of the Horizon-2020 ERA-PerMed call JTC-2019 Project Personalized Medicine: Multidisciplinary Research Towards Implementation (PersoRad). Z. B. was supported by the Doctoral College Resilient Embedded Systems, which is run jointly by the TU Wien's Faculty of Informatics and the UAS Technikum Wien.

References

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., van Ginneken, B., Bilello, M., Bilic, P., Christ, P.F., Do, R.K.G., Gollub, M.J., Heckers, S.H., Huisman, H., Jarnagin, W.R., McHugo, M.K., Napel, S., Pernicka, J.S.G., Rhode, K., Tobon-Gomez, C., Vorontsov, E., Meakin, J.A., Ourselin, S., Wiesenfarth, M., Arbeláez, P., Bae, B., Chen, S., Daza, L., Feng, J., He, B., Isensee, F., Ji, Y., Jia, F., Kim, I., Maier-Hein, K., Merhof, D., Pai, A., Park, B., Perslev, M., Rezaiifar, R., Rippel, O., Sarasua, I., Shen, W., Son, J., Wachinger, C., Wang, L., Wang, Y., Xia, Y., Xu, D., Xu, Z., Zheng, Y., Simpson, A.L., Maier-Hein, L., Cardoso, M.J., 2022. The medical segmentation decathlon. Nature Commun. 13 (1), 4128.

Baid, U., Ghodasara, S., Bilello, M., Mohan, S., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., Prevedello, L.M., Rudie, J.D., Sako, C., Shinohara, R.T., Bergquist, T., Chai, R., Eddy, J., Elliott, J., Reade, W., Schaffter, T., Yu, T., Zheng, J., Annotators, B., Davatzikos, C., Mongan, J., Hess, C., Cha, S., Villanueva-Meyer, J.E., Freymann, J.B., Kirby, J.S., Wiestler, B., Crivellaro, P., Colen, R.R., Kotrotsou, A., Marcus, D.S., Milchenko, M., Nazeri, A., Fathallah-Shaykh, H.M., Wiest, R., Jakab, A., Weber, M., Mahajan, A., Menze, B.H., Flanders, A.E., Bakas, S., 2021. The RSNA-ASNR-MICCAI brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. CoRR abs/2107.02314. arXiv:2107.02314. URL: https://arxiv.org/abs/2107.02314.

Bloch, N., Madabhushi, A., Huisman, H., Freymann, J., Kirby, J., Grauer, M., Enquobahrie, A., Jaffe, C., Clarke, L., Farahani, K., 2015. NCI-ISBI 2013 challenge: Automated segmentation of prostate structures. the cancer imaging archive. http://dx.doi.org/10.7937/K9/TCIA.2015.zF0vlOPv.

Caba, B., Cafaro, A., Lombard, A., Arnold, D.L., Elliott, C., Liu, D., Jiang, X., Gafson, A., Fisher, E., Belachew, S.M., Paragios, N., 2023. Single-timepoint low-dimensional characterization and classification of acute versus chronic multiple sclerosis lesions using machine learning. NeuroImage 265, 119787. http://dx.doi.org/10.1016/j.neuroimage.2022.119787, URL: https://www.sciencedirect.com/science/article/pii/S1053811922009089.

Choi, M.S., Choi, B.S., Chung, S.Y., Kim, N., Chun, J., Kim, Y.B., Chang, J.S., Kim, J.S., 2020. Clinical evaluation of atlas-and deep learning-based automatic segmentation of multiple organs and clinical target volumes for breast cancer. Radiother. Oncol. 153, 139–145.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 424–432, URL: https://lmb.informatik.uni-freiburg.de/Publications/2016/CABR16/.

Comelli, A., Dahiya, N., Stefano, A., Vernuccio, F., Portoghese, M., Cutaia, G., Bruno, A., Salvaggio, G., Yezzi, A., 2021. Deep learning-based methods for prostate segmentation in magnetic resonance imaging. Appl. Sci. 11 (2), http://dx.doi.org/10.3390/app11020782, URL: https://www.mdpi.com/2076-3417/11/2/782.

Consortium, M., 2020. MONAI: Medical open network for AI. http://dx.doi.org/10.5281/zenodo.6114127.

Crimi, A., Bakas, S. (Eds.), 2022. Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries. Springer International Publishing, http://dx.doi.org/10.1007/978-3-031-08999-2.

D'Aviero, A., Re, A., Catucci, F., Piccari, D., Votta, C., Piro, D., Piras, A., Di Dio, C., Iezzi, M., Preziosi, F., Menna, S., Quaranta, F., Boschetti, A., Marras, M., Miccichè, F., Gallus, R., Indovina, L., Bussu, F., Valentini, V., Cusumano, D., Mattiucci, G.C., 2022. Clinical validation of a deep-learning segmentation software in head and neck: An early analysis in a developing radiation oncology center. Int. J. Environ. Res. Public Health 19 (15), http://dx.doi.org/10.3390/ijerph19159057, URL: https://www.mdpi.com/1660-4601/19/15/9057.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–255.

Dias-Santagata, D., Heist, R.S., Bard, A.Z., da Silva, A.F.L., Dagogo-Jack, I., Nardi, V., Ritterhouse, L.L., Spring, L.M., Jessop, N., Farahani, A.A., Mino-Kenudson, M., Allen, J., Goyal, L., Parikh, A., Misdraji, J., Shankar, G., Jordan, J.T., Martinez-Lage, M., Frosch, M., Graubert, T., Fathi, A.T., Hobbs, G.S., Hasserjian, R.P., Raje, N., Abramson, J., Schwartz, J.H., Sullivan, R.J., Miller, D., Hoang, M.P., Isakoff, S., Ly, A., Bouberhan, S., Watkins, J., Oliva, E., Wirth, L., Sadow, P.M., Faquin, W., Cote, G.M., Hung, Y.P., Gao, X., Wu, C.-L., Garg, S., Rivera, M., Le, L.P., John Iafrate, A., Juric, D., Hochberg, E.P., Clark, J., Bardia, A., Lennerz, J.K., 2022. Implementation and clinical adoption of precision oncology workflows across a healthcare network. Oncologist 27 (11), 930–939.

Estienne, T., Lerousseau, M., Vakalopoulou, M., Alvarez Andres, E., Battistella, E., Carré, A., Chandra, S., Christodoulidis, S., Sahasrabudhe, M., Sun, R., Robert, C., Talbot, H., Paragios, N., Deutsch, E., 2020. Deep learning-based concurrent brain registration and tumor segmentation. Front. Comput. Neurosci. 14, 17, URL: https://github.com/TheoEst/joint_registration_tumor_segmentation.

Falkner, S., Klein, A., Hutter, F., 2018. BOHB: Robust and efficient hyperparameter optimization at scale. In: Proceedings of the 35th International Conference on Machine Learning. pp. 1436–1445, URL: https://github.com/automl/HpBandSter.

Ghavami, N., Hu, Y., Gibson, E., Bonmati, E., Emberton, M., Moore, C.M., Barratt, D.C., 2019. Automatic segmentation of prostate MRI using convolutional neural networks: Investigating the impact of network architecture on the accuracy of volume measurement and MRI-ultrasound registration. Med. Image Anal. 58, 101558. http://dx.doi.org/10.1016/j.media.2019.101558, URL: https://github.com/NifTK/NiftyNet.

Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D.I., Wang, G., Eaton-Rosen, Z., Gray, R., Doel, T., Hu, Y., et al., 2018. NiftyNet: a deep-learning platform for medical imaging. Comput. Methods Programs Biomed. 158, 113–122.

Gillespie, D., Kendrick, C., Boon, I., Boon, C., Rattay, T., Yap, M.H., 2020. Deep learning in magnetic resonance prostate segmentation: A review and a new perspective. arXiv preprint arXiv:2011.07795.

Goldenberg, S.L., Nir, G., Salcudean, S.E., 2019. A new era: artificial intelligence and machine learning in prostate cancer. Nature Rev. Urol. 16 (7), 391–403.

Gong, W., Yao, Y., Ni, J., Jiang, H., Jia, L., Xiong, W., Zhang, W., He, S., Wei, Z., Zhou, J., 2022. Deep learning-based low-dose CT for adaptive radiotherapy of abdominal and pelvic tumors. Front. Oncol. 12, http://dx.doi.org/10.3389/fonc.2022.968537, URL: https://www.frontiersin.org/articles/10.3389/fonc.2022.968537.

Gunashekar, D.D., Bielak, L., Hägele, L., Berlin, A., Oerther, B., Benndorf, M., Grosu, A., Zamboglou, C., Bock, M., 2022. Explainable AI for CNN-based prostate tumor segmentation in multi-parametric MRI correlated to whole mount histopathology. Radiat. Oncol. http://dx.doi.org/10.21203/rs.3.rs-1225229/v1.

Haga, A., Takahashi, W., Aoki, S., Nawa, K., Yamashita, H., Abe, O., Nakagawa, K., 2019. Standardization of imaging features for radiomics analysis. J. Med. Invest. 66 (1.2), 35–37.

Hall, W.A., Paulson, E., Li, X.A., Erickson, B., Schultz, C., Tree, A., Awan, M., Low, D.A., McDonald, B.A., Salzillo, T., Glide-Hurst, C.K., Kishan, A.U., Fuller, C.D., 2021. Magnetic resonance linear accelerator technology and adaptive radiation therapy: An overview for clinicians. CA Cancer J. Clin. 72 (1), 34–56.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D., 2022a. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I. Springer, pp. 272–284, URL: https://github.com/Project-MONAI/MONAI.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D., 2022b. Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584, URL: https://github.com/Project-MONAI/MONAI.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778, URL: https://github.com/pytorch/vision/blob/main/torchvision/models/resnet.py.

Hekler, A., Utikal, J.S., Enk, A.H., Solass, W., Schmitt, M., Klode, J., Schadendorf, D., Sondermann, W., Franklin, C., Bestvater, F., Flaig, M.J., Krahl, D., von Kalle, C., Fröhling, S., Brinker, T.J., 2019. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. Eur. J. Cancer 118, 91–96. http://dx.doi.org/10.1016/j.ejca.2019.06.012, URL: https://www.sciencedirect.com/science/article/pii/S0959804919303806.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on Machine Learning 37, 448–456.

Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18 (2), 203–211, URL: https://github.com/MIC-DKFZ/nnUNet.

Isensee, F., Ulrich, C., Wald, T., Maier-Hein, K.H., 2022. Extending nnu-net is all you need. arXiv preprint arXiv:2208.10791.

Jeong, H., Ntolkeras, G., Alhilani, M., Atefi, S.R., Zöllei, L., Fujimoto, K., Pourvaziri, A., Lev, M.H., Grant, P.E., Bonmassar, G., 2021. Development, validation, and pilot MRI safety study of a high-resolution, open source, whole body pediatric numerical simulation model. PLoS One 16 (1), e0241682.

Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al., 2022. AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. arXiv preprint arXiv:2206.08023.

Jia, H., Cai, W., Huang, H., Xia, Y., 2022. Learning multi-scale synergic discriminative features for prostate image segmentation. Pattern Recognit. 126, 108556. http://dx.doi.org/10.1016/j.patcog.2022.108556, URL: https://www.sciencedirect.com/science/article/pii/S0031320322000371.

Jia, H., Song, Y., Huang, H., Cai, W., Xia, Y., 2019. HD-net: Hybrid discriminative network for prostate segmentation in MR images. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer International Publishing, Cham, pp. 110–118.

Jin, Y., Yang, G., Fang, Y., Li, R., Xu, X., Liu, Y., Lai, X., 2021. 3D PBV-net: An automated prostate MRI data segmentation method. Comput. Biol. Med. 128, 104160. http://dx.doi.org/10.1016/j.compbiomed.2020.104160, URL: https://www.sciencedirect.com/science/article/pii/S0010482520304911.

Karimi, D., Samei, G., Kesch, C., Nir, G., Salcudean, S.E., 2018. Prostate segmentation in MRI using a convolutional neural network architecture and training strategy based on statistical shape models. Int. J. Comput. Assist. Radiol. Surg. 13 (8), 1211—1219.

Kawula, M., Hadi, I., Nierer, L., Vagni, M., Cusumano, D., Boldrini, L., Placidi, L., Corradini, S., Belka, C., Landry, G., Kurz, C., 2023. Patient-specific transfer learning for auto-segmentation in adaptive 0.35 T MRgRT of prostate cancer: a bi-centric evaluation. Med. Phys. 50 (3), 1573–1585. http://dx.doi.org/10.1002/mp.16056, URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.16056.

Korreman, S., Eriksen, J.G., Grau, C., 2021. The changing role of radiation oncology professionals in a world of AI–just jobs lost–or a solution to the under-provision of radiotherapy? Clin. Transl. Radiat. Oncol. 26, 104–107.

Kostyszyn, D., Fechter, T., Bartl, N., Grosu, A.L., Gratzke, C., Sigle, A., Mix, M., Ruf, J., Fassbender, T.F., Kiefer, S., Bettermann, A.S., Nicolay, N.H., Spohn, S., Kramer, M.U., Bronsert, P., Guo, H., Qiu, X., Wang, F., Henkenberens, C., Werner, R.A., Baltas, D., Meyer, P.T., Derlin, T., Chen, M., Zamboglou, C., 2020. Intraprostatic tumour segmentation on PSMA-PET images in patients with primary prostate cancer with a convolutional neural network. J. Nucl. Med. URL: https://gitlab.com/dejankostyszyn/prostate-gtv-segmentation.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. pp. 1097–1105, URL: https://github.com/pytorch/vision/blob/main/torchvision/models/alexnet.py.

Lee, D.K., Sung, D.J., Kim, C.-S., Heo, Y., Lee, J.Y., Park, B.J., Kim, M.J., 2020. Three-dimensional convolutional neural network for prostate mri segmentation and comparison of prostate volume measurements by use of artificial neural network and ellipsoid formula. Am. J. Roentgenol. 214 (6), 1229–1238. http://dx.doi.org/10.2214/AJR.19.22254, PMID: 32208009.

Lei, Y., Dong, X., Tian, Z., Liu, Y., Tian, S., Wang, T., Jiang, X., Patel, P., Jani, A.B., Mao, H., Curran, W.J., Liu, T., Yang, X., 2020. CT prostate segmentation based on synthetic MRI-aided deep attention fully convolution network. Med. Phys. 47 (2), 530–540.

Lei, Y., Tian, S., He, X., Wang, T., Wang, B., Patel, P., Jani, A.B., Mao, H., Curran, W.J., Liu, T., Yang, X., 2019. Ultrasound prostate segmentation based on multidirectional deeply supervised V-Net. Med. Phys. 46 (7), 3194–3206.

Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., Huisman, H., 2014a. Computer-aided detection of prostate cancer in MRI. IEEE Trans. Med. Imaging 33 (5), 1083–1092.

Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., Huisman, H., 2017a. Prostatex challenge data. The cancer imaging archive. http://dx.doi.org/10.7937/K9TCIA.2017.MURS5CL.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I., 2017b. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Litjens, G., Toth, R., (van de Ven), W., Hoeks, C., Kerkstra, S., (van Ginneken), B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., Strand, R., Malmberg, F., Ou, Y., Davatzikos, C., Kirschner, M., Jung, F., Yuan, J., Qiu, W., Gao, Q., Edwards, P.E., Maan, B., van-der Heijden, F., Ghose, S., Mitra, J., Dowling, J., Barratt, D., Huisman, H., Madabhushi, A., 2014b. Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge. Med. Image Anal. 18 (2), 359–373, URL: https://promise12.grand-challenge.org.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022, URL: https://github.com/microsoft/Swin-Transformer.

Liu, J., Zhang, Y., Chen, J.-N., Xiao, J., Lu, Y., Landman, B.A., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z., 2023. CLIP-driven universal model for organ segmentation and tumor detection. arXiv preprint arXiv:2301.00785. URL: https://github.com/ljwztc/CLIP-Driven-Universal-Model.

Machireddy, A., Meermeier, N., Coakley, F., Song, X., 2020. Malignancy detection in prostate multi-parametric MR images using U-net with attention. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society. EMBC, pp. 1520–1523. http://dx.doi.org/10.1109/EMBC44109.2020.9176050.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A.P., Carass, A., et al., 2018. Why rankings of biomedical image analysis competitions should be interpreted with care. Nature Commun. 9 (1), 1–13.

Maier-Hein, L., Reinke, A., Christodoulou, E., Glocker, B., Godau, P., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M.A., et al., 2022. Metrics reloaded: Pitfalls and recommendations for image analysis validation. arXiv preprint arXiv:2206.01653.

Marhold, M., Kramer, G., Krainer, M., Le Magnen, C., 2022. The prostate cancer landscape in europe: Current challenges, future opportunities. Cancer Lett. 526, 304–310. http://dx.doi.org/10.1016/j.canlet.2021.11.033, URL: https://www.sciencedirect.com/science/article/pii/S0304383521006066.

McBee, M.P., Awan, O.A., Colucci, A.T., Ghobadi, C.W., Kadom, N., Kansagra, A.P., Tridandapani, S., Auffermann, W.F., 2018. Deep learning in radiology. Acad. Radiol. 25 (11), 1472–1480.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 565–571, URL: https://github.com/faustomilletari/VNet.

Myronenko, A., 2019. 3D MRI brain tumor segmentation using autoencoder regularization. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. pp. 311–320, URL: https://github.com/Project-MONAI/MONAI.

Nemoto, T., Futakami, N., Yagi, M., Kumabe, A., Takeda, A., Kunieda, E., Shigematsu, N., 2020. Efficacy evaluation of 2D, 3D U-net semantic segmentation and atlas-based segmentation of normal lungs excluding the trachea and main bronchi. J. Radiat. Res. 61 (2), 257–264.

Nikolov, S., Blackwell, S., Zverovitch, A., Mendes, R., Livne, M., De Fauw, J., Patel, Y., Meyer, C., Askham, H., Romera-Paredes, B., et al., 2018. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. arXiv preprint arXiv:1809.04430.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M.C.H., Heinrich, M.P., Misawa, K., Mori, K., McDonagh, S.G., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention U-net: Learning where to look for the pancreas. Med. Imaging Deep Learn. URL: https://github.com/ozan-oktay/Attention-Gated-Networks.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32.

Pham, T.-C., Luong, C.-M., Hoang, V.-D., Doucet, A., 2021. AI outperformed every dermatologist in dermoscopic melanoma diagnosis, using an optimized deep-CNN architecture with custom mini-batch logic and loss function. Sci. Rep. 11 (1), 17485. http://dx.doi.org/10.1038/s41598-021-96707-8.

Punn, N.S., Agarwal, S., 2022. Modality specific U-net variants for biomedical image segmentation: a survey. Artif. Intell. Rev. http://dx.doi.org/10.1007/s10462-022-10152-1.

Qin, X., 2019. Transfer learning with edge attention for prostate MRI segmentation. arXiv preprint arXiv:1912.09847.

Radici, L., Ferrario, S., Borca, V.C., Cante, D., Paolini, M., Piva, C., Baratto, L., Franco, P., La Porta, M.R., 2022. Implementation of a commercial deep learning-based auto segmentation software in radiotherapy: Evaluation of effectiveness and impact on workflow. Life 12 (12), http://dx.doi.org/10.3390/life12122088, URL: https://www.mdpi.com/2075-1729/12/12/2088.

Reinke, A., Eisenmann, M., Tizabi, M.D., Sudre, C.H., Radsch, T., Antonelli, M., Arbel, T., Bakas, S., Cardoso, M.J., Cheplygina, V., Farahani, K., Glocker, B., Heckmann-Notzel, D., Isensee, F., Jannin, P., Kahn, C.E., Kleesiek, J., Kurç, T.M., Kozubek, M., Landman, B.A., Litjens, G.J.S., Maier-Hein, K.H., Menze, B.H., Muller, H., Petersen, J., Reyes, M., Rieke, N., Stieltjes, B., Summers, R.M., Tsaftaris, S.A., van Ginneken, B., Kopp-Schneider, A., Jager, P.F., Maier-Hein, L., 2021. Common limitations of image processing metrics: A picture story. arXiv arXiv:2104.05642.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. In: Lecture Notes in Computer Science, pp. 234–241.

Saha, A., Hosseinzadeh, M., Huisman, H., 2021. End-to-end prostate cancer detection in bpMRI via 3D CNNs: Effects of attention mechanisms, clinical priori and decoupled false positive reduction. Med. Image Anal. 73, 102155. http://dx.doi.org/10.1016/j.media.2021.102155, URL: https://github.com/DIAGNijmegen/prostateMR_3D-CAD-csPCa.

Saha, A., Twilt, J.J., Bosma, J.S., van Ginneken, B., Yakar, D., Elschot, M., Veltman, J., Fütterer, J., de Rooij, M., Huisman, H., 2022. Artificial Intelligence and Radiologists at Prostate Cancer Detection in MRI: The PI-CAI Challenge (Study Protocol). http://dx.doi.org/10.5281/zenodo.6522364, URL: https://pi-cai.grand-challenge.org.

Santomartino, S.M., Yi, P.H., 2022. Systematic review of radiologist and medical student attitudes on the role and impact of AI in radiology. Acad. Radiol..

Santoro, M., Strolin, S., Paolani, G., Della Gala, G., Bartoloni, A., Giacometti, C., Ammendolia, I., Morganti, A.G., Strigari, L., 2022. Recent applications of artificial intelligence in radiotherapy: Where we are and beyond. Appl. Sci. 12 (7), http://dx.doi.org/10.3390/app12073223, URL: https://www.mdpi.com/2076-3417/12/7/3223.

Shahedi, M., Halicek, M., Guo, R., Zhang, G., Schuster, D.M., Fei, B., 2018. A semiautomatic segmentation method for prostate in CT images using local texture classification and statistical shape modeling. Med. Phys. 45 (6), 2527–2541. http://dx.doi.org/10.1002/mp.12898, _eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.12898.

Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B.H., Ronneberger, O., Summers, R.M., Bilic, P., Christ, P.F., Do, R.K.G., Gollub, M., Golia-Pernicka, J., Heckers, S., Jarnagin, W.R., McHugo, M., Napel, S., Vorontsov, E., Maier-Hein, L., Cardoso, M.J., 2019. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063. URL: http://medicaldecathlon.com.

Singh, S.P., Wang, L., Gupta, S., Goli, H., Padmanabhan, P., Gulyás, B., 2020. 3D deep learning on medical images: A review. Sensors 20 (18).

Skup, M., 2010. Longitudinal fMRI analysis: A review of methods. Stat Interface 3 (2), 235–252.

Spohn, S.K., Bettermann, A.S., Bamberg, F., Benndorf, M., Mix, M., Nicolay, N.H., Fechter, T., Holscher, T., Grosu, R., Chiti, A., Grosu, A.L., Zamboglou, C., 2021. Radiomics in prostate cancer imaging for a personalized treatment approach - current aspects of methodology and a systematic review on validated studies. Theranostics 11, 8027–8042. http://dx.doi.org/10.7150/thno.61207.

Steenbergen, P., Haustermans, K., Lerut, E., Oyen, R., De Wever, L., Van den Bergh, L., Kerkmeijer, L.G., Pameijer, F.A., Veldhuis, W.B., Pos, F.J., et al., 2015. Prostate tumor delineation using multiparametric magnetic resonance imaging: Inter-observer variability and pathology validation. Radiother. Oncol. 115 (2), 186–190.

Stojnic, R., Taylor, R., Kardas, M., Kerkez, V., Viaud, L., Saravia, E., Cucurull, G., 2021. The latest in machine learning - papers with code. https://paperswithcode.com. Accessed: 2021-02-06.

Sudre, C.H., Li, W., Vercauteren, T.K.M., Ourselin, S., Cardoso, M.J., 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. pp. 240–248.

Syed, K., Sleeman IV, W., Ivey, K., Hagan, M., Palta, J., Kapoor, R., Ghosh, P., 2020. Integrated natural language processing and machine learning models for standardizing radiotherapy structure names. In: Healthcare, Vol. 8. MDPI, p. 120.

Thompson, J., van Leeuwen, P., Moses, D., Shnier, R., Brenner, P., Delprado, W., Pulbrook, M., Böhm, M., Haynes, A., Hayen, A., Stricker, P., 2016. The diagnostic performance of multiparametric magnetic resonance imaging to detect significant prostate cancer. J. Urol. 195 (5), 1428–1435. http://dx.doi.org/10.1016/j.juro.2015.10.140, URL: https://www.sciencedirect.com/science/article/pii/S0022534715051630.

Vayena, E., Blasimme, A., Cohen, I.G., 2018. Machine learning in medicine: addressing ethical challenges. PLoS Med. 15 (11), e1002689.

Wong, J., Fong, A., McVicar, N., Smith, S., Giambattista, J., Wells, D., Kolbeck, C., Giambattista, J., Gondara, L., Alexander, A., 2020. Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. Radiother. Oncol. 144, 152–158, URL: https://limbus.ai.

Wong, J., Huang, V., Giambattista, J.A., Teke, T., Kolbeck, C., Giambattista, J., Atrchian, S., 2021. Training and validation of deep learning-based auto-segmentation models for lung stereotactic ablative radiotherapy using retrospective radiotherapy planning contours. Front. Oncol. 11, 626499.

Zamboglou, C., Spohn, S.K.B., Adebahr, S., Huber, M., Kirste, S., Sprave, T., Gratzke, C., Chen, R.C., Carl, E.G., Weber, W.A., Mix, M., Benndorf, M., Wiegel, T., Baltas, D., Jenkner, C., Grosu, A.L., 2021. PSMA-PET/MRI-Based focal dose escalation in patients with primary prostate cancer treated with stereotactic body radiation therapy (HypoFocal-SBRT): Study protocol of a randomized, multicentric phase III trial. Cancers 13 (22), http://dx.doi.org/10.3390/cancers13225795, URL: https://www.mdpi.com/2072-6694/13/22/5795.

Zhong, Y., Yang, Y., Fang, Y., Wang, J., Hu, W., 2021. A preliminary experience of implementing deep-learning based auto-segmentation in head and neck cancer: a study on real-world clinical cases. Front. Oncol. 11, 638197.

Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2019. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imaging 39 (6), 1856–1867.

Zhu, Q., Du, B., Yan, P., 2019. Boundary-weighted domain adaptive neural network for prostate MR image segmentation. IEEE Trans. Med. Imaging http://dx.doi.org/10.1109/TMI.2019.2935018, URL: https://github.com/ahukui/BOWDANet.