Kaunas University of Technology

Faculty of Mathematics and Natural Sciences

Deep Learning Based Multi-Organ Segmentation of Organs-at-risk

Master's Final Degree Project

Reda Venskauskaitė

Project author

Assoc. Prof. Dr. Jurgita Laurikaitienė

Supervisor

Kaunas, 2025

Kaunas University of Technology

Faculty of Mathematics and Natural Sciences

Deep Learning Based Multi-Organ Segmentation of Organs-at-risk

Master's Final Degree Project

Medical Physics (6213GX001)

Reda Venskauskaitė

Project author

Assoc. Prof. Dr. Jurgita Laurikaitienė

Supervisor

Assoc. Prof. Dr. Darius Virbukas

Reviewer

Kaunas, 2025

Kaunas University of Technology

Faculty of Mathematics and Natural Sciences

Reda Venskauskaitė

Deep Learning Based Multi-Organ Segmentation of Organs-at-risk

Declaration of Academic Integrity

I confirm the following:

1. I have prepared the final degree project independently and honestly without any violations of the copyrights or other rights of others, following the provisions of the Law on Copyrights and Related Rights of the Republic of Lithuania, the Regulations on the Management and Transfer of Intellectual Property of Kaunas University of Technology (hereinafter – University) and the ethical requirements stipulated by the Code of Academic Ethics of the University;

2. All the data and research results provided in the final degree project are correct and obtained legally; none of the parts of this project are plagiarised from any printed or electronic sources; all the quotations and references provided in the text of the final degree project are indicated in the list of references;

3. I have not paid anyone any monetary funds for the final degree project or the parts thereof unless required by the law;

4. I understand that in the case of any discovery of the fact of dishonesty or violation of any rights of others, the academic penalties will be imposed on me under the procedure applied at the University; I will be expelled from the University and my final degree project can be submitted to the Office of the Ombudsperson for Academic Ethics and Procedures in the examination of a possible violation of academic ethics.

Reda Venskauskaitė

Confirmed electronically

Venskauskaitė Reda. Deep Learning Based Multi-Organ Segmentation of Organs-at-risk. Master's Final Degree Project / supervisor Assoc. Prof. Dr. Jurgita Laurikaitienė; Faculty of Mathematics and Natural Sciences, Kaunas University of Technology.

Study field and area (study field group): Medical technologies (Health Sciences).

Keywords: deep learning, organs at risk, auto-segmentation.

Kaunas, 2025. 80 p.

Summary

In recent years, deep learning (DL) methods have been increasingly applied to improve the automation and accuracy of organ-at-risk (OAR) segmentation in radiotherapy planning. Among these, convolutional neural networks (CNNs), particularly U-Net and its extensions, have shown significant advantages over manual contouring by offering greater consistency, reduced inter-observer variability, and faster execution in medical image segmentation tasks.

This study systematically evaluated the performance of three 3D DL-based segmentation models: U-Net, Residual Encoder U-Net (ResEncU-Net), and SwinUNETR, on three multi-organ CT datasets: AMOS (Abdominal Multi-Organ Segmentation), BTCV (Beyond the Cranial Vault), KC (dataset provided by The Hospital of Lithuanian University of Health Sciences Kauno Klinikos), with the goal of assessing DL segmentation model suitability for clinical use. Quantitative evaluation based on Dice Similarity Coefficient (DSC), Surface DSC (sDSC), and 95th percentile Hausdorff Distance (HD95) revealed that ResEncU-Net delivered significantly higher segmentation accuracy across datasets, achieving a DSC of up to 0.916, sDSC values exceeding 0.88.

U-Net demonstrated strong baseline performance, particularly in more homogeneous datasets such as KC, where its DSC (0.913) was close to that of ResEncU-Net (0.916). However, its segmentation accuracy declined slightly in datasets with greater variability. SwinUNETR, despite its Transformer-based architecture, showed the weakest performance, with largest mean HD95 values (up to 45.95 mm) and inconsistent sDSC scores, especially for small or low-contrast structures. These findings contrast with previously published results that reported strong SwinUNETR performance in large-scale studies, suggesting that its effectiveness is highly dependent on pretraining and dataset size – factors rarely feasible in clinical settings.

In conclusion, the findings demonstrated that open-source segmentation models can be effectively integrated into clinical radiotherapy workflows. With expert validation by radiation oncologists, these models can significantly reduce manual contouring time, maintain high segmentation quality, and support more efficient and reproducible treatment planning. This highlights the practical potential of open-source DL-based tools for routine clinical use in radiation oncology, especially when it is tailored to institutional data and workflows.

Venskauskaitė Reda. Giliuoju mokymusi grįsta kritinių organų segmentacija. Magistro baigiamasis projektas. vadovė doc. dr. Jurgita Laurikaitienė; Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas.

Studijų kryptis ir sritis (studijų krypčių grupė): Medicinos technologijos (Sveikatos mokslai).

Reikšminiai žodžiai: gilusis mokymasis, kritiniai organai, autosegmentacija.

Kaunas, 2025. Puslapių sk. 80 p.

Santrauka

Pastaraisiais metais giliojo mokymosi (angl. *deep learning* – DL) metodai vis dažniau yra taikomi siekiant pagerinti kritinių organų kontūravimo automatizavimą ir tikslumą planuojant spindulinį gydymą. Tarp kurių konvoliuciniai neuroniniai tinklai (CNN), tokie kaip U-Net ir kiti susiję variantai, parodė didelius pranašumus, lyginant su kontūravimu rankiniu būdu, nes padeda užtikrinti mažesnį rezultatų išsiskyrimą tarp stebėtojų, gaunamų rezultatų nuoseklumą ir greitesnį medicininių vaizdų segmentavimo vykdymą.

Šiame tyrime buvo įvertintas trijų 3D DL pagrindu sukurtų segmentavimo modelių U-Net, Residual Encoder U-Net (ResEncU-Net) ir SwinUNETR, veikimas su trimis kompiuterinės tomografijos duomenų rinkiniais: AMOS (angl. *Abdominal Multi-Organ Segmentation*), BTCV (angl. *Beyond the Cranial Vault*), KC (duomenų rinkinį pateikė Lietuvos sveikatos mokslų universiteto ligoninė Kauno klinikos), siekiant įvertinti DL segmentavimo modelių tinkamumą klinikiniam naudojimui. Kiekybinis vertinimas, pagrįstas *Dice* panašumo koeficientu (DSC), paviršiaus DSC (sDSC) ir 95-ojo procentilio Hausdorfo atstumu (HD95) parodė, kad ResEncU-Net modelis pasižymi žymiai didesniu segmentavimo tikslumu visuose duomenų rinkiniuose, pasiekdamas iki 0,916 DSC, o sDSC vertės viršijo 0,88.

U-Net modelio segmentacijos parodė gerus rezultatus, ypač homogeniškesniuose duomenų rinkiniuose, tokiuose kaip KC, kur DSC (0,913) buvo artimas ResEncU-Net (0,916). Tačiau jo segmentavimo tikslumas gautas šiek tiek mažesnis duomenų rinkiniuose, kuriuose pastebima didesnė duomenų įvairovė. SwinUNETR, nepaisant transformatoriais pagrįstos architektūros, pasižymėjo mažiausiu našumu, turėdamas didžiausias vidutines HD95 vertes (iki 45,95 mm) ir mažesnius sDSC balus, ypač mažo tūrio arba mažo kontrasto struktūroms. Šie rezultatai prieštarauja kitų tyrimų rezultatams, kurie parodė puikų SwinUNETR našumą tyrimuose su didelio masto duomenų rinkiniais, o tai rodo, kad jo veiksmingumas priklauso nuo išankstinio modelio apmokymo ir duomenų rinkinio imties – veiksnių, kuriuos yra sudėtinga sukontroliuoti klinikinėje aplinkoje.

Apibendrinant, galima teigti, kad rezultatai parodė, jog atvirojo kodo semantinio segmentavimo modelius galima sėkmingai integruoti į spindulinės terapijos darbo aplinką. Gydytojui onkologui radioterapeutui įvertinant šių modelių segmentavimo rezultatų patikimumą, galima žymiai sumažinti rankinio kontūravimo laiką (nuo kelių valandų iki poros minučių), išlaikyti aukštą segmentavimo kokybę bei užtikrinti efektyvesnį gydymo planavimą.

Table of contents

List of figures

List of tables

List of abbreviations

Abbreviations:

3D-CRT – Three-Dimensional Conformal Radiation Therapy

AI – Artificial Intelligence

CT – Computed Tomography

DICOM – Digital Imaging and Communications in Medicine

DL – Deep Learning

DLAS – Deep Learning-based Auto-Segmentation

IMRT – Intensity-Modulated Radiation Therapy

ML – Machine Learning

MRI – Magnetic Resonance Imaging

NifTI – Neuroimaging Informatics Technology Initiative

OAR – Organ at Risk

TV – Target Volume

VMAT – Volumetric Modulated Arc Therapy

Introduction

Accurate delineation of organs at risk (OARs) in planning Computed Tomography (CT) images is a crucial step in radiation treatment planning [1]. Precise segmentation ensures that therapeutic radiation doses are delivered effectively to target volumes (TVs) while minimizing exposure to surrounding healthy organs or tissues, thereby reducing the risk of radiation induced toxicity. Traditionally, this process relies on manual contouring by clinicians, a process which is not only time-consuming but also susceptible to inter-observer and intra-observer variability, potentially impacting treatment consistency and patient treatment outcomes [2, 3].

In recent years, deep learning (DL) approaches have emerged as promising tools to automate and enhance the accuracy of OAR segmentation. Convolutional neural networks (CNNs), such as U-Net and its variants, have demonstrated improved efficiency and consistency over manual methods in medical image segmentation tasks [4]. Open-source frameworks like MONAI, nnU-Net, and foundational models (MedSAM variants) have further facilitated the development and deployment of DL models in clinical settings.

Despite these advancements, integrating open-sourced DL-based segmentation models into clinical workflows presents challenges such as hardware and software compatibility issues, need for extensive annotated datasets for training and concerns regarding the generalizability of models across diverse patient populations and imaging modalities [5]. Additionally, while commercial deep learning-based auto-segmentation (DLAS) systems have also shown significant time savings and improved consistency in contouring, their implementation can be hindered by challenges such as performance degradation when pretrained models are applied to diverse clinical scenarios without institution-specific fine-tuning [6, 7]. Moreover, the cost-effectiveness of implementing such systems, open-sourced or commercial, remains a consideration for many healthcare institutions. All these factors stress the necessity for the auto-segmentation algorithms to be thoroughly evaluated before their implementation in terms of accuracy and robustness as well as impact to the clinical workflow (for instance, manual editing time of auto-segmented structures).

The aim of this study is to train and systematically evaluate deep-learning-based segmentation models for multi-class segmentation of organs at risk, quantitatively comparing them to determine their effectiveness and clinical applicability in radiotherapy planning workflows.

Tasks:

1. To implement and train 3D image segmentation models on selected datasets for multi-organ segmentation.
2. To evaluate the models using standard segmentation metrics, providing analysis per organ, model and dataset to assess consistency and performance across anatomical and data variations.
3. To identify strengths and limitations of each network architecture, examining factors such as model complexity, training efficiency and segmentation accuracy.
4. To evaluate factors that could influence automated segmentation performance of the models in clinical settings.

1. Literature review

This section outlines the current landscape of cancer incidence in Lithuania and the critical role of radiotherapy planning process, with a focus on medical imaging and contouring practices. Review of the automated segmentation in radiation therapy is assessed, from traditional image processing methods to advanced artificial intelligence and deep learning approaches. Additionally, evaluation methods along with factors influencing segmentation performance are addressed. Finally, the ethical considerations associated with automated segmentation implementation in clinical settings are reviewed.

1.1. Cancer statistics in Lithuania

According to World Health Organisation (WHO) International Agency for Research on Cancer, in 2022, over sixteen thousand people were diagnosed with cancer in Lithuania, with prostate, breast, colorectum, lungs and stomach cancer being the most frequent types (Fig. 1) [8].

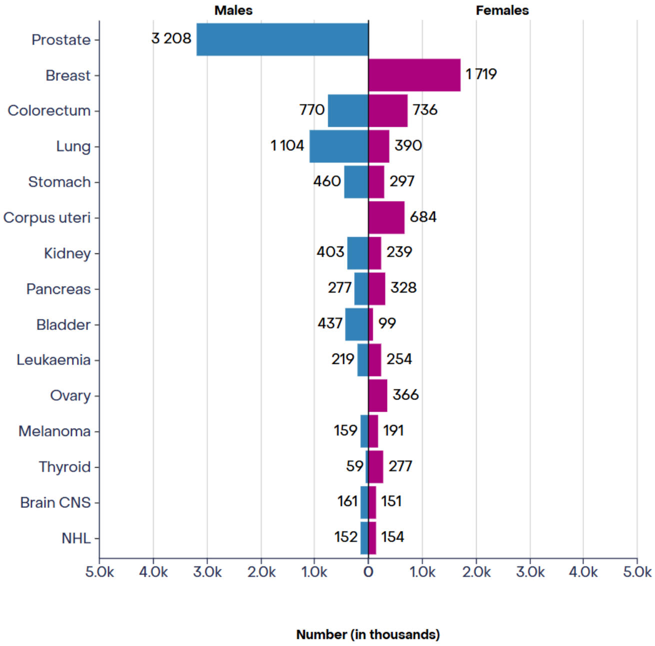


Fig. 1. Incidence of cancer in Lithuania, 2023 [8]

Cancer is one of the leading causes of death worldwide and the second most common cause of death in Lithuania after the cardiovascular diseases, accounting for 18.6 % of female deaths and 23.5 % of male deaths in 2023 [9]. According to R. Gaidelytė and M. Garbuvienė [10], incidence rate of cancer has been around 698 per one hundred thousand people in 2019, then significantly dropped in years 2020 and 2021, which could be attributed to reduced cancer screenings and delayed diagnoses during COVID-19 pandemic [11]. In the recent years, healthcare services gradually resumed their usual capacity and by 2023 the incidence had risen to pre-pandemic detection levels. As can be seen in Table 1, cancer mortality rates have remained stable, with minor fluctuations over the years.

Table 1. Incidence and mortality from malignant neoplasms per 100 000 population in Lithuania [10]

Year	**2019**	**2020**	**2021**	**2022**	**2023**
Incidence	697.9	555.9	566.9	600.8	658.1
Mortality	286.1	292.1	275.9	279.0	271.1

Radiotherapy is currently performed in five oncology centres across Lithuania – Vilnius University Hospital Santaros Clinics, National Cancer Institute, Hospital of Lithuanian University of Health Sciences Kauno Klinikos, Klaipėda University Hospital, Šiauliai Hospital, however, Vilnius University Hospital Santaros Clinics only offers brachytherapy [12].

1.2. Radiotherapy planning process

After the discovery of X-rays and radioactivity in the end of 19th century, there have been series of breakthroughs in the way cancer treatment is approached. Advances in computer technology have paved the way to transition from two-dimensional (2D) radiation treatment, for which treatment was prepared using a conventional X-ray simulator for setup and provided low conformity to the tumour, to three-dimensional Conformal Radiotherapy (3D-CRT), and Intensity Modulated Radiotherapy (IMRT), and Volumetric Modulated Arc Therapy (VMAT), which allows to refine the radiation volumes to patients’ anatomy and the target volumes [13]. Utilizing 2D radiation treatment, field arrangement, geometries and positions were set based on fluoroscopy and orthogonal radiographs during simulation [13]. Clinicians would shape the beams, typically square or rectangular-shaped, however, low conformity to the tumour resulted in inadequate amount of radiation and less effective treatment while delivering high doses to adjacent tissues and organs and increasing risk of side effects [14]. 3D conformal radiation treatment incorporated more advanced imaging modalities (Computed Tomography (CT) and Magnetic Resonance imaging (MRI)) and allowed for more accurate delineation of the tumour (size, shape, and location). Differently from 2D radiotherapy, in 3D-CRT, multiple radiation beams are shaped to match tumour’s 3D geometry employing multileaf collimators (MLC), but precision was still limited, especially for tumours with complex geometries. Furthermore, IMRT allowed for varying beam intensities as opposed to uniformly distributed dose delivered with 3D-CRT, thus ensuring better sparing of nearby organs while achieving high tumour control. VMAT took IMRT a step further by delivering radiation in a single continuous arc [15, 16].

Building on the advancements in radiation therapy techniques, the effectiveness of treatment heavily relies on a well-structured radiation treatment planning process (Fig. 2).

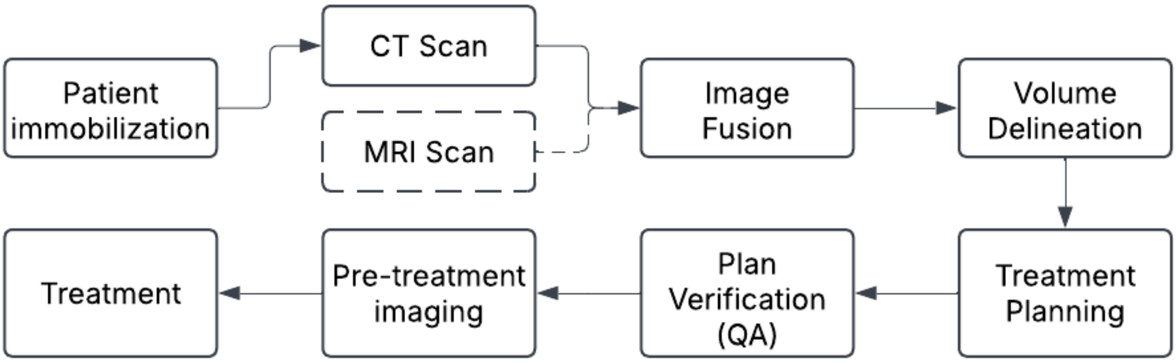


Fig. 2. Radiation treatment planning and delivery process [17]

After patient is diagnosed and it is decided to proceed with curative radiation therapy, initiation requires collaboration of various experts and could be constituted out of following steps: patient positioning and imaging, contouring of target volumes (TV) and organs-at-risk (OARs), treatment planning and evaluation, treatment delivery and follow up, while also performing quality assurance tests before, during and after treatment [17]. Altogether, this process is known as radiation treatment planning (RTP).

1.2.1. Medical imaging

Imaging is essential in RTP, since it enables precise tumour localization, accurate dose delivery and real-time treatment adaptation with minimal intervention. Various image acquisition techniques could be employed during RT, but CT has become the standard modality, providing data on electron density that is needed for dose calculation as well as 3D evaluation of the tumour and surrounding tissues [18]. During CT simulation, the machine rotates the X-ray tube around the immobilized patient while the table moves through a gantry, producing cross-sectional images that are reconstructed from measurements of X-ray attenuation coefficients. Pixels in the obtained images are displayed using Hounsfield units (HU), which are compared to known tissue density [19–21]. Water is set as 0 HU, air to -1000 HU at 0 degrees Celsius and a pressure of 100 kPa [21]. HU for the scanned tissues (Fig. 3) are calculated using equation

$$HU = 1000 \cdot \frac{(\mu_{tissue} - \mu_{water})}{\mu_{water}} \qquad (1)$$

where μ is the linear attenuation coefficient (cm^{-1}) [21].

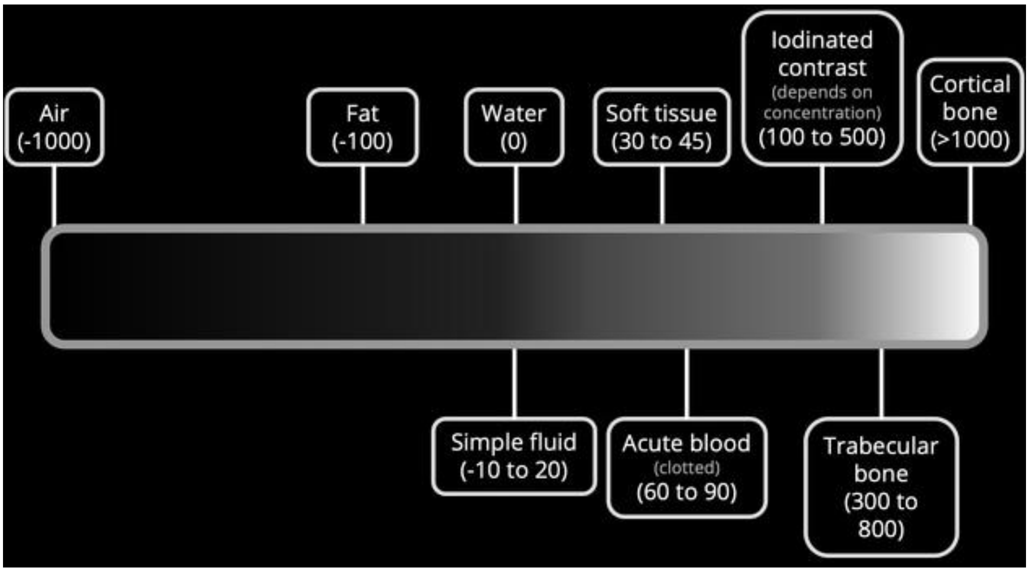
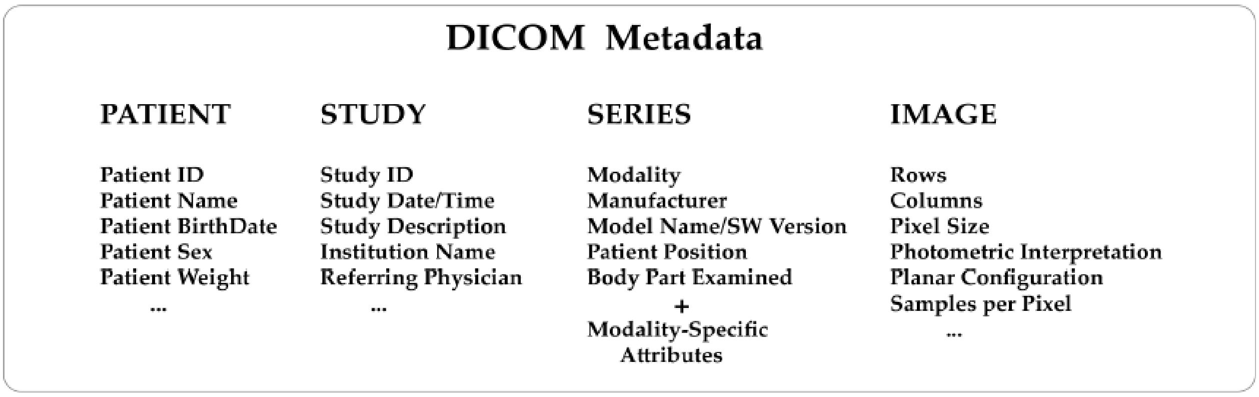

Air (-1000)
Fat (-100)
Water (0)
Soft tissue (30 to 45)
Iodinated contrast (depends on concentration) (100 to 500)
Cortical bone (>1000)
Simple fluid (-10 to 20)
Acute blood (clotted) (60 to 90)
Trabecular bone (300 to 800)

Fig. 3. Simplified HU scale [22]

As well as transverse slices, coronal, sagittal slice planes can be reconstructed from original data, which allows for better visualization of longitudinally oriented structures, such as musculoskeletal system. It also helps to evaluate whether the possible findings are surely present or could be attributed to image artifacts [21]. There are several CT scan types, categorized depending on the process of image acquisition and the technological advancements in construction of the scanner. Sequential (also known as step-and-shoot) involves incremental movement of the scanner table through the gantry – table moves after each slice is acquired; during the helical scan, patient is moved continuously through the gantry during X-ray source rotation and data acquisition [23]. Comparing the two scan modes, both provide high-quality images, however the preference between is dependable on the clinical scenario – for shorter scan lengths (≤ 16 cm) axial scanning is more practical, with shorter duration and lower doses, so it is preferred in routine head imaging, where images are not affected by motion respiration artifacts [24]. Conversely, helical scanning is more advantageous for longer scans and is associated with reduced scanning duration and consistent image quality due to continuous scanning. It is faster and reduces artifacts for patients who have difficulty holding still [25, 26].

CT is the standard imaging modality for radiation therapy planning in the current clinical practice due to fast acquisition times, wide availability, and electron density data for accurate dose calculation, although several challenges can impact its effectiveness [17]. Limited soft tissue contrast poses struggle to differentiate between tissues with similar densities. Additionally, metal implant caused artifacts and photon starvation further complicates accurate delineation of tumours from surrounding tissues [27]. Moreover, repeated CT scans for treatment planning and verification increases cumulative radiation doses. Dual-energy CT (DECT) addresses some of the key limitations of conventional CT, enhancing soft-tissue contrast with mono-energetic images (MEIs) at lower keV levels and reducing metal artifacts with high-energy MEIs, but remains of limited usage due to cost and technical constraints as well as occurrence of motion artifacts [27]. Limited soft tissue contrast issue is mitigated utilizing Magnetic Resonance Imaging (MRI) in radiotherapy planning, particularly for patients with head and neck (HN), brain, pancreatic, hepatobiliary malignancies, and prostate cancer, additionally offering dose reductions [17, 28]. Due to this reason CT images are sometimes fused with MRI or Positron Emission Tomography (PET) images, to obtain better soft tissue contrast for the delineation of the critical organs and target volumes [17, 29, 30]. Fusion of different modalities had proved to be useful in RTP for brain and HN cancers, cervical and prostatic cancers, central nervous system lymphoma [30–32]. Recently, there has also been research on synthetic computed tomography (sCT) generation from MRIs using deep learning (DL) algorithms, which could serve as a substitute for CT in patient positioning and radiotherapy planning and could be incorporated into MR-Linac workflows, saving time and solving image registration issues that emerge during CT-MRI fusion [33, 34]. However, issues like computation cost, lack of quality training data and proper quality assurance (QA) protocols might slow the process of its wider implementation in clinical practice [35].

After the acquisition of images, they are formatted and managed using DICOM (Digital Imaging and Communications in Medicine) standard. Along with the image data (layers of 2D slices), DICOM files also store metadata (Fig. 4), integrating patient information, imaging parameters and settings of acquisition within the files using tags. Tag is a code consisting of two numbers (for example, “0010,0010” code represents patient’s name) and relies on a DICOM dictionary – substantial number of standard defined variables, called public tags [36]. Additionally, manufacturers can define their private tags, encoding information that is specific to the manufacturer. Larobina [37] expressed that the purpose of such protocol is “to establish communication between diagnostic and sometimes therapy systems of different manufacturers and display, storage, and management devices on a network”. The DICOM standard allows sharing of images across hospital Picture Archiving and Communication Systems (PACS), supporting numerous services such as image storage, retrieval, printing for several imaging modalities – CT, MRI, ultrasound, mammograms, X-ray angiography [38]. Although DICOM is comprehensive and flexible, one of the cons of DICOM format is file size – a single scan could result in hundreds of separate DICOM files, each containing the image and metadata, raising data storage issues, and requiring higher computational power for image analysis. For example, a CT scan could contain hundreds of slices of 512 x 512 pixels and take up to a gigabyte of storage.



DICOM Metadata

PATIENT	STUDY	SERIES	IMAGE
Patient ID	Study ID	Modality	Rows
Patient Name	Study Date/Time	Manufacturer	Columns
Patient BirthDate	Study Description	Model Name/SW Version	Pixel Size
Patient Sex	Institution Name	Patient Position	Photometric Interpretation
Patient Weight	Referring Physician	Body Part Examined	Planar Configuration
...	...	+	Samples per Pixel
		Modality-Specific Attributes	...

Fig. 4. DICOM image metadata [37]

Another commonly used format in medical imaging is NIfTI (Neuroimaging Informatics Technology Initiative). Originally created to solve orientation problems in neuroimaging, it has been widely adopted by researchers of computer vision tasks, especially AI-driven solutions. Rather than multiple 2D slices as in DICOM, images are presented in a single file as a 3D volume, taking less storage space. Additionally, in contrast to DICOM, which is complex and verbose, NIfTI is a simple and minimalistic format, with less metadata, hence DICOM is a gold standard in clinical care, but NIfTI aids in acceleration of research for medical image analysis [36].

1.2.2. Contouring

Accurate delineation of target volumes (TVs) and organs-at-risk (OARs) is a crucial step in radiation therapy planning. This process requires a radiation oncologist to use a dedicated program to draw contours around the tumour and nearby organs adhering to the hospital or national guidelines [39]. Radiotherapy aims to deliver a precise dose of ionizing radiation to malignant tissues while sparing the healthy tissues, which can only be achieved if these structures are carefully delineated before planning the beams with treatment planning system (TPS) – dosimetric evaluation of the ROIs would be unavailable otherwise [40]. For precise treatment delivery and consistent reporting, definitions for the TVs and OARs have been provided and updated in ICRU (International Commission on Radiation Units and Measurements) Reports 50, 62, 71 and 83 [41–44].

Some of the volumes that are mentioned most often in clinical practice (Fig. 5) include:

- GTV (Gross Tumour Volume) – visible extent and location of the tumour.
- CTV (Clinical Target Volume) – demonstratable GTV and margins for microscopic spread of disease.
- PTV (Planning Target Volume) – volume expanded for internal motion of the organs (respiration or filling of the bladder) and geometrical uncertainties (patient positioning, beam alignment).
- TV (Treated Volume) – tissue that receives no less than 95 % of the prescribed dose.
- IV (Irradiated Volume) – tissue that receives a significant radiation dose (in relation to normal tissue tolerance).
- OARs (Organs-at-Risk) – healthy tissues and organs near the CTV, which can be damaged by radiation therapy if the tolerance level is exceeded [41–45].

Anatomically precise delineation of TVs and OARs gives way for accurate treatment delivery since different tissues have different tolerance to ionizing radiation. Including the voxels that represent healthy tissues in target volumes could cause side effects related to radiation toxicity due to unnecessary dose escalation; underestimating the TV would lead to geographic miss, and reduced tumour control probability (TCP) which increases the likelihood of recurrence [46]. Similarly, overestimated OARs volumes could enforce unwanted restrictions on tumour dose delivery; but not including the OARs voxels would increase risks of excessive radiation to these critical organs and underrepresentation of potential complications [47].

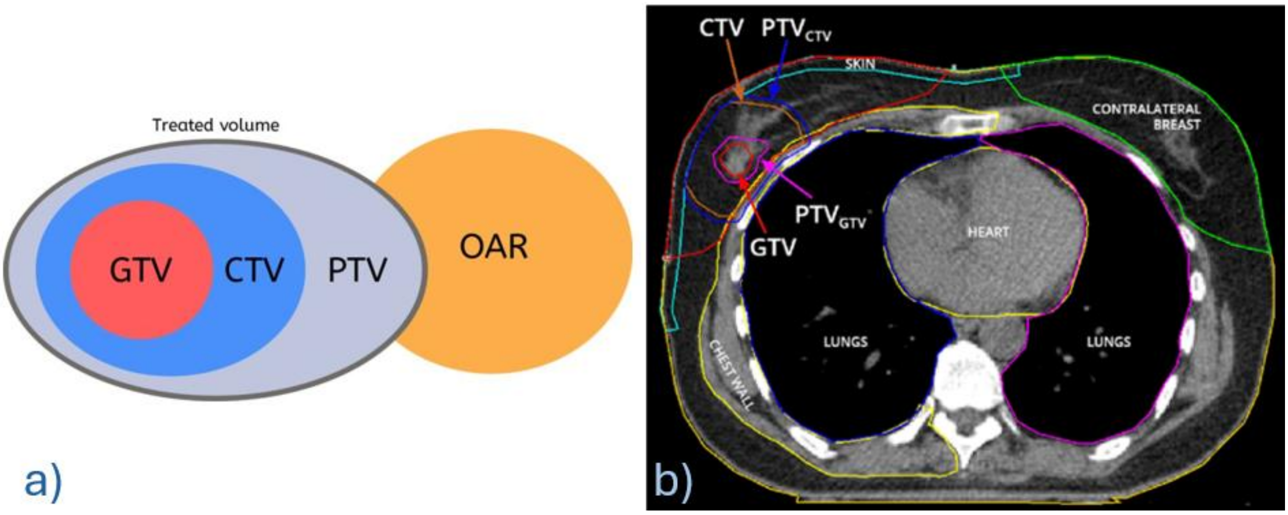


Fig. 5. a) Schematic representation of the target volumes; b) example of contoured structures on a planning CT scan [48, 49]

European Society for Radiotherapy and Oncology (ESTRO) has provided guidelines for contouring the TVs for several cases, some of which include the radiotherapy for early stage breast cancers [50, 51], glioblastomas [52], gastrointestinal malignancies [53], pancreatic [54], prostate [55, 56], cervical cancers [57], small and non-small cell lung cancers [58, 59]. Moreover, Global Harmonization Group has compiled a comprehensive reference list for delineation of 73 OARs by reviewing 157 radiation therapy clinical trials and incorporating expert feedback [60]. In most of the cases, radiation oncologist delineates the structures manually and they must be peer reviewed before the generation of treatment plans. However, despite the availability of guidelines, intra-observer (single specialist) and inter-observer (different practitioners) variations (IOVs) of contoured structures had been evident in research [47, 61–64]. Patrick et al. [63] evaluated IOV for delineations done for prostate cancer treatment plans, finding significant differences between contours. Authors successfully decreased contouring IOV with a reflective intervention in their institution, but study proves that IOV results in dose distribution differences between specialists' plans and could influence treatment outcomes, thus revealing a necessity of a systemic bias assessment in clinical practice [63]. Mercieca et al. [65] expressed that some of the ways to reduce IOV in target volume definition would include use of clearer protocols, multimodality images, rigorous training and correctly implemented peer review process, and automated segmentation. Some of these solutions were implemented by scientists in the University of Texas, who have proposed a real-time contouring feedback tool, making self-directed learning and immediate corrections available for the training ROs [66]. Study takes consensus-based training approach by using the STAPLE algorithm (Simultaneous Truth and Performance Level Estimation) which aids in generation of reference contours by combining multiple expert segmentations, eliminating the reliance on a single expert's contour to be a ground truth and ensures that it is robust and statistically meaningful [66].

Another issue that makes manual structure delineation a weak link in RTP is that it is usually a time-consuming process, depending on RO's experience and localization of malignancies. The additional process of peer review, although sometimes skipped, makes it even lengthier. Montague et al. [67] did a thorough study regarding the contouring times in United Kingdom (Fig. 6), showing that median delineation time for a single radiotherapy plan was 85 minutes, median time spent delineating GTV was 20 minutes, CTV – 30 minutes, 5 minutes for PTV and 20 minutes for the OARs. Median peer review time for all cases were similar (10-20 minutes), median contour approval time varied from 5 to 30 minutes [67]. Differences in median times were also found depending on the tumour location, treated volumes (time increases with inclusion of regional lymph nodes) and technology used (3D CRT usually takes shorter time than IMRT and VMAT planning) [67].

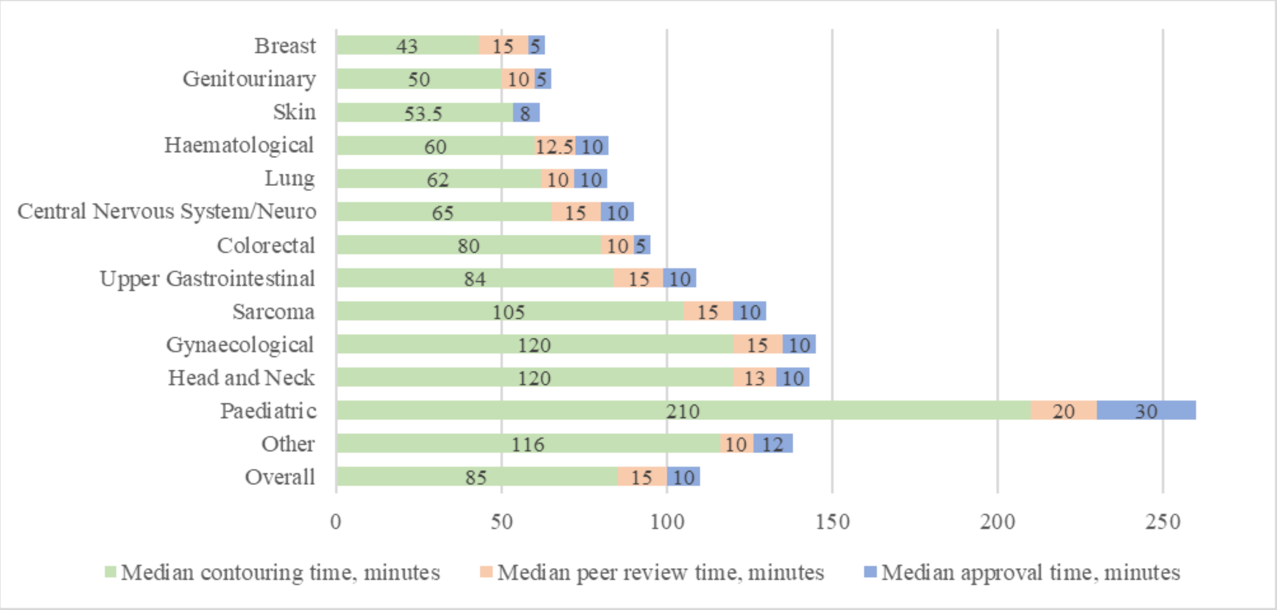


Fig. 6. Median duration of contouring, review and approval (per case) for different types of cancer [67]

Duration of contouring also vary if metal artifacts (MAs) are present in planning CT images, as shown in study by Katsura et al. [68]. Furthermore, researchers have found the significant statistical relationship between workload and variables like patients age, part of the day and the week, as well as the planning system used and ROs experience level (junior or senior) [69]. All these studies show that manual structure delineation is a lengthy process dependent on a range of factors, emphasizing the need for optimization strategies, since longer time spent on TV and OAR delineation increases treatment delivery waiting time for the patient, which influences treatment outcome. While automated segmentation algorithms, open-sourced or commercial, have shown promise in reducing the contouring time and IOV, the clinical adoption remains variable [70–72]. Many centres still generally rely on manual contouring, since human oversight is needed, there are regulatory considerations to comply with, and across different tumour sites, the results of artificial intelligence (AI) models may vary [73]. Nevertheless, these technologies are being gradually integrated into clinical workflows to support the role of ROs in radiation treatment planning.

1.3. Artificial Intelligence for the processing of medical images

Artificial intelligence (AI) has emerged as a promising tool in the field of medicine in recent years, aiding in treatment personalization, enhancing diagnostic accuracy, and streamlining clinical workflows through automation of tasks that are repetitive and time-consuming [74]. Modern clinical care generates vast amounts of data, including radiologic and pathologic images, electronic health records, treatment responses, which could not be processed with manual analysis alone. Machine learning (ML) and deep learning (DL) algorithms had revolutionized the way medical data is processed, by enabling automated pattern recognition, feature extraction and predictive modelling. ML and DL models, especially convolutional neural networks (CNNs), are applied growingly in key tasks related to medical imaging such as image classification, object detection and semantic segmentation (Fig. 7) [75]. Training of AI algorithms for these tasks additionally requires the researcher to pre-process the images accordingly. Not all frameworks support the same input image formats, so DICOM images may need to be converted to other lossless formats such as PNG or TIFF for 2D images and NIfTI for 3D. After that, depending on the task at hand, key steps could include background removal, denoising, resampling, registration, and intensity normalization of training images, which improve the quality and uniformity between images [76].

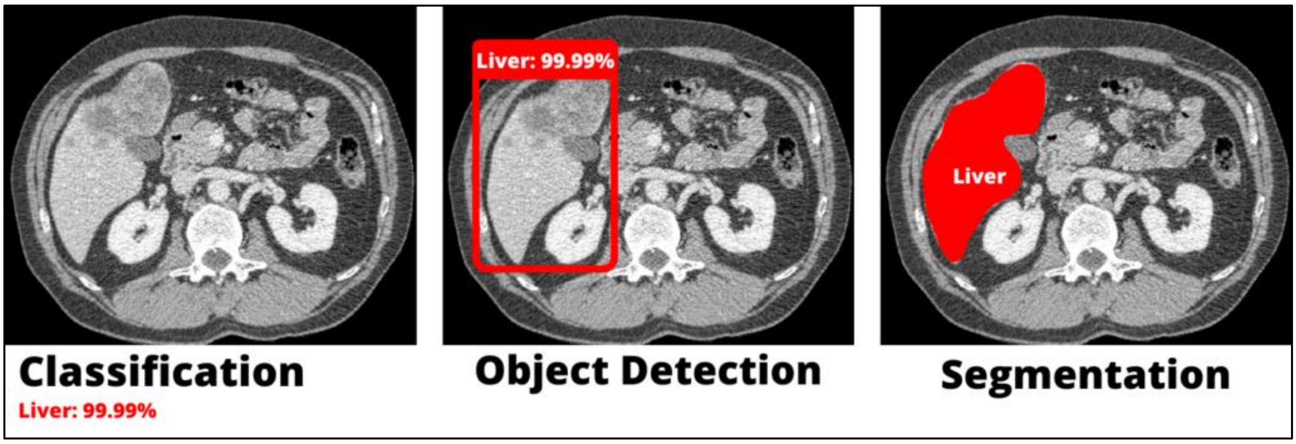


Fig. 7 Examples of computer vision tasks in medical imaging [77]

Classification of medical images refers to the process of assigning labels or categories to medical images based on their features, like modality, anatomical location or pathology, detection [70]. Traditional methods rely on shape, texture and colour features analysed manually, while CNNs can be used to extract the features from raw pixel or voxel data. Binary classification means categorizing images into two classes, like in Chaunzwa et al. [71] study, where researchers aimed to classify lung cancer subtypes (adenocarcinoma and squamous cell carcinoma) in standard CT scans using CNNs, aiming for a non-invasive method. Srinivasan et al. [72] have proposed hybrid classification model for the detection of brain tumours in magnetic resonance images (binary classification) with 99.53 % accuracy, categorization of images into five distinct types with accuracy of 93.81 % and classifying brain tumours into different grades (multi-class classification) with 98.56 % accuracy [72]. DL based classification models have been beneficial to Computer-Aided Diagnosis (CAD) systems, particularly in tasks of tumour type identification or recognizing malignant findings from benign, supporting radiologists with automatic analysis of medical images by assigning probability scores to suspected regions and offering diagnostic suggestions. Research by Ahmad et al. [73] and Hekal et al. [74] related to breast cancer diagnosis showed that integration of ML and DL-based classifiers into CAD significantly enhanced lesion detection accuracy and speed in clinical settings. Despite the made improvements of algorithms regarding their classification accuracy, one of the problems of AI-based classifiers limiting their implementation is their “black box” nature – the lack of explainability of how AI model made a particular decision [75]. This problem affects various AI-based algorithms, however, classifiers have a direct-role in diagnostic decision-making process, thus requiring clinicians trust, as understanding the reasoning behind the diagnosis is crucial in patient care. For this reason, explainable AI (XAI) models are being developed progressively, offering visual or textual explanations, highlighting most prominent features or regions contributing to the classification outcomes [75, 76].

Object detection in computer vision could be defined as a task of identifying and localizing instances of object within an image or a video – not only classifying the objects but also drawing bounding boxes around their location [78]. Most common frameworks like nnDetection, different versions of YOLO (You Only Look Once), or MedYOLO, had been developed for more efficient localization of pathologies, such as tumours, lung nodules, bone fractures in 2D and 3D medical images [79–82]. Billah et al. [82] compared YOLOv8 and YOLOv10 object detection algorithms for identification of kidney stones in CT images, showing that YOLOv10 improved accuracy, precision, and inference speed. Study highlighted the potential for deploying such frameworks in clinical workflows to reduce the time for interpretation and support earlier intervention for symptomatic patients. Advancements in real-time detection algorithms could serve as a promising solution in situations where immediate feedback is needed and can influence clinical decisions, like identification of fractures or haemorrhages in CT or X-ray images of trauma patients [83]. Moreover, it could be useful during interventional image-guided procedures, for example, in Smithmaitrie et al. [84], YOLOv7 DL framework was applied for anatomical landmark detection and generation of guiding dissection line during laparoscopic cholecystectomy with 95.71 % acceptance rate of supervising surgeons. Wijata et al. [85] proposed a CNN-based fully automatic method for detection of biopsy core needles in 2D ultrasound images, crucial in guiding biopsy procedures.

While object detection algorithms locate objects by draw bounding boxes around them, image segmentation - digital image processing technique that assigns labels to pixels (voxels) in the image - not only locates the object, but also its boundaries within the image. Segmentation process provides detailed outlines of structures, making it useful for further analysis of their shape, size, and spatial relationships. In medical imaging, it allows for quantitative analysis of pathological regions or anatomical landmarks, treatment planning and tracking the progression of disease over time [86]. Main types of segmentation could be categorized as semantic segmentation, instance segmentation, and panoptic segmentation, which combines both (Fig.8) [87].

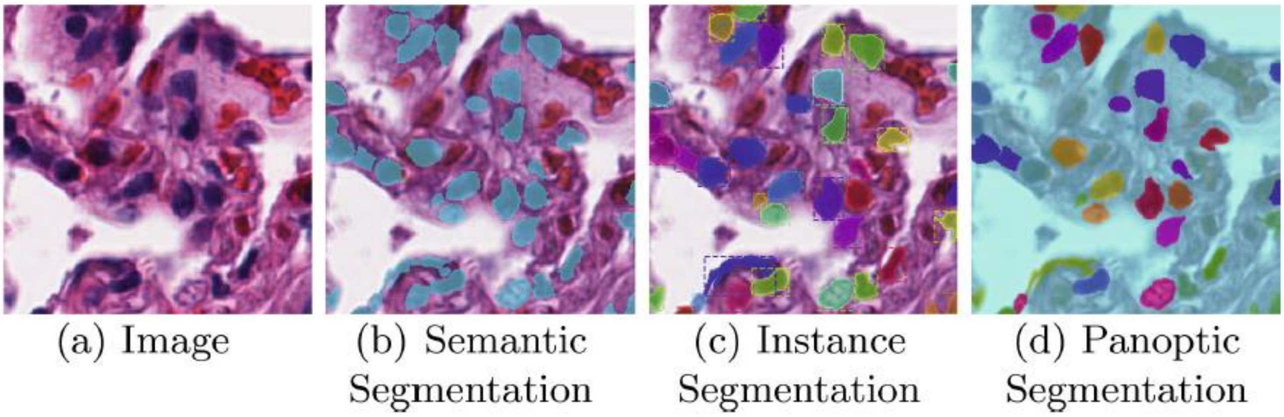

(a) Image
(b) Semantic Segmentation
(c) Instance Segmentation
(d) Panoptic Segmentation

Fig. 8. Different segmentation types [87]

An example of semantic segmentation could include tumour segmentation in MRI or CT scans, distinguishing the masks of tumours (foreground) from healthy tissue (background), segmentation of several OARs in a whole body CT would illustrate multi-class semantic segmentation; instance segmentation extends this task by registering individual instances of objects belonging to the same class, such as giving identification numbers for separate tumours for further analysis of their individual growth or shrinkage; while panoptic segmentation, though less frequent in medical image analysis, both produces the semantic label, and its instance ID, such as segmented organ mask and individually identified surrounding tumours [87].

After the segmentation, application of post-processing techniques can improve the results, as shown in the study of Furtado et al. [88]. Authors proposed that an organ should be represented as a single continuous volume, the largest connected label region, isolated or noisy pixels at the boundaries or further from the main region can be removed using erosion techniques, organs edges or surfaces should be smoothed, and any non-anatomical holes within the segmented regions should be filled. After applying the post-processing algorithms for liver segmentation task in CT and MRI scans, Intersection over Union (IoU) metric was increased by 4 % points in average over all patient sequences [88].

1.4. Traditional auto-segmentation approaches

Auto-segmentation techniques (Table 2) have been evolving along with increasing availability of computational power and memory throughout the years. In early stages, most techniques applied no or minimal prior knowledge during image processing, mostly judging by characteristics of pixels within the image [3]. With low-level medical image segmentation algorithms such as thresholding, edge detection, region growing, a lot of contextual factors such as typical shape and location of organs as well as their variability in a set of patient images would not be considered. These models were susceptible to issues related to image quality, presence of noise, intensity non-uniformity, imaging artifacts. Although relatively fast and easy to implement, these algorithms would only be useful in cases where organs in the images exhibit clear contours and are distinguishable from the background (such as bones and lungs in CT images) [89]. In later techniques, such as statistical model, atlas or machine learning-based approaches, due to the growth of computational resources, prior knowledge could be introduced, creating more robust solutions than low-level auto-segmentation techniques and broadening the set of OARs that could be contoured automatically [3].

Table 2. Techniques of traditional auto-segmentation [3, 90]

Segmentation method	**Description**	**Example algorithms**	**Strengths**	**Limitations**	**Use cases**
Thresholding	Grayscale images are converted to binary using set intensity threshold	Otsu’s method, iterative thresholding, Niblack, Sauvola, Bernsen	Simple and fast, computationally undemanding	Ineffective in cases where noise is abundant, and boundaries of ROIs are ambiguous	Bone, brain, lung segmentation in CT images
Edge-based	Boundaries of ROIs are detected based on intensity changes	Sobel, Laplacian of Gaussian, Prewitt	Effective for sharp boundaries	Noise-sensitive and dependent on image-quality	Mapping of blood vessels, organ boundary delineation

Table 2. Techniques of traditional auto-segmentation [3, 90] (continued)

Segmentation method	**Description**	**Example algorithms**	**Strengths**	**Limitations**	**Use cases**
Atlas-based	Previously segmented reference images are used to guide new segmentations	Single-Atlas and Multi-Atlas applications with deformable registration	Includes prior knowledge, can be effective with anatomical variability in the set of atlases, robust for structures with consistent morphology	Dependant on quality of deformable registration, and similarity between organ morphology of atlas and patient	HN, cardiac substructures, pelvic organ segmentation
Statistical model-based	Uses shape/appearance models to restrict segmentations to possible anatomical shapes	Principal Component Analysis (PCA)	Incorporates prior knowledge, prevents segmentation of impossible shapes	Less flexible, depends on the characteristics of training data	Pelvic organ (prostate, bladder) segmentation
Region-based	Pixels with similar intensities are grouped	Seeded region growing	Applicable for homogeneous ROIs, spatial continuity is considered	Sensitive to the selection of the seeds, issues with heterogeneous intensities	Analysis of tumour margin growth
Clustering-based	Divides pixels into clusters based on hierarchical or partitional similarity of regions	K-means, Fuzzy C-Means clustering	Unsupervised, adaptable to varying data	Sensitive to initialization, noise, shape assumptions	Segmentation of tumours, tissue classification, analysis of segmented cell clusters in histopathological images
Graph-based	Image is transformed into a graph and segmentation is done via graph partitioning	Minimum spanning tree, shortest path methods, Markov Random Fields	Global context is captured, robustness to the variation of noise and shape	Computationally intensive	Segmentation of brain MRI, retinal layer, complex anatomical structures

Traditional methods do not outperform the emerging deep-learning (DL) based techniques often but could be deemed ideal with acceptable accuracy for real-time applications or resource-constrained environments, where DL models would be impractical to use due to longer inference times when computing power is insufficient. Additionally, unlike with DL methods, traditional methods do not require large, labelled datasets for training, and could be useful in scenarios where data is scarce, for example with rare medical conditions. In Palazzo et al. [91] quantitative evaluation of different segmentation algorithms showed that while DL models consistently yielded better results in terms of segmentation accuracy and boundary delineation, some traditional methods still exhibit benefits in specific clinical scenarios, especially due to their faster inference times and reduced data requirements. Research by Marinov et al. [92] explored whether traditional auto-segmentation methods could outperform a modern, large-scale DL-based model. Authors compared thresholding, k-means clustering and shape-based interpolation techniques with a variant of DL-based foundation algorithm MedSAM across 11 imaging modalities, revealing that with several modalities (PET, OCT and microscopy images), classical methods exceeded the DL algorithm in terms of segmentation performance and computational efficiency [92]. Such findings could suggest that even though emerging DL models often become the state-of-the-art solutions for automated segmentation, researchers could benefit from comparing these new models with classical, simpler approaches to expose their limitations, consequently helping to increase their robustness and generalizability.

1.5. Deep Learning approaches

Deep Learning (DL) revolutionized segmentation of medical images, enabling models to learn complex, hierarchical features from complicated data structures and in turn overcome the image-related issues that were unresolvable applying the traditional segmentation methods, such as low accuracy when segmenting structures that exhibit ambiguous boundaries, and sensitivity to noise – which is unavoidable with most modalities of medical images.

In the last decade, a plethora of DL methods for image segmentation have emerged, one of the first being the proposal of Fully Convolutional Network (FCN) (Fig. 9) by Long et al. [93], where authors adapted existing classification networks to manage non-fixed sized images and output spatial segmentation maps instead of classification scores by replacing the fully-connected layers with fully-convolutional layers. This architectural shift enabled end-to-end learning for pixel-wise classification, where each pixel in the input image is assigned a label corresponding to a semantic category.

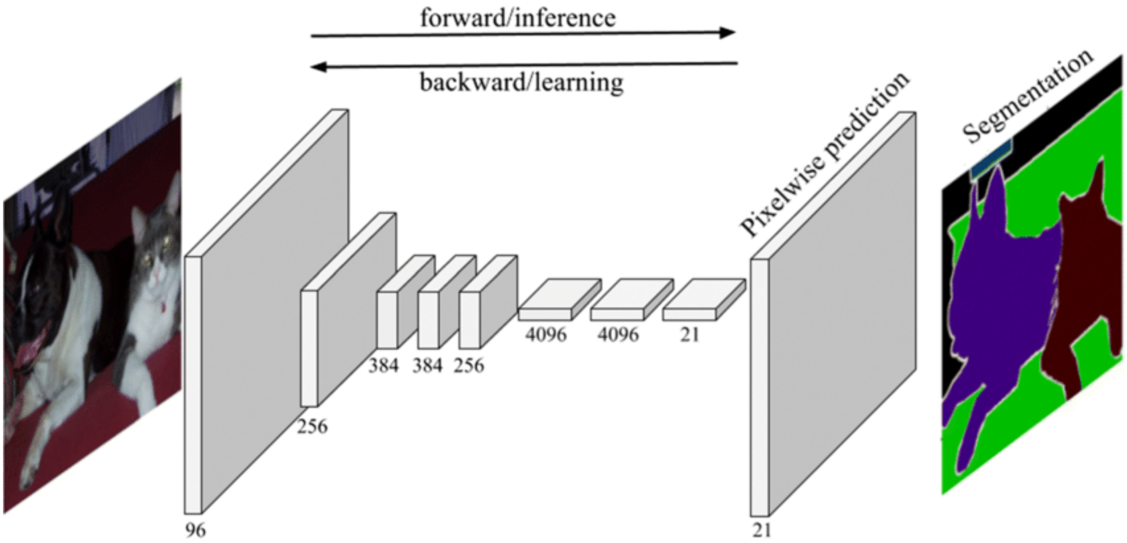


Fig. 9. Architecture of a Fully Convolutional Network (FCN) [94]

Although this established the groundwork for the subsequent DL networks, the conventional FCN model had limitations concerning the inference time, challenges with its applicability for 3D images and inability to consider the global context information efficiently [95]. Some of these limitations were addressed with U-Net, the work of Ronneberger et al. [96], who improved the FCN by employing the encoder-decoder structure with skip connections (Fig. 10).

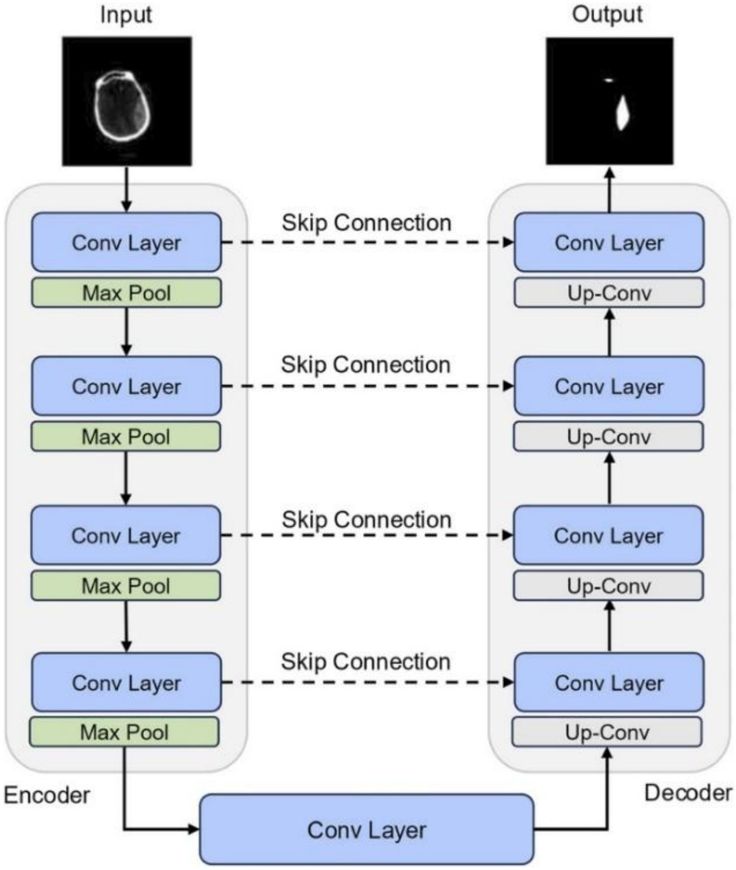


Fig. 10. Simplified architecture of U-Net [97]

Additionally, authors applied data augmentation for training images to improve performance, which proved to be indispensable in medical imaging and other fields where acquisition of labelled data is challenging [98]. Çiçek et al. [99] extended the U-Net for volumetric segmentation (3D U-Net) by replacing all 2D operations (convolutions, pooling, upsampling) with 3D operations – this allowed for the input images to be three-dimensional instead of requiring every slice of the image to be loaded separately, consequently leading to better segmentation quality of volumetric (for example, CT or MRI) images due to the spatial context being captured.

In Xia et al. [97], DL-based approaches to medical image segmentation were categorized into five groups – Convolutional Neural Network (CNN), Transformer, Mamba-based methods, semi-supervised and weakly-supervised learning methods. Aforementioned networks fall into the category of *CNNs* and are the backbone of subsequent CNN-based medical image segmentation networks developed since their emergence. Through extensions and changes in the network architecture, such as introduction of attention gates (Attention U-Net), dense connections (Dense-UNet), and multiscale strategies (UNet++) developers have been aiming to improve the segmentation accuracy without sacrificing the computational efficiency [99–101]. However, Isensee et al. [102] have argued that a lot of architectural modifications fail to enhance the performance of fully optimized networks and suggested that non-architectural aspects might be equally influential to the segmentation outcomes. Therefore, nnU-Net (“no-new-net”) framework has been proposed, optimizing several steps of image processing for segmentation tasks while adapting to the dataset at hand to train 2D and 3D U-Net based models [102]. The framework had been evaluated on a wide range of biomedical datasets since, scoring several first places in leaderboards of medical segmentation challenges, remaining competitive and providing basis for other researchers to develop their networks upon.

Segmentation methods based on *Transformers* (Vision Transformer - ViT, Swin Transformer) have emerged as an alternative to CNNs. Transformers have changed neural language processing and enabled development of large language models (LLMs) and lately have been successfully adapted to image processing tasks, including medical image segmentation [97, 103, 104]. However, researchers have underlined the computational cost and requirements for higher quantities of labelled data for most pure-transformer models to surpass the CNN-based, leading to the growing development of hybrid architectures that combine both approaches, such as integrating transformer blocks into encoder or decoder structures of CNNs, mainly with serial, parallel or skip connections [97, 105]. Prominent examples of such hybrid designs include SwinUNETR [106], which integrates Swin (Shifting-window) Transformer blocks into the encoder path of the U-Net like structure for 3D medical image segmentation, and nnFormer [107], which combines convolutional and transformer modules throughout the whole network architecture, outperforming several state of the art models on several public datasets [108].

Moreover, *Mamba* based networks, built upon State Space Models (SSM), have been reported to achieve competitive accuracy without significant difference from models like nnUNet, while using fewer parameters for training, however, taking significantly longer to train compared to CNN based models [109]. It is important to note, that Mamba-based segmentation networks are relatively new – first introduced in 2024 – and still require further research for their full implementation.

Altogether previously mentioned methods belong to the category of models that rely on supervised learning, meaning than large amounts of labelled data, such as manually contoured structures, is required to acquire accurate segmentation performance. While there are vast amounts of visual data collected in medical practice daily, annotated images are challenging to obtain, due to reasons such as data privacy regulations, diversity of necessitated delineations for different patient cases, as well as the time-consuming nature of the task. To address these challenges, researchers propose *semi-supervised* methods, utilizing both labelled and unlabelled data, and *weakly-supervised* methods using sparsely annotated datasets as well as employing visual prompts for image segmentation rather than depending on dense pixel-wise annotations. These prompts include text-prompts, image-level labels to provide categories of regions of interest (ROIs), bounding box labels to provide their localizations, point-level labels (foreground and background), scribbles or region labels to indicate the objects position (Fig. 11). Although less commonly used for medical segmentation, weakly supervised algorithms had been made of use to reduce annotation costs by generating pseudo labels from the prompts [110, 111].

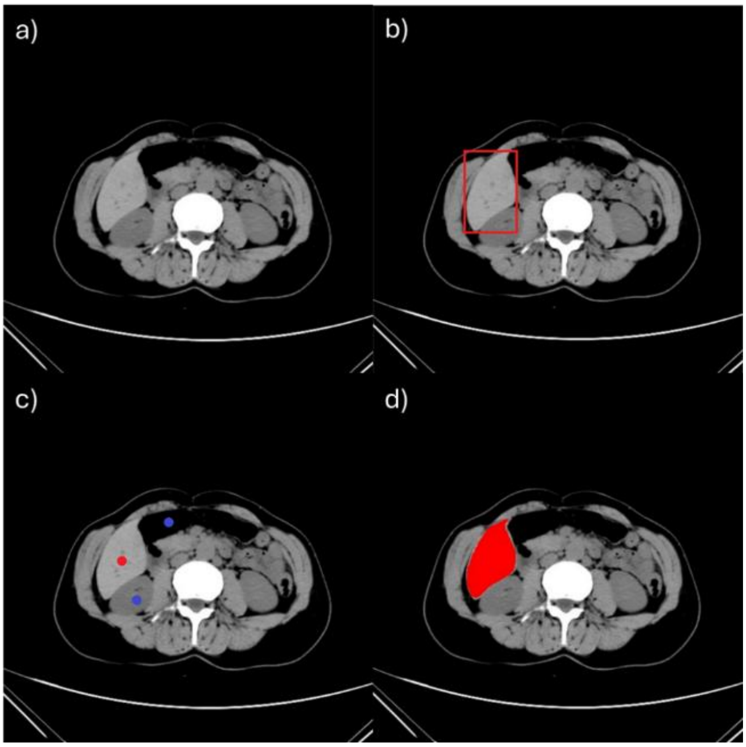


Fig. 11. Annotations for image segmentation. a) raw image; b) bounding box; c) point (red – foreground, blue – background); d) pixel-wise ground truth [112]

Moreover, building on recent advancements in image segmentation, foundation frameworks are being growingly applied for various biomedical image segmentation tasks. Foundation models refer to DL models, pretrained on diverse and massive-sized datasets, often with self-supervised or weakly supervised prompts. Thus, strong generalization across various domains and tasks can be offered. For example, Segment Anything Model (SAM) which was developed for general-purpose image segmentation by Meta AI researchers, was trained on eleven million images and over one billion masks [113]. However, research by He et al. [114] showed evidence that segmentation results of dataset-specific DL algorithms could not be surpassed by zero-shot application of SAM on twelve often benchmarked medical image datasets. Therefore, finetuning of the pretrained SAM model on a medical image segmentation dataset is necessary to improve zero- and one-shot segmentation results of the foundation model on unseen medical images. Aiming to expand generalizability of SAM and its recently improved variants for medical segmentation, several researchers have trained model on large-scale medical datasets and achieved state-of-the-art (SOTA) level segmentation performances on various anatomical structures, pathological conditions and medical imaging modalities [115–118]. Growing development and optimization of foundation models offers interactive auto-segmentation in radiotherapy workflows, when radiation oncologists provide real-time prompts (clicks or bounding boxes) to guide the segmentation of regions of interest - this approach offers the advantage of supervised refinement, improving accuracy through expert input. Nevertheless, while significant advances had been made in adapting prompt-based foundation models for medical image segmentation, several limitations persist regarding the implementation in radiotherapy planning process. For example, need for manual prompting increases the overall duration of OAR contouring, which might be less efficient compared to fully automatic segmentation methods; small organs, low contrast complex anatomical structures are often incorrectly merged with adjacent tissues with bounding box prompting [115]. In addition, integration into clinical workflows can be quite laborious and requires careful software engineering and computational resources, similarly to the integration of domain-specific deep learning models.

1.6. Evaluation of automated segmentation systems

Medical image segmentation models require thorough evaluation before their implementation in clinical settings, since the quality of automated segmentation model has direct influence on the workload of the clinician as well as the patient treatment outcomes. More importantly, accuracy variations of automated segmentation in radiation therapy results in clinically significant consequences for both treatment quality and patient safety, thus necessitating standardized methodologies for evaluation of model's applicability to segment unseen cases. Types of evaluation could be categorized into groups of quantitative, subjective and clinical assessment methods (Table 3), according to Yang et al. [119]. Quantitative methods are carried out calculating the level of similarity between the ground truth (manually contoured by an expert) and auto-segmented structure, subjective methods rely on a specialist's opinion regarding the acceptability of auto-segmented contour, and clinical assessment methods evaluate the auto-segmentations impact on the clinical workflow, usually regarding editing time and effort required from the clinician [119]. Application in radiation therapy planning would additionally benefit from the dosimetric evaluation – calculating the delivered dose to the regions of interest (ROI).

Table 3. Overview of automated segmentation assessment methods

Method	Associated Metrics	Benefits	Limitations
Quantitative	Dice Similarity Coefficient (DSC), Intersection over Union (IoU), Hausdorff Distance (HD), Surface DSC (sDSC), Added Path Length (APL)	Objective, reproducible, easy to compute, suitable for benchmarking	May not reflect subsequent editing effort of auto-segmented structures, some metrics do not represent clinical relevance [120]
Subjective	Likert scores, qualitative ratings	Captures expert judgement and clinical acceptability, shown to correlate with patient outcomes [121]	Requires expert involvement, prone to variability, time-consuming
Dosimetric	Dose-Volume histogram (DVH), Mean dose (Dmean)	Directly linked to treatment quality and safety [122]	Requires treatment planning system, time-consuming
Clinical Assessment	Contour quality classification (CQC) [123], editing time	Reflects impact on clinical workflow [122]	Dependent on institutional preferences and workflows, requires annotated edits

Sherer et al. [122] and Baroudi et al. [124] suggest that use of a single metric, such as Dice Similarity Coefficient (DSC) for contour-evaluation is insufficient to reflect on the performance and clinical acceptability of the algorithm and recommend combining multiple metrics and, if possible, methods of assessment. However, while strategies composed of multiple methods provide comprehensive estimation of the auto-segmentation systems applicability in clinical settings, most of non-quantitative methods require involvement of multidisciplinary team of experts (radiation oncologists, medical physicists) and additional training to conduct the evaluation, which is time and resource-intensive. Due to this reason, efforts have been directed to develop quantitative metrics – such as Surface DSC (sDSC) and Added Path Length (APL) – that correlate with clinically relevant factors, including the time required for manual contour correction [125]. While these metrics cannot fully replace the multi-method evaluation, they represent the clinical effectiveness of segmentation model better compared to the DSC metric. In turn, these metrics provide a scalable approach for benchmarking and comparing segmentation models in early phases of development, where implementation of non-quantitative evaluation methods is impractical due to limitations in time, resources or access to clinical infrastructure.

Furthermore, like training, evaluation of deep learning auto-segmentation algorithms for radiation therapy planning requires significant amounts of high-quality annotated data across diverse anatomical sites, modalities and patient populations [126]. Depending on whether the algorithm is intended for a specific segmentation task or a unified approach (as with foundation models), datasets used for evaluation must be appropriately tailored to reflect the targeted anatomical regions, imaging modalities, and diversity in patient cases to ensure reliable assessment of model performance and generalizability. However, the availability of comprehensive datasets remains limited, due to several reasons – institution-specific datasets may not have enough targeted cases, while cross-institutional or international sharing of the data is prevented by data privacy concerns; manual contouring for the ground truth labels as well as organization and anonymization of the medical images demands additional work from the personnel [127]. Shani et al. [128] describes it as a data-bottleneck problem, which emphasizes that the lack of accessible quality data is one of the primary barriers to advancing the AI-based automated segmentation. In response, several initiatives have been focused on developing publicly available benchmarking datasets, aimed to provide consistent, well-annotated resources for evaluating algorithm performance across different anatomical sites and imaging modalities. Most of these public datasets are provided during and after medical image segmentation challenges – including those hosted by Medical Image Computing and Computer Assisted Intervention (MICCAI) Society, Medical segmentation decathlon (MSD) [129] or are provided in databases by The Cancer Imaging Archive (TCIA) or Synapse. Commonly benchmarked datasets for multi-organ segmentation include:

- AMOS (Abdominal Multi-Organ Segmentation): a large-scale, clinically diverse dataset consisting of CT and MRI scans from multiple centres, vendors, disease types with voxel-wise annotations for 15 abdominal organs, making it valuable for evaluation of model robustness and ability to generalize [130];
- BTCV (Beyond the Cranial Vault): widely benchmarked abdominal CT dataset, with 13 annotated organs, suitable for evaluation of model’s segmentation performance when trained on limited data [131];
- CT-ORG: dataset with annotated structure sets consisting of 6 organ classes [132]. Includes large, easily segmented organs (lungs, liver) and smaller ones (bladder, kidneys)
- TotalSegmentator dataset: consisting of 1204 randomly sampled CT examinations, with 104 anatomic structures – although useful to assess the large-scale foundation models, includes structures that are rarely used in radiotherapy planning (small muscles, facial bones), which limits the utility for radiotherapy-specific tasks evaluation [133]

Isensee et al. [134] have assessed most broadly used benchmarking datasets for various segmentation tasks and stated that a dataset utilized for evaluation of the model should ensure statistical stability while capturing meaningful methodological differences. Research showed that AMOS dataset is one of the most suitable datasets for benchmarking 3D multi-organ segmentation models due to low statistical noise along with effective differentiation between methods.

1.7. Auto-segmentation in radiation therapy workflows

Automated segmentation, especially DL-based, has proved to be beneficial in RTP process, with several studies showing significant time-savings and reduction of inter-observer variability (IOV) when compared to manual contouring. Zabel et al. [135] evaluated different contouring approaches (manual, DL-based and atlas-based) of bladder and rectum contouring for RT regarding prostate cancer, showing that the workflow (including initial contour generation and radiation oncologist reviewing and editing) of DL based algorithm required 55 % less time than manual and 40 % than atlas-based delineation workflows. This research also showed that even though contour generation time were comparable between DL and atlas based algorithms (1.4 and 1.2 minutes, respectively), required editing times were longer and the geometric extent of these edits were consistently bigger for atlas-based workflows, while DL delineation reduced the time without negatively affecting the contour geometry, editing times or dose-volume metrics [135]. Similarly, study by Bordigoni et al. [136] compared Atlas, ML and DL based auto-segmentation tools to generate contours for pelvic radiotherapy, and for cervical cancer cases, DL tools reduced segmentation times from 30-45 minutes required by manual contouring to 0.7-1.1 minutes for AS, significantly lower than time required by atlas and ML-based tools (22 and 21 minutes, respectively). Manual correction times for DL generated auto-contours were 5-12 minutes compared to 30 minutes required by atlas and ML generated ones. Additionally, the segmentation accuracy, evaluated with Dice similarity coefficient (DSC) and Hausdorff distance (HD) was significantly better of DL tools [136]. In Kibudde et al. [137], time-saving impact of AI-based auto-segmentation tool in two low- and middle-income countries was assessed for prostate and Head and Neck (HN) cancer OAR delineation, reporting that auto-segmentation reduced the manual contouring time from around 60 to 2 minutes per case, resulting in annual savings of around 1000 hours. However, in this article, editing times of contours were not evaluated [137].

In terms of IOV, DL-based auto-segmentation improves the geometric contour agreement – in two-phase study by Choi et al. [138], variations between manual and auto-segmented OAR and TV structures were compared, with DSC scores improving from 0.69 in manual contours to 0.77 in auto-segmented and HD values lowering from 34.9 to 17.9 mm. Results of this research show lowered IOV, additionally evaluating clinical acceptability with qualitative measures, such as subjective evaluation of the DL generated contours, showing that OAR structures were mostly acceptable for 37.5 % of respondents or required minimal editing for 62.5 % [138]. However, contours of DL segmented target structures had lower ratings of acceptability – 40.5 % being unusable or needing major editing, raising doubts for DL implementation in target structure delineation [138]. Hoque et al. [6] evaluated clinical applicability of commercial AI-based auto-contouring software (research version of Limbus Contour, Limbus AI Inc.) by carrying out geometric and dosimetric comparisons between manual and AI assisted contouring of PTVs and OARs in HN and prostate cancer cases. Authors concluded that even though AI significantly speeds up the radiation therapy planning (RTP) process, results of DL-based algorithm is dependent on the structure sites and necessitates human oversight, suggesting a hybrid approach, during which clinicians would focus on areas that exhibit more variance or are prone to errors, while auto-segmentation is used to generate contours where AI consistently performs at acceptable level [6]. Additionally performed dosimetric analysis, based on the results of normalized plan quality metrics (nPQM), demonstrated no statistically significant differences between plans generated with manual contours and auto-segmented structures [6].

Since AI is changing the pace of image processing in medical fields, developers of commercial contouring systems have made effort to implement it in their software. In Doolan et al. [7] five commercial contouring systems were compared (Table 4). Although results of all these systems seem satisfactory, availability and connectivity with the existing clinical workflows, contouring software and TPS, should also be considered, along with the cost of implementation and the systems transparency to the end user.

Table 4. Comparison of commercial auto-contouring systems [7]

Auto-contouring system	**Mirada**	**MVision**	**Radformation**	**RayStation**	**TheraPanacea**
Number of contours offered	99	143	83	67	86
Evaluation results (DSC, a.u.)	0.82	0.88	0.86	0.87	0.88
Time savings reported, min.	39.8	43.6	36.6	43.2	45.2

In a similar study by Kim et al. [139], seven auto-contouring systems were investigated and additionally a method for clinics to evaluate the systems for quality assurance was proposed. Authors underlined that benchmarking these systems should be done using real patient data, calculation of quantitative metrics, such as DSC, HD, Surface DSC and statistical analysis should be performed, evaluating results based on the anatomical site, organ size and complexity. Moreover, findings revealed that no single system consistently outperformed other systems across the anatomical sites, and most small, complex or low-contrast anatomical structures were difficult for the systems to delineate, as can be judged by lower DSC and sDSC, and higher HD values [139]. Authors have also reported that systems trained with institution-specific data show better segmentation results [139]. It demonstrates that institutions should have their own datasets that correspond to their needs and adhere to institutional guidelines for the fine-tuning of auto-contouring algorithms before their implementation in clinical workflows.

1.8. Factors influencing segmentation outcomes

The accuracy and reliability of medical image segmentation are shaped by a combined influence of factors stemming from both the imaging data quality and the segmentation algorithms themselves [140]. Understanding of these influences is necessary to correctly interpret the model performance, guide methodological improvements and ensure robust application in clinical settings.

Many issues regarding deep learning (DL) segmentation performance arise from the large differences in datasets used for training and testing of the model and actual clinical applications [141]. To estimate basic model performance, difficult conditions (large artifacts, abnormal anatomies, rare cancer cases) are often not included the validation or testing datasets. However, models developed with datasets that are not inclusive of these irregularities or distortions, fail to segment regions of interest (ROIs) when images have missing structures due to their surgical removal or large artifacts (such as metal artifacts due to dental implants in head and neck CT images or hip implants in pelvic cancer cases) [141]. While metal artifacts are routinely managed in clinical practice by contouring the artifact and metal implant and manually adjusting the relative electron density values (assigning low density artefacts to water density, and implant volumes to a density appropriate for their materials), automated segmentation models face challenges in their handling [142]. Specialists recognize artifacts through clinical context and anatomical knowledge, allowing them to understand distortions that are not representative of the underlying anatomy, but auto-segmentation models rely solely on learned patterns from the training images. This makes them vulnerable to irregularities in clinical data if during training there was insufficient number of samples that represent a variety of cases commonly dealt with in-clinic – and even with their inclusion, most of the time these cases require manual editing after auto-segmentation [140]. Additionally, if the training data is inconsistent and present significant variations in the ground truth labels of the structures belonging to the same class or if certain biases exist in the manual annotation process (for example, some OARs are systematically under or over-contoured), model accuracy can be impacted. Sylolypavan et al. [143] describe four primary causes in manual annotation inconsistencies in clinical practice: 1) annotators may lack sufficient information to perform the labelling reliably (ambiguous guidelines, suboptimal data quality); 2) insufficient domain expertise, particularly dealing with more complex cases; 3) human error due fatigue, oversight, time constraints; 4) subjectivity regarding the task, leading to personal or interpretative bias. Related problems in manually delineated structures in training data can translate into the segmentation models and degrade the model reliability and clinical safety [144].

It is important to recognize that although high-quality training data is crucial for good performance of a deep learning-based auto-segmentation model, the design of neural networks architecture and optimization strategies applied to the training of these models are equally important. Advanced architectures, like U-Net variants, have become standard in the field of medical image processing due to their ability to capture both local and global features, and mechanisms like skip connections, attention gates or residual connections have shown to enhance feature learning and improve segmentation accuracy [145]. Nevertheless, architectural variations must align with specific requirements of the medical segmentation task – different imaging modalities (MRI, CT, PET) present unique characteristics including spatial resolution, contrast, noise patterns and acquisition artifacts that influence the way input data should be processed and what features should be extracted [146]. Furthermore, the nature of anatomical structures being segmented requires appropriate structure of the network – while contours of large and relatively uniform organs such as liver or lungs can be effectively predicted with standard convolutional architectures, smaller, more irregularly shaped structures such as blood vessels, tumours or nerves demand more advanced configurations [147, 148]. In this context, specialized models tailored to particular tasks or anatomical regions often outperform general-purpose architectures, as they can be fine-tuned to capture the unique spatial, morphological and contextual characteristics of specific targets [149]. However, it should be noted, that increased complexity of the architectural setup does not always lead to better segmentation performance – overly complex models may overfit the training data, and capture noise and irrelevant patterns rather than generalizable characteristics of the images, which would lead to poor performance on unseen datasets, particularly when training data is limited or imbalanced [150, 151]. Moreover, highly intricate and too deep architectures demand more computational resources and training time and can become difficult to optimize, prone to issues like vanishing gradients or unstable convergence, especially without appropriate regularization techniques [151]. Model optimization, referring to the adjustment of model’s hyperparameters to minimize error and improve accuracy, is an important part of deep learning model development – choice of loss function, optimizer, learning rates and their scheduling, batch sizes, number of training epochs influence if the model’s performance is improved or degraded [152]. Therefore, it is crucial to balance model complexity and optimization strategies with the size and quality of the training data – techniques such as dropout (disactivating random neurons during training), early stopping (halting the training process when the performance stops improving) and data augmentation (increasing number of training images by applying transformations) are commonly employed to mitigate overfitting and enhance model robustness and generalization [153]. In conclusion, an interplay of data and model design-based factors strongly influences the auto-segmentation results, all requiring attention when developing a system for automated contouring in clinical settings.

1.9. Ethical concerns in automated medical segmentation

Rapid adoption of artificial intelligence (AI) and deep learning (DL) algorithms has transformed clinical workflows, offering increased efficiency, standardized contouring in radiation oncology, where accurate segmentation is vital for further treatment planning. However, the integration of these technologies into clinical practice also introduces ethical considerations, ranging from concerns about data privacy and algorithmic bias to issues of accountability, transparency and other impacts of automation in healthcare.

First, training or fine-tuning and initial evaluation of DL models require large, diverse datasets, which often include sensitive patient information – in radiation therapy context, CT or MRI images acquired are in DICOM format, which stores patient specific metadata such as name, birth date, sex, acquisition date and medical institution name. Sharing and analysing this data, especially across institutions for AI training, poses privacy risks [154]. General Data Protection Regulation (GDPR) is designed to mitigate these risks by enforcing the protection of personal health information through pseudonymization, which replaces identifying metadata with artificial codes while preserving the analytical value of the data [154–156]. Anonymisation, when all identifiable elements (direct and indirect) are irreversibly removed or masked, is a stricter approach that makes re-identification very difficult, however, loss of certain contextual information can compromise relevance of the training data, and even without metadata, anatomical structures or presence of rare pathologies in the images can be matched to patient identities when cross-referenced with related datasets [157]. For example, in the case of head and neck imaging, facial reconstructions from these scans can be linked back to the individuals using facial recognition software, so to mitigate this, some publicly available datasets have facial de-identification (blurring, masking) algorithms applied to obscure or remove facial structures in the images while retaining relevant anatomical information [158]. Nevertheless, patient consent to use their private health information (medical records, scans) is crucial before any further processing, analysis and sharing of these images across institutions for AI model training.

Additionally, DL models often lack interpretability – which complicates efforts to audit or validate their performance, so transparent reporting of model training data, assumptions, performance metrics, limitations and failure cases is essential to establish trust in these systems and prevent biased outcomes that may affect underrepresented patient groups [159, 160]. Furthermore, automation bias poses another potential issue with DL-based segmentation implementation in clinical settings. As clinicians increasingly rely on AI generated outputs, there is a risk that they may place too much trust in these results, potentially overlooking segmentation errors or failing to identify anomalies. In radiation oncology domain, such overreliance could lead to suboptimal treatment planning results. This concern is reported by Bondi-Kelly et al. [161], who compared automation in aviation and healthcare and highlighted that automation bias can lead to unnecessary mistakes when clinicians follow AI suggestions without critical thinking, so proper training of the practitioners to work with automated systems is crucial, similarly to pilots undergoing extensive training in aircraft simulations to appropriately interact with automated systems. Moreover, authors stress that accountability must be clearly defined when AI tools are deployed in clinical settings [161]. In the event of an error, responsibility, whether it lies with the software developer, clinician or the institution, should be determined through regulatory framework, to properly address the occurred issues.

In conclusion, the reviewed literature suggests that while AI-based auto-segmentation systems, particularly those using deep learning, hold significant promise for improving efficiency and consistency in radiotherapy planning, their successful adoption requires careful consideration of technical, clinical, and ethical aspects.

2. Materials and methods

This section provides information about datasets used for deep learning-based image segmentation, highlighting their characteristics and use cases. Moreover, the process of artificial neural network training is explained, additionally describing the modular framework and architectures of evaluated networks. Further on, training settings along with the evaluation strategy are provided for reproducibility of the research.

2.1. Datasets

Three datasets were used for the training and evaluation of deep learning models to assess their performance in different settings:

1. Partial AMOS (Abdominal Multi-Organ Segmentation) dataset – made available in 2022 for MICCAI challenge that presented two tasks – multiple abdominal organ segmentation in CT images and cross-modality (CT and MRI) segmentation [130]. While the full dataset consists of 500 CT and 100 MRI scans, only 300 annotated CT scans are publicly available due to the rest being out-of-distribution and used for fair testing and ranking of novel segmentation algorithms. Annotated ground truth labels (15 organs) were automatically labelled by pretrained segmentation models, then refined by human annotators (5 junior radiologists and 3 senior specialists) [130]. It is one of the most benchmarked large-scale datasets, allowing to evaluate deep learning segmentation model's ability to generalize on a diverse image set of abdominal cancer cases
2. BTCV (Beyond the Cranial Vault) – dataset consisting of 30 patient abdominal CT scans obtained for routine clinical care. Data was acquired from CT scanners across the Vanderbilt University Medical Center, USA and was a part of MICCAI 2015 challenge, designed to push forward the development of algorithms for abdominal organ segmentation [131]. All scans were manually labelled (13 structures) by trained raters, and their accuracy was reviewed by a radiologist or radiation oncologist [162]. BTCV is a widely used dataset by researchers, allowing to compare the performance of their models trained on limited data.
3. Hospital of Lithuanian University of Health Sciences Kaunas Clinics (HLUHS KC) provided the third dataset (hereinafter referred to as KC), consisting of 19 anonymized DICOM sets of whole-body CT scans, captured with GE Lightspeed RT 16-slice scanner in helical scanning mode. Labels of 12 organs at risk (OARs) were delineated and reviewed by experienced radiation oncologists. Dataset can be utilized for testing the scalability of models beyond abdominal regions and allows to assess the performance for county specific clinical settings.

Together, these datasets provide a way to evaluate segmentation models under varying conditions of data size, image resolution, and anatomical coverage (Table 5). Datasets were split into training, validation and testing subsets patient-wise, to prevent data leakage and keep model performance evaluation unbiased. 5-fold cross-validation was performed for each model training, splitting the training dataset into five subsets, each fold is used as a validation set while model is trained on the rest. After training of all five folds, predictions are combined to generate ensemble prediction for every case in the validation set. Splits were made as follows:

- AMOS: 240 cases for training and validation, 60 for testing.
- BTCV: 25 for training and validation, 5 for testing.
- KC: 16 for training and validation, 3 for testing.

Table 5. Characteristics of the datasets

Dataset (Scope)	**Patient cases**	**CT slices**	**Average CT slices per case**	**Median voxel resolution**	**Labels**
AMOS (Abdominal)	300	41430	138	0.68×0.68×5.00 mm	Spleen, right kidney, left kidney, gall bladder, esophagus, liver, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, prostate/uterus
BTCV (Abdominal)	30	3779	126	0.76×0.76×3.00 mm	Spleen, right kidney, left kidney, gall bladder, esophagus, liver, stomach, aorta, inferior vena cava, portal and splenic veins, pancreas, right adrenal gland, left adrenal gland
KC (Whole-body)	19	9333	491	1.27×1.27×2.50 mm (identical throughout the dataset)	Bladder, bones, heart, left kidney, right kidney, left lung, right lung, thyroid, stomach, liver, spleen, esophagus.

2.2. Artificial neural network

Artificial neural network (ANN) (Fig. 12a) is a computational system consisting of layers of interconnected units called artificial neurons (Fig. 12b) which process the input data to recognize patterns and solve complex problems through adaptive learning.

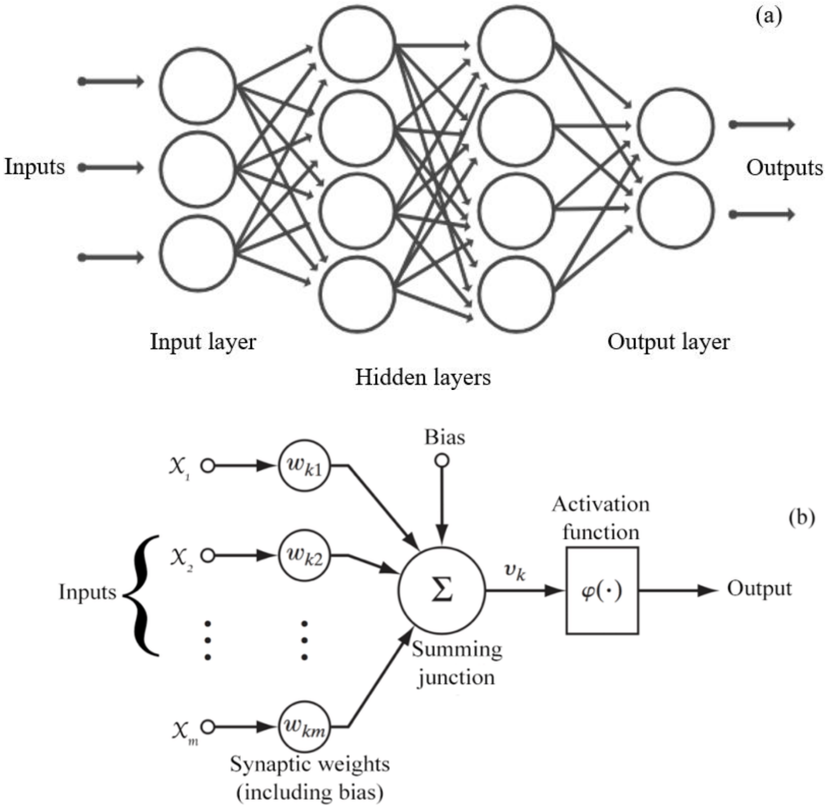


Fig. 12. (a) artificial neural network; (b) single artificial neuron (perceptron) [163, 164]

Neuron in ANN receives one or more input values – data features (for example, pixel intensities in an image), and multiplies it by a weight that determines the relative importance of the input to neurons output. Bias value b is added to the weighted sum of these inputs (Formula 2), allowing the neuron to shift the activation function (ReLU, Leaky ReLU).

$$z = \sum_{i=1}^{n} w_i x_i + b \tag{2}$$

Where x_i – input feature, w_i – weight associated with the input, b is the bias term.

Then, an activation function (Formulas 3 and 4) transforms the summed value into a non-linear output, which enables the network to model complex relationships that linear functions cannot capture. The result is passed to the next layer (in hidden layers) or is used as the networks prediction in the output layer (class probabilities in classification, pixel-wise masks in segmentation).

Rectified Linear Unit (ReLU) is an activation function that receives the input and allows only positive values to pass through unchanged, while setting negative values to zero.

$$ReLU = \max(0, z) \tag{3}$$

Based on the ReLU, Leaky ReLU allows for a small non-zero gradient to pass through, keeping neurons in the ANN from becoming permanently inactive, which happens when only positive values are allowed to pass through the activation function:

$$Leaky\ ReLU = \begin{cases} z & if\ z > 0 \\ \alpha z & if\ z \leq 0 \end{cases} \tag{4}$$

where α is a small constant, typically set to 0.01.

The network prediction is then compared to the ground truth using a loss function that acts as an error measure. Loss function calculated during the model training in this research consists of a sum of Dice loss and cross entropy (CE) loss, where Dice loss is a negative average Dice score across all classes [165]:

$$\mathcal{L}_{dc} = -\frac{2}{|K|} \sum_{k \in K} \frac{\sum_{i \in I} u_i^k v_i^k}{\sum_{i \in I} u_i^k + \sum_{i \in I} v_i^k} \tag{5}$$

where u - softmax output of the network, v - one-hot encoding of the ground truth segmentation map. Shape of u and v is $[I \times K]$, where $i \in I$ corresponds to the pixels in the training batch and $k \in K$ represents the segmentation classes [102]. Values of $\mathcal{L}_{dc}$ range from -1 (all classes perfectly overlap) to 0 (no overlap is observed).

Cross entropy (CE) *loss* measures classification error for every voxel in the segmentation map and is calculated via [166]:

$$\mathcal{L}_{CE} = -\frac{1}{|I|} \sum_{i \in I} \sum_{k \in K} v_k^i \log(u_k^i) \tag{6}$$

Perfect voxel-wise classification would produce $\mathcal{L}_{CE}$ equal to 0.

Finally (5) and (6) are added with equal weighing:

$$\mathcal{L} = \mathcal{L}_{dc} + \mathcal{L}_{CE} \tag{7}$$

Through repeated data exposure, the network learns to adjust the weights via loss optimization and backpropagation, enabling it to generalize to new, unseen data. Stochastic Gradient Descent (SGD) optimizer is popularly used in CNNs for medical image segmentation, which works by updating weights in the network based on partial derivatives of the loss function with respect to each weight [167]. SGD uses a subset of training data to compute the gradients instead of the whole training set, which is more efficient for processing of large medical datasets.

2.3. nnU-Net framework

Data preprocessing and model training was done utilizing open-source nnU-Net [102] framework, integrating different model architectures and datasets to have an equal ground for their evaluation. The modular nature of the framework allows user to implement different training algorithms and network architectures, additionally providing tools to preprocess the data used for training and validation while taking care of computational constraints. Dataset fingerprint (spacings of the voxels and overall shape of the images, intensity distribution and medical image modality) is extracted, allowing to further optimize the network training strategy, setting parameters such as patch and batch size, intensity normalization and the network topology to the computational power availability (Fig. 13).

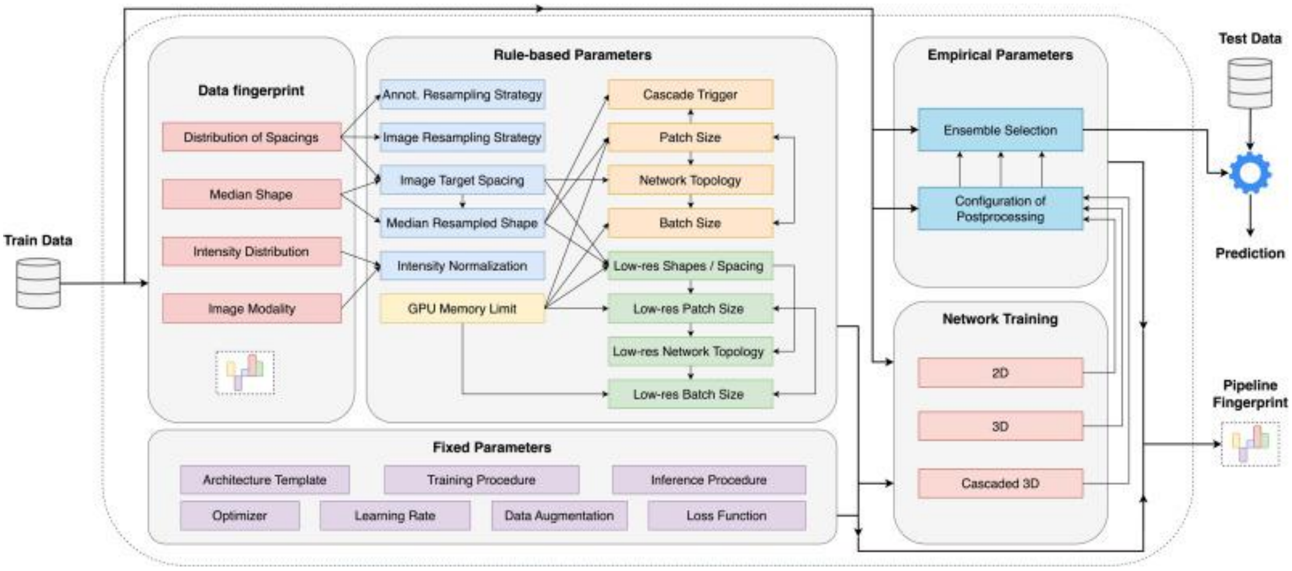


Fig. 13. nnU-Net configuration strategy [108]

Preprocessing was performed as follows:

1. DICOM images were converted into a NIfTI format.
2. Images in all datasets were resampled to consistent voxel spacings.
3. Z-score normalization was applied to the CT images
4. Images were cropped to the network patch size.

During normalization step, for each voxel intensity, the normalized pixel value is calculated as:

$$x' = \frac{x - \mu}{\sigma} \tag{8}$$

Where μ is the mean and σ is the standard deviation of the foreground intensities in the training cases.

nnU-Net pipeline also allows for k-fold cross-validation strategy, when training data is categorized into folds, each leaving out a different part of dataset for validation. This allows for the training of the model to be unbiased by dataset splitting choices.

2.4. Network architectures

Three network architectures were evaluated: 3D U-Net, 3D U-Net with Residual Encoder (ResEncU-Net) and a Transformer based Swin UNETR.

2.4.1. 3D U-Net

3D U-Net follows the same encoder-decoder structure as the original U-Net (Fig. 14), however, all operations (convolutional blocks, pooling, upsampling) are extended in three-dimensions to enable the data to be processed volumetrically. Each layer in the decoder consists of two convolutional blocks (with 3x3x3 kernels moved by 2x2x2 stride), followed by 3D batch normalization and *Leaky ReLU* activations [168]. Each step in the decoder doubles the feature channels allowing network to learn increasingly complex features. When the bottleneck layer is reached, two 3D convolutional layers capture the abstract features. Decoder stage mirrors the encoder, using 3D up-convolutions (transposed convolutions), upsampling the feature maps, halving the number of channels and concatenating the features with those sourced from corresponding encoder block via skip connections. Finally, the output layer uses 1x1x1 kernel for the convolution operation which maps features to the desired number of output classes, producing a volumetric segmentation map [168].

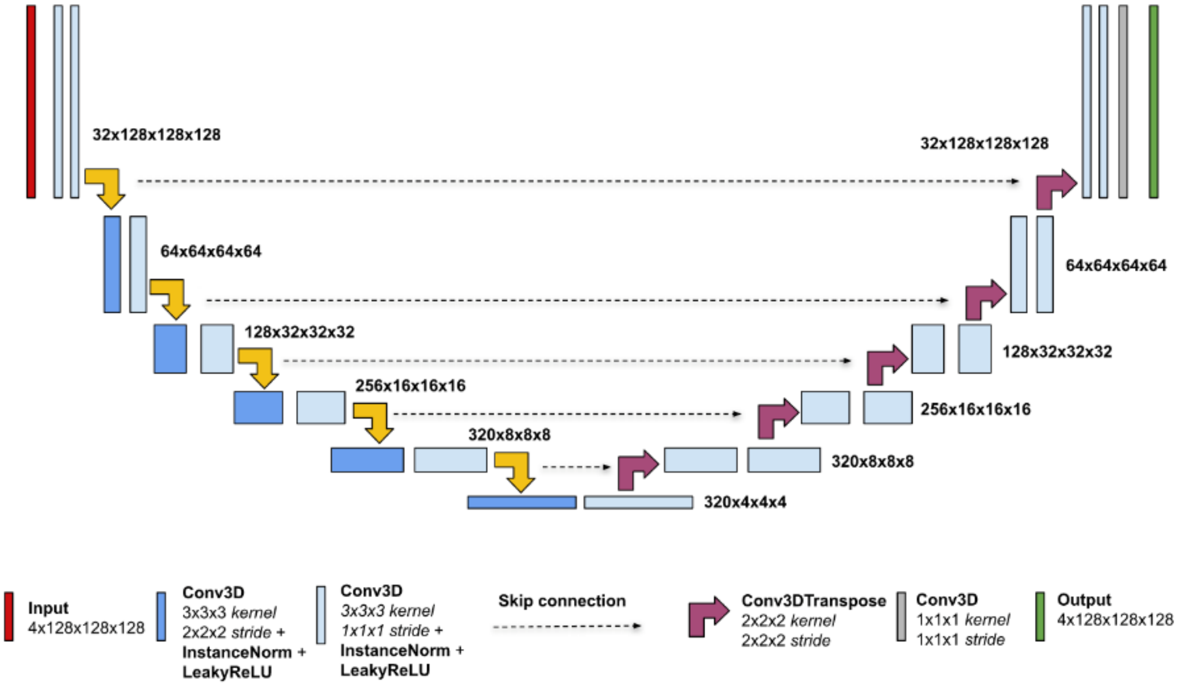


Fig. 14. 3D U-Net architecture [168]

On all datasets, 6-stage 3D U-Net is utilized with consistent feature progression from 32 to 320 channels across the encoding and decoding paths. Each stage includes two convolutional layers, instance normalization is applied throughout the network, along with Leaky ReLU as the activation functions. The models are trained with a batch size of 2. The configuration differences across the datasets are presented in Table 6.

Table 6. 3D U-Net configuration differences

Dataset	Patch size in voxels (depth x height x width)	Kernel sizes	Strides
AMOS	64x160x160	1x3x3 (input layer), 3x3x3 in the rest hidden layers	1x2x2 (input layer), 2x2x2 in the rest hidden layers
BTCV	48x192x192	1x3x3 (input layer), 3x3x3 in the rest hidden layers	1x2x2 (input layer) 2x2x2 in the rest hidden layers
KC	128x128x128	3x3x3 (all stages)	2x2x2 (all stages)

AMOS and BTCV have anisotropic voxel spacing, with much thicker slices along the depth axis compared to KC. Therefore, patch sizes differ, and the first layer of the model uses 1×3×3 convolutions with 1×2×2 strides to preserve detail in the low-resolution depth direction.

2.4.2. Residual Encoder UNet

The Residual Encoder U-Net (ResEncU-Net) is a variant of the U-Net architecture in which the convolutional blocks within the encoder path are replaced by series of residual blocks (Fig. 15), as introduced by He et al. [169] in ResNet architecture, while the decoder path consists of standard convolutional blocks with normalization and activation function without residual connections. This modification was proposed to address the vanishing gradient problem encountered in deep neural networks, introducing the identity skip connections. Residual connections allow the gradient to flow more directly through the network during backpropagation, which is offers more effective training of DL models and solves the "vanishing gradient" problem often encountered in convolutional neural network training [169–171].

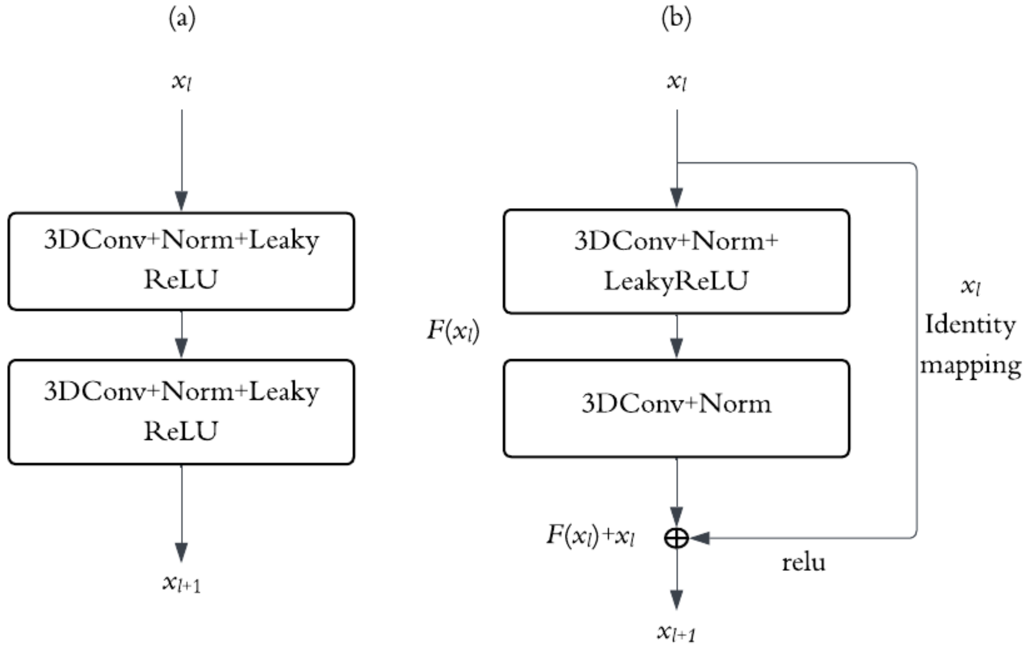
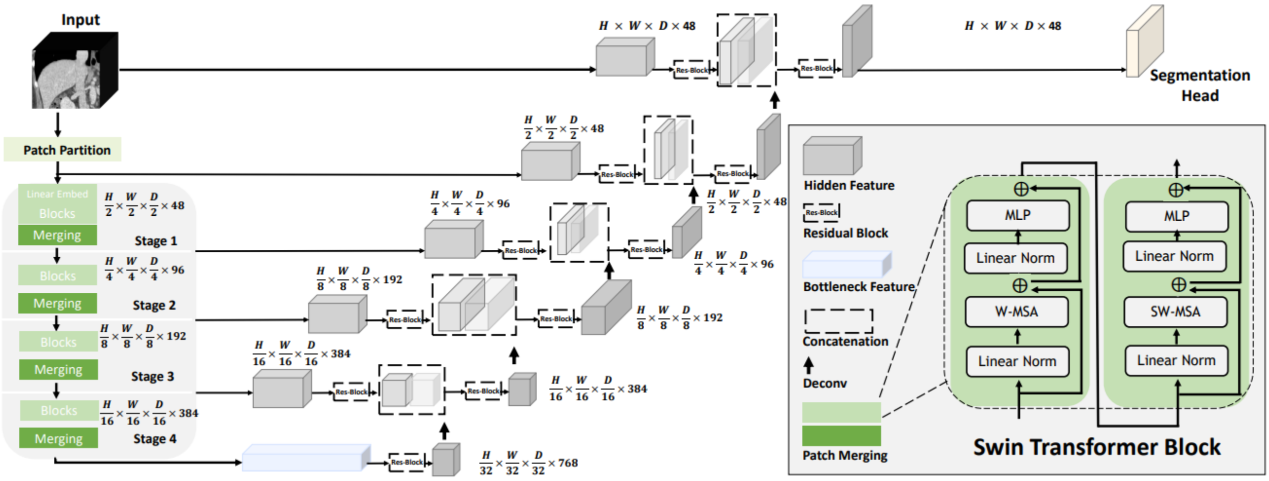


Fig. 15. (a) Convolutional block in U-Net; (b) Residual block in ResEncU-Net [172]

Regarding this research, all ResEncU-Net models trained on different datasets consist of 6 stages, and as in the 3D U-Net, features per stage range from 32 to 320, patch and kernel sizes along with strides remain as were provided in Table 6. In the encoder path, number of residual blocks are as follows for each stage: 1 block in the first stage, 3 in the second, 4 in the third and 6 in the rest of the encoder stages. This ensures that in the deeper stages – where spatial size of the features decrease, but complexity increases – network has enhanced capacity to model the contextual dependencies while maintaining efficient gradient propagation.

2.4.3. SwinUNETR (Transformer based)

SwinUNETR is a hybrid encoder-decoder segmentation model that integrates the Swin Transformer with the spatial decoding strategy of U-Net architecture (Fig. 16). Proposed by Hatamizadeh et al. [106], it is aimed to adress the limitations of traditional CNNs where long-range dependencies in volumetric data are not sufficiently captured, especially in 3D medical image segmentation tasks.



Input
Patch Partition
Linear Embed
Blocks
Merging
Stage 1
Stage 2
Stage 3
Stage 4
$H \times W \times D \times 48$
$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 48$
$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 96$
$\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 192$
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 384$
$\frac{H}{32} \times \frac{W}{32} \times \frac{D}{32} \times 768$
Res-Block
Segmentation Head
Hidden Feature
Residual Block
Bottleneck Feature
Concatenation
Deconv
Patch Merging
MLP
Linear Norm
W-MSA
SW-MSA
Swin Transformer Block

Fig. 16. SwinUNETR architecture [106]

Implemented in nnU-Net pipeline, SwinUNETR operates on the pre-processed 3D image patches (that were provided in Table 6). These patches are first processed by patch partition module (Fig. 16), that divides the volumetric input into smaller non-overlapping patches (2x2x2 voxels) and flattens them, forming a sequence of patch tokens that are input into the transformer encoder. Encoder comprises series of Swin Visual Transformer blocks organized into four stages. In each Swin Transformer block, window based multi-head self-attention (W-MSA) operation is performed, examining how features relate to each other, followed by shifted window mechanism (SW-MSA) in alternating layers, which lets the model see connections between the neighbouring windows [173]. Additionally, within each block, linear layer normalization is applied to stabilize the data, and residual connections are applied to ensure that original information is not lost during the transformations. Before the output is produced, model further processes the data through a multi-layer perceptron (MLP), which refines and enhances the extracted features, and applies the activation function to introduce the non-linearity and determine which neurons in the network should be activated to focus on the most relevant features of the data [106].

Between the encoder stages, spatial resolution is reduced via patch merging, which concatenates the neighbouring patch embeddings, reducing the spatial dimension and increasing the channel depth.

This design mimics the resolution hierarchy of the U-Net architecture and connects the encoder and decoder paths through skip connections to ensure preservation of spatial information and accurate image reconstruction during decoding [106]. Decoder of SwinUNETR is fully convolutional, performing transposed convolutions to increase the spatial resolution of feature maps, concatenating these features with corresponding encoder outputs from skip connections, applying instance normalization and Leaky ReLU activation. Lastly, output layer uses 1x1x1 convolution to reduce number of feature channels to the desired number of output classes (for instance, organ labels) [106].

2.5. Training and inference

All models (3 networks on 3 datasets – 9 models overall) were trained using the same set of hyperparameters (Table 7) to ensure a fair comparison. Training was performed for 100 epochs each, using a batch size of 2, meaning that two input samples are processed simultaneously during the forward and backward pass of the training loop to balance memory usage and training stability. Optimizer used was Stochastic Gradient Descent with an initial learning rate of 0.01, which is gradually reduced over time using a PolyLR (Polynomial Learning Rate) scheduler. The loss function combined Dice loss and Cross-Entropy loss; Leaky ReLu was used as the activation function throughout the networks. Models were also monitored calculating mean validation Dice score across all classes (OARs) after each epoch.

Table 7. Model learning hyperparameters

Parameter	**Setting**
Activation function	Leaky ReLU
Loss function	Dice loss + Cross-Entropy loss (DiceCE)
Optimizer	Stochastic Gradient Descent
Learning rate scheduler	PolyLR
Initial learning rate	0.01
Batch size	2
Epochs	100
Number of iterations per epoch	250

After finishing the training process, inference was performed on the testing dataset unseen by model during training, using predictions from all 5 cross-validation folds. Each fold generated softmax predictions for every voxel, and the final segmentation output was produced by averaging these predictions across all folds.

Post-processing of the predicted images is conducted by removing all but the largest connected component for each class that might benefit from it (comparing performance metrics before and after) – this eliminates small, irrelevant structures that may be incorrectly segmented due to noise.

Preparing of the datasets, training and evaluation of the models were conducted in a remote server that belongs to Kaunas University of Technology Artificial Intelligence Center, on NVIDIA H100 NVL graphical processing unit (GPU), with approximately 10 gigabytes of GPU allocated per training using CUDA version 12.7.

2.6. Evaluation

Final evaluation of the models was carried out comparing the predicted organ labels in the test sets with their ground truth masks voxel- and surface-wise. Voxel classification accuracy can be assessed with confusion matrix elements (true positives (TP) – when a voxel is correctly predicted as a part of the structure; true negatives (TN) – voxel is correctly predicted as not part of the structure; false positives (FP) - voxel is incorrectly predicted as part of the structure, also referred to as over-segmentation; and false negatives (FN) – when a voxel is missed by the model (Fig. 17).

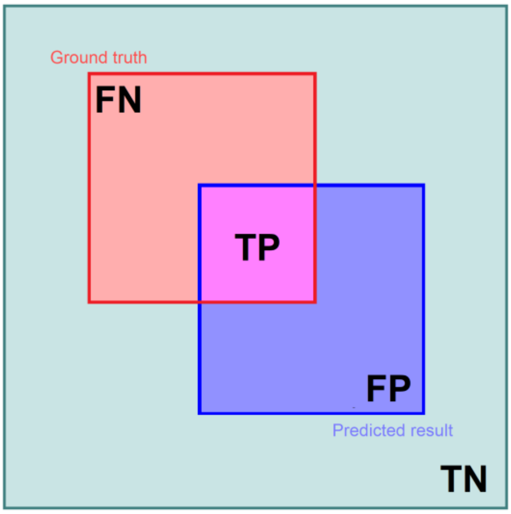


Fig. 17. Segmentation accuracy evaluation regarding the confusion matrix elements

From the obtained predicted voxels belonging to the particular organ class, and the corresponding ground truth masks, *Dice Similarity Coefficient (DSC)* can be calculated:

$$DSC = \frac{2(\mathcal{M}_{gt} \cap \mathcal{M}_{pr})}{\mathcal{M}_{gt} + \mathcal{M}_{pr}} = \frac{2TP}{2TP + FP + FN} \tag{9}$$

where $\mathcal{M}_{gt}$ – ground truth mask, $\mathcal{M}_{pr}$ – mask predicted by model. Values range from 0 to 1, where *DSC = 0* means no overlap of the masks is observed and *DSC = 1* means all voxels of the predicted mask overlap with the voxels of the ground truth structure.

Hausdorff distance (HD) measures the largest distance between surface points of predicted segmentation and the ground truth mask.

$$HD(S_{gt}, S_{pr}) = \max\left(h(S_{gt}, S_{pr}), h(S_{pr}, S_{gt})\right) \tag{10}$$

where S_{gt}, S_{pr} are the ground truth and prediction masks' surfaces, $h(S_{gt}, S_{pr})$ denotes the function to calculate the Euclidian distance between the surface points (s_{gt}, s_{pr}) of the masks:

$$h(S_{gt}, S_{pr}) = \max_{s_{gt} \in S_{gt}} \min_{s_{pr} \in S_{pr}} \left\| s_{gt} - s_{pr} \right\| \tag{11}$$

Since the *HD* measures the maximum surface distance, capturing the worst-case mismatch, it is extremely sensitive to outliers (small false-positive regions far from the ground truth surface) and might not always adequately reflect the clinical impact. Small incorrectly segmented regions can be easily erased during editing, hence, 95th percentile of the of all distances is computed to enhance the robustness of the used metric.

Surface DSC (sDSC), introduced by Nikolov et al. [174], assesses the accuracy of segmentation boundaries by considering how much of the predicted surface lies within a certain distance (denoted as tolerance τ) of the ground truth surface.

$$sDSC = \frac{\left|S_{gt} \cap B_{pr}^{(\tau)}\right| + \left|S_{pr} \cap B_{gt}^{(\tau)}\right|}{\left|S_{gt}\right| + \left|S_{pr}\right|} \tag{12}$$

Where $B_{gt}^{(\tau)}$ and $B_{pr}^{(\tau)}$ are border regions of surface areas S_{gt} and S_{pr} at a tolerance τ (mm).

Similarly to DSC, values lie in the range from 0 to 1, 1 meaning the perfect overlap of the two surfaces. Research by Vaassen et al. [120] has shown that values this metric to correlate better with the time-savings and clinical applicability of the auto-segmented contours. Rhee et al. [175] determined that surface DSC with a tolerance of 1-3 mm are the best-known metrics to predict the clinical acceptability of the automatically generated contours, since higher tolerances are not as accurate in predicting acceptability, and smaller than 1 mm introduces sensitivity to tiny discrepancies, smaller than typical resolution of CT images.

To evaluate statistical significances of the performance metrics results across models, Kruskall-Wallis's test was used to detect overall differences across groups followed by Mann-Whitney U (also known as Wilcoxon rank-sum test) post-hoc tests with Bonferroni correction. This approach was chosen because it does not assume a normal distribution of the data, making it well suited for performance metrics that are often skewed.

3. Results and Discussion

This section presents the experimental results obtained from training and evaluating three different neural network architectures: U-Net, ResEncU-Net, and SwinUNETR, on three distinct medical imaging datasets: AMOS, KC, and BTCV. The training process was monitored through loss metrics and validation performance to assess learning dynamics, while final model performance was evaluated using Dice Similarity Coefficient (DSC), Surface DSC (sDSC), and Hausdorff Distance (HD95) on testing datasets. Both training behaviour and quantitative results are analysed to compare effectiveness of the models in multi-organ segmentation tasks.

3.1. Analysis of the model training processes

Each dataset was used to train the model by iteratively passing data to the neural networks. Training and validation losses were calculated and logged each training epoch, allowing to follow the training progress and evaluate the learning behaviour of the models. Although each model was trained five times for k-fold cross-validation, in this section, training logs of one representational fold of each model will be presented.

AMOS dataset training and validation losses (Fig. 18) show that validation loss was consistently decreasing throughout the whole training, however, ResEncU-Net seems to converge more quickly than the other two models and achieves the lowest final loss values among the three models, indicating better learning. None of the networks show any rough jumps in the training and validation losses, indicating that the initial learning rate for the optimizer was chosen correctly. Notably, slope of SwinUNETR training might indicate that it needs more epochs to learn the dataset features.

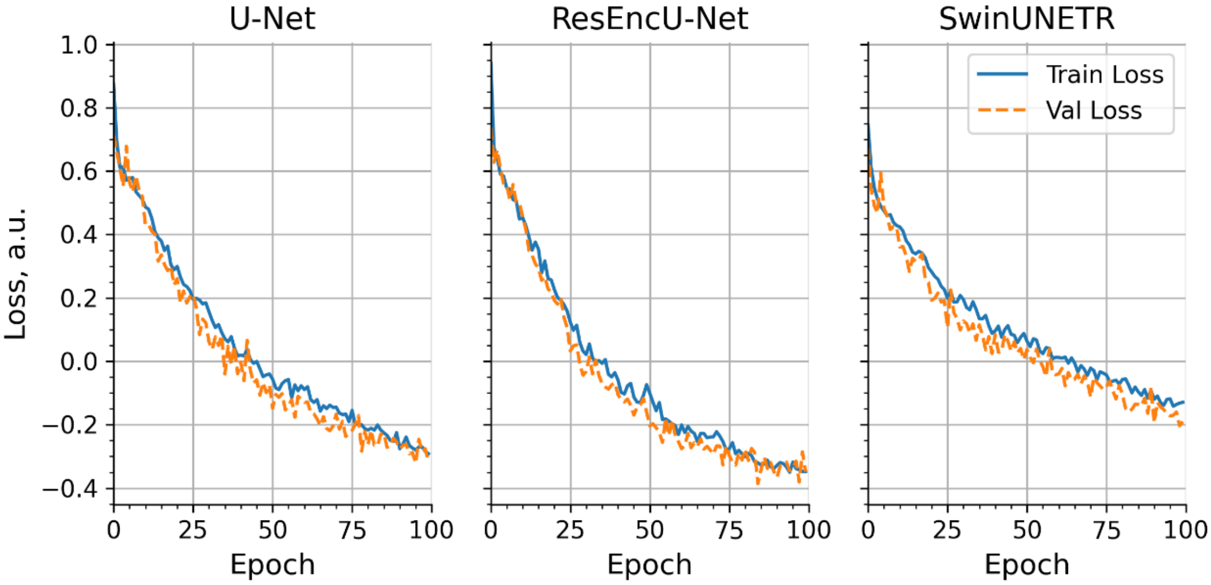


Fig. 18. Training and validation losses (AMOS dataset)

From Averaged Dice Similarity Coefficient values calculated during training, it can be observed more in more detailed manner that the ResEncU-Net predicted validation cases more accurately during the training process and learned the dataset features more quickly (Fig. 19 – orange curve), while SwinUNETR failed to follow the learning trends of other two models from 30th epoch till the end of the training (Fig. 19 – green curve). Validation Dice plots from the training with other two datasets are provided in the Appendix 1.

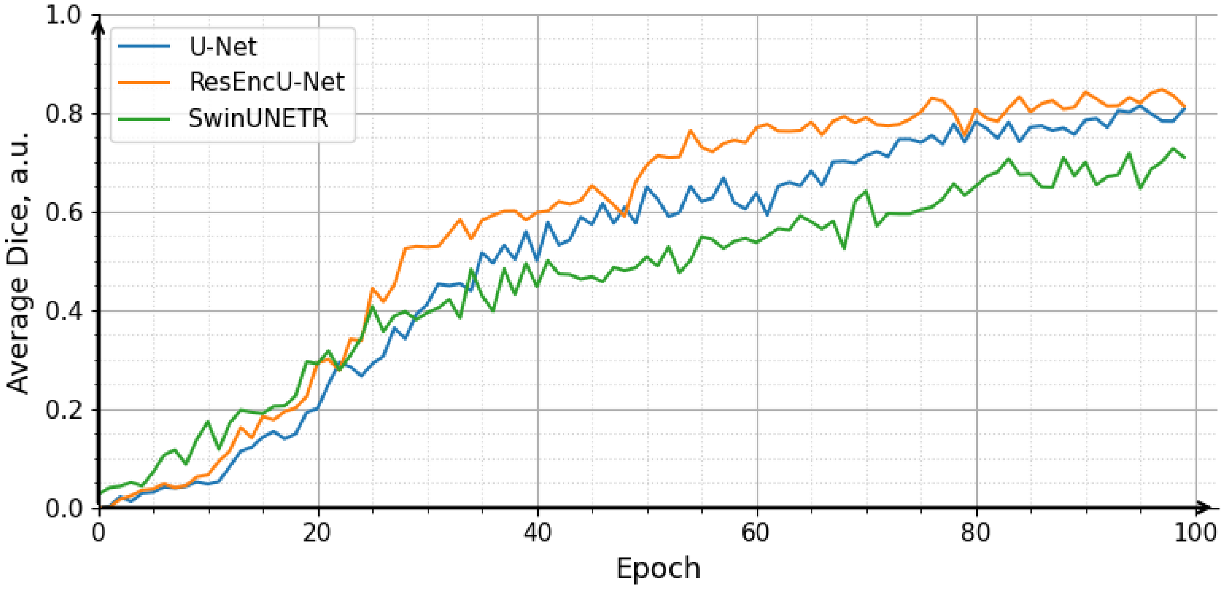


Fig. 19. Average Dice calculated on validation dataset during training

KC dataset is significantly smaller than the AMOS, less diverse in terms of patient cases and image acquisition protocols (all images in KC dataset were taken with the same CT scanner and have identical voxel spacings) – with fewer and more similar samples, models can process the entire dataset more frequently, leading to faster convergence in training loss across all models (Fig 20). Loss values of these trainings display that all models reach lower losses than during training with AMOS dataset. Both U-Net and ResEncU-Net show similar training behaviour and plateaus at around 75 epochs, while SwinUNETR loss reaches similar values 25 epochs later.

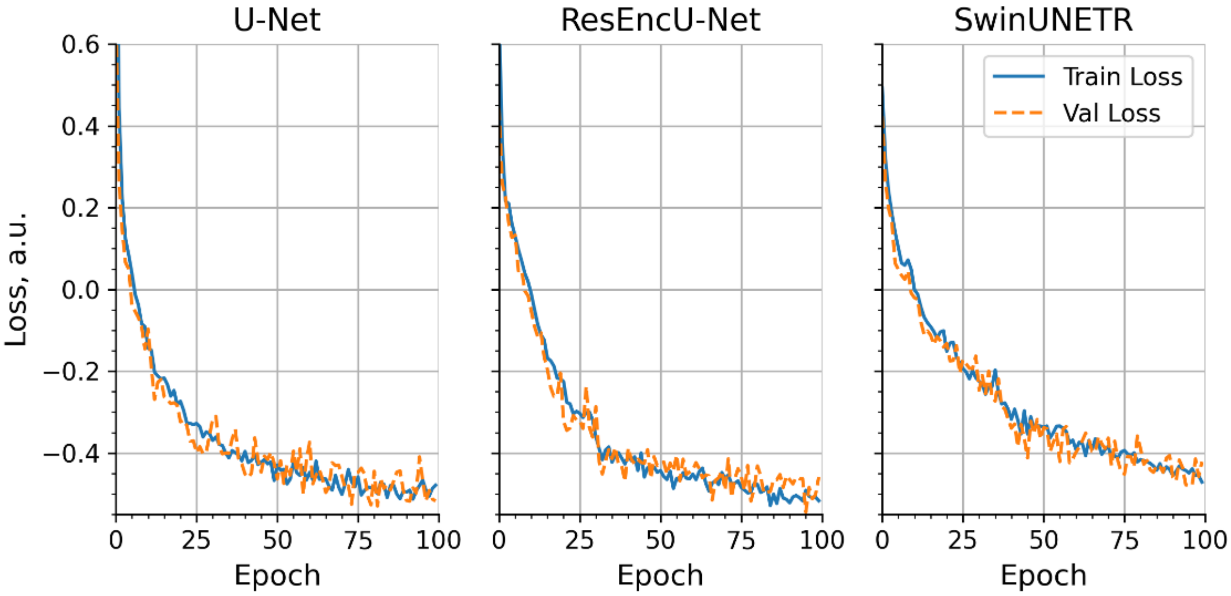


Fig. 20. Training and validation losses (KC dataset)

While training the models on BTCV dataset, all networks show gradually decreasing training and validation losses. Slight overfitting can be observed in U-Net and ResEncU-Net validation loss curves, as it stabilizes at around 75 epochs. BTCV dataset has more patient cases in the training dataset compared to the KC dataset, however it consists of fewer CT slices covering only abdominal regions and is comprised of the images acquired with several CT scanners. Smaller size of the dataset may have contributed to slightly worse loss optimization, although further evaluation is required. Additionally, annotated organs in this dataset include very small and detailed structures (both adrenal glands, portal and splenic veins) which can be too complex for auto-segmentation model to learn due to the class imbalances, or variances in manual delineations in the training set.

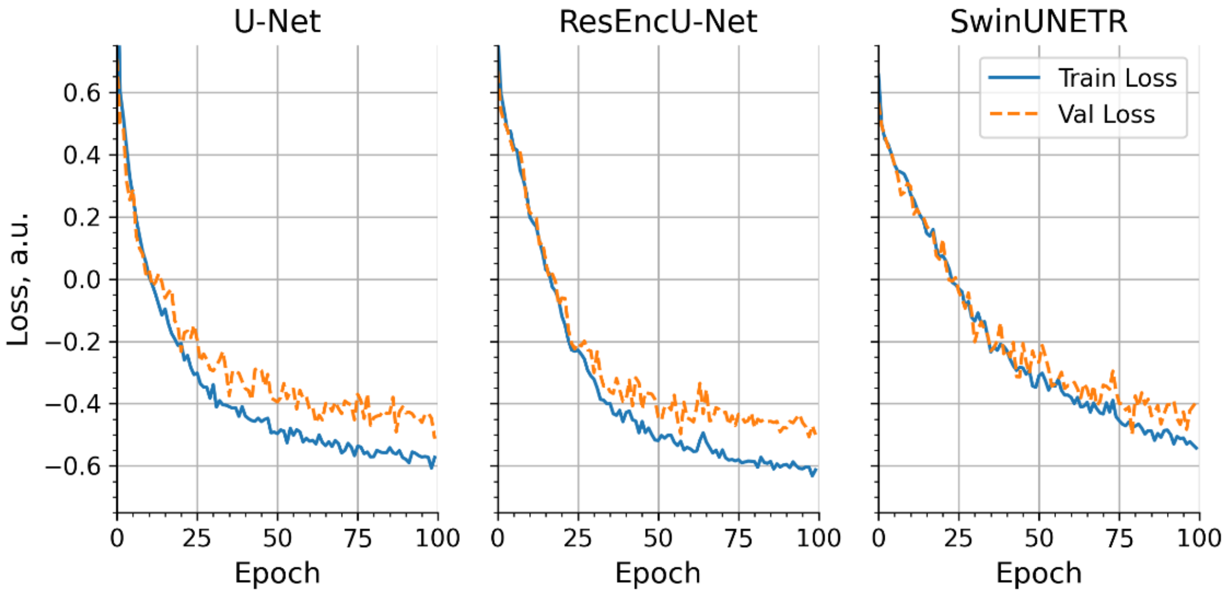


Fig. 21. Training and validation losses (BTCV dataset)

Judging from the training and validation loss curves, it could be concluded that SwinUNETR might benefit from additional training to observe if the loss values would decrease further, while U-Net and ResEncU-Net both converge it faster.

Overall training time of five folds (Table 8) also varied depending on the model architecture and dataset used. Among the evaluated models SwinUNETR consistently required the longest training time for 100 epochs across all datasets. This can be attributed to more complex transformer-based network architecture, which includes computationally intensive operations such as self-attention and hierarchical feature processing. In contrast, CNN-based models demonstrated faster training cycles with the same training parameters. Furthermore, KC dataset took the most time to train on regardless of model applied, which can be attributed to higher volumes of patches used for model training compared to other datasets.

Table 8. Overall training time for each model

	AMOS	**KC**	**BTCV**
ResEncU-Net	4.25 h	6 h	3.75 h
SwinUNETR	8 h	10 h	8 h
U-Net	2.5 h	4 h	3 h

It is important to note that overall training time is not indicative of model segmentation performance on unseen patient cases. However, it is useful to know in situations where model requires retraining or finetuning due to declined accuracy in clinical use or when new training data becomes available that includes additional organs at risk (OARs). In such cases, computational cost and time efficiency become significant factors in selecting and maintaining deep learning models for clinical deployment.

After training, inference (automated segmentation) on test subsets for all datasets was performed to assess each model's ability to segment OARs in computed tomography (CT) images of unseen patient cases (CT scans not encountered during training and validation). Inference time (Fig. 22) increases linearly with the number of slices in the CT volumes. Out of the three models, Residual Encoder U-Net (ResEncU-Net) required the most time for the auto-segmentation – scans with only abdominal regions present (up to 360 slices) took 2.5 minutes, while whole body scans took up to 4.5 minutes. The inference times for SwinUNETR models were slightly shorter, with whole body scans taking up to 4 minutes, while U-Net took the least time – largest CT volumes were segmented in less than 3 minutes. Longer times for auto-segmentation can be attributed to model architecture complexity, number of computational operations and higher amount of parameters that have to be learnt by the model.

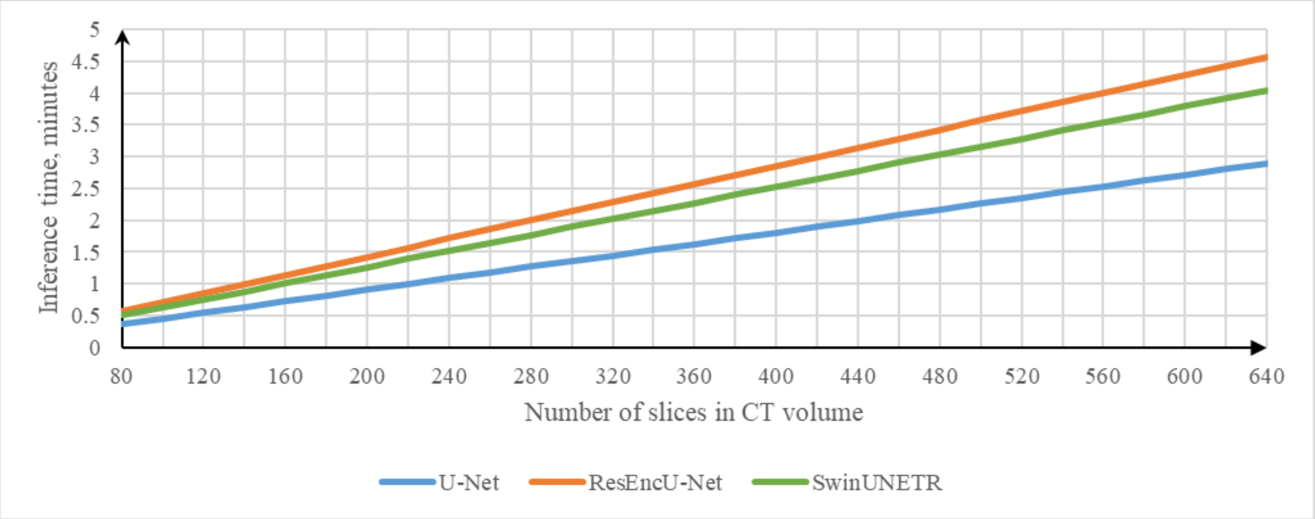


Fig. 22. Auto-segmentation duration for each model

3.2. Model evaluation

Evaluation of the models were conducted on unseen sets of patient cases in each dataset, following metrics were calculated: Dice Similarity Coefficient (DSC), Surface DSC (sDSC) with 0 and 2 mm tolerance along with Hausdorff Distance 95th percentile (HD95). Such tolerances for Surface DSC were chosen, because sDSC (τ = 0 mm) allows to see the level of absolute alignment between structure boundaries and sDSC with introduced 2 mm tolerance (τ = 2 mm) would show performance that correlates well with segmentation influence on contouring workflow and has been proven to be the best metric to predict clinical acceptability of the contours [175].

Averaged metrics (Table 9) show that across all datasets, ResEncU-Net consistently outperforms the other models, particularly in terms of DSC and sDSC at investigated tolerances (0 and 2 mm), suggesting that this model is highly effective at capturing both volumetric overlap and precise surface alignment. These performance differences are statistically significant ($p < 0.05$) in most comparisons, as shown by post-hoc Mann-Whitney U tests with Bonferroni corrections. Detailed statistical test results are provided in Appendices 2 and 3.

U-Net demonstrates competitive performance, with very close values in KC dataset and slightly lower values in the AMOS and BTCV. This indicates that U-Net remains a strong baseline model of medical image segmentation, particularly in scenarios where computational resources or implementation simplicity are prioritized. In contrast, SwinUNETR model performs the worst out of the three models in all datasets, especially in sDSC and HD95 metrics. This is the most pronounced in AMOS dataset, where averaged Hausdorff distance is nearly three times higher than of the ResEncU-Net's. These differences are statistically significant across all reported metrics ($p < 0.05$), highlighting SwinUNETR's relative weakness in precise surface alignment and boundary accuracy. This suggests challenges in delineating precise anatomical contours – an essential requirement for clinical applications. However, since Hausdorff distance represents the largest distances between points, it might not accurately indicate influence in clinical workflows, since erasing of these small over-segmented regions in the contouring system during editing is not necessarily time-consuming. Therefore, visual assessment might be beneficial to better interpret the practical relevance of these discrepancies.

Table 9. Averaged metrics throughout the datasets

	Network	**DSC, a.u. ↑**	**sDSC (τ = 0 mm), a.u. ↑**	**sDSC (τ = 2 mm), a.u. ↑**	**HD95, mm ↓**
AMOS	ResEncU-Net	0.867^{a}	0.629^{a}	0.884^{a}	16.607^{a}
	SwinUNETR	0.736^{c}	0.434^{c}	0.711^{c}	45.954^{c}
	U-Net	0.838^{b}	0.580^{b}	0.848^{b}	23.909^{b}
BTCV	ResEncU-Net	0.842^{a}	0.624^{a}	0.874^{a}	9.518^{a}
	SwinUNETR	0.797^{a}	0.549^{b}	0.805^{b}	22.260^{b}
	U-Net	0.832^{a}	0.607^{a}	0.860^{a}	11.299^{a}
KC	ResEncU-Net	0.916^{a}	0.669^{a}	0.884^{a}	7.436^{a}
	SwinUNETR	0.878^{a}	0.578^{a}	0.786^{b}	31.828^{b}
	U-Net	0.913^{a}	0.664^{a}	0.878^{a}	7.783^{a}

Superscript letters (^{a}, ^{b}, ^{c}) indicate statistically significant groupings based on post-hoc pairwise comparisons using the Mann-Whitney U test with Bonferroni correction ($p < 0.05$).
Models sharing the same letter are **not significantly different** from each other, while models with **different letters show statistically significant differences** in performance for the given metric.

Acceptability of the segmentations provided by all three models organ-wise were assessed utilizing Surface DSC metric with 2 mm tolerance, with set acceptable threshold of the metric to be 0.7 a.u. (Fig. 23). ResEncU-Net consistently achieves the highest sDSC (τ = 2 mm) values across all organs, indicating robust and clinically acceptable performance across various organ structures, with values close to threshold only indicated for heart structure and prostate/uterus organ label, which might be the result influenced by combining of these labels (prostate and uterus) in AMOS dataset, introducing uncertainties into the models. Compared to both U-Net and SwinUNETR, ResEncU-Net's performance is significantly better ($p < 0.05$) for most organs, especially smaller and mid-sized structures. Detailed results of Kruskall-Wallis statistical test along with the post-hoc analysis using Bonferroni correction method are provided in Appendices 4 and 5.

Similarly, U-Net performance is high across most organs, except prostate and uterus combination. Most values are lower compared to the ResEncU-Net model, however, for several larger organs like lungs and aorta, performance differences between U-Net and ResEncU-Net are not statistically significant.

Unlike other two models, SwinUNETR shows greater variability in performance and fails to achieve acceptable segmentation (sDSC ≥ 0.7) for several smaller organs and those with low contrast relative to surrounding tissues, such as adrenal glands, bladder, duodenum, gallbladder, pancreas, prostate and stomach. However, segmentations of OARs with bigger volumes, like aorta, bones, and both lungs showed comparable results of segmentation.

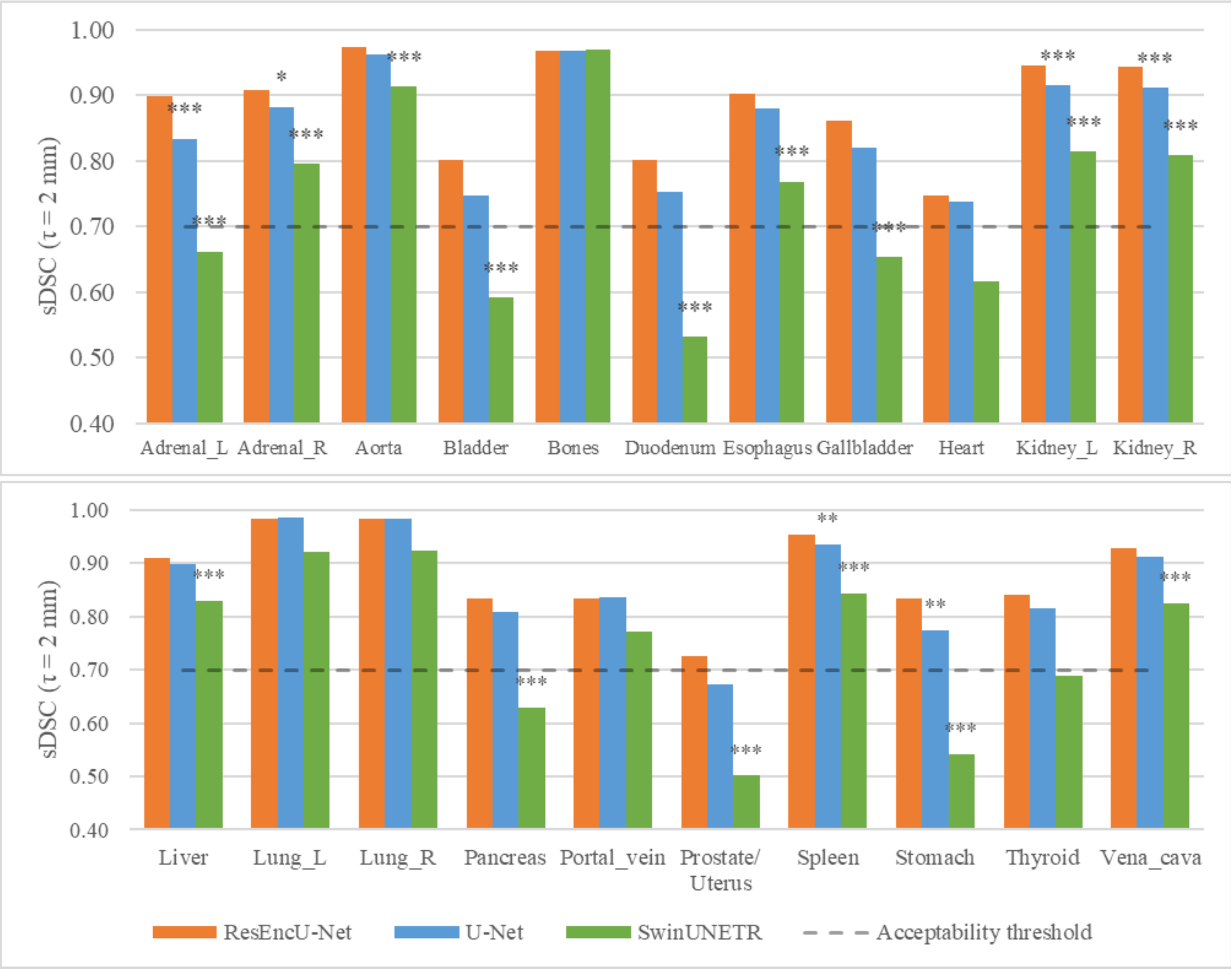


Fig. 23. Evaluation of segmentation organ-wise. Asterisks denote significant differences comparing other two models to ResEncU-Net (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$)

In general, ResEncU-Net provides the most reliable and robust segmentations across nearly all organs, making it the preferred choice for multi-organ segmentation tasks in CT images. Surface Dice values indicate that it produces contours requiring minimal manual correction – a critical factor in clinical workflows like radiotherapy planning, making this model superior for integration into semi- or fully-automated contouring pipelines, helping reduce inter-observer variability and clinician workload.

U-Net also delivers strong performance for most organs, though slightly less robust compared to ResEncU-Net in more anatomically complex or poorly contrasted regions, suggesting that manual refinement may be more frequently needed. Nonetheless, U-Net's architecture remains attractive in clinical environments that prioritize computational efficiency, faster deployment and lower hardware demands, due to shorter training durations and faster inference times compared to more complex architectures. However, improved performance of ResEncU-Net over U-Net further indicates that introduction of residual connections in U-Net structures enhances model generalizability and surface-level accuracy, which can be beneficial for adapting model to heterogeneous datasets or varying anatomical presentations across patient populations.

SwinUNETR, despite its advanced transformer-based architecture, demonstrates inconsistent results and is less suitable for smaller or low-contrast structures. While adequate to apply for large and contrasting organ (lungs, aorta, bones) segmentation, its low performance metric scores for OARs like pancreas, duodenum, stomach, hinders its clinical applicability. Therefore, SwinUNETR model may require further optimization and bigger datasets for training before clinical adoption for OAR segmentation in radiotherapy applications, where precise segmentation and boundary alignment is crucial for patient outcomes.

To conduct qualitative analysis of model segmentation, a case from each dataset was selected. For abdominal datasets, case *amos_0405* (Fig. 24) scored the highest mean sDSC ($\tau = 2$ mm) across all segmented cases - equal to 0.956 a.u. with ResEncU-Net model (compared to 0.924 and 0.782 a.u. for U-Net and SwinUNETR models, respectively). Visually it can be rated that the ground truth and U-Net and ResEncU-Net predicted classes are very similar, with few false positive voxels near the surface of the contour. Such auto-segmented structures could be evaluated as clinically acceptable with minor editing efforts. Conversely, SwinUNETR model struggled to correctly classify voxels to their corresponding OAR labels (especially for pancreas, stomach, Fig. 24 – brown and green masks, accordingly), and would take longer to edit to be considered as acceptable contour for further treatment planning.

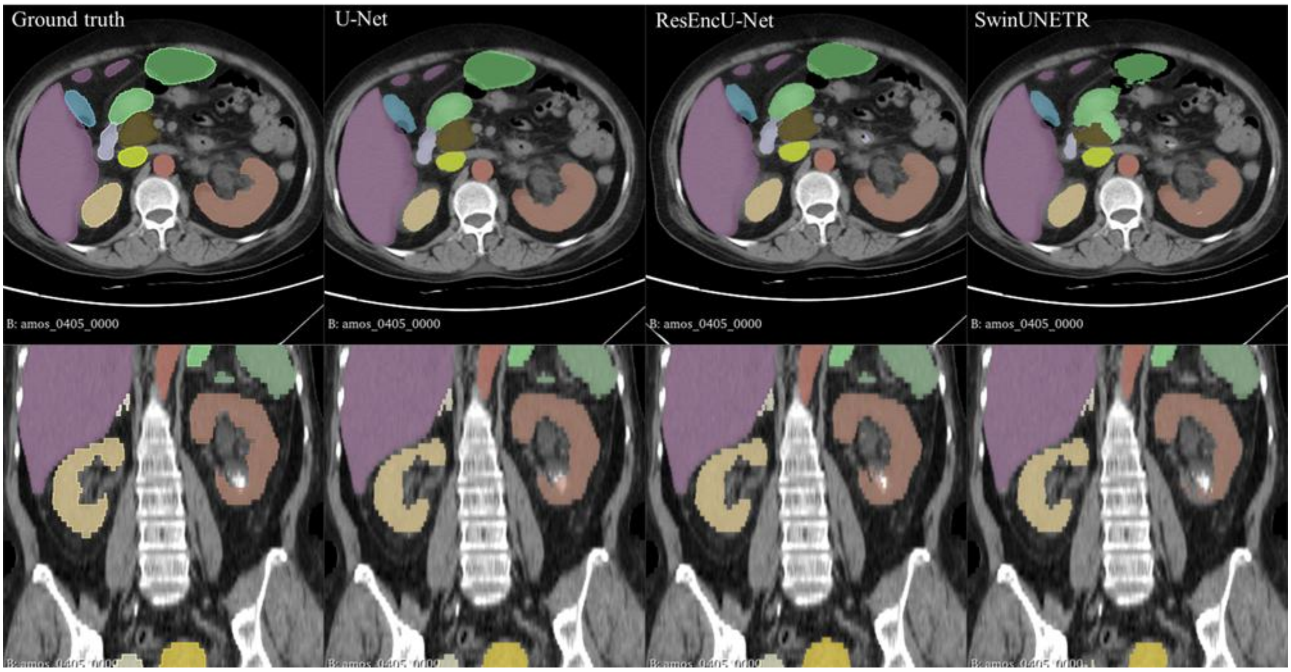


Fig. 24. Visual comparison of ground truth contours and segmentations of each model (AMOS dataset)

Similarly, for a case *image_014* in KC dataset, which consists of full-body scans, most organs segmented by U-Net and ResEncU-Net seem undistinguishable from the ground truth, both in axial and coronal planes. SwinUNETR, on the other hand, provided incorrect voxel classifications for the lungs, as can be seen in the rightmost bottom image in Fig. 25 – voxels of the right lung (coloured green) were incorrectly segmented as the left lung (coloured light brown), which indicates that the model struggles with spatial localization and distinguishing between anatomically symmetric structures. Notably, the lungs are among the most readily segmentable organs in medical imaging due to their distinct shape and contrast, which underscores the significance of such segmentation error.

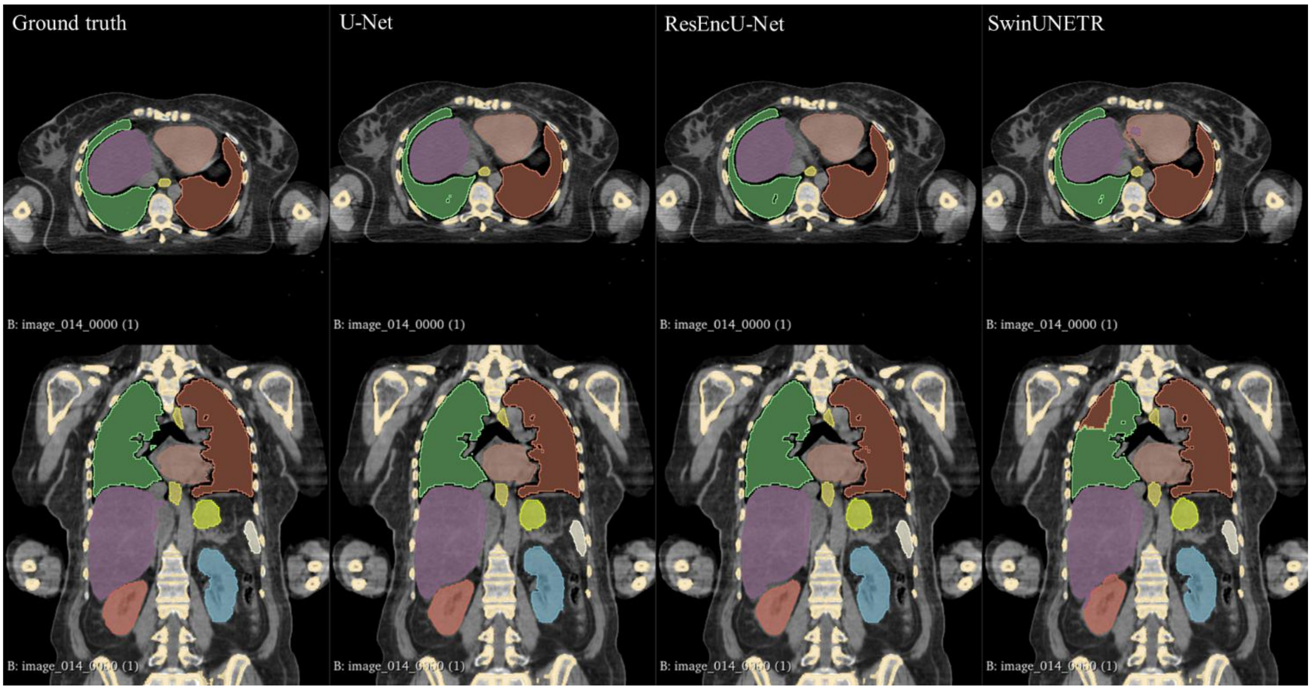


Fig. 25. Visual comparison of ground truth contours and segmentations of each model (KC dataset)

Comparison of segmentations in a case *image_0040* from BTCV dataset showed similarities between ground truth and auto-segmented volumes of all models for most volumes, although stomach and spleen (Fig. 26) exhibited false negative regions across all models. This may be partly due to the relatively small size of the BTCV training set compared to AMOS, as well as its greater variability in scanners and acquisition protocols compared to the more homogeneous KC dataset, making consistent feature learning more challenging for all models.

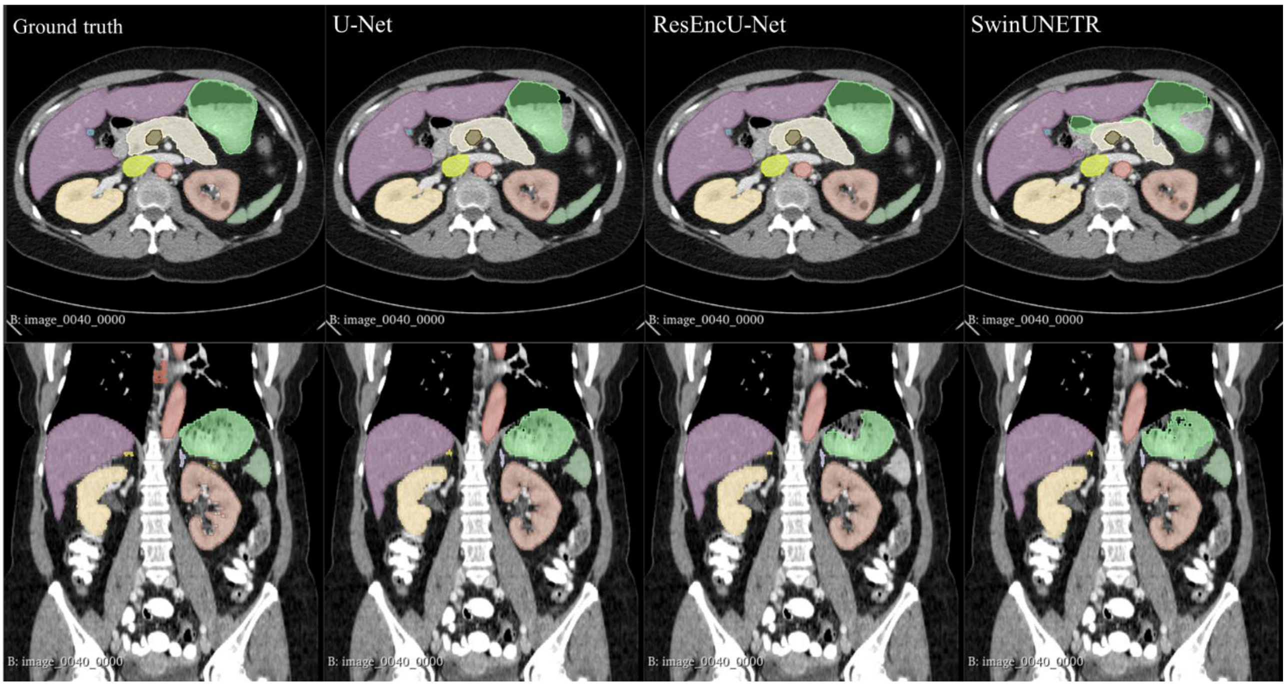


Fig. 26. Visual comparison of ground truth contours and segmentations of each model (BTCV dataset)

Based on the results obtained in this research, Visual Transformer based SwinUNETR model integrated into nnU-Net framework consistently underperformed when compared with the convolutional neural network (CNN) based architectures, particularly for smaller or low-contrast organs. However, it has demonstrated competitive or even state-of-the-art (SOTA) performance in some segmentation tasks reported in literature. In original study by Cao et al. [104] SwinUNETR outperformed original nnU-Net model when segmenting brain glioblastoma regions in BRaTS 2021 challenge dataset, which consists of multi-institutional multi-parametric magnetic resonance imaging scans (8160 scans from over 2000 patients) [176]. In the BTCV challenge, SwinUNETR achieved a reported DSC value of 0.918 and ranked among top-rated models, including U-Net variants in some studies, although this result is not consistent across implementations, and performance may depend on training strategy, dataset splitting and model configuration [177]. It has also been reported that self-supervised pretraining of this model using large scale CT datasets increased performance on BTCV dataset regarding DSC and HD metrics, stressing that Transformer based model achieves best results when pretrained on 5000 CT scans before finetuning on the BTCV dataset [178]. All training sets used in this research were relatively small (16, 25, 240 CT scans from KC, BTCV and AMOS datasets, respectively), which may have contributed to worse performance according to the segmentation metrics and visual analysis. Another problem reported in literature is prolonged training and inference runtimes compared to U-Net based architectures to achieve similar results, which also aligns with the findings of this research [179]. Therefore, SwinUNETR is inferior compared to the analysed CNN-based models for implementation in radiotherapy workflows, where contouring time, accuracy and adaptability of the auto-segmentation system altogether are important factors for efficient and precise treatment of the patients.

On the other hand, despite its relatively longer training and inference times, Residual Encoder U-Net consistently produced better segmentation results than the Transformer-based SwinUNETR and CNN-based U-Net – similarly to the results provided in Isensee et al. [134], although authors used bigger datasets for training and inference. While SwinUNETR is favoured in literature for large-scale tasks, ResEncU-Net demonstrated robust performance across datasets of varying size, suggesting that in the context of radiation treatment, where institutional datasets are often limited, model that maintains high performance under constrained training conditions is more suitable for clinical implementation. With further hyperparameter tuning, integration pipelines (loading CT scans, converting to NifTI, inference and exporting segmentations in DICOM format), and access to larger institution-specific data, model could be effectively integrated into treatment planning systems such as Eclipse (Varian Medical Systems, Palo Alto, USA).

Conclusions

1. Three 3D medical image segmentation models (U-Net, ResEncU-Net, and SwinUNETR) were successfully implemented and trained on the AMOS, KC, and BTCV datasets. Training results showed stable convergence across all models, with ResEncU-Net achieving the lowest validation loss values in each dataset (final validation loss on AMOS dataset was -0.35 for ResEncU-Net against -0.2 for SwinUNETR), indicating more efficient and effective and faster learning. ResEncU-Net also reached convergence more quickly across datasets, making it particularly suitable for clinical integration where reproducibility and computational efficiency are critical. Training times further reflected architectural differences: SwinUNETR required up to 10 hours for all cross-validation folds (KC), while ResEncU-Net needed 4.25–6 hours, and U-Net trained in as little as 2.5–4 hours, depending on the dataset.
2. Evaluation using standard segmentation metrics: Dice Similarity Coefficient (DSC), Surface DSC (sDSC), and 95th percentile Hausdorff Distance (HD95), showed that ResEncU-Net achieved the best overall performance across all datasets. For instance, in the AMOS dataset, ResEncU-Net scored 0.867 a.u. (DSC), 0.884 a.u. (sDSC $\tau = 2$ mm), and 16.6 mm (HD95). By contrast, SwinUNETR scored 0.736 a.u. (DSC) and 45.95 mm (HD95), significantly underperforming, especially in boundary alignment. These differences were statistically significant ($p < 0.05$). Organ-wise evaluation using sDSC ($\tau = 2$ mm) showed ResEncU-Net met the acceptability threshold (≥ 0.7 a.u.) in nearly all structures, making it the most reliable for clinical applications like radiotherapy treatment planning, where precise and consistent delineation of the OARs is critical.
3. Based on averaged metrics across datasets, ResEncU-Net outperformed U-Net by 2–5 % in DSC and sDSC and had lower HD95 values, particularly for anatomically complex or smaller structures. For example, in the BTCV dataset, ResEncU-Net achieved 0.842 a.u. (DSC) vs. 0.832 a.u. for U-Net, and 9.52 mm (HD95) vs. 11.3 mm, respectively. While U-Net remained competitive, especially in the KC dataset (0.913 a.u. DSC vs. 0.916 a.u. for ResEncU-Net), its performance decreased in more heterogeneous (multi-institutional, multi-vendor) datasets. SwinUNETR, despite its transformer-based design, showed the worst HD95 scores (31.8 mm in KC), and frequently failed to meet sDSC acceptability thresholds, especially for small or low-contrast organs. In contrast with published results where SwinUNETR achieved DSC ≥ 0.91 a.u. on BTCV with large-scale pretraining (for example, 5000 CT scans), findings in this research suggest that SwinUNETR's performance is highly dependent on dataset scale and pretraining, limiting its utility in typical institutional settings with smaller datasets.
4. Analysis revealed several key factors affecting model accuracy: dataset diversity, organ size, contrast, and scanner heterogeneity. For instance, homogeneity of KC (acquired with a single scanner) enabled faster convergence (plateau at around 75 epochs), whereas AMOS might require longer training due to its larger anatomical variability. Organs like the adrenal glands, bladder consistently yielded lower sDSC (often < 0.7), likely due to class imbalance and annotation variability. This highlights the importance of targeted refinement or expert-in-the-loop strategies for small OARs in radiotherapy workflows. Despite these challenges, ResEncU-Net maintained reliable performance under all conditions, suggesting strong generalizability and clinical robustness.

Acknowledgements

I would like to express my gratitude to the Head of the Radiation Therapy Department at Hospital of Lithuanian University of Health Sciences Kauno Klinikos, radiation oncologist Assoc. Prof. Dr. Laimonas Jaruševičius and his colleagues for providing access to anonymized patient imaging data and annotated structures.

My family, friends and coursemates for support, and J.R. for being all of the above.

Finally, I am sincerely grateful to my supervisor, Assoc. Prof. Dr. Jurgita Laurikaitienė, for her invaluable guidance, advice and consistent support throughout the course of this work.

List of references

1. YE, Xianghua, GUO, Dazhou, GE, Jia, YAN, Senxiang, XIN, Yi, SONG, Yuchen, YAN, Yongheng, HUANG, Bing shen, HUNG, Tsung Min, ZHU, Zhuotun, PENG, Ling, REN, Yanping, LIU, Rui, ZHANG, Gong, MAO, Mengyuan, CHEN, Xiaohua, LU, Zhongjie, LI, Wenxiang, CHEN, Yuzhen, HUANG, Lingyun, XIAO, Jing, HARRISON, Adam P., LU, Le, LIN, Chien Yu, JIN, Dakai and HO, Tsung Ying. Comprehensive and clinically accurate head and neck cancer organs-at-risk delineation on a multi-institutional study. Nature Communications 2022 13:1 [online]. 17 October 2022. Vol. 13, no. 1, p. 1–15. [Accessed 27 December 2024]. DOI 10.1038/s41467-022-33178-z. Available from: https://www.nature.com/articles/s41467-022-33178-z
2. VINOD, Shalini, MIN, Myo, JAMESON, Michael and HOLLOWAY, Lois. A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology. Journal of Medical Imaging and Radiation Oncology. 1 May 2016. Vol. 60. DOI 10.1111/1754-9485.12462.
3. CARDENAS, Carlos E., YANG, Jinzhong, ANDERSON, Brian M., COURT, Laurence E. and BROCK, Kristy B. Advances in Auto-Segmentation. Seminars in Radiation Oncology. July 2019. Vol. 29, no. 3, p. 185–197. DOI 10.1016/j.semradonc.2019.02.001.
4. HUANG, Lina, MIRON, Alina, HONE, Kate and LI, Yongmin. Segmenting Medical Images: From UNet to Res-UNet and nnUNet. Proceedings - IEEE Symposium on Computer-Based Medical Systems [online]. 5 July 2024. P. 483–489. [Accessed 28 December 2024]. DOI 10.1109/CBMS61543.2024.00086. Available from: https://arxiv.org/abs/2407.04353v1
5. SAVJANI, Ricky R., LAURIA, Michael, BOSE, Supratik, DENG, Jie, YUAN, Ye and ANDREARCZYK, Vincent. Automated Tumor Segmentation in Radiotherapy. Seminars in Radiation Oncology. 1 October 2022. Vol. 32, no. 4, p. 319–329. DOI 10.1016/J.SEMRADONC.2022.06.002.
6. HOQUE, S. M.Hasibul, PIRRONE, Giovanni, MATRONE, Fabio, DONOFRIO, Alessandra, FANETTI, Giuseppe, CAROLI, Angela, RISTA, Rahnuma Shahrin, BORTOLUS, Roberto, AVANZO, Michele, DRIGO, Annalisa and CHIOVATI, Paola. Clinical Use of a Commercial Artificial Intelligence-Based Software for Autocontouring in Radiation Therapy: Geometric Performance and Dosimetric Impact. Cancers [online]. 1 December 2023. Vol. 15, no. 24, p. 5735. [Accessed 3 April 2025]. DOI 10.3390/CANCERS15245735. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10741804/
7. DOOLAN, Paul J., CHARALAMBOUS, Stefanie, ROUSSAKIS, Yiannis, LECZYNSKI, Agnes, PERATIKOU, Mary, BENJAMIN, Melka, FERENTINOS, Konstantinos, STROUTHOS, Iosif, ZAMBOGLOU, Constantinos and KARAGIANNIS, Efstratios. A clinical evaluation of the performance of five commercial artificial intelligence contouring systems for radiotherapy. Frontiers in Oncology [online]. 2023. Vol. 13, p. 1213068. [Accessed 7 April 2025]. DOI 10.3389/FONC.2023.1213068/FULL. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10436522/
8. WORLD HEALTH ORGANISATION (WHO). Cancer Today. [online]. 8 February 2024. [Accessed 8 March 2025]. Available from: https://gco.iarc.who.int/today/en/dataviz/pie?mode=cancer&populations=440

9. MIRTIES ATVEJŲ IR JŲ PRIEŽASČIŲ STEBĖSENOS SKYRIUS. Mirties priežastys 2023 / Causes of death 2023 [online]. Vilnius, 2024. [Accessed 8 March 2025]. Available from: www.hi.lt
10. GAIDELYTĖ, Rita and GARBUVIENĖ, Milda. Health Statistics of Lithuania 2023 [online]. 2024. [Accessed 9 March 2025]. Available from: https://www.hi.lt/uploads/Statistikos_leidiniai_Sveikatos_statistika/la2023.pdf
11. PETRAUSKAS, Vidas, NARBUTAS, Šarūnas, ČIAKIENĖ, Neringa, GUDELYTĖ, Guoda and DULSKAS, Audrius. Access to Healthcare for Cancer Patients in Lithuania During the COVID-19 Pandemic. Acta Medica Lituanica [online]. 2021. Vol. 28, no. 2, p. 199. [Accessed 9 March 2025]. DOI 10.15388/AMED.2021.28.2.9. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC9133612/
12. OECD/EUROPEAN COMMISSION. EU Country Cancer Profile: Lithuania 2025. Paris : OECD Publishing, 2025.
13. HOL, Sandra and MOLLAERT, Isabelle. Treatment Planning for Breast/Chest Wall and Regional Lymph Nodes Including the Internal Mammary Chain. Breast Cancer Radiation Therapy: A Practical Guide for Technical Applications [online]. 1 January 2022. P. 167–173. [Accessed 9 March 2025]. DOI 10.1007/978-3-030-91170-6_23. Available from: https://link.springer.com/chapter/10.1007/978-3-030-91170-6_23
14. KNIPE, Henry, CAMPOS, Arlene and DI MUZIO, Bruno. Conventional radiation therapy. Radiopaedia.org. 8 April 2019.
15. CHO, Byungchul. Intensity-modulated radiation therapy: a review with a physics perspective. Radiation Oncology Journal. 31 March 2018. Vol. 36, no. 1, p. 1–10. DOI 10.3857/roj.2018.00122.
16. KOKA, Krishna, VERMA, Amit, DWARAKANATH, Bilikere S and VL PAPINENI, Rao. Technological Advancements in External Beam Radiation Therapy (EBRT): An Indispensable Tool for Cancer Treatment. [online]. 2022. [Accessed 11 March 2025]. DOI 10.2147/CMAR.S351744. Available from: https://doi.org/10.2147/CMAR.S351744
17. CHANDARANA, Hersh, WANG, Hesheng, TIJSSEN, Rob and DAS, Indra. Emerging Role of MRI in Radiation Therapy. Journal of Magnetic Resonance Imaging. 8 September 2018. Vol. 48. DOI 10.1002/jmri.26271.
18. GARCÍA-FIGUEIRAS, Roberto, BALEATO-GONZÁLEZ, Sandra, LUNA, Antonio, PADHANI, Anwar R., VILANOVA, Joan C., CARBALLO-CASTRO, Ana M., OLEAGA-ZUFIRIA, Laura, VALLEJO-CASAS, Juan Antonio, MARHUENDA, Ana and GÓMEZ-CAAMAÑO, Antonio. How Imaging Advances Are Defining the Future of Precision Radiation Therapy. Radiographics [online]. 1 February 2024. Vol. 44, no. 2. [Accessed 12 March 2025]. DOI 10.1148/RG.230152/ASSET/IMAGES/LARGE/RG.230152.FIG28.JPEG. Available from: https://pubs.rsna.org/doi/10.1148/rg.230152
19. BOTTARI, Antonio, CICERO, Giuseppe, SILIPIGNI, Salvatore, STAGNO, Alberto, CATANZARITI, Francesca, CINQUEGRANI, Antonella and ASCENTI, Giorgio. CT Scan. Anatomy for Urologic Surgeons in the Digital Era: Scanning, Modelling and 3D Printing [online]. 2 January 2023. P. 89–98. [Accessed 12 March 2025]. DOI 10.1007/978-3-030-59479-4_7. Available from: https://www.ncbi.nlm.nih.gov/books/NBK567796/

20. DENOTTER, Tami D. and SCHUBERT, Johanna. Hounsfield Unit. Radiopaedia.org [online]. 6 March 2023. [Accessed 12 March 2025]. DOI 10.53347/rid-38181. Available from: https://www.ncbi.nlm.nih.gov/books/NBK547721/

21. HERMENA, Shady and YOUNG, Michael. CT-scan Image Production Procedures. StatPearls [online]. 8 August 2023. [Accessed 12 March 2025]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK574548/

22. FORTIN, Francis. Hounsfield scale (diagram). In : Radiopaedia.org. Radiopaedia.org, 2020.

23. GARBA, I., FATIMA, A. M., ABBA, M., YAKUBU, M., MANSUR, Y., LAWAL, Y., ABUBAKAR, Auwal and USMAN, Aminu U. Analysis of image quality and radiation dose in routine adult brain helical and wide-volume computed tomography procedures. Journal of Medical Imaging and Radiation Sciences. 1 September 2022. Vol. 53, no. 3, p. 429–436. DOI 10.1016/J.JMIR.2022.05.008.

24. DOUSI, M, FATSI, A, SOTIRAKOU, K, GKATZIA, N, PATELAROU, M and THEODOSIOU, A. Helical vs Conventional CT in Routine Head Imaging. A Comparison of Dose and Image Quality using VGC Analysis. Journal of Radiology and Clinical Imaging. 2021. Vol. 04, no. 01. DOI 10.26502/JRCI.2809041.

25. LAMBERT, Jack W., PHILLIPS, Elizabeth D., VILLANUEVA-MEYER, Javier E., NARDO, Lorenzo, FACCHETTI, Luca and GOULD, Robert G. Axial or Helical? Considerations for wide collimation CT scanners capable of volumetric imaging in both modes. Medical physics [online]. 1 November 2017. Vol. 44, no. 11, p. 5718–5725. [Accessed 12 March 2025]. DOI 10.1002/MP.12525. Available from: https://pubmed.ncbi.nlm.nih.gov/28833277/

26. LI, Y., LI, X., LI, J., YANG, J. and GUO, J. Axial or Helical? CT imaging of the thorax for dyspnoea patients with free-breathing using 16 cm wide-detector CT. Clinical Radiology. 1 October 2020. Vol. 75, no. 10, p. 797.e21-797.e26. DOI 10.1016/J.CRAD.2020.05.014.

27. NOID, George, ZHU, Justin, TAI, An, MISTRY, Nilesh, SCHOTT, Diane, PRAH, Douglas, PAULSON, Eric, SCHULTZ, Christopher and LI, X. Allen. Improving Structure Delineation for Radiation Therapy Planning Using Dual-Energy CT. Frontiers in Oncology [online]. 28 August 2020. Vol. 10, p. 555030. [Accessed 13 March 2025]. DOI 10.3389/FONC.2020.01694/BIBTEX. Available from: www.frontiersin.org

28. DE PIETRO, Simona, DI MARTINO, Giulia, CAROPRESE, Mara, BARILLARO, Angela, COCOZZA, Sirio, PACELLI, Roberto, CUOCOLO, Renato, UGGA, Lorenzo, BRIGANTI, Francesco, BRUNETTI, Arturo, CONSON, Manuel and ELEFANTE, Andrea. The role of MRI in radiotherapy planning: a narrative review "from head to toe." Insights into Imaging [online]. 1 December 2024. Vol. 15, no. 1, p. 1–13. [Accessed 13 March 2025]. DOI 10.1186/S13244-024-01799-1/FIGURES/8. Available from: https://insightsimaging.springeropen.com/articles/10.1186/s13244-024-01799-1

29. YAN, Qi, YAN, Xia, YANG, Xin, LI, Sijin and SONG, Jianbo. The use of PET/MRI in radiotherapy. Insights into Imaging 2024 15:1 [online]. 27 February 2024. Vol. 15, no. 1, p. 1–15. [Accessed 13 March 2025]. DOI 10.1186/S13244-024-01627-6. Available from: https://insightsimaging.springeropen.com/articles/10.1186/s13244-024-01627-6

30. DECAZES, Pierre, HINAULT, Pauline, VERESEZAN, Ovidiu, THUREAU, Sébastien, GOUEL, Pierrick and VERA, Pierre. Trimodality PET/CT/MRI and Radiotherapy: A Mini-Review. Frontiers in Oncology [online]. 4 February 2021. Vol. 10, p. 614008. [Accessed 13 March 2025]. DOI 10.3389/FONC.2020.614008/BIBTEX. Available from: www.frontiersin.org

31. LI, D., LI, Y., SUN, Y., XU, B. and WANG, W. Effect of MRI/CT Image Fusion on Radiotherapy Planning for Central Nervous System Lymphoma. International Journal of Radiation Oncology*Biology*Physics [online]. 1 September 2019. Vol. 105, no. 1, p. E474. [Accessed 13 March 2025]. DOI 10.1016/j.ijrobp.2019.06.1452. Available from: https://www.redjournal.org/action/showFullText?pii=S0360301619322874
32. POLINATI, Srinivasu, BAVIRISETTI, Durga Prasad, RAJESH, Kandala N V P S, NAIK, Ganesh R and DHULI, Ravindra. The Fusion of MRI and CT Medical Images Using Variational Mode Decomposition. Applied Sciences. 19 November 2021. Vol. 11, no. 22, p. 10975. DOI 10.3390/app112210975.
33. VILLEGAS, Fernanda, DAL BELLO, Riccardo, ALVAREZ-ANDRES, Emilie, DHONT, Jennifer, JANSSEN, Tomas, MILAN, Lisa, ROBERT, Charlotte, SALAGEAN, Ghizela-Ana-Maria, TEJEDOR, Natalia, TRNKOVÁ, Petra, FUSELLA, Marco, PLACIDI, Lorenzo and CUSUMANO, Davide. Challenges and opportunities in the development and clinical implementation of artificial intelligence based synthetic computed tomography for magnetic resonance only radiotherapy. Radiotherapy and Oncology. September 2024. Vol. 198, p. 110387. DOI 10.1016/j.radonc.2024.110387.
34. PALMÉR, Emilia, NORDSTRÖM, Fredrik, KARLSSON, Anna, PETRUSON, Karin, LJUNGBERG, Maria and SOHLIN, Maja. Head and neck cancer patient positioning using synthetic CT data in MRI-only radiation therapy. Journal of Applied Clinical Medical Physics. 19 April 2022. Vol. 23, no. 4. DOI 10.1002/acm2.13525.
35. BAHLOUL, Mohamed A., JABEEN, Saima, BENOUMHANI, Sara, ALSALEH, Habib Abdulmohsen, BELKHATIR, Zehor and AL-WABIL, Areej. Advancements in synthetic CT generation from MRI: A review of techniques, and trends in radiation therapy planning. Journal of Applied Clinical Medical Physics. 26 November 2024. Vol. 25, no. 11. DOI 10.1002/acm2.14499.
36. LI, Xiangrui, MORGAN, Paul S., ASHBURNER, John, SMITH, Jolinda and RORDEN, Christopher. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. Journal of Neuroscience Methods. May 2016. Vol. 264, p. 47–56. DOI 10.1016/j.jneumeth.2016.03.001.
37. LAROBINA, Michele. Thirty Years of the DICOM Standard. Tomography. 1 October 2023. Vol. 9, no. 5, p. 1829–1838. DOI 10.3390/TOMOGRAPHY9050145.
38. MOORE, Michael, PATTERSON, Brandon, SAMUEL, Sara, SHERIDAN, Helenmary and CHRIS, Sorensen. Neuroimaging DICOM and NIfTI Data Curation Primer [online]. 2020. Available from: http://hdl.handle.net/11299/216582
39. STIEB, Sonja, MCDONALD, Brigid, GRONBERG, Mary, ENGESETH, Grete May, HE, Renjie and FULLER, Clifton David. Imaging for Target Delineation and Treatment Planning in Radiation Oncology: Current and Emerging Techniques. Hematology/oncology clinics of North America [online]. 1 December 2019. Vol. 33, no. 6, p. 963. [Accessed 16 May 2025]. DOI 10.1016/J.HOC.2019.08.008. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC7217094/
40. YE, Xianghua, GUO, Dazhou, GE, Jia, YAN, Senxiang, XIN, Yi, SONG, Yuchen, YAN, Yongheng, HUANG, Bing-shen, HUNG, Tsung-Min, ZHU, Zhuotun, PENG, Ling, REN, Yanping, LIU, Rui, ZHANG, Gong, MAO, Mengyuan, CHEN, Xiaohua, LU, Zhongjie, LI, Wenxiang, CHEN, Yuzhen, HUANG, Lingyun, XIAO, Jing, HARRISON, Adam P., LU, Le, LIN, Chien-Yu, JIN, Dakai and HO, Tsung-Ying. Comprehensive and clinically accurate head

and neck cancer organs-at-risk delineation on a multi-institutional study. Nature Communications. 17 October 2022. Vol. 13, no. 1, p. 6137. DOI 10.1038/s41467-022-33178-z.

41. ICRU Report 62, Prescribing, Recording and Reporting Photon Beam Therapy (Supplement to ICRU 50) – ICRU. [online]. 1999. [Accessed 9 January 2024]. Available from: https://www.icru.org/report/prescribing-recording-and-reporting-photon-beam-therapy-report-62/
42. ICRU Report 71, Prescribing, Recording, and Reporting Electron Beam Therapy – ICRU. [online]. 2004. [Accessed 9 January 2024]. Available from: https://www.icru.org/report/prescribing-recording-and-reporting-electron-beam-therapy-report-71/
43. ICRU Report 83, Prescribing, Recording, and Reporting Intensity-Modulated Photon-Beam Therapy (IMRT) – ICRU. [online]. 2010. [Accessed 9 January 2024]. Available from: https://www.icru.org/report/prescribing-recording-and-reporting-intensity-modulated-photon-beam-therapy-imrticru-report-83/
44. ICRU Report 50, Prescribing, Recording, and Reporting Photon Beam Therapy – ICRU. [online]. 1993. [Accessed 9 January 2024]. Available from: https://www.icru.org/report/prescribing-recording-and-reporting-photon-beam-therapy-report-50/
45. THE ROYAL COLLEGE OF RADIOLOGISTS. Radiotherapy Target Volume Definition and Peer Review: Second Edition: RCR guidance. 2022. London.
46. CHANG, Amy Tien Yee, TAN, Li Tee, DUKE, Simon and NG, Wai-Tong. Challenges for Quality Assurance of Target Volume Delineation in Clinical Trials. Frontiers in Oncology. 25 September 2017. Vol. 7. DOI 10.3389/fonc.2017.00221.
47. VAN DER VEEN, J., GULYBAN, A., WILLEMS, S., MAES, F. and NUYTS, S. Interobserver variability in organ at risk delineation in head and neck cancer. Radiation Oncology. 28 December 2021. Vol. 16, no. 1, p. 120. DOI 10.1186/s13014-020-01677-2.
48. PAVEL, Tomas. Feasibility of magnetic resonance imaging-based radiation therapy for brain tumour treatment. 2017.
49. CHARAGHVANDI, Ramona, VAN ASSELEN, Bram, PHILIPPENS, Marielle, VERKOOIJEN, H, GILS, C, DIEST, Paul, PIJNAPPEL, R, HOBBELINK, Monique, WITKAMP, Arjen, DALEN, Thijs, WALL, E, VAN HEIJST, Tristan, KOELEMIJ, R, VAN VULPEN, Marlou and BONGARD, Desiree. Redefining radiotherapy for early-stage breast cancer with single dose ablative treatment: A study protocol. BMC Cancer. 9 March 2017. Vol. 17, p. 181. DOI 10.1186/s12885-017-3144-5.
50. JETHWA, Krishan R., KAHILA, Mohamed M., HUNT, Katie N., BROWN, Lindsay C., CORBIN, Kimberly S., PARK, Sean S., YAN, Elizabeth S., BOUGHEY, Judy C. and MUTTER, Robert W. Delineation of Internal Mammary Nodal Target Volumes in Breast Cancer Radiation Therapy. International Journal of Radiation Oncology*Biology*Physics. March 2017. Vol. 97, no. 4, p. 762–769. DOI 10.1016/j.ijrobp.2016.11.037.
51. KAIDAR-PERSON, Orit, VROU OFFERSEN, Birgitte, HOL, Sandra, ARENAS, Meritxell, ARISTEI, Cynthia, BOURGIER, Celine, CARDOSO, Maria Joao, CHUA, Boon, COLES, Charlotte E., ENGBERG DAMSGAARD, Tine, GABRYS, Dorota, JAGSI, Reshma, JIMENEZ, Rachel, KIRBY, Anna M., KIRKOVE, Carine, KIROVA, Youlia, KOULOULIAS, Vassilis, MARINKO, Tanja, MEATTINI, Icro, MJAALAND, Ingvil, NADER MARTA, Gustavo, WITT NYSTROM, Petra, SENKUS, Elzbieta, SKYTTÄ, Tanja, TVEDSKOV, Tove F., VERHOEVEN,

Karolien and POORTMANS, Philip. ESTRO ACROP consensus guideline for target volume delineation in the setting of postmastectomy radiation therapy after implant-based immediate reconstruction for early stage breast cancer. Radiotherapy and Oncology. August 2019. Vol. 137, p. 159–166. DOI 10.1016/j.radonc.2019.04.010.

52. NIYAZI, Maximilian, ANDRATSCHKE, Nicolaus, BENDSZUS, Martin, CHALMERS, Anthony J, ERRIDGE, Sara C, GALLDIKS, Norbert, LAGERWAARD, Frank J, NAVARRIA, Pierina, MUNCK AF ROSENSCHÖLD, Per, RICARDI, Umberto, VAN DEN BENT, Martin J, WELLER, Michael, BELKA, Claus and MINNITI, Giuseppe. ESTRO-EANO guideline on target delineation and radiotherapy details for glioblastoma. Radiotherapy and Oncology. July 2023. Vol. 184, p. 109663. DOI 10.1016/j.radonc.2023.109663.
53. VALENTINI, Vincenzo, CELLINI, Francesco, RIDDELL, Angela, BRUNNER, Thomas B., ROEDER, Falk, GIULIANTE, Felice, ALFIERI, Sergio, MANFREDI, Riccardo, ARDITO, Francesco, FIORILLO, Claudio, PORZIELLA, Venanzio, MORGANTI, Alessio G., HAUSTERMANS, Karin, MARGARITORA, Stefano, DE BARI, Berardino, MATZINGER, Oscar, GKIKA, Eleni, BELKA, Claus, ALLUM, William and VERHEIJ, Marcel. ESTRO ACROP guidelines for the delineation of lymph nodal areas in upper gastrointestinal malignancies. Radiotherapy and Oncology. November 2021. Vol. 164, p. 92–97. DOI 10.1016/j.radonc.2021.08.026.
54. BRUNNER, Thomas B., HAUSTERMANS, Karin, HUGUET, Florence, MORGANTI, Alessio G., MUKHERJEE, Somnath, BELKA, Claus, KREMPIEN, Robert, HAWKINS, Maria A., VALENTINI, Vincenzo and ROEDER, Falk. ESTRO ACROP guidelines for target volume definition in pancreatic cancer. Radiotherapy and Oncology. January 2021. Vol. 154, p. 60–69. DOI 10.1016/j.radonc.2020.07.052.
55. SALEMBIER, Carl, VILLEIRS, Geert, DE BARI, Berardino, HOSKIN, Peter, PIETERS, Bradley R., VAN VULPEN, Marco, KHOO, Vincent, HENRY, Ann, BOSSI, Alberto, DE MEERLEER, Gert and FONTEYNE, Valérie. ESTRO ACROP consensus guideline on CT- and MRI-based target volume delineation for primary radiation therapy of localized prostate cancer. Radiotherapy and Oncology. April 2018. Vol. 127, no. 1, p. 49–61. DOI 10.1016/j.radonc.2018.01.014.
56. DAL PRA, Alan, DIRIX, Piet, KHOO, Vincent, CARRIE, Christian, COZZARINI, Cesare, FONTEYNE, Valérie, GHADJAR, Pirus, GOMEZ-ITURRIAGA, Alfonso, PANEBIANCO, Valeria, ZAPATERO, Almudena, BOSSI, Alberto and WIEGEL, Thomas. ESTRO ACROP guideline on prostate bed delineation for postoperative radiotherapy in prostate cancer. Clinical and Translational Radiation Oncology. July 2023. Vol. 41, p. 100638. DOI 10.1016/j.ctro.2023.100638.
57. MAHANTSHETTY, Umesh, POETTER, Richard, BERIWAL, Sushil, GROVER, Surbhi, LAVANYA, Gurram, RAI, Bhavana, PETRIC, Primoz, TANDERUP, Kari, CARVALHO, Heloisa, HEGAZY, Neamat, MOHAMED, Sandy, OHNO, Tatsuya and AMORNWICHET, Napapat. IBS-GEC ESTRO-ABS recommendations for CT based contouring in image guided adaptive brachytherapy for cervical cancer. Radiotherapy and Oncology. July 2021. Vol. 160, p. 273–284. DOI 10.1016/j.radonc.2021.05.010.
58. LE PECHOUX, Cecile, FAIVRE-FINN, Corinne, RAMELLA, Sara, MCDONALD, Fiona, MANAPOV, Farkhad, PUTORA, Paul Martin, SLOTMAN, Ben, DE RUYSSCHER, Dirk, RICARDI, Umberto, GEETS, Xavier, BELDERBOS, José, PÖTTGEN, Christoph, DZIADIUSZKO, Rafal, PEETERS, Stephanie, LIEVENS, Yolande, HURKMANS, Coen, VAN

HOUTTE, Paul and NESTLE, Ursula. ESTRO ACROP guidelines for target volume definition in the thoracic radiation treatment of small cell lung cancer. Radiotherapy and Oncology. November 2020. Vol. 152, p. 89–95. DOI 10.1016/j.radonc.2020.07.012.

59. NESTLE, Ursula, DE RUYSSCHER, Dirk, RICARDI, Umberto, GEETS, Xavier, BELDERBOS, Jose, PÖTTGEN, Christoph, DZIADIUSZKO, Rafal, PEETERS, Stephanie, LIEVENS, Yolande, HURKMANS, Coen, SLOTMAN, Ben, RAMELLA, Sara, FAIVRE-FINN, Corinne, MCDONALD, Fiona, MANAPOV, Farkhad, PUTORA, Paul Martin, LEPÉCHOUX, Cécile and VAN HOUTTE, Paul. ESTRO ACROP guidelines for target volume definition in the treatment of locally advanced non-small cell lung cancer. Radiotherapy and Oncology. April 2018. Vol. 127, no. 1, p. 1–5. DOI 10.1016/j.radonc.2018.02.023.

60. MIR, Romaana, KELLY, Sarah M., XIAO, Ying, MOORE, Alisha, CLARK, Catharine H., CLEMENTEL, Enrico, CORNING, Coreen, EBERT, Martin, HOSKIN, Peter, HURKMANS, Coen W., ISHIKURA, Satoshi, KRISTENSEN, Ingrid, KRY, Stephen F., LEHMANN, Joerg, MICHALSKI, Jeff M., MONTI, Angelo F., NAKAMURA, Mitsuhiro, THOMPSON, Kenton, YANG, Huiqi, ZUBIZARRETA, Eduardo, ANDRATSCHKE, Nicolaus and MILES, Elizabeth. Organ at risk delineation for radiation therapy clinical trials: Global Harmonization Group consensus guidelines. Radiotherapy and Oncology. September 2020. Vol. 150, p. 30–39. DOI 10.1016/j.radonc.2020.05.038.

61. NIELSEN, Camilla Panduro, LORENZEN, Ebbe L., JENSEN, Kenneth, ERIKSEN, Jesper Grau, JOHANSEN, Jørgen, GYLDENKERNE, Niels, ZUKAUSKAITE, Ruta, KJELLGREN, Martin, MAARE, Christian, LØNKVIST, Camilla Kjær, NOWICKA-MATUS, Kinga, SZEJNIUK, Weronika Maria, FARHADI, Mohammad, UJMAJURIDZE, Zaza, MARIENHAGEN, Kirsten, JOHANSEN, Tanja Stagaard, FRIBORG, Jeppe, OVERGAARD, Jens and HANSEN, Christian Rønn. Interobserver variation in organs at risk contouring in head and neck cancer according to the DAHANCA guidelines. Radiotherapy and Oncology. August 2024. Vol. 197, p. 110337. DOI 10.1016/j.radonc.2024.110337.

62. PODOBNIK, Gašper, IBRAGIMOV, Bulat, PETERLIN, Primož, STROJAN, Primož and VRTOVEC, Tomaž. vOARiability: Interobserver and intermodality variability analysis in OAR contouring from head and neck CT and MR images. Medical Physics. 17 March 2024. Vol. 51, no. 3, p. 2175–2186. DOI 10.1002/mp.16924.

63. PATRICK, H. M., SOUHAMI, L. and KILDEA, J. Reduction of inter-observer contouring variability in daily clinical practice through a retrospective, evidence-based intervention. Acta Oncologica. 1 February 2021. Vol. 60, no. 2, p. 229–236. DOI 10.1080/0284186X.2020.1825801.

64. DE LA LLANA, Victor, MAÑERU, Fernando, LIBRERO, Julián, PELLEJERO, Santiago and ARIAS, Fernando. Interobserver Variability in a Spanish Society of Radiation Oncology (SEOR) Head and Neck Course. Is Current Contouring Training Sufficient? Advances in Radiation Oncology. November 2024. Vol. 9, no. 11, p. 101591. DOI 10.1016/j.adro.2024.101591.

65. MERCIECA, Susan, BELDERBOS, José S. A. and VAN HERK, Marcel. Challenges in the target volume definition of lung cancer radiotherapy. Translational Lung Cancer Research. April 2021. Vol. 10, no. 4, p. 1983–1998. DOI 10.21037/tlcr-20-627.

66. NELSON, Christopher L., NGUYEN, Callistus, FANG, Raymond, COURT, Laurence E., CARDENAS, Carlos E., RHEE, Dong Joo, NETHERTON, Tucker J., MUMME, Raymond P., GAY, Skylar, GAY, Casey, MARQUEZ, Barbara, EL BASHA, Mohammad D., ZHAO, Yao, GRONBERG, Mary, HERNANDEZ, Soleil, NEALON, Kelly A., MARTEL, Mary K. and

YANG, Jinzhong. A real-time contouring feedback tool for consensus-based contour training. Frontiers in Oncology. 13 September 2023. Vol. 13. DOI 10.3389/fonc.2023.1204323.

67. MONTAGUE, E., ROQUES, T., SPENCER, K., BURNETT, A., LOURENCO, J. and THORP, N. How Long Does Contouring Really Take? Results of the Royal College of Radiologists Contouring Surveys. Clinical Oncology. June 2024. Vol. 36, no. 6, p. 335–342. DOI 10.1016/j.clon.2024.03.005.
68. KATSURA, Kouji, TANABE, Satoshi, NAKANO, Hisashi, SAKAI, Madoka, OHTA, Atsushi, KAIDU, Motoki, SOGA, Marie, KOBAYASHI, Taichi, TAKAMURA, Masaki and HAYASHI, Takafumi. The Relationship between the Contouring Time of the Metal Artifacts Area and Metal Artifacts in Head and Neck Radiotherapy. Tomography. 11 January 2023. Vol. 9, no. 1, p. 98–104. DOI 10.3390/tomography9010009.
69. ANDRIANARISON, V. A., LAOUITI, M., FARGIER-BOCHATON, O., DIPASQUALE, G., WANG, X., NGUYEN, N. P., MIRALBELL, R. and VINH-HUNG, V. Contouring workload in adjuvant breast cancer radiotherapy. Cancer/Radiothérapie. 1 December 2018. Vol. 22, no. 8, p. 747–753. DOI 10.1016/J.CANRAD.2018.01.008.
70. YE, Xianghua, GUO, Dazhou, TSENG, Chen-kan, GE, Jia, HUNG, Tsung-Min, PAI, Ping-Ching, REN, Yanping, ZHENG, Lu, ZHU, Xinli, PENG, Ling, CHEN, Ying, CHEN, Xiaohua, CHOU, Chen-Yu, CHEN, Danni, YU, Jiaze, CHEN, Yuzhen, JIAO, Feiran, XIN, Yi, HUANG, Lingyun, XIE, Guotong, XIAO, Jing, LU, Le, YAN, Senxiang, JIN, Dakai and HO, Tsung-Ying. Multi-institutional Validation of Two-Streamed Deep Learning Method for Automated Delineation of Esophageal Gross Tumor Volume using planning-CT and FDG-PETCT. . 11 October 2021.
71. SHIROKIKH, Boris, DALECHINA, Alexandra, SHEVTSOV, Alexey, KRIVOV, Egor, KOSTJUCHENKO, Valery, DURGARYAN, Amayak, GALKIN, Mikhail, GOLANOV, Andrey and BELYAEV, Mikhail. Systematic Clinical Evaluation of A Deep Learning Method for Medical Image Segmentation: Radiosurgery Application. . 21 August 2021.
72. MORAN, Keeva, POOLE, Claire and BARRETT, Sarah. Evaluating deep learning auto-contouring for lung radiation therapy: A review of accuracy, variability, efficiency and dose, in target volumes and organs at risk. Physics and Imaging in Radiation Oncology. January 2025. Vol. 33, p. 100736. DOI 10.1016/j.phro.2025.100736.
73. MEYER, Céline, HUGER, Sandrine, BRUAND, Marie, LEROY, Thomas, PALISSON, Jérémy, RÉTIF, Paul, SARRADE, Thomas, BARATEAU, Anais, RENARD, Sophie, JOLNEROVSKI, Maria, DEMOGEOT, Nicolas, MARCEL, Johann, MARTZ, Nicolas, STEFANI, Anaïs, SELLAMI, Selima, JACQUES, Juliette, AGNOUX, Emma, GEHIN, William, TRAMPETTI, Ida, MARGULIES, Agathe, GOLFIER, Constance, KHATTABI, Yassir, CRAVÉREAU, Olivier, RENAN, Alizée, PY, Jean-François and FAIVRE, Jean-Christophe. Artificial intelligence contouring in radiotherapy for organs-at-risk and lymph node areas. Radiation Oncology. 21 November 2024. Vol. 19, no. 1, p. 168. DOI 10.1186/s13014-024-02554-y.
74. AMISHA, MALIK, Paras, PATHANIA, Monika and RATHAUR, VyasKumar. Overview of artificial intelligence in medicine. Journal of Family Medicine and Primary Care. 2019. Vol. 8, no. 7, p. 2328. DOI 10.4103/jfmpc.jfmpc_440_19.
75. ALI, Shatha. Segmentation+and+Classification+of+Medical+Images+Using+Artific. . 21 July 2024. Vol. 3, p. 299–320. DOI 10.46649/fjiece.v3.2.20a.29.5.2024.

76. MASOUDI, Samira, HARMON, Stephanie A., MEHRALIVAND, Sherif, WALKER, Stephanie M., RAVIPRAKASH, Harish, BAGCI, Ulas, CHOYKE, Peter L. and TURKBEY, Baris. Quick guide on radiology image pre-processing for deep learning applications in prostate cancer research. Journal of Medical Imaging. 6 January 2021. Vol. 8, no. 01. DOI 10.1117/1.JMI.8.1.010901.
77. Automatic Liver Segmentation — Part 1/4: - PYCAD - Your Medical Imaging Partner. PYCAD - Your Medical Imaging Partner [online]. [Accessed 25 March 2025]. Available from: https://pycad.co/liver-segmentation-part-1/
78. SOLOVYEV, Roman, WANG, Weimin and GABRUSEVA, Tatiana. Weighted boxes fusion: Ensembling boxes from different object detection models. Image and Vision Computing. March 2021. Vol. 107, p. 104117. DOI 10.1016/j.imavis.2021.104117.
79. SOBEK, Joseph, INOJOSA, Jose R. Medina, INOJOSA, Betsy J. Medina, RASSOULINEJAD-MOUSAVI, S. M., CONTE, Gian Marco, LOPEZ-JIMENEZ, Francisco and ERICKSON, Bradley J. MedYOLO: A Medical Image Object Detection Framework. . 12 December 2023. DOI 10.1007/s10278-024-01138-2.
80. BAUMGARTNER, Michael, JÄGER, Paul F., ISENSEE, Fabian and MAIER-HEIN, Klaus H. nnDetection: A Self-configuring Method for Medical Object Detection. In : . 2021. p. 530–539.
81. EWAIDAT, Haytham Al and BRAG, Youness El. Identification of lung nodules CT scan using YOLOv5 based on convolution neural network. . 31 December 2022.
82. BILLAH, Muhammad Maruf, AL RAKIB, Abdullah, HAQUE, Md Imamul, AHAMED, Asif Shakil, HOSSAIN, Md Shakawat and BORSHA, Kamrun Nahar. Real-Time Object Detection in Medical Imaging Using YOLO Models for Kidney Stone Detection. European Journal of Computer Science and Information Technology. 26 July 2024. Vol. 12, no. 7, p. 54–65. DOI 10.37745/ejcsit.2013/vol12n75465.
83. WEI, Wanmian, HUANG, Yan, ZHENG, Junchi, RAO, Yuanyong, WEI, Yongping, TAN, Xingyue and OUYANG, Haiyang. YOLOv11-based multi-task learning for enhanced bone fracture detection and classification in X-ray images. Journal of Radiation Research and Applied Sciences. March 2025. Vol. 18, no. 1, p. 101309. DOI 10.1016/j.jrras.2025.101309.
84. SMITHMAITRIE, Pruittikorn, KHAONUALSRI, Methasit, SAE-LIM, Wannipa, WANGKULANGKUL, Piyanun, JEARANAI, Supakool and CHEEWATANAKORNKUL, Siripong. Development of deep learning framework for anatomical landmark detection and guided dissection line during laparoscopic cholecystectomy. Heliyon. February 2024. Vol. 10, no. 3, p. e25210. DOI 10.1016/j.heliyon.2024.e25210.
85. WIJATA, Agata, ANDRZEJEWSKI, Jacek and PYCIŃSKI, Bartłomiej. An Automatic Biopsy Needle Detection and Segmentation on Ultrasound Images Using a Convolutional Neural Network. Ultrasonic Imaging. 28 September 2021. Vol. 43, no. 5, p. 262–272. DOI 10.1177/01617346211025267.
86. MA, Jun, HE, Yuting, LI, Feifei, HAN, Lin, YOU, Chenyu and WANG, Bo. Segment anything in medical images. Nature Communications. 22 January 2024. Vol. 15, no. 1, p. 654. DOI 10.1038/s41467-024-44824-z.
87. ZHANG, Donghao, SONG, Yang, LIU, Dongnan, JIA, Haozhe, LIU, Siqi, XIA, Yong, HUANG, Heng and CAI, Weidong. Panoptic Segmentation with an End-to-End Cell R-CNN for Pathology Image Analysis. In : . 2018. p. 237–244.

88. FURTADO, Pedro. Loss, post-processing and standard architecture improvements of liver deep learning segmentation from Computed Tomography and magnetic resonance. Informatics in Medicine Unlocked. 1 January 2021. Vol. 24, p. 100585. DOI 10.1016/J.IMU.2021.100585.

89. GER, Rachel B., NETHERTON, Tucker J., RHEE, Dong Joo, COURT, Laurence E., YANG, Jinzhong and CARDENAS, Carlos E. Auto-contouring for Image-Guidance and Treatment Planning. Machine and Deep Learning in Oncology, Medical Physics and Radiology, Second Edition [online]. 1 January 2022. P. 231–293. [Accessed 6 January 2024]. DOI 10.1007/978-3-030-83047-2_11/FIGURES/4. Available from: https://link.springer.com/chapter/10.1007/978-3-030-83047-2_11

90. XU, Yan, QUAN, Rixiang, XU, Weiting, HUANG, Yi, CHEN, Xiaolong and LIU, Fengyuan. Advances in Medical Image Segmentation: A Comprehensive Review of Traditional, Deep Learning and Hybrid Approaches. [online]. 2024. [Accessed 12 April 2025]. DOI 10.3390/bioengineering11101034. Available from: https://doi.org/10.3390/bioengineering11101034

91. PALAZZO, Stefano, ZAMBETTA, Giovanni and CALBI, Roberto. An overview of segmentation techniques for CT and MRI images: Clinical implications and future directions in medical diagnostics. Medical Imaging Process & Technology. 28 November 2024. Vol. 7, no. 1, p. 7227. DOI 10.24294/mipt7227.

92. MARINOV, Zdravko, JAUS, Alexander, KLEESIEK, Jens and STIEFELHAGEN, Rainer. Taking a Step Back: Revisiting Classical Approaches for Efficient Interactive Segmentation of Medical Images. In : . 2025. p. 101–125.

93. LONG, Jonathan, SHELHAMER, Evan and DARRELL, Trevor. Fully convolutional networks for semantic segmentation. 2015.

94. BRAHIMI, Sourour, BEN AOUN, Najib, BENOIT, Alexandre, LAMBERT, Patrick and BEN AMAR, Chokri. Semantic segmentation using reinforced fully convolutional densenet with multiscale kernel. Multimedia Tools and Applications [online]. 15 August 2019. Vol. 78, no. 15, p. 22077–22098. [Accessed 29 April 2025]. DOI 10.1007/S11042-019-7430-X. Available from: https://www.researchgate.net/publication/331700105_Semantic_Segmentation_using_Reinforced_Fully_Convolutional_DenseNet_with_Multiscale_Kernel

95. MINAEE, Shervin, BOYKOV, Yuri, PORIKLI, Fatih, PLAZA, Antonio, KEHTARNAVAZ, Nasser and TERZOPOULOS, Demetri. Image Segmentation Using Deep Learning: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence [online]. 15 January 2020. Vol. 44, no. 7, p. 3523–3542. [Accessed 26 April 2025]. DOI 10.1109/TPAMI.2021.3059968. Available from: https://arxiv.org/abs/2001.05566v5

96. RONNEBERGER, Olaf, FISCHER, Philipp and BROX, Thomas. U-Net: Convolutional Networks for Biomedical Image Segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) [online]. 2015. Vol. 9351, p. 234–241. [Accessed 1 June 2024]. DOI 10.1007/978-3-319-24574-4_28. Available from: https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28

97. XIA, Qingling, ZHENG, Hong, ZOU, Haonan, LUO, Dinghao, TANG, Hongan, LI, Lingxiao and JIANG, Bin. A comprehensive review of deep learning for medical image segmentation. Neurocomputing [online]. 14 January 2025. Vol. 613, p. 128740. [Accessed 27 April 2025]. DOI 10.1016/J.NEUCOM.2024.128740. Available from: https://www.sciencedirect.com/science/article/pii/S092523122401511X

98. WENG, Weihao and ZHU, Xin. U-Net: Convolutional Networks for Biomedical Image Segmentation. IEEE Access [online]. 18 May 2015. Vol. 9, p. 16591–16603. [Accessed 24 May 2024]. DOI 10.1109/ACCESS.2021.3053408. Available from: https://arxiv.org/abs/1505.04597v1

99. ÇIÇEK, Özgün, ABDULKADIR, Ahmed, LIENKAMP, Soeren S., BROX, Thomas and RONNEBERGER, Olaf. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In : . 2016. p. 424–432.

100. OKTAY, Ozan, SCHLEMPER, Jo, FOLGOC, Loic, LEE, Matthew, HEINRICH, Mattias, MISAWA, Kazunari, MORI, Kensaku, MCDONAGH, Steven, HAMMERLA, Nils, KAINZ, Bernhard, GLOCKER, Ben and RUECKERT, Daniel. Attention U-Net: Learning Where to Look for the Pancreas. . 11 April 2018. DOI 10.48550/arXiv.1804.03999.

101. ZHOU, Zongwei, RAHMAN SIDDIQUEE, Md Mahfuzur, TAJBAKHSH, Nima and LIANG, Jianming. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In : . 2018. p. 3–11.

102. ISENSEE, Fabian, PETERSEN, Jens, KLEIN, Andre, ZIMMERER, David, JAEGER, Paul F., KOHL, Simon, WASSERTHAL, Jakob, KOEHLER, Gregor, NORAJITRA, Tobias, WIRKERT, Sebastian and MAIER-HEIN, Klaus H. nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. Informatik aktuell [online]. 27 September 2018. P. 22. [Accessed 24 May 2024]. DOI 10.1007/978-3-658-25326-4_7. Available from: https://arxiv.org/abs/1809.10486v1

103. HATAMIZADEH, Ali, TANG, Yucheng, NATH, Vishwesh, YANG, Dong, MYRONENKO, Andriy, LANDMAN, Bennett, ROTH, Holger R. and XU, Daguang. UNETR: Transformers for 3D Medical Image Segmentation. Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022 [online]. 18 March 2021. P. 1748–1758. [Accessed 3 May 2025]. DOI 10.1109/WACV51458.2022.00181. Available from: https://arxiv.org/pdf/2103.10504

104. CAO, Hu, WANG, Yueyue, CHEN, Joy, JIANG, Dongsheng, ZHANG, Xiaopeng, TIAN, Qi and WANG, Manning. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) [online]. 2023. Vol. 13803 LNCS, p. 205–218. [Accessed 3 May 2025]. DOI 10.1007/978-3-031-25066-8_9. Available from: https://link.springer.com/chapter/10.1007/978-3-031-25066-8_9

105. PU, Qiumei, XI, Zuoxin, YIN, Shuai, ZHAO, Zhe and ZHAO, Lina. Advantages of transformer and its application for medical image segmentation: a survey. BioMedical Engineering OnLine. 3 February 2024. Vol. 23, no. 1, p. 14. DOI 10.1186/s12938-024-01212-4.

106. HATAMIZADEH, Ali, NATH, Vishwesh, TANG, Yucheng, YANG, Dong, ROTH, Holger R. and XU, Daguang. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) [online]. 4 January 2022. Vol. 12962 LNCS, p. 272–284. [Accessed 15 May 2025]. DOI 10.1007/978-3-031-08999-2_22. Available from: https://arxiv.org/pdf/2201.01266

107. ZHOU, Hong-Yu, GUO, Jiansen, ZHANG, Yinghao, HAN, Xiaoguang, YU, Lequan, WANG, Liansheng and YU, Yizhou. nnFormer: Interleaved Transformer for Volumetric Segmentation.

IEEE TRANSACTIONS ON MEDICAL IMAGING [online]. 7 September 2021. Vol. XX, p. 1. [Accessed 23 May 2025]. Available from: https://arxiv.org/pdf/2109.03201

108. MANKO, Maksym and RAMÍREZ, Javier. 2D and 3D segmentation of organs using artificial intelligence. Advances in Artificial Intelligence: Biomedical Engineering Applications in Signals and Imaging [online]. 1 January 2024. P. 437–490. [Accessed 14 May 2025]. DOI 10.1016/B978-0-443-19073-5.00010-0. Available from: https://www.sciencedirect.com/science/article/pii/B9780443190735000100

109. KAZAJ, Pooya Mohammadi, BAJ, Giovanni, SALIMI, Yazdan, STARK, Anselm W, VALENZUELA, Waldo, GEORGE, ;, SIONTIS, C M, ZAIDI, ; Habib, REYES, Mauricio, GRÄNI, Christoph and SHIRI, Isaac. From Claims to Evidence: A Unified Framework and Critical Analysis of CNN vs. Transformer vs. Mamba in Medical Image Segmentation. [online]. 3 March 2025. [Accessed 3 May 2025]. Available from: https://arxiv.org/pdf/2503.01306

110. FAN, Zhaoxin, JIANG, Runmin, WU, Junhao, HUANG, Xin, WANG, Tianyang, HUANG, Heng and XU, Min. Enhancing Weakly Supervised 3D Medical Image Segmentation through Probabilistic-aware Learning. [online]. 4 March 2024. [Accessed 3 May 2025]. Available from: https://arxiv.org/pdf/2403.02566v1

111. XIE, Yuxin, ZHOU, Tao, ZHOU, Yi and CHEN, Geng. SimTxtSeg: Weakly-Supervised Medical Image Segmentation with Simple Text Cues. [online]. 27 June 2024. [Accessed 3 May 2025]. DOI 10.1007/978-3-031-72111-3_60. Available from: https://arxiv.org/pdf/2406.19364

112. LIU, Yuanpeng, HUI, Qinglei, PENG, Zhiyi, GONG, Shaolin and KONG, Dexing. Automatic CT Segmentation from Bounding Box Annotations using Convolutional Neural Networks. [online]. 29 May 2021. [Accessed 3 May 2025]. Available from: http://arxiv.org/abs/2105.14314

113. KIRILLOV, Alexander, MINTUN, Eric, RAVI, Nikhila, MAO, Hanzi, ROLLAND, Chloe, GUSTAFSON, Laura, XIAO, Tete, WHITEHEAD, Spencer, BERG, Alexander C., LO, Wan Yen, DOLLÁR, Piotr and GIRSHICK, Ross. Segment Anything. Proceedings of the IEEE International Conference on Computer Vision [online]. 5 April 2023. P. 3992–4003. [Accessed 3 May 2025]. DOI 10.1109/ICCV51070.2023.00371. Available from: https://arxiv.org/pdf/2304.02643

114. HE, Sheng, BAO, Rina, LI, Jingpeng, STOUT, Jeffrey, BJORNERUD, Atle, GRANT, P. Ellen and OU, Yangming. Computer-Vision Benchmark Segment-Anything Model (SAM) in Medical Images: Accuracy in 12 Datasets. [online]. 18 April 2023. [Accessed 5 May 2025]. Available from: https://arxiv.org/pdf/2304.09324

115. MA, Jun, HE, Yuting, LI, Feifei, HAN, Lin, YOU, Chenyu and WANG, Bo. Segment Anything in Medical Images. Nature Communications [online]. 24 April 2023. Vol. 15, no. 1. [Accessed 6 May 2025]. DOI 10.1038/s41467-024-44824-z. Available from: http://arxiv.org/abs/2304.12306

116. ZHU, Jiayuan, HAMDI, Abdullah, QI, Yunli, JIN, Yueming and WU, Junde. Medical SAM 2: Segment medical images as video via Segment Anything Model 2. [online]. 1 August 2024. [Accessed 6 May 2025]. Available from: https://arxiv.org/pdf/2408.00874

117. DONG, Haoyu, GU, Hanxue, CHEN, Yaqian, YANG, Jichen, CHEN, Yuwen and MAZUROWSKI, Maciej A. Segment anything model 2: an application to 2D and 3D medical images. [online]. 1 August 2024. [Accessed 6 May 2025]. Available from: https://arxiv.org/pdf/2408.00756

118. WU, Junde, WANG, Ziyue, HONG, Mingxuan, JI, Wei, FU, Huazhu, XU, Yanwu, XU, Min and JIN, Yueming. Medical SAM adapter: Adapting segment anything model for medical image

segmentation. Medical Image Analysis [online]. 1 May 2025. Vol. 102, p. 103547. [Accessed 6 May 2025]. DOI 10.1016/J.MEDIA.2025.103547. Available from: https://www.sciencedirect.com/science/article/pii/S1361841525000945

119. YANG, Jinzhong, SHARP, Gregory C. and GOODING, Mark J. Auto-Segmentation for Radiation Oncology [online]. CRC Press, 2021. ISBN 9780429323782. Available from: https://www.taylorfrancis.com/books/9781000376302

120. VAASSEN, Femke, HAZELAAR, Colien, VANIQUI, Ana, GOODING, Mark, VAN DER HEYDEN, Brent, CANTERS, Richard and VAN ELMPT, Wouter. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. Physics and Imaging in Radiation Oncology. 1 January 2020. Vol. 13, p. 1–6. DOI 10.1016/J.PHRO.2019.12.001.

121. WEBER, Damien C., TOMSEJ, Milan, MELIDIS, Christos and HURKMANS, Coen W. QA makes a clinical trial stronger: Evidence-based medicine in radiation therapy. Radiotherapy and Oncology [online]. 1 October 2012. Vol. 105, no. 1, p. 4–8. [Accessed 10 May 2025]. DOI 10.1016/j.radonc.2012.08.008. Available from: https://www.thegreenjournal.com/action/showFullText?pii=S0167814012003593

122. SHERER, Michael V., LIN, Diana, ELGUINDI, Sharif, DUKE, Simon, TAN, Li Tee, CACICEDO, Jon, DAHELE, Max and GILLESPIE, Erin F. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. Radiotherapy and Oncology. 1 July 2021. Vol. 160, p. 185–191. DOI 10.1016/J.RADONC.2021.05.003.

123. ZHANG, Ying, AMJAD, Asma, DING, Jie, SAROSIEK, Christina, ZARENIA, Mohammad, CONLIN, Renae, HALL, William A, ERICKSON, Beth and PAULSON, Eric. Comprehensive Clinical Usability-oriented Contour Quality Evaluation for Deep learning Auto-segmentation: Combining Multiple Quantitative Metrics through Machine Learning. Practical radiation oncology [online]. January 2024. Vol. 15, no. 1, p. 93. [Accessed 10 May 2025]. DOI 10.1016/J.PRRO.2024.07.007. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11711007/

124. BAROUDI, Hana, BROCK, Kristy K., CAO, Wenhua, CHEN, Xinru, CHUNG, Caroline, COURT, Laurence E., EL BASHA, Mohammad D., FARHAT, Maguy, GAY, Skylar, GRONBERG, Mary P., GUPTA, Aashish Chandra, HERNANDEZ, Soleil, HUANG, Kai, JAFFRAY, David A., LIM, Rebecca, MARQUEZ, Barbara, NEALON, Kelly, NETHERTON, Tucker J., NGUYEN, Callistus M., REBER, Brandon, RHEE, Dong Joo, SALAZAR, Ramon M., SHANKER, Mihir D., SJOGREEN, Carlos, WOODLAND, McKell, YANG, Jinzhong, YU, Cenji and ZHAO, Yao. Automated Contouring and Planning in Radiation Therapy: What Is 'Clinically Acceptable'? Diagnostics 2023, Vol. 13, Page 667 [online]. 10 February 2023. Vol. 13, no. 4, p. 667. [Accessed 24 May 2024]. DOI 10.3390/DIAGNOSTICS13040667. Available from: https://www.mdpi.com/2075-4418/13/4/667

125. KISER, Kendall J., BARMAN, Arko, STIEB, Sonja, FULLER, Clifton D. and GIANCARDO, Luca. Novel Autosegmentation Spatial Similarity Metrics Capture the Time Required to Correct Segmentations Better Than Traditional Metrics in a Thoracic Cavity Segmentation Workflow. Journal of Digital Imaging [online]. 1 June 2021. Vol. 34, no. 3, p. 541–553. [Accessed 24 May 2024]. DOI 10.1007/S10278-021-00460-3/FIGURES/6. Available from: https://link.springer.com/article/10.1007/s10278-021-00460-3

126. SHAN, Guoping, YU, Shunfei, LAI, Zhongjun, XUAN, Zhiqiang, ZHANG, Jie, WANG, Binbing and GE, Yun. A Review of Artificial Intelligence Application for Radiotherapy. Dose-

Response [online]. 1 April 2024. Vol. 22, no. 2, p. 15593258241263688. [Accessed 11 May 2025]. DOI 10.1177/15593258241263687. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11193352/

127. MOSQUEIRA-REY, Eduardo, HERNÁNDEZ-PEREIRA, Elena, BOBES-BASCARÁN, José, ALONSO-RÍOS, David, PÉREZ-SÁNCHEZ, Alberto, FERNÁNDEZ-LEAL, Ángel, MORET-BONILLO, Vicente, VIDAL-ÍNSUA, Yolanda and VÁZQUEZ-RIVERA, Francisca. Addressing the data bottleneck in medical deep learning models using a human-in-the-loop machine learning approach. Neural Computing and Applications [online]. 1 February 2024. Vol. 36, no. 5, p. 2597–2616. [Accessed 11 May 2025]. DOI 10.1007/S00521-023-09197-2/FIGURES/3. Available from: https://link.springer.com/article/10.1007/s00521-023-09197-2

128. SHANI, Chen, ZARECKI, Jonathan and SHAHAF, Dafna. The Lean Data Scientist: Recent Advances toward Overcoming the Data Bottleneck. Communications of the ACM [online]. 20 January 2023. Vol. 66, no. 2, p. 92–102. [Accessed 11 May 2025]. DOI 10.1145/3551635/SUPPL_FILE/P92-SHANI-SUPP.PDF. Available from: https://dl.acm.org/doi/pdf/10.1145/3551635

129. ANTONELLI, Michela, REINKE, Annika, BAKAS, Spyridon, FARAHANI, Keyvan, KOPP-SCHNEIDER, Annette, LANDMAN, Bennett A., LITJENS, Geert, MENZE, Bjoern, RONNEBERGER, Olaf, SUMMERS, Ronald M., VAN GINNEKEN, Bram, BILELLO, Michel, BILIC, Patrick, CHRIST, Patrick F., DO, Richard K.G., GOLLUB, Marc J., HECKERS, Stephan H., HUISMAN, Henkjan, JARNAGIN, William R., MCHUGO, Maureen K., NAPEL, Sandy, PERNICKA, Jennifer S.Golia, RHODE, Kawal, TOBON-GOMEZ, Catalina, VORONTSOV, Eugene, MEAKIN, James A., OURSELIN, Sebastien, WIESENFARTH, Manuel, ARBELÁEZ, Pablo, BAE, Byeonguk, CHEN, Sihong, DAZA, Laura, FENG, Jianjiang, HE, Baochun, ISENSEE, Fabian, JI, Yuanfeng, JIA, Fucang, KIM, Ildoo, MAIER-HEIN, Klaus, MERHOF, Dorit, PAI, Akshay, PARK, Beomhee, PERSLEV, Mathias, REZAIIFAR, Ramin, RIPPEL, Oliver, SARASUA, Ignacio, SHEN, Wei, SON, Jaemin, WACHINGER, Christian, WANG, Liansheng, WANG, Yan, XIA, Yingda, XU, Daguang, XU, Zhanwei, ZHENG, Yefeng, SIMPSON, Amber L., MAIER-HEIN, Lena and CARDOSO, M. Jorge. The Medical Segmentation Decathlon. Nature Communications [online]. 1 December 2022. Vol. 13, no. 1, p. 1–13. [Accessed 11 May 2025]. DOI 10.1038/s41467-022-30695-9. Available from: https://www.nature.com/articles/s41467-022-30695-9

130. JI, Yuanfeng, BAI, Haotian, GE, Chongjian, YANG, Jie, ZHU, Ye, ZHANG, Ruimao, LI, Zhen, ZHANG, Lingyan, MA, Wanling, WAN, Xiang and LUO, Ping. AMOS: A Large-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation. Advances in Neural Information Processing Systems [online]. 16 June 2022. Vol. 35. [Accessed 11 May 2025]. Available from: https://arxiv.org/pdf/2206.08023

131. LANDMAN, Bennett, XU, Zhoubing, IGELSIAS, J, STYNER, Martin, LANGERAK, T and KLEIN, Arno. 2015 miccai multi-atlas labeling beyond the cranial vault workshop and challenge. In: Proc. MICCAI Multi Atlas Labeling Beyond Cranial Vault—Workshop Challenge. 2015.

132. RISTER, Blaine, YI, Darvin, SHIVAKUMAR, Kaushik, NOBASHI, Tomomi and RUBIN, Daniel L. CT-ORG, a new dataset for multiple organ segmentation in computed tomography. Scientific Data 2020 7:1 [online]. 11 November 2020. Vol. 7, no. 1, p. 1–9. [Accessed 11 May 2025]. DOI 10.1038/s41597-020-00715-8. Available from: https://www.nature.com/articles/s41597-020-00715-8

133. WASSERTHAL, Jakob, BREIT, Hanns-Christian, MEYER, Manfred T, PRADELLA, Maurice, HINCK, Daniel, SAUTER, Alexander W, HEYE, Tobias, BOLL, Daniel T, CYRIAC, Joshy, YANG, Shan, BACH, Michael and SEGEROTH, Martin. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. Radiology: Artificial Intelligence [online]. 2023. Vol. 5, no. 5, p. e230024. DOI 10.1148/ryai.230024. Available from: https://doi.org/10.1148/ryai.230024

134. ISENSEE, Fabian, WALD, Tassilo, ULRICH, Constantin, BAUMGARTNER, Michael, ROY, Saikat, MAIER-HEIN, Klaus and JÄGER, Paul F. nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation. In : . 2024. p. 488–498.

135. ZABEL, W. Jeffrey, CONWAY, Jessica L., GLADWISH, Adam, SKLIARENKO, Julia, DIDIODATO, Giulio, GOORTS-MATTHEWS, Leah, MICHALAK, Adam, REISTETTER, Sarah, KING, Jenna, NAKONECHNY, Keith, MALKOSKE, Kyle, TRAN, Muoi N. and MCVICAR, Nevin. Clinical Evaluation of Deep Learning and Atlas-Based Auto-Contouring of Bladder and Rectum for Prostate Radiation Therapy. Practical radiation oncology [online]. 1 January 2021. Vol. 11, no. 1, p. e80–e89. [Accessed 2 April 2025]. DOI 10.1016/J.PRRO.2020.05.013. Available from: https://pubmed.ncbi.nlm.nih.gov/32599279/

136. BORDIGONI, B., TRIVELLATO, S., PELLEGRINI, R., MEREGALLI, S., BONETTO, E., BELMONTE, M., CASTELLANO, M., PANIZZA, D., ARCANGELI, S. and DE PONTI, E. Automated segmentation in pelvic radiotherapy: A comprehensive evaluation of ATLAS-, machine learning-, and deep learning-based models. Physica Medica. 1 September 2024. Vol. 125, p. 104486. DOI 10.1016/J.EJMP.2024.104486.

137. KIBUDDE, Solomon, KAVUMA, Awusi, HAO, Yao, ZHAO, Tianyu, GAY, Hiram, VAN RHEENEN, Jacaranda, JHAVERI, Pavan Mukesh, MINJGEE, Minjmaa, VANCHINBAZAR, Enkhsetseg, NANSALMAA, Urdenekhuu and SUN, Baozhou. Impact of Artificial Intelligence-Based Autosegmentation of Organs at Risk in Low- and Middle-Income Countries. Advances in Radiation Oncology [online]. 1 November 2024. Vol. 9, no. 11, p. 101638. [Accessed 2 April 2025]. DOI 10.1016/j.adro.2024.101638. Available from: https://www.advancesradonc.org/action/showFullText?pii=S245210942400201X

138. CHOI, Min Seo, CHANG, Jee Suk, KIM, Kyubo, KIM, Jin Hee, KIM, Tae Hyung, KIM, Sungmin, CHA, Hyejung, CHO, Oyeon, CHOI, Jin Hwa, KIM, Myungsoo, KIM, Juree, KIM, Tae Gyu, YEO, Seung Gu, CHANG, Ah Ram, AHN, Sung Ja, CHOI, Jinhyun, KANG, Ki Mun, KWON, Jeanny, KOO, Taeryool, KIM, Mi Young, CHOI, Seo Hee, JEONG, Bae Kwon, JANG, Bum Sup, JO, In Young, LEE, Hyebin, KIM, Nalee, PARK, Hae Jin, IM, Jung Ho, LEE, Sea Won, CHO, Yeona, LEE, Sun Young, CHANG, Ji Hyun, CHUN, Jaehee, LEE, Eung Man, KIM, Jin Sung, SHIN, Kyung Hwan and KIM, Yong Bae. Assessment of deep learning-based auto-contouring on interobserver consistency in target volume and organs-at-risk delineation for breast cancer: Implications for RTQA program in a multi-institutional study. Breast [online]. 1 February 2024. Vol. 73, p. 103599. [Accessed 2 April 2025]. DOI 10.1016/j.breast.2023.103599. Available from: https://www.thebreastonline.com/action/showFullText?pii=S0960977623007257

139. KIM, Young Woo, BIGGS, Simon and CLARIDGE MACKONIS, Elizabeth. Investigation on performance of multiple AI-based auto-contouring systems in organs at risks (OARs) delineation. Physical and engineering sciences in medicine [online]. 1 September 2024. Vol. 47, no. 3. [Accessed 10 April 2025]. DOI 10.1007/S13246-024-01434-9. Available from: https://pubmed.ncbi.nlm.nih.gov/39222214/

140. AMJAD, Asma, XU, Jiaofeng, THILL, Dan, LAWTON, Colleen, HALL, William, AWAN, Musaddiq J., SHUKLA, Monica, ERICKSON, Beth A. and LI, X. Allen. General and Custom Deep Learning Auto-Segmentation Models for Organs in Head and Neck, Abdomen, and Male Pelvis. Medical physics [online]. 1 March 2022. Vol. 49, no. 3, p. 1686. [Accessed 12 May 2025]. DOI 10.1002/MP.15507. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8917093/

141. VEERARAGHAVAN, Harini. Clinical Commissioning Guidelines. Auto-Segmentation for Radiation Oncology [online]. 23 February 2021. P. 189–200. [Accessed 12 May 2025]. DOI 10.1201/9780429323782-16. Available from: https://www.researchgate.net/publication/349534569_Clinical_Commissioning_Guidelines

142. KING, John, WHITTAM, Shona, SMITH, David and AL-QAISIEH, Bashar. The impact of a metal artefact reduction algorithm on treatment planning for patients undergoing radiotherapy of the pelvis. Physics and Imaging in Radiation Oncology [online]. 1 October 2022. Vol. 24, p. 138. [Accessed 11 May 2025]. DOI 10.1016/J.PHRO.2022.11.007. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC9674537/

143. SYLOLYPAVAN, Aneeta, SLEEMAN, Derek, WU, Honghan and SIM, Malcolm. The impact of inconsistent human annotations on AI driven clinical decision making. NPJ Digital Medicine [online]. 1 December 2023. Vol. 6, no. 1, p. 26. [Accessed 12 May 2025]. DOI 10.1038/S41746-023-00773-3. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC9944930/

144. VĂDINEANU, Şerban, PELT, D., DZYUBACHYK, O. and BATENBURG, K. An Analysis of the Impact of Annotation Errors on the Accuracy of Deep Learning for Cell Segmentation. International Conference on Medical Imaging with Deep Learning. 2022.

145. SIDDIQUE, Nahian, PAHEDING, Sidike, ELKIN, Colin P. and DEVABHAKTUNI, Vijay. U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications. IEEE Access. 2021. Vol. 9, p. 82031–82057. DOI 10.1109/ACCESS.2021.3086020.

146. GÓRKA, Monika, JAWOREK, Daniel and WODZINSKI, Marek. Deep Learning-Based Segmentation of Tumors in PET/CT Volumes: Benchmark of Different Architectures and Training Strategies. [online]. 2024. [Accessed 24 May 2025]. Available from: https://arxiv.org/pdf/2404.09761v1

147. HUMADY, Khaled, AL-SAEED, Yasmeen, ELADAWI, Nabila, ELGARAYHI, Ahmed, ELMOGY, Mohammed and SALLAH, Mohammed. Efficient liver segmentation with 3D CNN using computed tomography scans. . 28 August 2022.

148. DING, Jiaqi, ZHANG, Zehua, TANG, Jijun and GUO, Fei. A Multichannel Deep Neural Network for Retina Vessel Segmentation via a Fusion Mechanism. Frontiers in Bioengineering and Biotechnology [online]. 19 August 2021. Vol. 9, p. 697915. [Accessed 24 May 2025]. DOI 10.3389/FBIOE.2021.697915. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8417313/

149. JAVIER GIL-TERRÓN, F., FERRI, Pablo, MONTOSA-I-MICÓ, Víctor, GÓMEZ MAHIQUES, María, LOPEZ-MATEU, Carles, MARTÍ, Pau, GARCÍA-GÓMEZ, Juan M. and FUSTER-GARCIA, Elies. Exploring the Trade-Off between generalist and specialized Models: A center-based comparative analysis for glioblastoma segmentation. International Journal of Medical Informatics [online]. 1 November 2024. Vol. 191. [Accessed 24 May 2025]. DOI 10.1016/j.ijmedinf.2024.105604. Available from: https://pubmed.ncbi.nlm.nih.gov/39154600/

150. MUTASA, Simukayi, SUN, Shawn and HA, Richard. Understanding artificial intelligence based radiology studies: What is overfitting? Clinical Imaging [online]. 1 September 2020. Vol.

65, p. 96–99. [Accessed 24 May 2025]. DOI 10.1016/J.CLINIMAG.2020.04.025. Available from: https://www.sciencedirect.com/science/article/pii/S0899707120301376

151. LIU, Xiangbin, SONG, Liping, LIU, Shuai and ZHANG, Yudong. A Review of Deep-Learning-Based Medical Image Segmentation Methods. Sustainability 2021, Vol. 13, Page 1224 [online]. 25 January 2021. Vol. 13, no. 3, p. 1224. [Accessed 6 January 2024]. DOI 10.3390/SU13031224. Available from: https://www.mdpi.com/2071-1050/13/3/1224/htm

152. XU, Chuhan, COEN-PIRANI, Pablo and JIANG, Xia. Empirical Study of Overfitting in Deep Learning for Predicting Breast Cancer Metastasis. Cancers [online]. 1 April 2023. Vol. 15, no. 7, p. 1969. [Accessed 24 May 2025]. DOI 10.3390/CANCERS15071969/S1. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10093528/

153. YING, Xue. An Overview of Overfitting and its Solutions. Journal of Physics: Conference Series. 12 March 2019. Vol. 1168, no. 2. DOI 10.1088/1742-6596/1168/2/022022.

154. ZHU, Yanming, YIN, Xuefei, WEE, Alan, LIEW, Chung and TIAN, Hui. Privacy-Preserving in Medical Image Analysis: A Review of Methods and Applications. [online]. 5 December 2024. [Accessed 25 May 2025]. Available from: https://arxiv.org/pdf/2412.03924v1

155. SURI, Abhinav and SUMMERS, Ronald M. Privacy, Please: Safeguarding Medical Data in Imaging AI Using Differential Privacy Techniques. Radiology: Artificial Intelligence [online]. 1 January 2024. Vol. 6, no. 1, p. e230560. [Accessed 25 May 2025]. DOI 10.1148/RYAI.230560. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10831504/

156. RUPP, Valentin and VON GRAFENSTEIN, Max. Clarifying “personal data” and the role of anonymisation in data protection law: Including and excluding data from the scope of the GDPR (more clearly) through refining the concept of data protection. Computer Law & Security Review [online]. 1 April 2024. Vol. 52, p. 105932. [Accessed 25 May 2025]. DOI 10.1016/J.CLSR.2023.105932. Available from: https://www.sciencedirect.com/science/article/pii/S0267364923001425

157. Anonymization and GDPR compliance; an overview - GDPR Summary. [online]. [Accessed 25 May 2025]. Available from: https://www.gdprsummary.com/anonymization-and-gdpr/

158. STEEG, Katharina, BOHRER, Evelyn, SCHÄFER, Stefan Benjamin, VU, Viet Duc, SCHERBERICH, Jan, WINDFELDER, Anton George and KROMBACH, Gabriele Anja. Re-identification of anonymised MRI head images with publicly available software: investigation of the current risk to patient privacy. eClinicalMedicine [online]. 1 December 2024. Vol. 78, p. 102930. [Accessed 25 May 2025]. DOI 10.1016/J.ECLINM.2024.102930. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11617779/

159. GAGGION, Nicolás, ECHEVESTE, Rodrigo, MANSILLA, Lucas, MILONE, Diego H. and FERRANTE, Enzo. Unsupervised bias discovery in medical image segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) [online]. 1 September 2023. Vol. 14242 LNCS, p. 266–275. [Accessed 25 May 2025]. DOI 10.1007/978-3-031-45249-9_26. Available from: https://arxiv.org/pdf/2309.00451

160. UEDA, Daiju, KAKINUMA, Taichi, FUJITA, Shohei, KAMAGATA, Koji, FUSHIMI, Yasutaka, ITO, Rintaro, MATSUI, Yusuke, NOZAKI, Taiki, NAKAURA, Takeshi, FUJIMA, Noriyuki, TATSUGAMI, Fuminari, YANAGAWA, Masahiro, HIRATA, Kenji, YAMADA, Akira, TSUBOYAMA, Takahiro, KAWAMURA, Mariko, FUJIOKA, Tomoyuki and NAGANAWA, Shinji. Fairness of artificial intelligence in healthcare: review and

recommendations. Japanese Journal of Radiology [online]. 1 January 2023. Vol. 42, no. 1, p. 3. [Accessed 25 May 2025]. DOI 10.1007/S11604-023-01474-3. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10764412/

161. BONDI-KELLY, Elizabeth, HARTVIGSEN, Tom, SANNEMAN, Lindsay M, SANKARANARAYANAN, Swami, HARNED, Zach, WICKERSON, Grace, GICHOYA, Judy Wawira, OAKDEN-RAYNER, Lauren, CELI, Leo Anthony, LUNGREN, Matthew P, SHAH, Julie A and GHASSEMI, Marzyeh. Taking Off with AI: Lessons from Aviation for Healthcare. In : Equity and Access in Algorithms, Mechanisms, and Optimization. New York, NY, USA : ACM, 30 October 2023. p. 1–14. ISBN 9798400703812.

162. Multi-Atlas Labeling Beyond the Cranial Vault - Workshop and Challenge - syn3193805 - Wiki. [online]. [Accessed 13 May 2025]. Available from: https://www.synapse.org/Synapse:syn3193805/wiki/89480

163. Deep Learning - MATLAB & Simulink. [online]. [Accessed 24 May 2024]. Available from: https://ch.mathworks.com/discovery/deep-learning.html

164. What is a Neural Network? | IBM. [online]. [Accessed 24 May 2024]. Available from: https://www.ibm.com/topics/neural-networks

165. JADON, Shruti. A survey of loss functions for semantic segmentation. 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2020. 27 October 2020. DOI 10.1109/CIBCB48159.2020.9277638.

166. HUGHES, Chris. A Brief Overview of Cross Entropy Loss. [online]. 25 September 2024. [Accessed 10 December 2024]. Available from: https://medium.com/@chris.p.hughes10/a-brief-overview-of-cross-entropy-loss-523aa56b75d5

167. NAGENDRAM, Sanam, SINGH, Arunendra, HARISH BABU, Gade, JOSHI, Rahul, PANDE, Sandeep Dwarkanath, AHAMMAD, S. K.Hasane, DHABLIYA, Dharmesh and BISHT, Aadarsh. Stochastic gradient descent optimisation for convolutional neural network for medical image segmentation. Open Life Sciences [online]. 1 January 2023. Vol. 18, no. 1, p. 20220665. [Accessed 14 May 2025]. DOI 10.1515/BIOL-2022-0665. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10426722/

168. nnU-Net for PyTorch | NVIDIA NGC. [online]. 2022. [Accessed 10 December 2024]. Available from: https://catalog.ngc.nvidia.com/orgs/nvidia/teams/dle/resources/nnunet_pyt

169. HE, Kaiming, ZHANG, Xiangyu, REN, Shaoqing and SUN, Jian. Deep Residual Learning for Image Recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition [online]. 10 December 2015. Vol. 2016-December, p. 770–778. [Accessed 15 May 2025]. DOI 10.1109/CVPR.2016.90. Available from: https://arxiv.org/pdf/1512.03385

170. CHEN, Sirui, ZHAO, Shengjie and LAN, Quan. Residual Block Based Nested U-Type Architecture for Multi-Modal Brain Tumor Image Segmentation. Frontiers in Neuroscience [online]. 9 March 2022. Vol. 16, p. 832824. [Accessed 15 May 2025]. DOI 10.3389/FNINS.2022.832824/BIBTEX. Available from: www.frontiersin.org

171. JONNALA, Naga Surekha, SIRAAJ, Shaik, PRASTUTI, Y., CHINNABABU, P., PRAVEEN BABU, B., BANSAL, Shonak, UPADHYAYA, Prashant, PRAKASH, Krishna, FARUQUE, Mohammad Rashed Iqbal and AL-MUGREN, K. S. AER U-Net: attention-enhanced multi-scale residual U-Net structure for water body segmentation using Sentinel-2 satellite images. Scientific Reports [online]. 1 December 2025. Vol. 15, no. 1, p. 1–12. [Accessed 15 May 2025]. DOI

https://doi.org/10.1038/s41598-025-99322-z. Available from: https://www.nature.com/articles/s41598-025-99322-z

172. PEÑARROYA, Pelayo, CENTUORI, Simone, SANJURJO, Manuel and HERMOSÍN, Pablo. A LiDAR-less approach to autonomous hazard detection and avoidance systems based on semantic segmentation. Celestial Mechanics and Dynamical Astronomy [online]. 1 June 2023. Vol. 135, no. 3, p. 1–26. [Accessed 15 May 2025]. DOI 10.1007/S10569-023-10140-9/FIGURES/25. Available from: https://link.springer.com/article/10.1007/s10569-023-10140-9

173. LIU, Ze, LIN, Yutong, CAO, Yue, HU, Han, WEI, Yixuan, ZHANG, Zheng, LIN, Stephen and GUO, Baining. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Proceedings of the IEEE International Conference on Computer Vision [online]. 25 March 2021. P. 9992–10002. [Accessed 15 May 2025]. DOI 10.1109/ICCV48922.2021.00986. Available from: https://arxiv.org/pdf/2103.14030

174. NIKOLOV, Stanislav, BLACKWELL, Sam, ZVEROVITCH, Alexei, MENDES, Ruheena, LIVNE, Michelle, DE FAUW, Jeffrey, PATEL, Yojan, MEYER, Clemens, ASKHAM, Harry, ROMERA-PAREDES, Bernardino, KELLY, Christopher, KARTHIKESALINGAM, Alan, CHU, Carlton, CARNELL, Dawn, BOON, Cheng, D'SOUZA, Derek, MOINUDDIN, Syed Ali, GARIE, Bethany, MCQUINLAN, Yasmin, IRELAND, Sarah, HAMPTON, Kiarna, FULLER, Krystle, MONTGOMERY, Hugh, REES, Geraint, SULEYMAN, Mustafa, BACK, Trevor, HUGHES, Cían O, LEDSAM, Joseph R and RONNEBERGER, Olaf. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. [online]. 12 September 2018. [Accessed 6 January 2024]. Available from: https://arxiv.org/abs/1809.04430v3

175. RHEE, Dong Joo, AKINFENWA, Chidinma P.Anakwenze, RIGAUD, Bastien, JHINGRAN, Anuja, CARDENAS, Carlos E., ZHANG, Lifei, PRAJAPATI, Surendra, KRY, Stephen F., BROCK, Kristy K., BEADLE, Beth M., SHAW, William, O'REILLY, Frederika, PARKES, Jeannette, BURGER, Hester, FAKIE, Nazia, TRAUERNICHT, Chris, SIMONDS, Hannah and COURT, Laurence E. Automatic contouring QA method using a deep learning–based autocontouring system. Journal of Applied Clinical Medical Physics [online]. 1 August 2022. Vol. 23, no. 8, p. e13647. [Accessed 16 May 2025]. DOI 10.1002/ACM2.13647. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC9359039/

176. BAID, U., GHODASARA, S., MOHAN, S., BILELLO, M., CALABRESE, E., COLAK, E., FARAHANI, K., KALPATHY-CRAMER, J., KITAMURA, F. C., PATI, S. and BAKAS, S. RSNA-ASNR-MICCAI-BraTS-2021 - The Cancer Imaging Archive (TCIA). 2023. The Cancer Imaging Archive. 1.

177. Novel Transformer Model Achieves State-of-the-Art Benchmarks in 3D Medical Image Analysis | NVIDIA Technical Blog. [online]. [Accessed 26 May 2025]. Available from: https://developer.nvidia.com/blog/novel-transformer-model-achieves-state-of-the-art-benchmarks-in-3d-medical-image-analysis/

178. TANG, Yucheng, YANG, Dong, LI, Wenqi, ROTH, Holger R., LANDMAN, Bennett, XU, Daguang, NATH, Vishwesh and HATAMIZADEH, Ali. Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition [online]. 29 November 2021. Vol. 2022-June, p. 20698–20708. [Accessed 26 May 2025]. DOI 10.1109/CVPR52688.2022.02007. Available from: https://arxiv.org/pdf/2111.14791

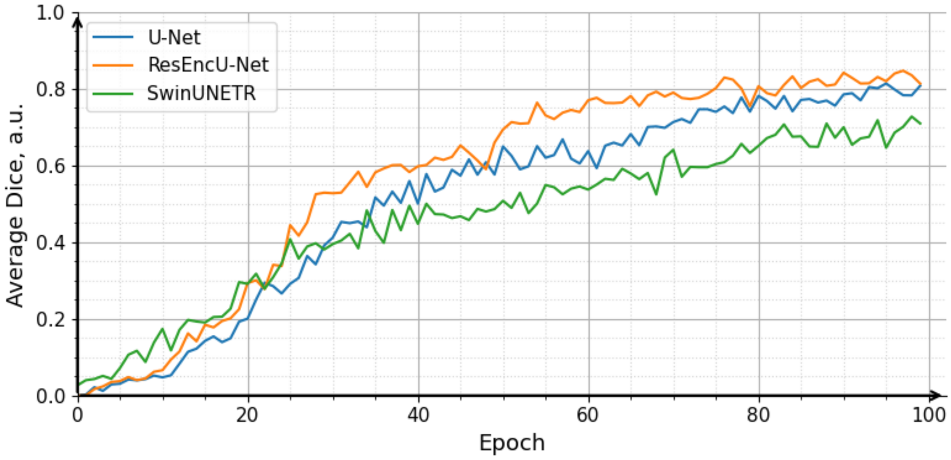
179. LEE, Soyeon and LEE, Minhyeok. MetaSwin: a unified meta vision transformer model for medical image segmentation. PeerJ Computer Science [online]. 2024. Vol. 10, p. e1762.

[Accessed 26 May 2025]. DOI 10.7717/PEERJ-CS.1762. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10773825/

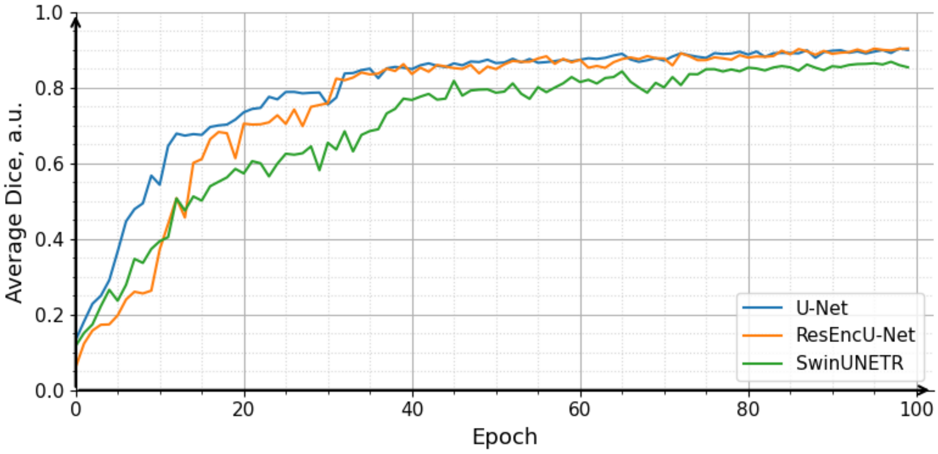
Appendices

Appendix 1. Average validation Dice values during the training process

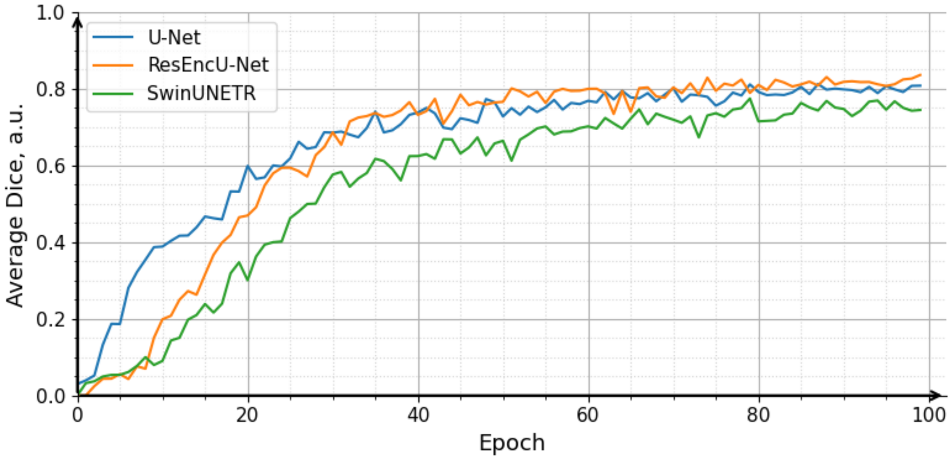
AMOS



KC



BTCV



Appendix 2. Statistical test results for metric comparison across models (Kruskall-Wallis test)

Dataset	Metric	Kruskal-Wallis H	p-value
AMOS	Dice	311.7782	1.99E-68
AMOS	sDice_0mm	593.1072	1.62E-129
AMOS	sDice_2mm	589.8473	8.25E-129
AMOS	HD95	524.8552	1.07E-114
BTCV	Dice	4.260828	0.118788081
BTCV	sDice_0mm	11.97084	0.002515152
BTCV	sDice_2mm	13.22148	0.001345839
BTCV	HD95	14.68777	0.000646532
KC	Dice	6.908314	0.031613951
KC	sDice_0mm	7.429324	0.024363677
KC	sDice_2mm	11.83322	0.002694318
KC	HD95	23.39196	8.33E-06

Appendix 3. Post-hoc test results for metric pair-wise comparison across models (Mann-Whitney U test with Bonferroni correction)

Dataset	Metric	Model 1	Model 2	Raw p-value	Corrected p-value	Significant
AMOS	Dice	UNET	ResEnc	2.20E-07	6.59E-07	TRUE
AMOS	Dice	UNET	SwinUNETR	3.85E-36	1.15E-35	TRUE
AMOS	Dice	ResEnc	SwinUNETR	1.21E-63	3.64E-63	TRUE
AMOS	sDice_0mm	UNET	ResEnc	7.89E-13	2.37E-12	TRUE
AMOS	sDice_0mm	UNET	SwinUNETR	5.79E-71	1.74E-70	TRUE
AMOS	sDice_0mm	ResEnc	SwinUNETR	1.03E-115	3.10E-115	TRUE
AMOS	sDice_2mm	UNET	ResEnc	8.28E-13	2.48E-12	TRUE
AMOS	sDice_2mm	UNET	SwinUNETR	6.77E-71	2.03E-70	TRUE
AMOS	sDice_2mm	ResEnc	SwinUNETR	9.23E-115	2.77E-114	TRUE
AMOS	HD95	UNET	ResEnc	4.01E-13	1.20E-12	TRUE
AMOS	HD95	UNET	SwinUNETR	7.29E-58	2.19E-57	TRUE
AMOS	HD95	ResEnc	SwinUNETR	4.87E-106	1.46E-105	TRUE
BTCV	sDice_0mm	UNET	ResEnc	0.407205	1	FALSE
BTCV	sDice_0mm	UNET	SwinUNETR	0.010162	0.030485	TRUE
BTCV	sDice_0mm	ResEnc	SwinUNETR	0.001135	0.003405	TRUE
BTCV	sDice_2mm	UNET	ResEnc	0.394156	1	FALSE
BTCV	sDice_2mm	UNET	SwinUNETR	0.007216	0.021649	TRUE
BTCV	sDice_2mm	ResEnc	SwinUNETR	0.00059	0.001769	TRUE
BTCV	HD95	UNET	ResEnc	0.421751	1	FALSE
BTCV	HD95	UNET	SwinUNETR	0.004179	0.012536	TRUE
BTCV	HD95	ResEnc	SwinUNETR	0.000297	0.00089	TRUE
KC	Dice	UNET	ResEnc	0.791268	1	FALSE
KC	Dice	UNET	SwinUNETR	0.026895	0.080685	FALSE
KC	Dice	ResEnc	SwinUNETR	0.021271	0.063813	FALSE
KC	sDice_0mm	UNET	ResEnc	0.905866	1	FALSE
KC	sDice_0mm	UNET	SwinUNETR	0.019443	0.058329	FALSE
KC	sDice_0mm	ResEnc	SwinUNETR	0.018302	0.054906	FALSE
KC	sDice_2mm	UNET	ResEnc	0.73971	1	FALSE
KC	sDice_2mm	UNET	SwinUNETR	0.004154	0.012461	TRUE
KC	sDice_2mm	ResEnc	SwinUNETR	0.002316	0.006947	TRUE
KC	HD95	UNET	ResEnc	0.608663	1	FALSE
KC	HD95	UNET	SwinUNETR	7.39E-05	0.000222	TRUE
KC	HD95	ResEnc	SwinUNETR	1.49E-05	4.47E-05	TRUE

Appendix 4. Statistical test results for sDSC metric pair-wise model comparison across different organs (Kruskall-Wallis test)

Organ	Kruskal-Wallis H	p-value
Bladder	41.23761	1.11E-09
Bones	0.8	0.67032
Heart	5.422222	0.066463
Kidney_L	90.5216	2.21E-20
Kidney_R	69.58829	7.75E-16
Lung_L	5.955556	0.050906
Lung_R	5.422222	0.066463
Thyroid	1.688889	0.429796
Stomach	90.69046	2.03E-20
Liver	50.72371	9.67E-12
Spleen	69.67526	7.42E-16
Esophagus	60.73567	6.48E-14
Gallbladder	48.09319	3.60E-11
Aorta	64.62585	9.26E-15
Vena_cava	77.29717	1.64E-17
Pancreas	85.89256	2.23E-19
Adrenal_R	44.57042	2.10E-10
Adrenal_L	95.50554	1.82E-21
Duodenum	84.23617	5.11E-19
Prostate/Uterus	43.69526	3.25E-10
Portal_vein	3.38	0.18452

Appendix 5. Post-hoc analysis (Mann-Whitney U test with Bonferroni correction)

Organ	Model 1	Model 2	Raw p-value	Corrected p-value	Significant
Bladder	UNET	ResEnc	0.060984	0.182952	FALSE
Bladder	UNET	SwinUNETR	2.68E-06	8.05E-06	TRUE
Bladder	ResEnc	SwinUNETR	1.79E-09	5.36E-09	TRUE
Kidney_L	UNET	ResEnc	1.63E-06	4.90E-06	TRUE
Kidney_L	UNET	SwinUNETR	2.80E-11	8.39E-11	TRUE
Kidney_L	ResEnc	SwinUNETR	3.38E-17	1.01E-16	TRUE
Kidney_R	UNET	ResEnc	0.000112	0.000336	TRUE
Kidney_R	UNET	SwinUNETR	3.84E-09	1.15E-08	TRUE
Kidney_R	ResEnc	SwinUNETR	6.76E-14	2.03E-13	TRUE
Stomach	UNET	ResEnc	0.000978	0.002933	TRUE
Stomach	UNET	SwinUNETR	9.72E-14	2.92E-13	TRUE
Stomach	ResEnc	SwinUNETR	3.38E-17	1.01E-16	TRUE
Liver	UNET	ResEnc	0.04694	0.140821	FALSE
Liver	UNET	SwinUNETR	2.91E-08	8.72E-08	TRUE
Liver	ResEnc	SwinUNETR	1.17E-10	3.50E-10	TRUE
Spleen	UNET	ResEnc	0.001089	0.003267	TRUE
Spleen	UNET	SwinUNETR	1.47E-09	4.41E-09	TRUE
Spleen	ResEnc	SwinUNETR	2.74E-14	8.22E-14	TRUE
Esophagus	UNET	ResEnc	0.075083	0.22525	FALSE
Esophagus	UNET	SwinUNETR	3.72E-09	1.12E-08	TRUE
Esophagus	ResEnc	SwinUNETR	3.45E-13	1.03E-12	TRUE
Gallbladder	UNET	ResEnc	0.231138	0.693414	FALSE
Gallbladder	UNET	SwinUNETR	5.86E-08	1.76E-07	TRUE
Gallbladder	ResEnc	SwinUNETR	1.26E-10	3.78E-10	TRUE
Aorta	UNET	ResEnc	0.097389	0.292167	FALSE
Aorta	UNET	SwinUNETR	6.85E-10	2.05E-09	TRUE
Aorta	ResEnc	SwinUNETR	7.82E-14	2.34E-13	TRUE
Vena_cava	UNET	ResEnc	0.016909	0.050726	FALSE
Vena_cava	UNET	SwinUNETR	3.26E-12	9.79E-12	TRUE
Vena_cava	ResEnc	SwinUNETR	3.31E-15	9.94E-15	TRUE
Pancreas	UNET	ResEnc	0.029668	0.089004	FALSE
Pancreas	UNET	SwinUNETR	1.32E-14	3.95E-14	TRUE
Pancreas	ResEnc	SwinUNETR	5.30E-16	1.59E-15	TRUE
Adrenal_R	UNET	ResEnc	0.010579	0.031736	TRUE
Adrenal_R	UNET	SwinUNETR	1.82E-06	5.45E-06	TRUE
Adrenal_R	ResEnc	SwinUNETR	7.93E-10	2.38E-09	TRUE
Adrenal_L	UNET	ResEnc	4.43E-05	0.000133	TRUE

Adrenal_L	UNET	SwinUNETR	1.47E-12	4.40E-12	TRUE
Adrenal_L	ResEnc	SwinUNETR	1.11E-18	3.32E-18	TRUE
Duodenum	UNET	ResEnc	0.035546	0.106639	FALSE
Duodenum	UNET	SwinUNETR	6.33E-13	1.90E-12	TRUE
Duodenum	ResEnc	SwinUNETR	5.81E-17	1.74E-16	TRUE
Prostate/Uterus	UNET	ResEnc	0.052456	0.157368	FALSE
Prostate/Uterus	UNET	SwinUNETR	8.86E-07	2.66E-06	TRUE
Prostate/Uterus	ResEnc	SwinUNETR	8.91E-10	2.67E-09	TRUE