Norwegian University of Life Sciences
Faculty of Science and Technology

Philosophiae Doctor (PhD)
Thesis 2023:12

Assessment of machine learning methods for automatic tumor segmentation

Evaluering av maskinlæringsmetoder for automatisk tumorsegmentering

Aurora Rosvoll Grøndahl

Assessment of machine learning methods for automatic tumor segmentation

Evaluering av maskinlæringsmetoder for automatisk tumorsegmentering

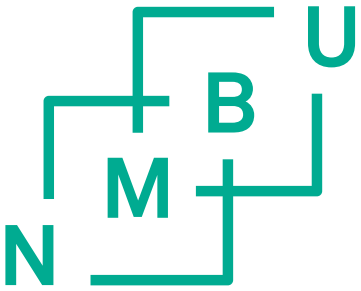
Philosophiae Doctor (PhD) Thesis

Aurora Rosvoll Grøndahl

Norwegian University of Life Sciences

Faculty of Science and Technology

Ås, 2023



Thesis number: 2023:12
ISSN: 1894-6402
ISBN: 978-82-575-2041-0

Abstract

The definition of target volumes and organs at risk (OARs) is a critical part of radiotherapy planning. In routine practice, this is typically done manually by clinical experts who contour the structures in medical images prior to dosimetric planning. This is a time-consuming and labor-intensive task. Moreover, manual contouring is inherently a subjective task and substantial contour variability can occur, potentially impacting on radiotherapy treatment and image-derived biomarkers. Automatic segmentation (auto-segmentation) of target volumes and OARs has the potential to save time and resources while reducing contouring variability. Recently, auto-segmentation of OARs using machine learning methods has been integrated into the clinical workflow by several institutions and such tools have been made commercially available by major vendors. The use of machine learning methods for auto-segmentation of target volumes including the gross tumor volume (GTV) is less mature at present but is the focus of extensive ongoing research.

The primary aim of this thesis was to investigate the use of machine learning methods for auto-segmentation of the GTV in medical images. Manual GTV contours constituted the ground truth in the analyses. Volumetric overlap and distance-based metrics were used to quantify auto-segmentation performance. Four different image datasets were evaluated. The first dataset, analyzed in papers I–II, consisted of positron emission tomography (PET) and contrast-enhanced computed tomography (ceCT) images of 197 patients with head and neck cancer (HNC). The ceCT images of this dataset were also included in paper IV. Two datasets were analyzed separately in paper III, namely (i) PET, ceCT, and low-dose CT (ldCT) images of 86 patients with anal cancer (AC), and (ii) PET, ceCT, ldCT, and $T2$ and diffusion-weighted (T2W and DW, respectively) MR images of a subset ($n = 36$) of the aforementioned AC patients. The last dataset consisted of ceCT images of 36 canine patients with HNC and was analyzed in paper IV.

In paper I, three approaches to auto-segmentation of the GTV in patients with HNC were evaluated and compared, namely conventional PET thresholding, classical machine learning algorithms, and deep learning using a 2-dimensional (2D) U-Net convolutional neural network (CNN). For the latter two approaches the effect of imaging modality on auto-segmentation performance was also assessed. Deep learning based on multimodality PET/ceCT image input resulted in superior agreement with the manual ground truth contours, as quantified by geometric overlap and distance-based performance evaluation metrics calculated on a per patient basis. Moreover, only deep learning provided adequate performance for segmenta-

tion based solely on ceCT images. For segmentation based on PET-only, all three approaches provided adequate segmentation performance, though deep learning ranked first, followed by classical machine learning, and PET thresholding. In paper II, deep learning-based auto-segmentation of the GTV in patients with HNC using a 2D U-Net architecture was evaluated more thoroughly by introducing new structure-based performance evaluation metrics and including qualitative expert evaluation of the resulting auto-segmentation quality. As in paper I, multimodal PET/ceCT image input provided superior segmentation performance, compared to the single modality CNN models. The structure-based metrics showed quantitatively that the PET signal was vital for the sensitivity of the CNN models, as the superior PET/ceCT-based model identified 86 % of all malignant GTV structures whereas the ceCT-based model only identified 53 % of these structures. Furthermore, the majority of the qualitatively evaluated auto-segmentations ($\sim$ 90 %) generated by the best PET/ceCT-based CNN were given a quality score corresponding to substantial clinical value. Based on papers I and II, deep learning with multimodality PET/ceCT image input would be the recommended approach for auto-segmentation of the GTV in human patients with HNC.

In paper III, deep learning-based auto-segmentation of the GTV in patients with AC was evaluated for the first time, using a 2D U-Net architecture. Furthermore, an extensive comparison of the impact of different single modality and multimodality combinations of PET, ceCT, ldCT, T2W, and/or DW image input on quantitative auto-segmentation performance was conducted. For both the 86-patient and 36-patient datasets, the models based on PET/ceCT provided the highest mean overlap with the manual ground truth contours. For this task, however, comparable auto-segmentation quality was obtained for solely ceCT-based CNN models. The CNN model based solely on T2W images also obtained acceptable auto-segmentation performance and was ranked as the second-best single modality model for the 36-patient dataset. These results indicate that deep learning could prove a versatile future tool for auto-segmentation of the GTV in patients with AC.

Paper IV investigated for the first time the applicability of deep learning-based auto-segmentation of the GTV in canine patients with HNC, using a 3-dimensional (3D) U-Net architecture and ceCT image input. A transfer learning approach where CNN models were pre-trained on the human HNC data and subsequently fine-tuned on canine data was compared to training models from scratch on canine data. These two approaches resulted in similar auto-segmentation performances, which on average was comparable to the overlap metrics obtained for ceCT-based auto-segmentation in human HNC patients. Auto-segmentation in canine HNC patients appeared particularly promising for nasal cavity tumors, as the average overlap with manual contours was 25 % higher for this subgroup, compared to the average for all included tumor sites.

In conclusion, deep learning with CNNs provided high-quality GTV auto-segmentations for all datasets included in this thesis. In all cases, the best-performing deep learning models resulted in an average overlap with manual contours which was comparable to the reported interobserver agreements between human experts performing manual GTV contouring for the given cancer type and imaging modality. Based on these findings, further investigation of deep learning-based auto-segmentation of the GTV in the given diagnoses would be highly warranted.

Sammendrag

Definisjon av målvolum og risikoorganer er en kritisk del av planleggingen av strålebehandling. I praksis gjøres dette vanligvis manuelt av kliniske eksperter som tegner inn strukturenes konturer i medisinske bilder før dosimetrisk planlegging. Dette er en tids- og arbeidskrevende oppgave. Manuell inntegning er også subjektiv, og betydelig variasjon i inntegnede konturer kan forekomme. Slik variasjon kan potensielt påvirke strålebehandlingen og bildebaserte biomarkører. Automatisk segmentering (auto-segmentering) av målvolum og risikoorganer kan potensielt spare tid og ressurser samtidig som konturvariasjonen reduseres. Auto-segmentering av risikoorganer ved hjelp av maskinlæringsmetoder har nylig blitt implementert som del av den kliniske arbeidsflyten ved flere helseinstitusjoner, og slike verktøy er kommersielt tilgjengelige hos store leverandører av medisinsk teknologi. Auto-segmentering av målvolum inkludert tumorvolumet *gross tumor volume* (GTV) ved hjelp av maskinlæringsmetoder er per i dag mindre teknologisk modent, men dette området er fokus for omfattende pågående forskning.

Hovedmålet med denne avhandlingen var å undersøke bruken av maskinlæringsmetoder for auto-segmentering av GTV i medisinske bilder. Manuelle GTV-inntegninger utgjorde grunnsannheten (*the ground truth*) i analysene. Mål på volumetrisk overlapp og avstand mellom sanne og predikerte konturer ble brukt til å kvantifisere kvaliteten til de automatisk genererte GTV-konturene. Fire forskjellige bildedatasett ble evaluert. Det første datasettet, analysert i artikkel I–II, bestod av positronemisjonstomografi (PET) og kontrastforsterkede computertomografi (ceCT) bilder av 197 pasienter med hode/halskreft. ceCT-bildene i dette datasettet ble også inkludert i artikkel IV. To datasett ble analysert separat i artikkel III, nemlig (i) PET, ceCT og lavdose CT (ldCT) bilder av 86 pasienter med analkreft, og (ii) PET, ceCT, ldCT og $T2$- og diffusjonsvektet (henholdsvis T2W og DW) MR-bilder av en undergruppe ($n = 36$) av de ovennevnte analkreft-pasientene. Det siste datasettet, som bestod av ceCT-bilder av 36 hunder med hode/halskreft, ble analysert i artikkel IV.

I artikkel I ble følgende tre tilnærminger til auto-segmentering av GTV evaluert og sammenlignet for humane pasienter med hode/halskreft: (i) konvensjonell PET-terskling, (ii) klassiske maskinlæringsalgoritmer og (iii) dyp læring ved bruk av et 2-dimensjonalt (2D) U-Net konvolusjonelt nevralt nettverk (CNN). For de to sistnevnte tilnærmingene ble effekten av bildemodalitet på auto-segmenteringsytelsen også undersøkt. Dyp læring basert på multimodale PET/ceCT-bilder resulterte i signifikant bedre samsvar med de manuelle GTV-konturene, sammenlignet med de øvrige tilnærmingene. Videre resulterte dyp læring i akseptabel segmenter-

ingsytelse kun basert på ceCT-bilder. For segmentering kun basert på PET ga alle tre tilnærmingene adekvat segmenteringsytelse, selv om dyp læring ble rangert som den beste tilnærmingen, etterfulgt av klassisk maskinlæring og PET-terskling. I artikkel II ble auto-segmentering av GTV hos humane pasienter med hode/halskreft ved bruk av en 2D U-Net-arkitektur evaluert mer grundig ved å introdusere nye strukturbaserte ytelsesmål og inkludere kvalitativ ekspert-evaluering av de automatisk genererte GTV-konturene. Som i artikkel I ga CNN-modellen basert på multimodale PET/ceCT-bilder den beste segmenteringsytelsen, sammenlignet med CNN-modeller basert på enten ceCT eller PET-bilder. De strukturbaserte ytelsesmålene viste kvantitativt at PET-signalet var avgjørende for sensitiviteten til CNN-modellene, ettersom den høyest rangerte PET/ceCT-baserte modellen identifiserte 86 % av alle GTV-strukturene, mens den ceCT-baserte modellen bare identifiserte 53 % av disse strukturene. Videre ble majoriteten av de kvalitativt evaluerte auto-segmenteringene ($\sim$ 90 %) generert av den høyest rangerte PET/ceCT-baserte CNN-modellen gitt en kvalitetsskår tilsvarende betydelig klinisk verdi. Basert på artikkel I og II, er dyp læring med kombinert PET/ceCT-bildedata den anbefalte tilnærmingen til auto-segmentering av GTV hos humane pasienter med hode/halskreft.

I artikkel III ble dyp læring-basert auto-segmentering av GTV hos pasienter med analkreft evaluert for første gang ved bruk av en 2D U-Net CNN-arkitektur. Videre ble det utført en omfattende sammenligning av effekten ulike enkeltmodaliteter og multimodalitets-kombinasjoner av PET, ceCT, ldCT, T2W og/eller DW bildedata har på den kvantitative auto-segmenteringsytelse. For både 86-pasient og 36-pasient-datasettene ga modellene basert på PET/ceCT høyest gjennomsnittlige overlapp med de manuelle GTV-konturene. I disse analysene ble det imidlertid oppnådd sammenlignbar auto-segmenteringskvalitet for utelukkende ceCT-baserte CNN-modeller. CNN-modellen basert utelukkende på T2W-bilder oppnådde også akseptabel auto-segmenteringsytelse og ble rangert som den nest beste enkelt-modalitets-modellen for datasettet med 36 pasienter. Disse resultatene indikerer at dyp læring potensielt kan være et allsidig fremtidig verktøy for auto-segmentering av GTV hos pasienter med analkreft.

I artikkel IV ble dyp læring-basert auto-segmentering av GTV hos hunder med hode/halskreft evaluert for første gang ved å anvende en 3-dimensjonal (3D) U-Net CNN-arkitektur og ceCT-bildedata. En overføringslæring-tilnærming der CNN-modeller ble forhåndstrent på humane hode/halskreft data og deretter finjustert på hundedata ble sammenlignet med å trene modeller fra grunnen av på hundedata. Disse to tilnærmingene resulterte i lignende auto-segmenteringsytelser, som i gjennomsnitt var sammenlignbare med ceCT-basert auto-segmentering hos humane hode/halskreft-pasienter. Auto-segmentering av hode/halskreft i hund virket spesielt lovende for svulster i nesehulen, ettersom gjennomsnittlig overlapp med manuelle GTV-konturer var 25 % høyere for denne undergruppen, sammenlignet med gjennomsnittet for alle inkluderte tumorlokaliteter.

Dyp læring-basert auto-segmentering resulterte i GTV-konturer av høy kvalitet for alle datasett inkludert i denne avhandlingen. I alle tilfeller resulterte de høyest rangerte dyplæringsmodellene i gjennomsnittlig overlapp med manuelle konturer som var sammenlignbar med rapporterte tall for overensstemmelsen mellom menneskelige eksperter som utfører manuell GTV-inntegning for den gitte krefttypen og avbildningsmodaliteten. Basert på disse funnene vil ytterligere undersøkelse av dyp læring-basert auto-segmentering av GTV i de gitte diagnosene være svært berettiget.

Acknowledgements

The work presented in this thesis was conducted at the Faculty of Science and Technology at the Norwegian University of Life Sciences (NMBU), in collaboration with partners at the University of Oslo, Oslo University Hospital and the Faculty of Veterinary Medicine at NMBU. The funding for the PhD position was provided by NMBU, for which I am grateful.

I would like to thank all my supervisors Cecilia Marie Futsæther, Eirik Malinen, Oliver Tomic, Hege Kippenes Skogmo, Åste Søvik and Ulf Geir Indahl for their support and their valuable and thorough feedback. A special thanks to my main supervisor Cecilia for her enthusiastic and supportive guidance throughout this project. This journey would not have been the same without her. She has been the best supervisor I could have hoped for and has always been willing to discuss minor and major aspects related to both scientific work and teaching duties.

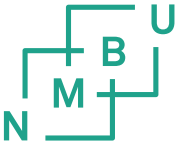
This work would not have been possible without the valuable contributions of all my co-authors, and everyone involved in data collection, and I thank them all. A special thanks to Bao Ngoc Huynh and Yngve Mardal Moe, for always being willing to share their extensive knowledge on machine learning and programming. I would also like to express my thankfulness to all the talented master students who have been part of our research group during my PhD.

A huge thanks to my present and past office mates at NMBU. Dear Maren Anna, Johanne and Eivind, thanks for adopting me when I first started my PhD! I always smile when I look back at the first year we teached FYS103 together. A special thanks to Maren Anna for always listening when I needed someone and sharing useful PhD life hacks.

Finally, I would like to thank my dear family and friends for their support and love throughout this process, and last but not least, a special thanks to Endre for always being so patient, encouraging and supportive.

Aurora Rosvoll Grøndahl

Ås, January 2023



Norwegian University
of Life Sciences

Contents

List of papers

Paper I

[1] Groendahl AR, Knudtsen IS, Huynh BN, Mulstad M, Moe YM, Knuth F, Tomic O, Indahl UG, Torheim T, Dale E, Malinen E, and Futsaether CM. A comparison of methods for fully automatic segmentation of tumors and involved nodes in PET/CT of head and neck cancers. *Physics in Medicine and Biology*, 6:065012, 2021

Paper II

[2] Moe YM, Groendahl AR, Tomic O, Dale E, Malinen E, and Futsaether CM. Deep learning-based auto-delineation of gross tumour volumes and involved nodes in PET/CT images of head and neck cancer patients. *European Journal of Nuclear Medicine and Molecular Imaging*, 9:2782–2792, 2021

Paper III

[3] Groendahl AR, Moe YM, Kaushal CK, Huynh BN, Rusten E, Tomic O, Hernes E, Hanekamp B, Undseth C, Guren MG, Malinen E, and Futsaether CM. Deep learning-based automatic delineation of anal cancer gross tumour volume: a multimodality comparison of CT, PET and MRI. *Acta Oncologica*, 61:1:89–96, 2021

Paper IV

[4] Groendahl AR, Huynh BN, Tomic O, Søvik Å, Dale E, Malinen E, Skogmo HK, and Futsaether CM. Automatic gross tumor segmentation of canine head and neck cancer using deep learning and cross-species transfer learning. To be submitted to *Frontiers in Veterinary Science*

Additional scientific work

Papers

[5] Knuth F, Adde IA, Huynh BN, Groendahl AR, Winter RM, Negård A, Holmedal SH, Meltzer S, Ree AH, Flatmark K, Dueland S, Hole KH, Seierstad T, Redalen KR, and Futsaether CM. MRI-based automatic segmentation of rectal cancer using 2D U-Net on two independent cohorts. *Acta Oncologica*, 61:2:255–263, 2022

[6] Knuth F, Groendahl AR, Winter RM, Torheim T, Negård A, Holmedal SH, Bakke KM, Meltzer S, Futsaether CM, and Redalen KR. Semi-automatic tumor segmentation of rectal cancer based on functional magnetic resonance imaging. *Physics and imaging in radiation oncology*, 22:77–84, 2022

Popular science papers

[7] Grøndahl AR, Kuttner S, Elschot M, Bathen TF, Sundset R, Knudtsen IS, and Futsæther CM. Maskinlæring og medisinske bilder for bedre diagnostikk og persontilpasset kreftbehandling. *HMT*, volume 5, 2018. Published in Norwegian.

Conference proceedings

[8] Ren J, Huynh BN, Groendahl AR, Tomic O, Futsaether CM, and Korreman SS. PET normalizations to improve deep learning auto-segmentation of head and neck tumors in 3D PET/CT. In *Head and Neck Tumor Segmentation and outcome prediction*. HECKTOR 2021. Lecture Notes in Computer Science, vol 13209. Springer, 2022

[9] Huynh BN, Ren J, Groendahl AR, Tomic O, Korreman SS, and Futsaether CM. Comparing deep learning and conventional machine learning for outcome prediction of head and neck cancer in PET/CT. In *Head and Neck Tumor Segmentation and outcome prediction*. HECKTOR 2021. Lecture Notes in Computer Science, vol 13209. Springer, 2022

Oral presentations

⋆ Groendahl AR, Huynh BN, Moe YM, Kaushal CK, Rusten E, Tomic O, Hernes E, Hanekamp B, Undseth C, Guren MG, Malinen E, and Futsaether CM. Deep learning-based automatic delineation of anal cancer gross tumour volume: A multimodality comparison of CT, PET and MRI. Presented at: *BiGART 2021* 2021-10-05–2021-10-06

⋆ Huynh BN, Groendahl AR, Moe YM, Tomic O, Dale E, Malinen E, and Futsaether CM. Deep learning for automatic segmentation of head and neck cancers in PET/CT images: the simpler, the better. Presented at: *BiGART 2021* 2021-10-05–2021-10-06

⋆ Knuth F, Adde IA, Huynh BN, Groendahl AR, Winter RM, Negård A, Holmedal SH, Meltzer S, Ree AH, Flatmark K, Dueland S, Holde KH, Seierstad T, Redalen KR, and Futsaether CM. MRI-based automatic segmentation of rectal cancer using 2D U-Net on two independent cohorts. Presented at: *BiGART 2021* 2021-10-05–2021-10-06

⋆ Groendahl AR, Huynh BN, Moe YM, Kaushal CK, Rusten E, Tomic O, Hernes E, Hanekamp B, Undseth C, Guren MG, Dale E, Malinen E, and Futsaether CM. Deep learning for automatic target volume delineation. Presented at: *NACP 2020/21* 2021-04-11–2021-04-13

⋆ Knuth F, Groendahl AR, Winter RM, Torheim T, Negård A, Holmedal SH, Bakke KM, Meltzer S, Futsaether CM, and Redalen K. Influence of functional MRI sequences on automatic segmentation of rectal cancer. Presented at: *NACP 2020/21* 2021-04-11–2021-04-13

⋆ Groendahl AR, Moe Y, Kaushal CK, Tomic O, Dale E, Guren MG, Malinen E, and CM Futsæther. Machine learning for automatic tumor segmentation. Presented at: *Mini-MedFys 2020* 2020-02-03

⋆ Knuth F, Groendahl AR, Torheim T, Negård A, Holmedal SH, Bakke KM, Meltzer S, Futsaether CM, and Redalen K. Functional MR-based automatic tumor segmentation of rectal cancer. Presented at: *BiGART 2019* 2019-05-22–2019-05-24

⋆ Knuth F, Groendahl AR, Torheim T, Negård A, Holmedal SH, Bakke KM, Meltzer S, Futsaether CM, and Redalen K. Automatic tumour delineation in rectal cancer using functional MRI and machine learning. Presented at: *ESTRO 2019* 2021-04-26–2021-04-30

⋆ Knuth F, Groendahl AR, Torheim T, Negård A, Holmedal SH, Bakke KM, Meltzer S, Futsaether CM, and Redalen K. Automatic tumor delineation in rectal cancer using functional MRI and machine learning. Presented at: *MedFys 2019* 2019-02-04–2019-02-06

Poster presentations

⋆ Groendahl AR, Huynh BN, Moe YM, Kaushal CK, Rusten E, Tomic O, Hernes E, Hanekamp B, Undseth C, Guren MG, Malinen E, and Futsaether CM. Deep learning-based automatic delineation of anal cancer gross tumour volume: A multimodality comparison of CT, PET and MRI. Presented at: *BiGART 2021* 2021-10-06

⋆ Huynh BN, Groendahl AR, Moe YM, Tomic O, Dale E, Malinen E, and Futsaether CM. Tuning deep learning models for automatic segmentation of head and neck cancers in PET/CT images. Presented at: *ESTRO 2021* 2021-08-27–2021-08-31

⋆ Moe YM, Groendahl AR, Mulstad M, Tomic O, Indahl UG Dale E, Malinen E, and Futsaether CM. Deep learning for automatic tumour segmentation in PET/CT images of patients with head and neck cancers. Presented at: *MIDL 2019* 2019-07-08–2019-07-10

⋆ Groendahl AR, Knudtsen IS, Mulstad M, Tomic O, Moe YM, Indahl UG, Torheim T, Dale E, Malinen E, and Futsaether CM. Automatic tumour delineation of head and neck cancers in PET/CT images using thresholding and machine learning methods. Presented at: *BiGART 2019* 2019-05-22–2019-05-24

⋆ Langberg GS, Groendahl AR, Midtfjord AD, Tomic O, Liland KH, Knudtsen IS, Dale E, Malinen E, and Futsaether CM. Establishing a complete radiomics framework for biomarker identification and outcome prediction using PET/CT images of head neck cancers. Presented at: *BiGART 2019* 2019-05-22–2019-05-24

⋆ Groendahl AR, Mulstad M, Moe YM, Knudtsen IS, Torheim T, Tomic O, Indahl UG, Malinen E, Dale E, and Futsaether CM. Comparison of automatic tumour segmentation approaches for head and neck cancers in PET/CT images. Presented at: *ESTRO 2019* 2021-04-26–2021-04-30

⋆ Groendahl AR, Midtfjord AD, Langberg GS, Tomic O, Indahl UG, Knudtsen IS, Malinen E, Dale E, and Futsaether CM. Prediction of treatment outcome for head and neck cancers using radiomics of PET/CT images. Presented at: *ESTRO 2019* 2021-04-26–2021-04-30

List of abbreviations

2D	2-dimensional
3D	3-dimensional
3D-CRT	3-dimensional conformal radiotherapy
AC	anal cancer
Adam	adaptive moment estimation
ADC	apparent diffusion coefficient
ASCC	anal squamous cell carcinoma
ASD	average surface distance
auto-segmentation	automatic segmentation
ceCT	contrast-enhanced computed tomography
CNNs	convolutional neural networks
CT	computed tomography
CTV	clinical target volume
DW	diffusion-weighted
FDG	^{18}F-fluorodeoxyglucose
FN	false negative
FP	false positive
GNB	Gaussian naïve Bayes
GTV	gross tumor volume
HD	Hausdorff distance
HD_{95}	95th percentile Hausdorff distance
HECKTOR	head and neck tumor segmentation and outcome prediction
HNC	head and neck cancer
HNSCC	head and neck squamous cell carcinoma
HPV	human papillomavirus
IMRT	intensity-modulated radiotherapy
IV	intravenous
LASSO	least absolute shrinkage and selection operator
LDA	linear discriminant analysis
ldCT	low-dose computed tomography
LoG	Laplacian of Gaussian
LR	logistic regression
MICCAI	medical image computing and computer assisted intervention

MR	magnetic resonance
MRI	magnetic resonance imaging
MSD	median surface distance
OARs	organs at risk
PET	positron emission tomography
PPV	positive predictive value
PTV	planning target volume
QDA	quadratic discriminant analysis
radiotracer	radioactive tracer
ReLU	rectified linear unit
RF	random forest
SUV	standardized uptake value
SVM	support vector machines
T2W	$T2$-weighted
TN	true negative
TP	true positive
TPR	true positive rate
VMAT	volumetric modulated arch therapy

Chapter 1

Introduction and aims

Radiotherapy is one of the most widely used cancer treatments [26, 27]. The primary aim and challenge of radiotherapy is to accurately deliver a sufficiently high radiation dose to the target volume, eradicating the cancer cells within, while minimizing concurrent damage to surrounding normal tissues and organs [28, 29]. Modern advances in radiotherapy delivery allow highly conformal doses to the target volume, thereby reducing radiation-induced normal tissue toxicities [30–34]. However, optimal high-precision radiotherapy requires more accurate definition of both target volumes and critical normal tissue structures, known as organs at risk (OARs), compared to conventional radiotherapy techniques. Accurate definition of target volumes and OARs is, therefore, a critical step of high-precision radiotherapy planning [35].

In routine clinical practice, the definition of target volumes and OARs is typically done manually by clinical experts who outline the structure boundaries in medical images. This process, which is commonly referred to as contouring or delineation, is time-consuming and labor-intensive [36]. The time spent on contouring has increased substantially with the advent of high-precision radiotherapy, as multimodality image interpretation is used more extensively and more accurate volume definitions are required [37–39]. Another limitation of manual contouring is its subjective nature. Considerable lack of contour agreement has been reported in a range of diagnoses for human experts contouring the same volume (interobserver variability) [37, 40], and, though less frequently studied, lack of agreement also occurs for the same human expert contouring the same volume at different occasions (intraobserver variability) [35]. Contouring variability could potentially impact on treatment outcome and quality of life, as inadequate contour definition has been associated with poorer disease control and increased toxicity [37, 41, 42]. Integration of automatic segmentation (auto-segmentation) methods in the radiotherapy planning workflow can reduce interobserver variability and contouring time [43, 44],

thereby potentially improving clinical outcomes and providing the clinical experts with more time for patient consultations and other tasks. An expected increase in cancer incidence of almost 50 % within 2040 [45], and the increased focus on adaptive radiotherapy strategies, where re-contouring of structures is required during the course of treatment [46], both emphasize the relevancy of contouring automation.

The general concept of auto-segmentation for radiotherapy has been studied for more than two decades, and span methods of varying complexity such as intensity thresholding, atlases, and machine learning algorithms [47–50]. Advances in computational resources and algorithms, along with increasing dataset sizes has led to several breakthroughs within the field of artificial intelligence during this time-period, with deep learning algorithms achieving human-level performance on various complex tasks outside the medical domain, as summarized in [51]. Deep learning, which is a subfield of artificial intelligence and machine learning, has also quickly obtained a central position within medical image analysis over the past few years [50]. Though still a field of extensive research, clinically acceptable auto-segmentation of several OARs can be achieved using deep learning with convolutional neural networks (CNNs) [52, 53] and such tools are currently commercially available [54, 55]. Auto-segmentation of target volumes including the gross tumor volume (GTV) is to the best of our knowledge not yet commercially available but is the focus of extensive ongoing research efforts including several public challenges [56, 57].

The overall aim of this thesis was to investigate the use of machine learning methods for automatic GTV segmentation in medical images. The specific aims were as follows:

1. To compare and evaluate thresholding methods, classical machine learning algorithms and deep learning with CNNs for auto-segmentation of the gross primary tumor volume (GTV-T) and involved nodal volume (GTV-N) in patients with head and neck cancer (HNC) based on either positron emission tomography (PET), contrast-enhanced computed tomography (ceCT), or combined PET/ceCT images (paper I).

2. To further assess CNNs for auto-segmentation of the GTV-T and GTV-N in HNC patients based on either PET, ceCT, or combined PET/ceCT images using larger image regions of interest than in paper I, also providing new quantitative structure-based performance evaluation metrics and qualitative human performance evaluation of automatically generated contours (paper II).

3. To evaluate the use of CNNs for automatic GTV segmentation in patients with anal cancer (AC), comparing the effects of different single modality and multimodality combinations of PET, CT and/or MR images on auto-segmentation performance (paper III).

4. To evaluate the applicability of CNNs for auto-segmentation of the GTV-T and GTV-N in canine HNC patients based on ceCT images, also assessing the impact of transfer learning from human HNC patients on auto-segmentation performance (paper IV).

Chapter 2

Theoretical background

2.1 Cancer

Cancer refers to a broad group of diseases that originates from mutations in the genetic material (DNA) of normal cells [58,59]. Cancer progression is characterized by continual unregulated cell proliferation, which for solid cancers results in the formation of a malignant primary tumor, also referred to as a malignant neoplasm, capable of destructive invasion of nearby tissues and subsequent spread to other parts of the body (metastases) [58, 59]. Cancers can be classified according to the cell type from which they originate and the anatomical site of initial development [60]. The majority of solid cancers can be broadly classified as either carcinomas, sarcomas, or lymphomas [58,59]. Carcinomas, which originate from epithelial cells, account for approximately 80–90 % of all human cancers [58, 60, 61].

Available cancer treatments include surgery, radiotherapy, and systemic treatment, including chemotherapy, targeted therapy, hormonal therapy, and immunotherapy. The selected treatment(s) commonly depends on the cancer type and site, as well as the stage of the disease, i.e., to what extent the disease has spread at the time of diagnosis [26]. The internationally accepted tumor–node–metastasis (TNM) system is widely used for staging of solid cancers [62]. Briefly, the TNM system uses alphanumeric codes to describe the extent of the primary tumor (T1–T4), the absence or presence and degree of regional lymph node involvement (N0–N3), and the absence or presence of distant metastasis (M0-M1). The rules used for these categorizations may vary between cancer sites [63].

2.1.1 Cancer in humans

Cancer ranks as a leading cause of death in the human population [64]. In 2020, an estimated number of approximately 19 million new cancer cases and close to 10

million cancer deaths occurred worldwide. The global cancer incidence is expected to increase by 47 % within 2040 solely due to aging, which is associated with increased risk of cancer, and growth of the population. Changes in the prevalence and geographical distribution of known cancer risk factors may aggravate this trend, potentially also leading to more abrupt increases in cancer burden and associated strain on the healthcare systems, especially in lower income countries experiencing social and economic transition [45].

Head and neck cancer

HNC is a common term for malignancies that originate from the anatomical sites of the upper aerodigestive tract [65]. An illustration of the different head and neck cancer regions is shown in figure 2.1. As of 2020, HNC was the seventh leading type of cancer by incidence worldwide accounting for approximately 900 000 cancer cases [45]. Most HNCs (90 %) are head and neck squamous cell carcinomas (HNSCCs) of the oral cavity, oropharynx, hypopharynx, and larynx [66]. Historically, smoking, and heavy alcohol consumption have been the major risk factors for head and neck SCCs in humans. In recent years, there has been a marked rise in oropharyngeal SCCs associated with carcinogenic human papillomavirus (HPV) infection [67]. HNC is a diverse group of malignancies, and the recommended treatment strategy depends on the primary tumor site, in addition to cancer stage and relevant patient factors. Radiotherapy, often in combination with concomitant chemotherapy (cisplatin), is an integral part of the management of most patients [65, 68].

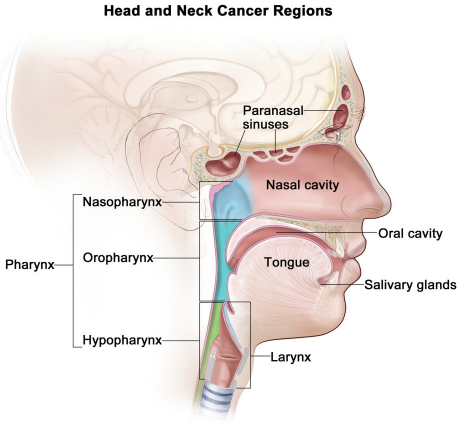


Figure 2.1: Illustration of the different head and neck cancer regions. Reproduced with permission from Terese Winslow. For the National Cancer Institute ©2012 Terese Winslow LLC, U.S. Govt. has certain rights.

Anal cancer

AC is a relatively rare form of cancer defined as malignant neoplasms of the anus and anal canal [69]. Anal SCC (ASCC) is the most common histological subtype, accounting for approximately 80 % of all cases [59, 69]. The incidence of ASCC has been increasing in many Western populations over the past years, primarily as a consequence of increased prevalence of HPV infection [69, 70]. An estimated 50 865 people were diagnosed with AC worldwide in 2020 [45]. Radiotherapy with concomitant chemotherapy (mitomycin C and 5-fluorouracil) is the standard of care for ASCC patients with local or locoregional disease [70].

2.1.2 Cancer in dogs

As for humans, cancer ranks as a leading cause of death in domestic dogs, and the estimated incidence rates of canine cancer is similar to that of humans [71]. Accurate worldwide cancer statistics are lacking in veterinary medicine, but it has been estimated that roughly 6 million dogs are diagnosed with cancer annually in the U.S. alone [72]. The treatment of veterinary cancer patients has evolved alongside human cancer treatment [71], and the main treatment options are in principle the same as for humans. In clinical practice, however, treatments such as radiotherapy and chemotherapy are generally less available for dogs and the treatment selection is to a greater extent influenced by economic and practical factors than for humans.

Companion dogs share the human environment, and spontaneously developed canine cancers share key features with the equivalent human cancers, including clinical presentation, biology, and treatment response. Moreover, the disease progression is often considerably faster, and certain rare cancers are more frequent in dogs than in humans [71]. Therefore, dogs with naturally occurring cancers have been used to model human cancers within the field of comparative oncology, expanding the knowledge of both canine and human cancers [71, 73–76].

Canine head and neck cancer

Canine HNC is less formally defined than human HNC and often refers to more anatomical primary tumor sites of the head and neck region than those included in the human HNC definition, such as the thyroid gland and ear [77–79]. Several studies have reported the oral cavity as the most frequent canine HNC location and malignancies of the oral cavity have been ranked as the 4th most common cancer in dogs [78, 80, 81]. The nasal cavity has been reported as the second most common canine HNC location in a study on Danish dogs [78]. Compared to its human counterparts, canine HNCs display a greater variety of cancer cell subtypes and SCCs are less predominant [82, 83]. For most canine HNCs, surgery is the primary treatment. Radiotherapy is, however, increasingly available in veterinary medicine

[84] and constitutes the primary treatment of choice for canine nasal cancers [77, 82]. Multimodal treatments including surgery, radiotherapy, and chemotherapy may also be considered for the treatment of canine HNCs [77].

2.2 Medical imaging

Medical imaging is an integral part of the management of patients with cancer. Various medical imaging modalities are used for diagnosis and initial staging, planning and delivery of treatment, and evaluation of treatment response [59,85]. Computed tomography (CT), positron-emission tomography (PET), and magnetic resonance imaging (MRI) are the three main imaging modalities used in the process of radiotherapy planning [86]. All three modalities are used to generate cross-sectional images of the interior anatomy or tissue function of the patient, based on the detection of electromagnetic radiation. The resulting 2D images each represent a slice of the scanned tissue with a given thickness and position in space, which can be combined to form a 3D representation of the imaged region. Thus, each 2D image pixel represents a volume element (voxel) of tissue with a given location on a 3D grid [87].

2.2.1 Computed tomography

CT provides anatomical images with high spatial resolution and low acquisition time and is the most commonly used cross-sectional imaging modality [85]. CT takes advantage of the variation in attenuation of x-ray photons according to tissue density [59]. In a CT examination, a rotating source and detector assembly is used to emit x-ray beams through the patient from different angles and subsequently detect the transmitted attenuated beams. The patient is advanced through the scanner, until the desired region is covered. The collected raw data are processed with a tomographic reconstruction method to form a series of 2D images of the internal anatomy of the patient [59].

CT image intensities are commonly expressed using the Hounsfield unit (HU) scale which is a dimensionless scale for radiodensity, i.e., the opacity to x-ray photons [88], typically having a range of [–1 024, 3 071] HU [89]. For the purpose of image interpretation and analysis, the high dynamic range of the CT images is typically reduced using windowing, where a selected range (i.e., a “window”) of HU values is mapped to a gray scale. By adjusting the midpoint and range of the window, referred to as the window center and width, the brightness and contrast of the image can be altered to highlight different anatomical structures. Narrow window widths (50–350 HU) are well suited to examine areas of similar radiodensities, such as soft tissues [90]. An example is shown in figure 2.2.

Depending on the anatomical region and the indication of the CT examination, contrast-enhanced CT (ceCT) may be appropriate to increase the visibility of low-contrast normal tissues or pathologies [91]. Intravenous (IV) iodine-based CT contrast agents enhance vascular structures and are used routinely for a wide range of cancer diagnoses [85,92,93]. Iodinated agents act by increasing the radiodensity, and thereby the CT image intensity, by an amount proportional to the iodine concentration [94]. Following an IV injection, the contrast agent is distributed by the cardiovascular system, and the CT acquisition is timed to coincide with optimal contrast differences in the target region or organ [94].

2.2.2 Positron emission tomography

PET is a molecular nuclear imaging technique that provides functional information based on the distribution of a radioactive tracer (radiotracer) containing a positron emitting radionuclide. In a PET examination, the radiotracer is administered to the patient, whereupon it distributes within the body according to its biochemical properties before the radionuclide in the tracer decays [59]. Positrons emitted by the radionuclide have short lifetimes in biological tissues and travel a short distance (typically $\leq$ 1 mm, depending on the radionuclide) before they interact with electrons by annihilation [95]. The majority of annihilations result in two 0.511 keV gamma photons that move at an angle of approximately 180° relative to each other [96]. Such annihilation photon-pairs form the basis of the PET signal and are registered by a ring of detectors surrounding the imaged region of the patient. Based on this, the line connecting the two detector elements, also known as the line of response, and the position of the annihilation event can be determined [95]. The PET detectors are made up of scintillation crystals that convert high energy gamma photons to visible light, which is further converted to an electrical signal in a photosensor [97]. The collected raw data are subject to error corrections before reconstruction to images where the intensity is proportional to the tracer uptake [95].

The spatial resolution of PET is limited by several factors including detector size [98], and PET, therefore, has lower spatial resolution than CT and MRI [99]. This has contributed to the development of hybrid imaging methods combining PET and CT or PET and MRI [99]. Currently, PET imaging is most commonly performed with a hybrid PET/CT scanner [100]. The CT provides tissue density information used for attenuation correction of the PET images, resulting in improved PET image quality, as well as high-resolution anatomical information for improved localization of the PET signal upon image interpretation [101].

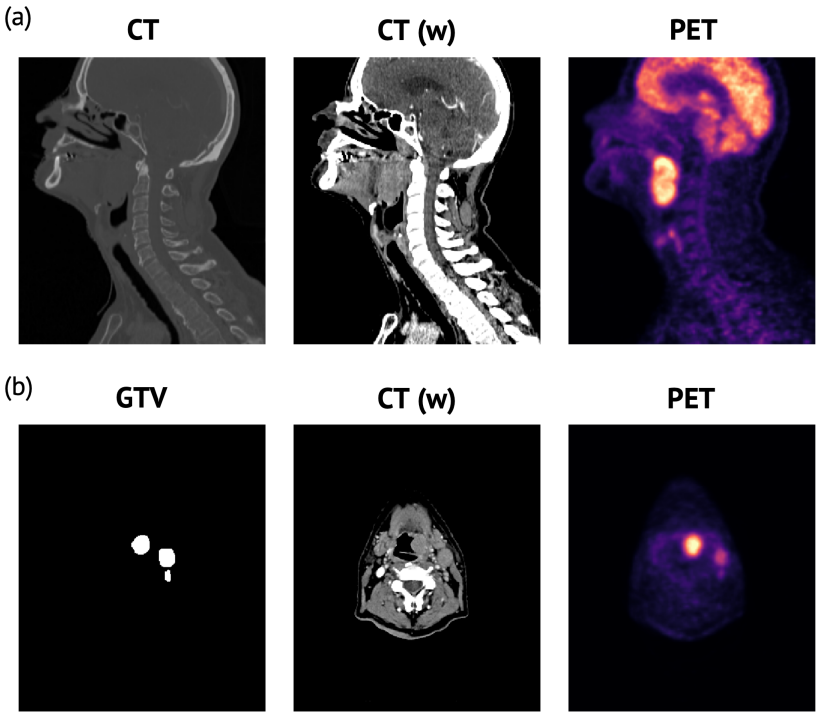


Figure 2.2: (a) Coronal image slice of a patient with head and neck cancer. Left: original contrast-enhanced CT image. Center: windowed (w) contrast-enhanced CT image (center: 70 HU; width: 200 HU). Right: FDG PET image. (b) Axial image slice of the same patient. Left: gross tumor volume (GTV-T) and involved nodal volume (GTV-N) shown as a binary mask. Center: windowed (w) contrast-enhanced CT image (center: 70 HU; width: 200 HU). Right: FDG PET image.

The PET signal is commonly expressed as the standardized uptake value (SUV) prior to image interpretation, using the following definition:

$$SUV = \frac{r}{a'/w}, \tag{2.1}$$

where r is the radioactivity concentration [kBq/ml] measured in a volume of interest, usually an image voxel, a' is the activity administered to the patient [kBq], and w is the body weight of the patient [g] [102].

Radionuclides used for PET imaging must have appropriate physical half-lives, preferably a low maximum positron decay energy limiting the positron range prior to annihilation, a high fraction of decays occurring via positron emission,

as well as suitable chemical characteristics [95, 101]. The glucose analogue ^{18}F-fluorodeoxyglucose (FDG) is by far the most commonly used tracer in clinical oncologic PET imaging [103]. Due to its construction, where one of the glucose hydroxyl groups is replaced with radioactive ^{18}F, FDG is taken up as unlabeled glucose but cannot be fully metabolized and is concentrated in the cells until ^{18}F decays. As a result, FDG PET imaging provides a map of the glucose consumption in the patient [103], which is elevated in most cancers [103, 104]. An FDG PET image of a patient with HNC is shown in figure 2.2.

FDG PET imaging is shown to have high sensitivity for detection of regional lymph node metastases, distant metastases and second primary tumors in a range of cancer diagnoses [104]. FDG PET is also used routinely for assessment of cancer treatment response and long-term patient monitoring for recurrence detection [59]. A limitation of FDG PET is that the tracer uptake is not fully cancer specific. For example, high glucose metabolism and associated elevated FDG uptake is not restricted to cancer cells, but can also be present in certain organs, and in areas of infection or inflammation [103, 105]. On the other hand, not all malignancies display increased metabolic activity [105]. This can potentially lead to false positive or false negative FDG PET image interpretation [104, 105].

2.2.3 Magnetic resonance imaging

MRI is based on the concept of nuclear spin and the fact that atomic nuclei with non-zero spin possess a magnetic dipole moment. In principle, all nuclei with non-zero spin may be used for MRI [106], but the vast majority of clinical MRI examinations are based on the ^{1}H nucleus (the proton), which is highly abundant in biological tissues [107]. Under the influence of a strong external magnetic field, the magnetic dipole moment of each proton in the imaged tissue will align itself with the external field while precessing with a frequency ω_0 known as the Larmor frequency, given by [108]:

$$\omega_0 = \gamma B_0, \tag{2.2}$$

where γ is the gyromagnetic ratio, which is specific to the type of nucleus [107], and B_0 is the magnetic field strength. In sum, the above alignments result in a net equilibrium magnetization parallel to the external field $\boldsymbol{B_0}$ (by convention defined as the z direction of the system) [108].

The net magnetization is vanishingly small compared to B_0 and is, thus, not practically feasible to detect while in equilibrium [108]. By exposing the protons to electromagnetic radiation with the Larmor frequency, which for clinical field strengths is within the radio-frequency region, nuclear resonance occurs, and protons are excited to a higher energy level [107]. As a result, all the nuclear spins

are brought into phase and the net magnetization is flipped into the transversal xy-plane while precessing about $\boldsymbol{B_0}$. This precession movement represents a time-variant magnetic field which is detected in a receiver coil [108]. After the electromagnetic radio frequency pulse is turned off, the protons return to their initial energy state and the nuclear spins dephase in a process known as relaxation. This results in an exponential recovery of longitudinal equilibrium magnetization and decay of transversal magnetization which are characterized by two independent time constants $T1$ and $T2$, respectively, both of which are tissue specific [107]. The change in transversal magnetization following a radio frequency pulse is registered as a time-variant decaying signal by the receiver coil [108].

To localize signals in space, secondary gradient fields must be applied in addition to $\boldsymbol{B_0}$ causing the magnetic field strength, and thereby the Larmor frequency, to vary systematically with position [109]. The intrinsic contrast of MRI primarily depends on the proton density and the $T1$ and $T2$ relaxation times of the imaged tissue, along with other tissue specific parameters. The image contrast can be manipulated by placing emphasis on, i.e., weighting one of the tissue parameters over the others [107]. The desired weighting and spatial information is achieved by applying a suitable MRI sequence consisting of specific combinations of radio frequency pulses and gradients [109].

Compared to CT, MRI provides superior soft tissue contrast and does not involve the use of ionizing radiation [110]. The limitations of MRI include long acquisition time and the fact that it is subject to unique image artefacts that may degrade image quality [107, 110]. The possibility of weighting different tissue parameters makes MRI highly versatile [108]. Furthermore, MRI can be used to image tissue function using for example dynamic contrast enhanced MRI to assess perfusion and permeability [111], or diffusion-weighted (DW) MRI to assess the thermal diffusion of water molecules which can be altered by pathologic conditions such as cancer [112].

DW MRI is relatively easy to implement and can provide clinically relevant information for a range of cancer types [113]. In a DW acquisition, dephasing and rephasing gradients which are equal in strength but exactly opposite, are applied as part of the MRI sequence. These two gradients will have no net effect on stationary protons, whereas diffusing protons will acquire a phase change as their positions change between dephasing and rephasing [107]. This results in a diffusion-dependent attenuation of the MR signal, i.e., a signal loss [112]. The diffusion-weighted MR signal S can be expressed as:

$$S(b) = S_0 e^{-bADC}, \tag{2.3}$$

where b is the degree of diffusion-weighting [s/mm^2], and depends on the strength, duration, and timing of the applied gradients, S_0 is the signal without any diffusion-

weighting (i.e., $b = 0$), and ADC is the apparent diffusion coefficient [mm^2/s] [112] which is an intrinsic tissue specific parameter [107].

DW images are commonly obtained for several different b values [113]. Higher b values result in images where the areas of high diffusion will appear hypointense, whereas the low-diffusion regions appear hyperintense [112]. Cancerous tissues, which can have high cellularity and associated restricted diffusion, will typically have low ADC values and appear hyperintense on DW images with high b values. To quantify the diffusion information in the DW images, ADC maps can be calculated based on two or more DW images acquired with different b values [108]. As implicit in its name, ADC values not only depend on water diffusion, though this is the largest contributor, but are also influenced by other forms of molecular motion such as capillary perfusion [108, 114]. The latter effects can potentially be minimized by selecting appropriate b values for the ADC calculations [115].

2.3 Radiotherapy and target volumes

2.3.1 Radiotherapy

Radiotherapy is one of the main treatment modalities for cancer and may be used alone or in combination with other treatments, such as surgery, chemotherapy, and immunotherapy [26, 27]. The most commonly used radiotherapy approach is external beam irradiation with high energy (MeV) photons generated by a linear accelerator [116]. Briefly, radiotherapy is based on the use of ionizing radiation to damage exposed cancer cells, with cancer cell death as a desired end result [117, 118]. The mechanisms of radiation-induced cell damage are not selective to cancer cells, and irradiation of normal tissues during radiotherapy may result in acute and/or late toxicity [30]. Thus, the fundamental aim and challenge of radiotherapy is to ensure sufficiently high radiation dose to the cancer cells while keeping the dose to surrounding normal tissues at acceptable levels [28, 29].

Radiotherapy has undergone considerable technical developments over the past decades [119, 120]. Intensity modulated radiotherapy (IMRT), which is an advanced form of 3D conformal radiotherapy (3D-CRT), is considered the state-of-the-art technology for external beam photon radiotherapy [121, 122]. In IMRT, the intensities of multiple small photon beams (beamlets) are modulated to accurately conform the high dose to the shape of the target volume while minimizing the dose to critical normal tissue structures (OARs) [122]. Volumetric modulated arc therapy (VMAT) is a further refined form of IMRT where the radiation source is moved in an arc around the patient [119]. Such high-precision techniques are particularly relevant for cancers where the target volume is located in close proximity of radiation sensitive OARs. IMRT/VMAT is shown to reduce normal tissue toxicity compared to conventional 3D-CRT for both HNC and AC [32, 34] and

is recommended for both these groups of patients [68, 70]. High precision techniques including IMRT are also highly relevant for the treatment of canine HNC patients [123, 124].

Radiotherapy must be carefully planned to maximize its therapeutic ratio. After diagnosis and referral, a simulation CT scan of the patient immobilized in the treatment position is acquired and imported to the treatment planning system. The simulation CT usually forms the basis for definition of target volumes and OARs. Other imaging modalities, such as PET and MRI, may be used to support volume definitions [85]. As CT images hold intrinsic information on tissue densities and, thus, can be converted to 3D density maps, the simulation CT also forms the basis for calculation of radiotherapy dose distributions used in treatment plan design [125, 126]. Following volume definitions, the treatment objectives must be defined, including the desired doses to target volumes and dose constraints for OARs. An optimization process is initialized in the treatment planning system, and the expected dose distribution associated with the above objectives is calculated. The resulting dose plan is evaluated and, if necessary, the objectives are adjusted and a new optimization process performed, until the dose plan is found to be satisfactory [127]. The approved treatment plan is imported to the treatment system and delivered to the patient after the patient positioning and treatment setup have been verified [85].

The total prescribed radiation dose is conventionally delivered in smaller fractions over multiple weeks. As normal tissue cells typically have a better ability to repair radiation-induced damages compared to cancer cells, this allows healthy cells to recover to a greater extent than cancer cells between fractions. Fractionation also increases the probability of irradiating cancer cells at their most radiosensitive, as parameters linked to radiosensitivity, such as cell cycle stage and re-oxygenation, vary over the course of treatment [85]. Factors such as tumor and normal tissue anatomy may also change over the course of treatment [46]. Adaptive radiotherapy is an advancing treatment strategy, where the treatment plan is re-evaluated and, if deemed necessary, adjusted during the course of treatment [46, 128].

2.3.2 Target volume definitions

Traditionally the following three main target volumes are used in radiotherapy [35, 129–131]: the GTV, the clinical target volume (CTV), and the planning target volume (PTV). In general, the GTV is defined as the macroscopic volume of the tumor that can be determined by clinical examination (observation, palpation) and medical imaging, and may include the primary tumor (GTV-T), involved regional lymph nodes (GTV-N) and/or distant metastasis (GTV-M) [35]. The GTV usually corresponds to the regions of highest cancer cell density, and sufficient radiation dose must be delivered to the entire GTV to achieve local tumor control [35, 132]. In addition, the regression of the GTV may affect treatment plan

adaptation during the course of treatment and could potentially also predict treatment outcome [35]. The CTV includes the GTV and regions outside the GTV with suspected microscopic cancer infiltration, or with a high risk of involvement according to clinical experience, that should be treated adequately [35, 129]. The PTV is defined by adding a margin to the CTV to ensure that the prescribed dose is delivered to all parts of the CTV with an acceptable probability, taking uncertainties and variations in CTV size, shape, and position, as well as patient and beam positioning into account [35, 85]. The GTV, CTV and PTV are illustrated in figure 2.3.

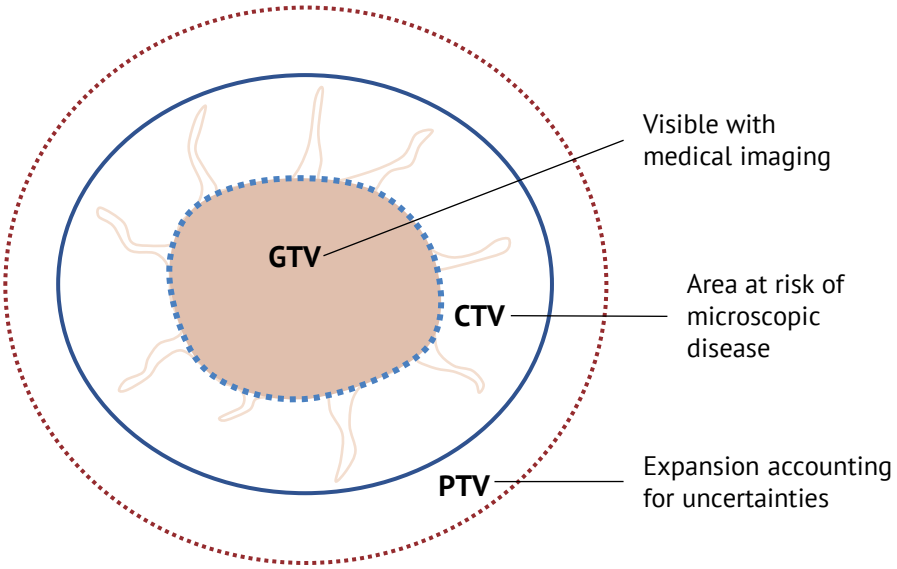


Figure 2.3: Illustration of the three main target volumes used in radiotherapy. GTV: gross tumor volume; CTV: clinical target volume; PTV: planning target volume. Based on a figure from [85].

2.3.3 Contouring of target volumes

The definition of target volumes and OARs is a critical part of radiotherapy planning, as all subsequent steps follow from their definition [40]. Due to the resulting reduced target volume margins and steep dose gradients, optimal high-precision radiotherapy requires more accurate definition of both target volumes and OARs than conventional radiotherapy [35, 40, 41]. In clinical practice, target volumes and OARs are typically defined manually by clinical experts who outline the given structures in medical images within the treatment planning system. This process is referred to as contouring or delineation. Cancer site-specific contouring guidelines

supplement the general volume definitions (cf. section 2.3.2) and give recommendations on the optimal imaging modalities for contouring [38,39,133–135]. Modern radiotherapy techniques have been accompanied by increasing use of multimodal image interpretation [37], particularly for contouring of the GTV [38,39,135,136].

Manual contouring is known to be time-consuming and labor-intensive, particularly for cancer sites with complex anatomy and many OAR structures such as HNC [36, 37]. As manual contouring is a subjective task, it is also inherently prone to intra- and interobserver variability. Contouring guidelines, atlases, and peer review strategies are used to increase contouring quality and consistency [41]. However, considerable interobserver variabilities still occur [37, 40], and manual contouring is recognized as a main source of geometric uncertainty within the process of radiotherapy planning and delivery [40]. Inaccurate contour definitions could result in reduced dose to the target volume or unacceptably high dose to normal tissues, potentially increasing the risk of locoregional failure or normal tissue toxicity [40]. Uncertainties introduced by contour variability may further be a confounder in single and multi-center clinical radiotherapy trials [40,137]. Auto-segmentation methods hold the potential to address the above challenges [43,44,50,137] and could facilitate adaptive radiotherapy strategies that rely on fast and accurate volume definition [46]. Consequently, auto-segmentation methods and their potential application within the radiotherapy workflow have received significant attention over the past years [50]. In addition, the application of auto-segmentation could facilitate large scale multi-centric radiomics studies [138].

Chapter 3

Materials and methods

3.1 Patient cohorts

3.1.1 Human head and neck cancer dataset

Papers I and II were based on a dataset consisting of FDG PET/ceCT images and corresponding manual GTV delineations of 197 patients with HNC. The ceCT images and GTV delineations of the same set of patients were also included and used for the transfer learning analysis in paper IV. The data was collected as part of a retrospective study including patients with HNSCC of the oral cavity, oropharynx, hypopharynx, and larynx, scheduled for radiotherapy at Oslo University Hospital in the period from 2007 to 2013 [139]. All patients were treated with primary radiotherapy (IMRT; 68–70 Gy in 2 Gy fractions). Most patients also received concurrent chemotherapy (cisplatin, 40 mg/m^2 per week) and the hypoxic radiosensitizer nimorazole.

FDG PET/ceCT imaging was performed at baseline with a Siemens Biograph 16 scanner. Both the visible primary tumor volume (GTV-T) and involved lymph nodes (GTV-N) were included in the GTV. The manual GTV delineations were made prospectively at the time of initial treatment planning and were used for radiotherapy planning. Delineations were based on FDG PET and ceCT image information as well as relevant clinical information such as the endoscopy report [139].

3.1.2 Anal cancer dataset

Paper III was based on pre-treatment images and corresponding manual GTV delineations of in total 86 patients with ASCC. Two datasets consisting of the

following image data were analyzed separately: (i) FDG PET, low-dose CT (ldCT) and ceCT images of all 86 patients, and (ii) FDG PET, ldCT, ceCT, $T2$-weighted (T2W) MR and ADC images of a subset of 36 patients. The 36 patients in the latter dataset had consented to a study-specific 3.0 T MR examination including T2W and DW imaging. For paper III, regression analysis [115] based on b-values of 200, 400, 600 and 800 mm^2/s was used to condense the DW series to an ADC map.

The included patients were part of the prospective ANCARAD observational trial (NCT01937780) [140] and were scheduled for chemoradiotherapy at Oslo University Hospital between 2013 and 2016. All patients received radiotherapy (IMRT/VMAT (67 %) or 3D-CRT (33 %); 54 or 58 Gy in 2 Gy fractions), and the majority of patients were given concurrent chemotherapy (one or two cycles of mitomycin C and 5-fluorouracil).

PET/ldCT images were obtained with a Siemens Biograph mCT 40 scanner and ceCT images were obtained with a General Electric LightSpeed Pro 16 scanner. The study-specific MR examination was performed with a Phillips Ingenia 3.0 T scanner. According to current practice for radiotherapy of anal cancer, the GTV was defined to include the visible tumor tissue and the entire anal canal and/or rectum circumference when tumor involvement was present. Delineations of involved lymph nodes (GTV-N) were not included in paper III. For all patients, the manual delineations were based on ceCT, FDG PET and standard (i.e., not study-specific) T2W images. As for the human HNC dataset, the delineations were made prospectively and used for radiotherapy planning.

3.1.3 Canine head and neck cancer dataset

The canine dataset analyzed in paper IV consisted of ceCT images and corresponding manual GTV delineations of 36 dogs with malignant neoplasms of the head and neck region. The ceCT data were collected retrospectively by reviewing the imaging database and patient record system at the University Animal Hospital at the Norwegian University of Life Sciences. ceCT images were acquired with a General Electric BrightSpeed S scanner. Manual delineations were made retrospectively based on the ceCT image information. As for the human HNC dataset, the canine GTV included both the GTV-T and GTV-N.

3.2 Automatic segmentation methods

Auto-segmentation for radiotherapy has been studied for more than two decades and numerous methods have been proposed for various cancer diagnoses and imaging modalities [47–50]. For automatic target volume segmentation, most methods

fall into one of three main approaches listed in the order of increasing complexity: intensity thresholding, classical machine learning and deep learning. In paper I, PET thresholding, classical machine learning and deep learning methods for automatic GTV segmentation were compared, whereas papers II–IV focused on automatic GTV segmentation using deep learning only.

The relationship between machine learning, deep learning, and the discipline of artificial intelligence, is often visualized as shown in figure 3.1. Artificial intelligence may be defined as "the effort to automate intellectual tasks normally performed by humans" [141]. One approach to artificial intelligence, termed the knowledge base approach, is to hard-code all the formal rules a computer system would need to solve a given task automatically. Machine learning, on the other hand, is the capability to perform a task by learning from input data without such hard-coded knowledge [51]. Deep learning refers to the subfield of machine learning where a hierarchy of concepts, also referred to as layered representations of the data, is learned jointly to perform a task [51, 141]. These layered representations are in most cases learned by deep learning models that use neural network architectures [141].

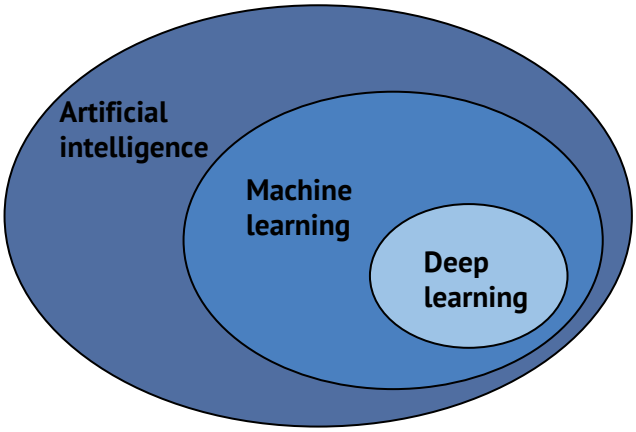


Figure 3.1: Euler diagram illustrating the relationship between artificial intelligence, machine learning and deep learning.

The classical machine learning and deep learning approaches to artificial intelligence are visualized schematically in figure 3.2. Deep learning algorithms inherently learn increasingly complex representations (features) of the input data, whereas solving complex tasks using classical machine learning algorithms generally requires a feature engineering step, where handcrafted features are created from the initial input before data is fed to the algorithm [51].

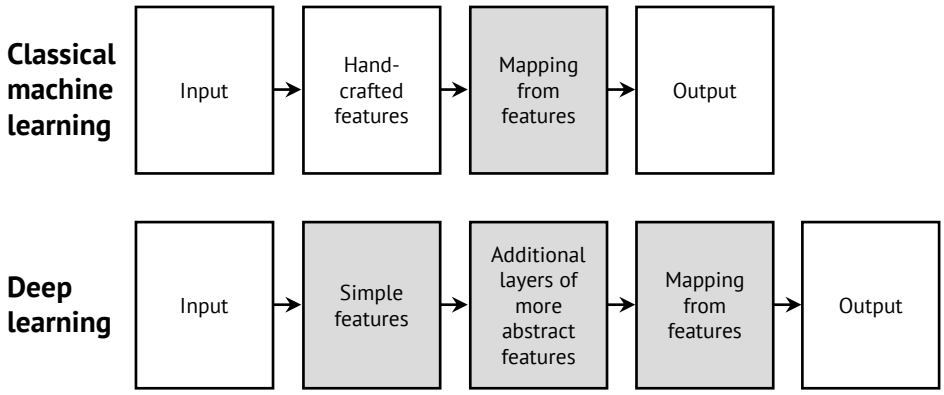


Figure 3.2: Flowcharts showing the main components of the classical machine learning (top row) and deep learning (bottom row) disciplines. Shaded boxes indicate components capable of learning from data. Based on a figure from [51].

Most machine learning algorithms belong to one of two branches: supervised and unsupervised learning [51]. Both supervised and unsupervised learning, or hybrids of the two approaches, may be used for segmentation tasks. In this thesis, the focus is on supervised classification algorithms, using the clinical expert GTV delineations as a binary outcome variable, also referred to as the ground truth. Supervised learning relies on the following three fundamental components: (1) input data, (2) labeled examples of the outcome, and (3) an objective (loss) function that measures how well the output of the algorithm matches the example (ground-truth) outcome. The objective function is optimized in the learning process, i.e., during model training [141].

3.2.1 Thresholding

Thresholding is a simple form of image segmentation, where the pixel intensity values of an entire image, or a region of interest within an image, are separated into foreground and background based on a threshold value q [142, 143]. The thresholding operation f_T can be defined as follows [143]:

$$f_T(a) = \begin{cases} a_0 & \text{for } a < q \\ a_1 & \text{for } a \geq q \end{cases}, \tag{3.1}$$

where a denotes an original pixel value, a_0 and a_1 are the two fixed foreground and background intensity values the pixels are mapped to, and the threshold value q is within the range of the original pixel intensity values. Numerous approaches to selecting the optimal threshold value q exist [142–144].

In paper I, thresholding of the PET images was done using either an absolute SUV threshold, a percentage of the maximum SUV (SUV_{max}) threshold, or a multi-step thresholding method based on the Laplacian of Gaussian (LoG) transformation [142] of the PET images (referred to as LoG thresholding). Each of the above thresholding methods was optimized by maximizing the mean overlap between the segmentations and the manually delineated GTV structures, as measured by the Sørensen-Dice similarity coefficient (cf. section 3.4.1 below). In addition, thresholding with 41 % of the SUV_{max} threshold, which has previously been recommended for PET thresholding [104, 145], was included for comparison.

3.2.2 Classical machine learning methods

In paper I, we evaluated the following six classical machine learning classification algorithms: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), Gaussian naïve Bayes (GNB), logistic regression (LR), linear support vector machines (SVM), and random forest (RF). Segmentation was performed by using different combinations of voxel-wise features derived from the original PET and/or CT images as input to the classification algorithms.

Classical machine learning algorithms expect the data to be structured as a 2D feature matrix and a response vector. Thus, prior to classification, the 3D image stacks were unfolded into 2D matrices where each row vector contained the input feature(s) of one unique voxel, and the ground truth delineations were unfolded into a vector with the corresponding class membership $y \in \{0, 1\}$, using a procedure previously described in Torheim et al. [146]. As the number of voxels in the foreground (class 1; voxels belonging to the GTV) and the background (class 0) were highly imbalanced, with the majority of voxels (94 %) belonging to the background, the majority class was randomly under-sampled to achieve 50–50 class-balance for each patient in the training data.

LDA, QDA, and GNB can all be categorized as generative classification algorithms, which model the joint probability distribution $p(X, y_k)$, or equivalently the conditional probability distribution $p(X \mid y_k)$ and the prior class probability $p(y_k)$, of the given observable input X for each class y_k [147]. Based on this, the above models use Bayes' theorem to determine the conditional probability $p(y_k \mid X)$ for each class [148]. Both LDA and QDA model the conditional class density of each class $p(X \mid y_k)$ as a multivariate Gaussian distribution [149]. LDA further assumes that the classes share the same covariance matrix, which results in linear decision boundaries separating pairs of classes. QDA relaxes the assumption of one shared covariance matrix, resulting in quadratic decision boundaries between pairs of classes. Naïve Bayes classifiers are based on the naïve assumption that the features of X are independent within each class. This simplifies the estimates needed to model each class density $p(X \mid y_k)$ which for GNB is found as the

product of univariate Gaussians (one for each input feature in X) [149]. Unless the covariance matrices of the classes are identical, GNB will result in quadratic decision boundaries between pairs of classes.

LR is a discriminative classification algorithm, which models the conditional probability $p(y_k \mid X)$ for each class directly [147]. LR is constructed to achieve this via functions that are linear in X under the constraint that the class probabilities $p(y_k \mid X)$ sum to one [149]. Thus, similarly to LDA, LR results in linear decision boundaries between classes. LR, however, relies on fewer assumptions and is thus more general. SVM and RF can also be defined as discriminative algorithms, as they perform a direct mapping from input X to class membership y. However, contrary to LR, SVM and RF are not developed from a probabilistic view. SVM [150, 151] is developed from a geometric perspective and identifies an optimal hyperplane separating the classes by maximizing the margin between the training observations of each class in the feature space. SVM can be kernelized to generate non-linear decision boundaries [152]. In paper I, however, the linear SVM classifier was used, which results in linear decision boundaries between pairs of classes. RF [153] is an ensemble-based classification method which repeatedly ($b = 1, 2, \ldots, B$ times) selects a random bootstrap sample of N training observations with replacement and then grows a random-forest decision tree T_b to these data. For each terminal node of the tree, m of P predictors (features) are selected at random, and the best variable/split point is determined before the node is split into two daughter nodes, until the pre-selected minimum number of observations per tree leaf (minimum node size) is reached. The class membership of a new observation is predicted as the majority vote of the B trees [149].

Regularization refers to techniques used to constrain a model to make it simpler and reduce the risk of overfitting to the training data, thereby improving the model generalizability [51, 154]. In paper I, both LR and SVM were trained using LASSO (least absolute shrinkage and selection operator; L1) [155] or Ridge (L2) regularization. Details on the above classifiers and regularization techniques, and their mathematical descriptions, can be found in for example Hastie et al. [149].

3.2.3 Deep learning methods

The fundamental components of a neural network are illustrated in figure 3.3. Layers are combined to form a network. Simply put, the layers, which consist of units called (artificial) neurons, are characterized by a set of weights and bias terms that are trainable parameters. The layers take one or more tensor(s) as input, process the input, and output one or more tensor(s). A suitable loss function is used to measure the predictive performance of the network. Deep leaning models are usually trained iteratively on smaller batches of the full training data, i.e., minibatches, selected at random. During training, the network maps the input to predictions, and the loss score of the given training iteration is calculated [141].

Based on this, a feedback signal is passed to an optimization algorithm (optimizer) which updates the weights and biases using some variant of the stochastic gradient descent algorithm [51]. A description of the mathematical concepts underpinning deep learning can be found in for example Goodfellow et al. [51].

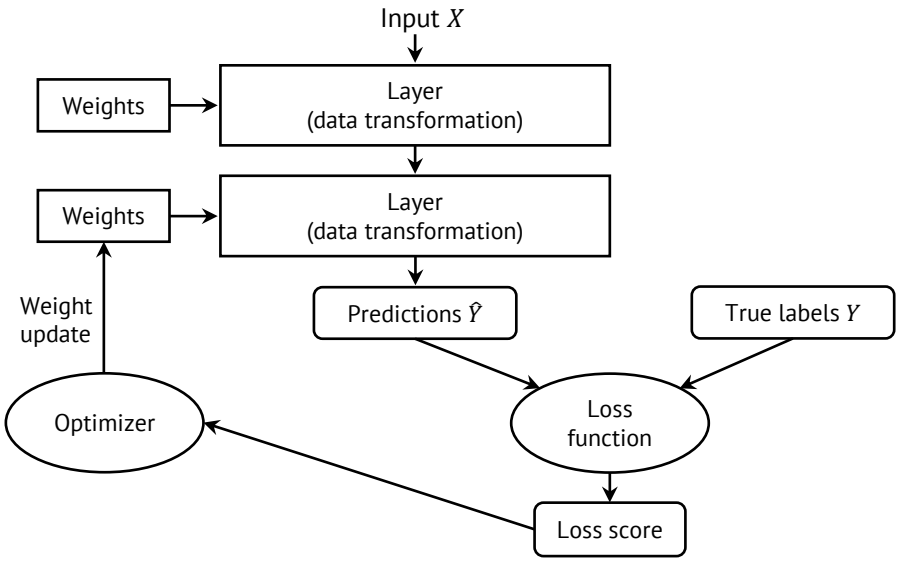


Figure 3.3: Fundamental concepts of a neural network. Based on a figure from [141].

Convolutional neural networks (CNNs) [156] are a class of neural networks used for tasks where the input data has a grid-like format, such as digital images. With CNNs images can be processed directly. As indicated by its name, a CNN employs the mathematical convolution operator in its layers. CNNs are, however, often implemented using the related, but conceptually simpler, cross-correlation operator, which differs from convolution in that it does not involve flipping of the operator kernel. The operator kernel values (weights) are learned during training, and the above distinction is not important for the application to CNNs. Thus, within the field of deep learning both above operations are commonly referred to as convolution [51]. The same convention will be used in the remainder of this text.

An example convolution of a 2D image matrix I and a 2D kernel K (without flipping of the kernel) is given in figure 3.4. As convolution will shrink the output dimensions, the image is commonly padded along each axis prior to convolution to let the output have the same size as the image. In the case of images with multiple channels, separate convolutions are performed on each image channel and the final

output is obtained by adding the result of each channel.

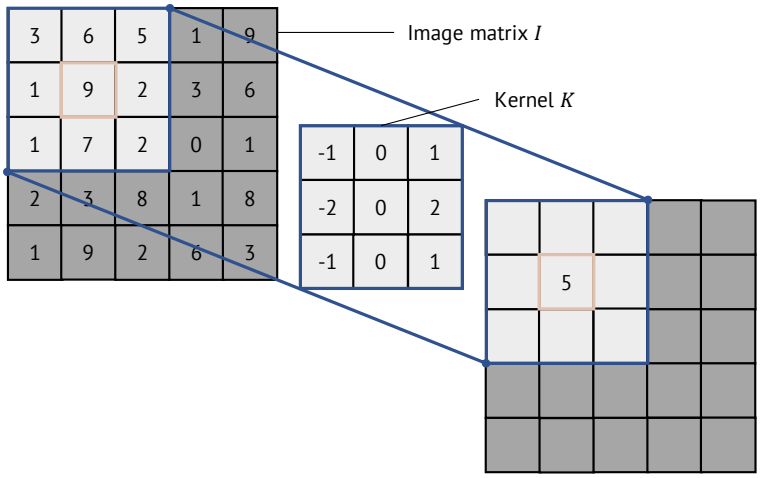


Figure 3.4: Illustration of a 2D convolution of an image matrix I and a 3×3 kernel K. The output pixel value is calculated as follows: $-1 \times 3 + (-2) \times 1 + (-1) \times 1 + 0 \times 6 + 0 \times 9 + 0 \times 7 + 1 \times 5 + 2 \times 2 + 1 \times 2 = 5$.

A convolutional layer commonly consists of multiple convolutions performed in parallel. As convolution is a linear mapping of the input, each element of the convolution output is commonly passed to a non-linear activation function, such as the rectified linear unit (ReLU), to allow more complex models to be learned [51]. The size of the convolution kernel is smaller than the input, a typical choice in 2D CNNs is a 3×3 kernel, which enables CNNs to learn local patterns in the input. To increase the receptive field of the network, and efficiently build a hierarchy of features, down-sampling, or pooling, operations are applied in between convolutional layers. One of the most frequently used pooling techniques is the max pooling operation [157], which returns the maximum output value within a sliding window.

Network architecture and training scheme

In the deep learning experiments of papers I–IV, we used the U-Net [158] CNN architecture, which is one of several existing CNN architectures designed for semantic segmentation. As illustrated in figure 3.5, the U-Net is made up of a contracting path (left) and an expanding path (right), also referred to as the decoder and encoder parts of the network, along with long-distance skip connections. The contracting path takes an input image and performs repeated convolutions followed by max pooling [157] operations for down-sampling. The expanding path

performs up-sampling operations to recover the original image size, also referred to as up-convolutions or transposed convolutions, each followed by convolutions on concatenated feature maps created via the skip connections. The use of skip connections preserves features learned in the contracting path, that otherwise would have been discarded in the down-sampling operations. All convolutions, except the last, are followed by the ReLU activation function. The final network output is a segmentation mask.

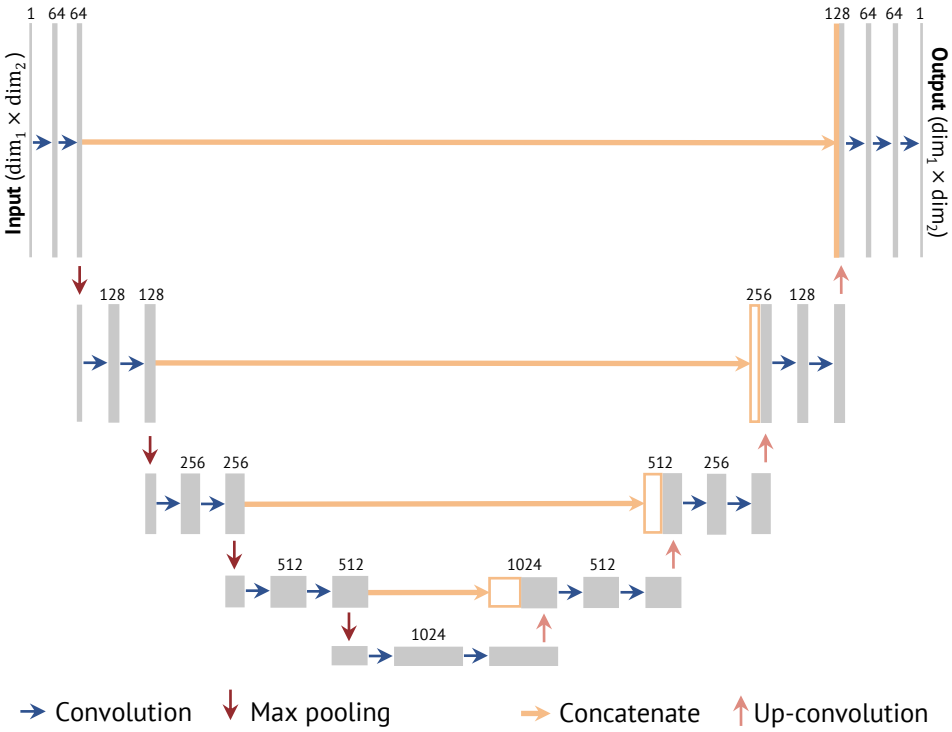


Figure 3.5: Illustration of the 2D U-Net architecture [158]. The input and output layers (feature maps) are shown as boxes, where the box length represents the number of channels (also indicated with a number) and the box height represent the spatial size. In the illustration, $\dim_1$ and $\dim_2$ refer to the x and y-dimensions of the input image.

In this thesis, 2D (papers I–III) and 3D (paper IV) U-Net [158, 159] architectures were applied using convolutional kernels of size 3×3 and $3 \times 3 \times 3$, respectively. In all papers, the ReLU activation function was applied after each convolution, followed by batch normalization [160]. Briefly, batch normalization adaptively normalizes the output from one layer before it is fed to the next layer during training [51]. The application of batch normalization is shown to give faster and more stable training, thereby allowing for deeper networks [141, 161]. All architectures

of papers I–IV used a sigmoidal activation function after the final convolutional layer, which outputs a number in the range [0,1] and thus can be regarded as the class 1 probability in the case of two classes $y \in \{0, 1\}$ [154]. The resulting probability maps can be converted to binary segmentation maps using a threshold value of, e.g., 0.50.

All CNNs were trained using the Adam (adaptive moment estimation) algorithm [162], which performs adaptive learning rate optimization. The CNN models trained from scratch were initialized using He weight initialization [163]. Regularization in the form of early stopping, i.e., stopping training if the validation set loss does not improve for a given number of training epochs, was used for all CNN models of papers I–III, and for the pre-training of CNN models in paper IV. Paper III also included regularization in the form of dropout [164]. With dropout regularization, a specified fraction of randomly selected neurons in one or more layers are set to zero at each training iteration and thus "dropped out". Both above techniques are among the most commonly used regularization techniques for neural networks [51, 154].

Loss functions

Different loss functions may be used, depending on the application. In medical image segmentation, it is often useful that the loss function can handle severe imbalance between the foreground and background classes. This motivated the Dice loss function (L_{Dice}) [165], which is based on the Sørensen-Dice similarity coefficient (cf. section 3.4.1 below), and is defined as:

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}_i}{\sum_{i=1}^{N} y_i^2 + \sum_{i=1}^{N} \hat{y}_i^2}, \tag{3.2}$$

where $\hat{y}_i$ denotes the predicted probability of pixel i belonging to the foreground (class 1), y_i is the true pixel class, and the summation runs over all N input image pixels.

In papers I, III, and IV, L_{Dice} was used as the objective function in all evaluated deep learning models. In paper II, the L_{Dice} was compared to the following three other losses: the f_β loss [106] with weighting parameter $\beta = \{2, 4\}$ and the binary cross-entropy loss (L_{BCE}). The f_β loss is a generalization of L_{Dice}, which allows for different weighting of false positive and false negative predictions. The weighting parameter β is a positive real-valued number. For $\beta = 1$, f_β is equal to the L_{Dice}, whereas higher β values ($\beta > 1$) increasingly penalize false negative classifications more than false positive classifications [106]. L_{BCE} provides an unweighted measure of the difference between the probability distributions of the ground truth and the predictions [141, 166], and was included in paper II due to its widespread use in classification tasks in particular but also in semantic segmentation tasks [166].

Image augmentation

CNNs generally require a substantial number of training samples to generalize well. Image augmentation, i.e., creating new modified images from the original ones, may be viewed as a form of regularization which aims at increasing the generalizability of the models by expanding the available image data [51, 167]. Multiple image augmentation methods exist [167]. In paper III, models trained using image augmentation in the form of 2D elastic deformation [168] of training set images was compared to models trained without the use of image augmentation. In paper IV, models were trained using image augmentation in the form of 3D rotation, zooming, and flipping [169], or 3D elastic deformation applied to training data.

Transfer learning

Transfer learning has been proposed as another means of addressing the challenge of limited training samples. In transfer learning, the knowledge gained in solving one problem is subsequently used to solve a separate problem, referred to as the source and target problem, respectively [170]. These problems can each be characterized using the concepts of domain and task. According to transfer learning definitions, a domain consists of a feature space and its probability distribution, while a task consists of a label space and a prediction function where the latter is learned from training data [171]. Using this terminology, the aim of transfer learning is to improve the learning of the target prediction function based on the information available in the source domain and task (see [171] for details). Different transfer learning approaches have been used in training deep learning models for a range of applications. For semantic segmentation and other vision applications, the most common transfer learning approach is to pretrain a model on a more data rich source domain and task, and subsequently fine–tune this model on the target domain and task, with the aim of improved target task learning [170]. In paper IV, this approach was used for cross-species transfer learning. CNN models were pretrained on the larger human HNC dataset and fine-tuned on the smaller canine HNC dataset, allowing all layers to be updated during fine-tuning, i.e., without freezing any of the pretrained layers.

3.3 Model evaluation strategies

When evaluating different supervised machine learning models there are generally two major goals: (1) To estimate the performance of all the trained models in order to select one best model, or in some cases a selection of the best models, and (2) To estimate the performance of the selected model(s) on previously unseen data, that is, to estimate the generalizability [149]. Both of the above goals can be met by dividing the available data into a training, validation, and test set, where the validation set is used for model selection and the test set is used to assess

generalizability. This is commonly the preferred approach in a data rich situation. Another commonly used approach to estimate prediction performance is K-fold cross-validation, where the data is first divided into K folds. For the kth fold of data, the remaining K–1 folds are used to train the model, and the kth fold is used to estimate performance. This procedure is repeated for $k = 1, 2, \ldots, K$ [149].

In paper I, we used five-fold cross-validation for model selection and a separate hold-out test set to assess model generalizability, whereas partitioning of data into a training, validation, and test set was used in paper II. All the models of paper III were evaluated using five-fold cross-validation. In paper IV, the human dataset was partitioned into a training, validation, and test set, whereas a modified four-fold cross-validation procedure was used to evaluate the models trained with canine data. The latter procedure consisted in using two folds of data for training, and the remaining two folds as a validation and test set, repeating the procedure until each fold had been used once as a validation set and once as a test set.

3.4 Performance measures

Auto-segmentation performance may be evaluated using quantitative geometric performance measures, qualitative assessment, dosimetric analysis, and/or quantification of time-savings [172], of which the former two were used in this thesis. The geometric performance measures can be subdivided into overlap-based metrics and distance-based metrics and are most often calculated patient-wise (i.e., on a per patient basis). Multiple patient-wise geometric performance metrics were included in papers I–IV of this thesis, with the aim of providing complementary information on the auto-segmentation quality, as recommended in [144]. Paper II also introduced quantitative structure-based performance metrics and included qualitative assessment.

3.4.1 Overlap-based metrics

Overlap metrics are calculated based on the two sets of voxels in the predicted auto-segmentation P and the ground truth delineation G, which reside in a voxel space. The most widely used overlap metric is the Sørensen-Dice similarity metric ($Dice$) [173, 174], defined as:

$$Dice = \frac{2|P \cap G|}{|P| + |G|}, \tag{3.3}$$

where $|P|$ and $|G|$ denote the cardinalities of P and G, and $|P \cap G|$ denotes the cardinality of their intersection. $Dice$ ranges from 0 (no overlap between P and G) to 1 (perfect overlap between P and G). As seen from equation 3.3, $Dice$ can

be calculated without knowing which set of voxels constitutes the ground truth. Thus, *Dice* does not separate between false positive and false negative predictions. To distinguish between these two error types, the positive predictive value (PPV, also known as precision) and the true positive rate (TPR, also known as sensitivity or recall) can be reported [144]. PPV and TPR are defined as [175]:

$$PPV = \frac{TP}{TP + FP}, \tag{3.4}$$

and

$$TPR = \frac{TP}{TP + FN}, \tag{3.5}$$

where TP, FP, and FN are the true positive, false positive, and false negative voxel predictions, respectively. Thus, PPV is the fraction of the predicted segmentation that overlaps with the ground truth delineation, whereas TPR is the fraction of the ground truth delineation that overlaps with the predicted segmentation. The above overlap-based metrics are illustrated in figure 3.6.

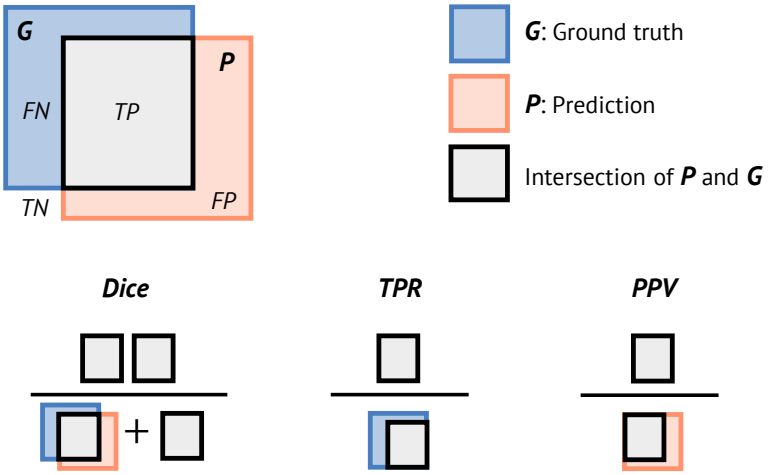


Figure 3.6: Illustration of the *Dice*, TPR and PPV performance metrics measuring the agreement between a reference (ground truth) volume G and a predicted volume P. *Dice*: Sørensen-Dice similarity coefficient; TPR: true positive rate; PPV: positive predictive value; TP: true positive; FP: false positive; TN: true negative; FN: false negative. Based on a figure from [144].

3.4.2 Distance-based metrics

If we now let P and G denote the sets of surface boundary voxels in the prediction and ground truth delineation, respectively, the Hausdorff distance (HD) between the two sets may be defined as [175]:

$$HD(G, P) = \max(h(P, G), h(G, P)), \tag{3.6}$$

where $h(P, G) = \max_{p \in P} \min_{g \in G} ||p - g||$ is the directed Hausdorff distance from P to G, which identifies the maximum of all (Euclidian) distances from the voxel points $p \in P$ to their nearest voxel point $g \in G$, and, conversely, $h(G, P) = \max_{g \in G} \min_{p \in P} ||g - p||$ is the directed Hausdorff distance from G to P.

As HD measures the maximum mismatch between P and G, it is sensitive to outliers. To mitigate this, the max operator in $h(P, G)$ and $h(G, P)$ can be replaced by a quantile [176] to exclude the most extreme observations. Commonly, the 95th percentile is used for this purpose, and the resulting metric is then referred to as the 95th percentile HD (HD_{95}).

The average surface distance (ASD) measures the typical distance between P and G and may be defined as [175]:

$$ASD(G, P) = \max(d(P, G), d(G, P)), \tag{3.7}$$

where $d(P, G) = \frac{1}{P} \sum_{p \in P} \min_{g \in G} ||p - g||$ is the directed average Hausdorff distances from P to G, which identifies the average of all (Euclidian) distances from the voxel points $p \in P$ to their nearest voxel point $g \in G$, and, conversely, $d(G, P) = \frac{1}{G} \sum_{g \in G} \min_{p \in P} ||g - p||$ is the directed average Hausdorff distance from G to P.

The surface distance metrics reported in papers I, III, and IV were calculated between sets P and G in accordance with the definitions in equations 3.6 and 3.7 (i.e., the metrics were defined as the maximum of the two calculations performed from P to G and vice versa), whereas the surface distance metrics reported in paper II were calculated solely based on the distances from P to G. Papers I–IV all included HD_{95} and ASD. In addition, the median (50th percentile HD) surface distance (MSD) was included in papers II and III as a supplement to ASD, since the median is more robust to outliers than the average.

The HD, HD_{95} and/or ASD metrics are commonly reported in the literature, and often accompany *Dice*, as they potentially provide complementary information [49, 172]. It should, however, be noted that the exact definitions and computational implementations of the distance-based metrics may vary between studies, potentially affecting the resulting metric values.

3.4.3 Structure-based metrics

The use of structure-based quantitative performance metrics may be an adequate supplement to the conventional geometric patient-wise metrics, particularly in the case of multiple ground truth structures. For the human HNC dataset, a large proportion of the patients (76 %) had nodal involvement and, therefore, several ground truth structures. This can complicate the interpretation of the patient-wise distance-based metrics. In general, the distance-based metrics can be skewed if the model falsely predicts or misses structures, as illustrated in 3.7. To allow for a more in-depth assessment of the auto-segmentation performance, particularly when multiple ground truth structures are present, a framework for structure-based performance evaluation was introduced in paper II.

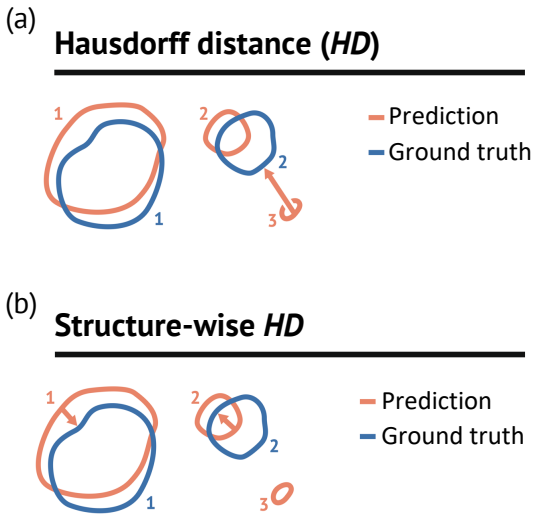


Figure 3.7: Example of how the patient-wise surface distance metrics can be affected by a falsely predicted structure (structure 3) (a), and how this issue can be alleviated by calculating the surface distance metrics only for predicted structures that have an overlap with the ground truth above a certain threshold (structures 1 and 2) (b). Adapted from paper II [2].

Briefly, the structure-based performance evaluation in paper II included the following: First, the predicted structures where 50 % of the predicted voxels overlapped with the ground truth, were defined as correctly identified by the CNN model. For each such structure, the surface distance metrics HD_{95}, MSD and ASD were calculated separately. Second, for each patient, structure-based sensitivity ($Sens_{GT}$) and positive predictive value (PPV_{CNN}) were calculated. $Sens_{GT}$ was defined as the per patient fraction of ground truth structures where $>$ 50 % of the ground truth voxels overlapped with a predicted structure. PPV_{CNN} was defined as the per patient fraction of predicted structures where $>$ 50 % of the predicted voxels

overlapped with a ground truth structure. Third, the volume of each predicted structure having > 50 % overlap with ground truth voxels ($Volume_{true}$) and the volume of each predicted structure having ≤ 50 % overlap with ground truth voxels ($Volume_{false}$) were calculated.

3.4.4 Qualitative assessment

Qualitative assessment of auto-segmentations by one or more clinical expert(s) can convey information about the degree of clinical acceptability that is not captured by quantitative metrics [172]. In paper II, auto-segmentations generated by the best-performing model in terms of mean per patient *Dice* were qualitatively assessed by an experienced oncologist (> 7 years' experience with manual contouring for radiotherapy in HNC patients). The oncologist was presented with the ground truth and CNN-generated contours of 15 randomly selected test set patients. For each patient, the oncologist was asked to identify which of the presented contours was generated by the CNN model, if possible. A scoring system ranging from 1 (no to little clinical value) to 10 (impossible to identify as CNN-generated, i.e., high clinical value) was used.

3.5 Statistical analysis

In papers I and III, the effect of algorithm and/or imaging modality on the quantitative segmentation performance was evaluated using the non-parametric Friedman test [177], which performs one-way repeated measures analysis of variance on ranks. If significant differences were detected, the Friedman test was followed by Nemenyi's one-sided many-to-one test or two-sided multiple pairwise comparisons [178]. In paper III, the effect of image augmentation on *Dice* performance was also assessed using the paired Wilcoxon signed-rank test [178]. A significance level of 0.05 was used for all statistical tests.

Chapter 4

Summary of papers

4.1 Paper I

A comparison of methods for fully automatic segmentation of tumors and involved nodes in PET/CT of head and neck cancers

In paper I, we evaluated and compared three different approaches to auto-segmentation of the GTV in human HNC patients based on PET/ceCT images, namely (i) PET thresholding, (ii) classical machine learning including feature engineering and various classification algorithms, and (iii) deep learning using CNNs. For the latter two approaches, the effect of imaging modality on auto-segmentation quality was assessed by comparing auto-segmentations obtained using ceCT, PET, or PET/ceCT image input. In addition, the effect of reducing the dynamic range of the ceCT images (windowing) was assessed. The segmentation task was considered a two-class problem (class 1: GTV-T and GTV-N; class 0: unaffected tissues).

Images were cropped from full size to a volume of interest including a 20 mm edge around the GTV-T and GTV-N in the axial plane and 1 mm in the z-direction. The 197 included patients were divided into a training set (157 patients) and a hold-out test set (40 patients). Five-fold cross-validation on the training set was used to tune hyper-parameters and compare models. For each approach, models were ranked separately based on the per patient cross-validation *Dice* performances and one model was selected for hold-out test set evaluation.

Four PET thresholding methods, six classical machine learning classifiers, and one deep learning CNN architecture (2D U-Net with L_{Dice} loss function) were evaluated. The resulting mean cross-validation *Dice* scores of the best-performing models were 0.62 for PET thresholding, 0.24 (ceCT) and 0.66 (PET; PET/ceCT) for classical machine learning, and 0.66 (ceCT), 0.68 (PET) and 0.74 (PET/ceCT) for deep learning. For the deep learning models that included ceCT image input, there was a small increase in mean *Dice* when the ceCT images were pre-processed with windowing (center: 60 HU; width: 100 HU). For the selected thresholding (absolute SUV threshold of 3.25), classical machine learning (SVM with PET voxel intensities in 3D neighborhoods), and deep learning (2D U-Net with PET/ceCT input) models, the mean per patient *Dice* on the hold-out test set was 0.63, 0.68, and 0.75, respectively. The PET/ceCT-based CNN model resulted in significantly higher cross-validation *Dice* than the single modality CNN models ($p \leq 0.0001$) and significantly better cross-validation and test set *Dice*, TPR, PPV, and surface distance metrics (ASD, HD_{95}), than the best-performing thresholding and classical machine learning models ($p \leq 0.0001$).

Our results show that deep learning was able to take advantage of the complementary information in the molecular PET and anatomical ceCT images to improve

auto-segmentation quality. In addition, the deep learning approach provided acceptable auto-segmentations based solely on ceCT images. This was not the case for the classical machine learning approach, which resulted in low-quality ceCT-based segmentation and no added benefit of combining PET and ceCT compared to using PET only. For solely PET-based segmentation, however, all three segmentation approaches provided more similar results. In conclusion, deep learning with multimodality PET/ceCT image input resulted in superior target coverage and less inclusion of unaffected tissues, strongly suggesting that this is the most appropriate approach, of those evaluated in paper I, for GTV auto-segmentation for radiotherapy of HNC.

4.2 Paper II

Deep learning-based auto-delineation of gross tumour volumes and involved nodes in PET/CT images of head and neck cancer patients

In paper II, deep learning-based auto-segmentation using the 2D U-Net architecture with PET, ceCT, or PET/ceCT input was evaluated more in-depth for the same task and patients as in paper I. As in paper I, the effect of reducing the dynamic range of the ceCT images (windowing) was also assessed. For each image input, four different loss functions were compared. To allow for thorough evaluation of auto-segmentation performance for all patients, including those with nodal involvement where the ground truth consists of multiple structures, we introduced new structure-based performance metrics to supplement existing conventional metrics. A selection of the resulting auto-segmentations was also assessed qualitatively by an experienced oncologist.

Compared to paper I, all images were cropped to a larger axial region measuring 191×265 mm^2. All image slices without ground truth contours were excluded. For model training, evaluation and comparison, the patients were divided into training (142 patients), validation (15 patients), and test (40 patients) sets, where the test set patients were the same as in paper I. In the case of ceCT windowing, a narrow soft tissue window with center value corresponding to the median HU value within the GTV-T and GTV-N of the training data was used (center: 70 HU; width: 200 HU). For each image input, the model with the highest mean per patient *Dice* on the validation set was selected for test set evaluation.

Choice of loss function had minor effect on validation performance as all performance metrics showed little variation for the different losses, whereas choice of image input had a substantial effect. Furthermore, there was a positive effect of ceCT windowing, particularly for solely ceCT-based auto-segmentation. The selected models obtained mean per patient *Dice* scores on the test set of 0.56 (ceCT; f_2 loss), 0.69 (PET; L_{Dice} loss), and 0.71 (PET/ceCT; f_2 loss). The ceCT, PET, and PET/ceCT-based CNN models identified on average 53 %, 77 % and 86 % of the ground truth structures in the test set patients, as measured by $Sens_{GT}$. Thus, the PET uptake was more important than the ceCT signal for the detection of ground truth structures. However, a considerable increase in $Sens_{GT}$ was seen when combining both PET and ceCT images, compared to using either of the single modality inputs. The PET/ceCT model also resulted in the lowest distance metrics, both on a per patient and a per structure basis. For all model inputs, the structure-based distance metrics (ASD, MSD, HD_{95}) were substantially lower than the corresponding patient-wise distance metrics, indicating that the patient-wise distance metrics were affected negatively by false positive structures predicted by the CNN models. As an example, the mean patient and structure-wise test set ASD of the PET/ceCT model was 4.7 mm and 1.0 mm, respectively. The ten-

dency to include false positive structures in the predicted auto-segmentations was also reflected by poor mean PPV_{CNN} test set scores of 28 % (ceCT), 45 % (PET) and 33 % (PET/ceCT). However, the average volume of falsely predicted structures ($Volume_{false}$) in the test set was small for all three evaluated models (ceCT: 0.56 cm^3; PET: 0.45 cm^3; PET/ceCT: 0.54 cm^3). The oncologist gave 13 of 15 randomly selected test set auto-segmentations predicted by the PET/ceCT-based CNN model a score of 8 or higher on a scale from 1 to 10.

In summary, as in paper I, the best overall segmentation performance was seen for the CNN model based on multimodality PET/ceCT input. The majority of the PET/ceCT-based auto-segmentations evaluated by the oncologist had considerable clinical value and could be used for radiotherapy with only minor to moderate revision. Furthermore, the inclusion of our newly introduced structure-based metrics allowed for a more in-depth quantitative analysis of the results.

4.3 Paper III

Deep learning-based automatic delineation of anal cancer gross tumour volume: a multimodality comparison of CT, PET and MRI

In paper III, deep learning-based auto-segmentation of the AC GTV was assessed for the first time, with emphasis on comparing single modality and multimodality combinations of PET, ceCT, ldCT and/or MR input. The impact of image augmentation consisting of 2D excessive random elastic deformations applied to 50 % of the training set images was also assessed. As in paper I and II, we used the 2D U-Net architecture.

Images were cropped to encompass approximately the same pelvic region (median in plane dimensions: 188 × 188 mm^2). 80 % of the image slices without ground truth contours were removed by random sampling from the datasets. Five-fold cross-validation was used to train models and evaluate performance. The ldCT and ceCT images were pre-processed using a narrow soft tissue window with {center, width} of {32, 220} HU and {70, 300} HU, respectively.

For both datasets, the highest mean cross-validation *Dice* (86-patient dataset: 0.76; 36-patient dataset: 0.83) was observed for the multimodality models based on PET and ceCT images with the inclusion of image augmentation. Contrary to our findings for HNC in paper I and II, similar mean *Dice* performances were obtained for the single modality models based on ceCT (86-patient dataset: 0.74; 36 patient dataset: 0.81). The models with the highest mean and median *Dice* generally also resulted in the lowest distance-based metrics (HD_{95}, ASD and MSD). The overall lowest distance metrics were observed for the 36-patient dataset, but all of the above models resulted in median $MSD \leq 2.50$ mm. For both datasets, there was a significant gain in *Dice* performance when using the radiotherapy planning ceCT images over the ldCT images, either alone or in combination with PET.

The somewhat poorer performance metrics for the 86-patient dataset was explained by a higher incidence of difficult to segment patients, not present in the 36-patient dataset. For most models, the inclusion of image augmentation moderately increased the mean and median *Dice*, but the effect was only statistically significant for the smallest dataset ($p < 0.001$). Based on our results for the 36-patient dataset, there was no added benefit of including MR information, compared to using PET and ceCT or solely ceCT as model input. However, the CNN model based solely on T2W MR images was the second-best single modality model and provided a mean cross-validation *Dice* of 0.77. In conclusion, the evaluated CNN approach provided high-quality automatic GTV segmentations based on either single modality or multimodality image input.

4.4 Paper IV

Automatic gross tumor segmentation of canine head and neck cancer using deep learning and cross-species transfer learning

In paper IV, deep learning-based auto-segmentation of the GTV in canine patients with HNC was evaluated for the first time. Two main approaches were assessed: (i) training CNN models from scratch based on ceCT images of canine patients, and (ii) using cross-species transfer learning where models were pretrained on ceCT images of human HNC patients and thereafter fine-tuned on ceCT images of canine patients. In this study, we used the 3D U-Net architecture. The impact of varying network complexity, image augmentation scheme, and different ceCT window settings of the input images was assessed.

Images were pre-processed to include a 191 $\times$ 265 $\times$ 173 mm^3 volume of interest. The human dataset was divided into training (126 patients), validation (31 patients) and test (40 patients) sets, whereas the canine dataset was divided into four equally sized folds. For each unique model configuration, the training of canine models was repeated four times using a cross-validation strategy where each fold of data was used once as a validation set and once as a test set in separate model runs. This strategy was selected to obtain a robust performance evaluation under the constraint of limited canine data. Based on the *Dice* performances on validation and test data, we selected one model trained from scratch and one transfer learning model for more in-depth performance evaluation.

Models trained from scratch or by using transfer learning resulted in relatively similar performances with mean validation (test) *Dice* scores in the range of 0.45–0.62 (0.39–0.55) and 0.52–0.57 (0.46–0.52), respectively. Despite breed-related variation in the canine head and neck anatomy and size, the average overlap with the expert ground truth contours was comparable to what has been obtained for solely ceCT-based GTV segmentation in human HNC patients, as exemplified by our results in paper II. The models selected for further performance evaluation had the lowest complexity in terms of U-Net depth and number of filters in the first layer, used image augmentation in the form of zooming, rotation and flipping, and included pre-processing of the ceCT images using a narrow soft tissue window. Auto-segmentation appeared particularly promising for canine patients with nasal cavity tumors, which was the most frequent tumor site in our dataset. For this subgroup of canine patients, both approaches resulted in mean *Dice* scores of 0.69. In conclusion, our results show promise for future application of deep learning-based auto-segmentation for canine HNC patients.

Chapter 5

Discussion

The overall contribution of this thesis is to expand the scholarly knowledge of the applicability of thresholding, classical machine learning and deep learning methods for automatic definition of target volumes used for radiotherapy in patients with cancer. Prior to the investigations in papers I and II, most studies on auto-segmentation of target volumes in patients with HNC, with the exception of [43, 179, 180], were based solely on FDG PET images, used semi- or fully automatic classical machine learning methods and/or included a low number of patients ($<$ 30) [181–188]. Solely FDG PET-based auto-segmentation methods will generally be limited by false positive and false negative uptake regions, as the FDG uptake is not fully cancer specific. The majority of the above PET-only studies relied on an operator to draw a region around or within the tumor [182–185] or focused on small pre-defined regions only including the tumor and its immediate background [186]. In contrast, papers I–II included a considerable number of patients, used larger image regions of interest, compared the segmentation performance obtained using single vs. multimodality PET and ceCT image input, and included deep learning methods. Papers III and IV are furthermore the first studies to investigate deep learning-based auto-segmentation of the GTV in patients with AC and canine HNC.

For all datasets investigated in this thesis, the performances of the best-performing auto-segmentation models were comparable to reported interobserver agreements between human experts performing the corresponding task manually, as measured by the *Dice* overlap metric. For ceCT and PET/ceCT-based manual contouring of the GTV-T in HNC, the mean interobserver *Dice* agreement has been reported to be 0.56–0.57 (ceCT) [189, 190] and 0.69 (PET/ceCT) [190]. In comparison, the best-performing deep learning models of papers I–II and paper IV obtained mean test set *Dice* scores of 0.55–0.56 (ceCT), 0.69 (PET), and 0.71–0.75 (PET/ceCT) for auto-segmentation of both the GTV-T and GTV-N. For AC, the median *Dice* interobserver agreement has been reported to be 0.80 and 0.74 for GTV-contouring based on PET/ceCT and T2W/DW/ceCT images [191]. In comparison, the deep learning model of paper III resulted in median cross-validation Dice scores of 0.78 (PET/ceCT; 86-patient dataset), 0.85 (PET/ceCT; 36-patient dataset) and 0.82 (T2W/ceCT; 36-patient dataset), with the inclusion of image augmentation. Though these interobserver studies were not conducted on the same datasets as in our analyses, with the exception of [191] which included a subset of the AC patients of paper III, and were limited by a relatively low number of patients (n = 10–19), the above comparisons give an indication of the performance of the given auto-segmentation models relative to human experts. As such, the above comparisons show promise for the future application of deep learning-based auto-segmentation of the GTV for the given cancers. However, caution is required not to over-interpret such comparisons in terms of clinical readiness as the errors made by human experts and a machine learning algorithm may differ substantially due to differences in contextual knowledge, which are not captured by the *Dice* overlap metric. It should also be noted that substantial inter-patient variability

in auto-segmentation performance occurred for all the investigated datasets.

5.1 Comparison of automatic segmentation methods

The segmentation method comparison of paper I identified deep learning as superior to PET thresholding and classical machine learning for the task of GTV segmentation in patients with HNC. This was particularly the case for ceCT and PET/ceCT image input, where deep learning substantially outperformed the classical machine learning approach. The differences between all three approaches were only moderate for PET-based segmentation, though deep learning ranked highest, followed by the SVM classical machine learning classifier and PET thresholding optimized with respect to *Dice*. Similar results were found in the first MICCAI (medical image computing and computer assisted intervention) challenge on PET tumor segmentation [186], which included both simulated, phantom, and clinical images of HNC and lung cancer patients. In [186] a CNN was ranked highest followed closely by classical machine learning methods. Hatt et al. [186] also included fixed thresholds of 40 % and 50 % of the SUV_{max} for comparison purposes, which they found to be among the poorest performing algorithms. The latter is also in line with our results for the fixed 41 % threshold. Thus, in the case of simple PET thresholding, optimization of the threshold value for the given task appears critical.

To our knowledge, no other studies but paper I and the above MICCAI challenge have compared thresholding, classical machine learning, and deep learning using data from HNC patients. The dataset used in [186] was created to fulfill specific criteria for a general benchmark dataset used to compare PET-based auto-segmentation algorithms [192], and the images were restricted to regions of interest that only included the tumor and its immediate background. In contrast, paper I focused on and thereby expanded the knowledge on the three aforementioned auto-segmentation approaches for a larger clinical HNC image dataset, using less restricted image regions of interest including both tumor and afflicted lymph nodes, and single vs. multimodality PET and/or ceCT image input.

The main focus of medical image segmentation studies and public challenges has shifted towards deep learning using CNNs over the past years [50, 138, 186, 193]. As CNNs extract complex latent features that would be difficult for a human to engineer, they are often superior to other methods for difficult tasks. This is supported by the success of CNNs for CT-based segmentation in a range of diagnoses and applications, as summarized in for example [50], as well as in our findings for CT-based segmentation in paper I. However, depending on the task and resources at hand, less complex methods could suffice, as exemplified by our results for solely PET-based segmentation in paper I. Thus, even though CNNs are

the current state-of-the-art approach for semantic segmentation, it can be argued that method comparisons as the one conducted in paper I remain relevant to document differences in performance and aid informed method selection. Ideally, such comparisons should be done on an openly available dataset.

5.2 Deep learning experiments

5.2.1 Impact of imaging modality

As for manual contouring, the optimal imaging modalities for auto-segmentation may depend on the task at hand, as well as image quality. For the deep learning analyses of papers I and II, molecular PET information in combination with anatomical and morphological ceCT information provided the highest quality segmentations. This is in line with the two other studies comparing single- and multimodality PET/ceCT-based segmentation of the GTV-T and GTV-N in HNC using deep learning [179, 180]. According to current delineation guidelines [38], MR should now be used in addition to PET/ceCT for manual target volume contouring in some of the HNC sites included in this thesis. The potential benefit of combining ceCT, PET, and anatomical MR images for deep learning-based auto-segmentation of the GTV-T and GTV-N in HNC was assessed recently by Ren et al. [194]. They found that all multimodal input combinations including PET information (PET/ceCT, PET/MRI, PET/ceCT/MRI) achieved comparable performances, suggesting that the inclusion of MR input was redundant, whereas combining anatomical MR and ceCT resulted in markedly lower performance. Though we did not evaluate MR images in our HNC studies, the structure-based performance evaluation in paper II also identified PET as vital for detecting a high proportion of the GTV structures, reflecting the sensitivity of FDG PET imaging in detecting FDG avid malignancies [104]. It should be noted that several studies have evaluated solely MR-based auto-segmentation of the primary tumor (GTV-T) in patients with either oropharyngeal or nasopharyngeal cancer [43, 195–197], resulting in particularly promising results for nasopharyngeal cancer [43, 197]. As nasopharyngeal cancer has different biology and clinical behavior than the other HNC sites [139], patients with nasopharyngeal cancer were not included in the human HNC dataset of this thesis, nor in Ren et al. [194].

CNN models based on PET/ceCT also ranked first for auto-segmentation of AC in paper III. However, for this task there was no statistically significant difference between PET/ceCT and ceCT-based segmentations. As in [194], there was no gain in segmentation performance when combining all three modalities (PET, ceCT and MR), whereas the CNN model based on T2W MR images ranked as the second-best single modality model. Apart from paper III, no other studies to date have evaluated deep learning-based automatic target volume segmentation in AC. Our results indicate that a high degree of overlap with manual delineations can

be obtained for several single modality or multimodality inputs. Particularly MR and PET/ceCT-based auto-segmentation are highly clinically relevant, as these modalities constitute state-of-the-art imaging for patients with AC in the U.S. and Europe [70,198]. Even though single modality imaging with ceCT is not considered state-of-the-art, ceCT-based auto-segmentation could be relevant in some clinical settings due to the associated cost and efficiency benefits. For the more frequent pelvic cancers, cervical and rectal cancer, auto-segmentation of the visible tumor tissue using CNNs has been evaluated for PET, T2W or T2W/DW image input [5, 199–205]. However, comparative studies of the effects different imaging modalities and modality combinations have on auto-segmentation performance as performed in paper III of this thesis have, to the best of our knowledge, not been conducted.

One limitation of paper III is the low number of patients in the dataset including MR images. The fact that the manual GTV delineations were based on routine rather than the study-specific MR images included for analysis in paper III may also have impacted on the auto-segmentation results for this modality. Furthermore, as the GTV according to clinical practice for AC included the anal canal and/or rectum circumference when involved [39, 135], the ranking of modalities in paper III may not be valid for segmentation of visible tumor tissue only. A comparison with visible tumor tissue segmentation would have been of interest, as the macroscopic tumor tissue could be a candidate for dose escalation in the context of dose painting [191] and is relevant for the derivation of image-based biomarkers. This was, however, outside the scope of paper III and would require manual re-contouring of a considerable number of patients.

5.2.2 Network architecture and configurations

The field of semantic segmentation is developing rapidly, and several new CNN architectures have been proposed after the U-Net [158] was first introduced, as reviewed in [206,207]. The U-Net is, however, a mature network with documented strong performance for a wide range of medical applications and diagnoses. Thus, the U-Net was a natural method of choice in our deep learning experiments, where the focus was on applicability rather than method development. Recent work by Isensee et al. [208] show that the U-Net surpasses specialized deep learning pipelines in a range of biomedical segmentation challenges, when a standardized framework for the experimental set-up, including pre- and post-processing, is used. This suggests, in line with the review by Litjens et al. [209], that the exact network architecture is not the driving force for obtaining good results. The use of 3D over 2D convolutions is, however, typically associated with a measurable increase in model performance and 3D convolutions are considered state-of-the-art for medical image segmentation [138, 180, 193]. However, there are other factors to consider when selecting between a 2D and 3D architecture. As 3D CNNs have more trainable parameters than their 2D counterparts, the model complexity, and

thus time, memory, and power requirements, are increased accordingly. This was the main reason for selecting a 2D architecture in papers I and II. The use of 3D CNNs could potentially also rely on more comprehensive image pre-processing due to e.g. out-of-plane voxel anisotropy. Furthermore, the surge in complexity of 3D over 2D CNNs could potentially impact on stability and reproducibility, expanding the required number of training samples. In paper III, the low number of patients in the smallest dataset as well as field of view limitations of the study-specific MR images also contributed to the choice of a 2D architecture. Even though the number of samples was equally low in paper IV, we opted for a 3D architecture with a desire to maximize the model performance on a difficult task. However, the inclusion of a reference 2D network could have been appropriate.

The 3D U-Net architecture with or without modifications, combined loss functions, image augmentation, and ensembles of models were used by the majority of participants, including the winning teams, in recent MICCAI challenges focusing on PET/ceCT-based segmentation of the GTV-T in oropharyngeal cancer patients collected from multiple centers [138, 193]. Measurable improvements in quantitative performance metrics would be expected for our PET/ceCT-based HNC experiments using similar configurations. The general positive effect of image augmentation observed in papers III and IV of this thesis are also in favor of the inclusion of augmentation schemes to increase auto-segmentation performance. The type of augmentation scheme is, however, highly relevant for the effect, and several augmentation options not included in this thesis could also be investigated as outlined and thoroughly reviewed in [206] and [167], respectively. One relevant option to consider, which can be viewed as a form of training and test-time image augmentation, is to include multiplanar image slices as input to a 2D network, thereby combining the benefits of 3D image information and the parameter efficiency of 2D CNNs [210].

5.2.3 Transfer learning experiments

For the transfer learning analysis of paper IV, we opted for a strategy where all layers of the pretrained models were fine-tuned on the target task. Another transfer learning approach is to fine-tune only certain layers, which is commonly done by having shared (i.e., frozen) lower layers and task-dependent (i.e., trainable) higher layers [51, 154]. The latter approach has been identified as superior to fine-tuning all layers for some medical applications, e.g., in [211]. However, according to a recent extensive study on transfer learning where several different medical segmentation tasks in humans were investigated, the former approach consistently gave better or equally good segmentation results as partial fine-tuning [170].

The results of Karimi et al. [170] further suggest that the effect of transfer learning on medical segmentation performance is highly task/data dependent and that significant effects of transfer learning, besides increased convergence speed, are

most likely when the number of training samples is low ($\sim$ 3–15), and/or the segmentation task is difficult. In Ghafoorian et al. [211], the improvement in white matter hyperintensity segmentation performance using transfer learning compared to training from scratch saturated at $\sim$ 75 target training samples. The segmentation task of paper IV is considered challenging and only 18 canine target samples were used for training in each model run. Thus, a transfer learning approach might be expected to be beneficial, based on the findings in the above studies. However, the pronounced dissimilarities between the human source data and canine target data, summarized in detail in paper IV, as well as the heterogeneity within the datasets, could be contributing factors to why transfer learning did not outperform training from scratch in paper IV.

5.3 Data cleaning and image pre-processing

Data cleaning and image pre-processing can have substantial impact on auto-segmentation performance. Even though all image and contour data were screened prior to the experiments in papers I–IV, we did not clean our datasets by removing anomalous cases such as patients with CT beam hardening artefacts or atypical ground truth structures. In a clinical setting, more careful attention to the quality and representativeness of the data used for training would be recommended to promote more stable segmentation performance. Furthermore, as HNC is a heterogeneous group of malignancies with different characteristics and incidence rates, several auto-segmentation studies, and public challenges on HNC focus solely on a single high-incidence and/or highly distinct primary tumor site [43, 138, 193, 195–197, 212, 213], potentially leading to improved results.

In this thesis, all images were pre-processed by defining a region or volume of interest encompassing the GTV structure(s), and in papers I–III most or all image slices without ground truth delineations were disregarded for both training and validation/test data. This pre-processing step was guided by the ground truth mask and/or the patient anatomy and was not fully automized. The definition of a region or volume of interest is customary in medical image segmentation studies, as it reduces the computational costs and makes the segmentation task less imbalanced, thereby limiting the number of false positive predictions. The potential effect of the region of interest on auto-segmentation performance is exemplified by the differences in performances for the PET/ceCT-based CNN models of papers I and II, where the mean $Dice$, HD_{95}, and ASD were 0.75, 5.79 mm and 1.36 mm in paper I (most restricted region of interest) and 0.71, 21.2 mm and 4.7 mm in paper II (less restricted region of interest).

To truly have a fully automatic segmentation pipeline the volume of interest definition should also be completely automatized. Furthermore, to avoid any "information leakage" the data used for model evaluation should ideally not be pre-

processed using ground truth information other than what can be inferred from the training data. With computational resources being more readily available and the increased use of 3D convolutions, there is a general trend towards having larger 3D image volumes as model input, as in paper IV of this thesis. However, the definition of a volume of interest is generally still required and associated with increased performance. Multi-step "course-to-fine" auto-segmentation pipelines, where automatic volume of interest definition is included prior to semantic segmentation, were presented by several top-ranked teams in the latest MICCAI head and neck tumor segmentation and outcome prediction (HECKTOR) challenge [214]. When using an existing target volume auto-segmentation tool in clinical practice, volume of interest definitions in new patients could alternatively also be done manually by a human expert as part of a semi-automatic contouring routine.

In papers I–II and IV, the union of the GTV-T and GTV-N were treated as a single class and the segmentation task was, thus, considered a two-class problem. This was motivated by the associated improved class-balance between the minority class, i.e. the union of the GTV-T and GTV-N, and the majority class, i.e. the background, compared to treating the GTV-T and GTV-N as two separate classes. Furthermore, the potentially similar image characteristics between the GTV-T and GTV-N, e.g., high FDG uptake, were considered a potential complicating factor for successful three-class segmentation. Preliminary experiments not included in paper II give some credence to this hypothesis, as inferior results were obtained for three-class segmentation compared to two-class segmentation. The fact that the GTV-N according to current practice is prescribed the same radiotherapy dose as the GTV-T also contributed to having the union of the GTV-T and GTV-N as one class. However, separation of the GTV-T and GTV-N segmentations in two different classes could still be advantageous, for example in the context of radiomics studies where it may be desirable to extract separate image-based features from the GTV-T and GTV-N [215]. Three-class segmentation of the GTV-T and GTV-N was addressed in the 2022 MICCAI HECKTOR challenge [216], where the top-ranked team obtained a mean *Dice* score of 0.80 and 0.78 for auto-segmentation of the GTV-T and GTV-N, respectively [214]. The high number of patients ($n = 883$, multi-centric data) and the fact that only oropharyngeal cancers were included in the associated dataset may have contributed to the successful use of a three-class approach in this case, along with methodological choices.

5.4 Model performance assessment

Dice and the other quantitative performance metrics included in this thesis provide information on the geometric agreement between contours but are limited by their inability to capture contextual differences, such as proximity to OARs, that may impact on clinical applicability [172, 217]. Several studies have demonstrated poor correlation between *Dice* or its related metric, the Jaccard similarity coefficient,

and clinical applicability in terms of qualitative scoring and dose plan quality, as summarized recently in [172]. The newer performance metrics surface Dice [52] ($Dice_S$) and added path length (APL) [218] are hypothesized to be more clinically meaningful, as they quantify the degree of revision required for an auto-segmented contour to match the contour it is compared against, within a predefined tolerance. Still, $Dice_S$ and APL are also global geometric metrics limited by their inability to provide information on the location of errors [217]. Vaassen et al. [218] found $Dice_S$ and APL to correlate better with time spent on manual revision of auto-generated OAR contours in patients with lung cancer, compared to the volumetric $Dice$ and HD metrics. Conflicting findings have, however, been reported for OAR and CTV segmentation in prostate cancer [219].

The structure-based performance metrics introduced in paper II do not capture contextual information. However, they allow for a more in-depth quantitative assessment of the auto-segmentation results, particularly in the case of multiple ground truth structures, and could be used to supplement conventional patient-wise metrics for such tasks. This is inferred from, and exemplified by, the performance analysis of paper II, as briefly detailed in the following. As the structure-based distance metrics were on average lower than the patient-wise distance metrics, it could be deduced that the patient-wise distances were in fact skewed by false positive structures. The structure-based sensitivity metric further showed quantitatively that the PET signal was vital for the sensitivity of the CNN models in detecting malignant structures, which is critical for radiotherapy. All evaluated models were, however, prone to including false positive structures in the predicted auto-segmentation. However, the structure-based volumetric size metric showed that the falsely predicted structures on average were small, and thus potentially could be filtered out in a post-processing step to reduce the need for manual revision.

Qualitative evaluation is generally highly valued in the medical community, as it provides an overall assessment of the clinical applicability of the contours. Such evaluation is, however, subjective, and resource-demanding. Furthermore, the design of the qualitative evaluation procedure requires attention to details. Among the factors to keep in mind is that human experts could be biased against contours they identify as generated by an AI algorithm. Regardless, the majority of the auto-generated contours (87 %) randomly selected for qualitative evaluation in paper II were found to have high clinical value (score of 8 or higher). Moreover, 13 % of the auto-generated contours could not be identified as generated by an AI algorithm (score of 10). In paper II both the manually delineated and auto-generated contours were shown simultaneously to the expert for each patient, and the colors of the two contours were assigned at random. To avoid the potential bias introduced by evaluating both contours simultaneously, a better strategy could have been to randomly draw either an auto-generated or a manually delineated contour from each patient. The latter approach was taken in Gooding et al. [220],

where both manual and automatic contours were assessed on a slice-by-slice basis by several experts.

Similarly to our initial qualitative assessment in paper II, the evaluation in [220] was inspired by the Turing test [221] and consisted in determining the origin of each contour (manual or automatic). Gooding et al. [220] found a better correspondence between time savings and the contour misclassification rates than the quantitative $Dice$, HD_{95}, and ASD metrics. In Lustberg et al. [222], there were only moderate differences in the actual time needed to revise contours receiving scores of 1–3 on a four-point scale. This questions the ability of humans to estimate the degree of revision required. In paper II, we used a ten-point scale where only the extremes were formally defined (score 1: little to no clinical value; score 10; not possible to separate automatic from manual contour). In retrospect, we acknowledge that this might be an unnecessarily wide and too loosely defined scale. Furthermore, due to the subjective nature of qualitative scoring, it would have been preferable to have multiple experts to evaluate the contours.

In summary, selection of appropriate performance measures is vital to properly assess auto-segmentation algorithms. As pointed out in [172], it can be instructive to consider the main goals of auto-segmentation, time savings and reduced contour variability, when designing and selecting performance evaluation metrics for future auto-segmentation studies. Using metrics that correlate with clinical usefulness will likely be increasingly important as auto-segmentation of target volumes evolves. However, the conventional geometric metrics are simpler and less resource-demanding to obtain compared to direct quantification of time savings, qualitative assessment and dosimetric evaluation, and will likely remain important for inter-study comparisons. Furthermore, the geometric measures are generally well suited to measure contour agreement [172].

Another important methodological choice is the model evaluation strategy. If the amount of data is too limited to have a separate hold-out test set as we had in papers I–II, a cross-validation procedure as the one used in paper IV would be recommended to obtain independent validation and test set results, even though this reduces the number of training samples compared to the simple cross-validation procedure used in paper III. Furthermore, the test set evaluations of this thesis only included internal test data. To fully evaluate the model generalizability an external test set with data from one or several other institutions would be required.

Chapter 6

Conclusions and future perspectives

The overall aim of this thesis was to investigate the use of machine learning methods for automatic GTV segmentation in medical images. The results presented in this thesis consolidates deep learning as a method of choice for auto-segmentation within the medical domain and adds to a growing body of literature showing that CNNs can provide high quality GTV segmentations in various cancer diagnoses. For all cancer types investigated in papers I–IV of this thesis, namely HNC, AC, and canine HNC, the highest ranked CNN models resulted in automatic segmentations which on average had an overlap with manual ground truth contours that was on par with reported interobserver agreements for manual contouring. Based on these findings, it is inferred that further investigation of CNNs for automatic GTV segmentation in the given cancer diagnoses is highly warranted. The inter-institutional generalizability of our models was, however, not assessed, as all papers only included single center data. Thus, inter-institutional generalizability would be relevant to assess in any related future work.

The comparison of methods for auto-segmentation of the GTV-T and GTV-N in patients with HNC, conducted in paper I, showed that the added benefit of using CNNs over less complex algorithms can depend on the imaging modality in question. For segmentations based solely on PET images, all investigated methods, including conventional PET thresholding, classical machine learning algorithms, and deep learning with CNNs, provided fair overall quantitative segmentation performance. For ceCT and PET/ceCT-based segmentation, on the other hand, CNNs outperformed classical machine learning algorithms. The combination of PET and ceCT image input resulted in superior CNN performance for the given task in both papers I and II. Thus, of the evaluated approaches, deep learning with CNNs using multimodality PET/ceCT image input would be recommended

for segmentation of the GTV-T and GTV-N in HNC. Furthermore, the proposed structure-based performance metrics introduced in paper II provided a more in-depth assessment of the CNN model characteristics and can with advantage be used as a supplement to existing conventional patient-wise metrics, particularly in the case of multiple ground truth structures. A possible extension of the analyses of papers I and II would be to evaluate the potential added benefit of using a 3D rather than a 2D CNN, and to assess the effect of different image augmentation schemes.

Combined PET/ceCT input also resulted in the highest overlap with the ground truth GTV delineations for deep learning-based segmentation of the AC GTV in paper III. For this task, however, comparable CNN performance was obtained for single modality ceCT image input. Several multimodality models incorporating T2W MR images also performed well, and the CNN based solely on T2W images ranked as the second-best single modality model. Thus, the findings in paper III suggest that the given segmentation task could be performed satisfactorily based on several different image inputs. Paper III was, however, limited by the low number of patients with MR images, and the lack of test set evaluation. A natural extension of paper III would, therefore, be to evaluate and compare MR, PET, and ceCT-based auto-segmentation for a larger number of AC patients, also including test set evaluation to assess model generalizability. Ideally, such an analysis would include two different segmentation tasks, namely segmentation of the GTV as defined in routine clinical practice/paper III, and segmentation of the visible tumor tissue only.

The analysis of canine HNC patients in paper IV indicated that cross-species transfer learning from a larger human HNC cohort could increase segmentation quality for individual patients. On average, however, the best segmentation performance was achieved training canine HNC models from scratch. Differences in source and target domains as well as the heterogeneous nature of HNC may have complicated the transfer learning analyses. In paper IV only ceCT-based segmentation was investigated as this was the common modality available for both humans and dogs. Pre contrast imaging is, however, part of the routine CT imaging protocol for the canine patients. Thus, a relevant next step for automatic GTV segmentation in this group of patients would be to investigate the potential added benefit of including both pre and post contrast CT images. Furthermore, a subgroup analysis of dogs with nasal cavity tumors would be warranted, as this was the most numerous subgroup that on average obtained the most promising auto-segmentations. Due to the low number of canine patients it could be reasonable to opt for a 2D CNN in further analyses of this dataset, or alternatively compare the performances obtained with 2D and 3D CNNs. Implicitly including 3D information as input to a 2D network by the use of image augmentation as in [210] could also be relevant.

Though deep learning-based segmentation of the GTV provided highly promising results, substantial between-patient variability in segmentation quality could

occur. The poorest performance was generally seen for patients with atypical image and/or GTV characteristics, indicating that the network was limited by the number of representative training samples. Such variability is not unique for the auto-segmentation analyses included in this thesis. The relatively low number of training samples available within the medical domain has been pointed out as one major unresolved issue for the application of deep learning-based medical auto-segmentation [223]. Strategies that could lead to more stable auto-segmentation performance with a limited number of training samples is thus an important field of study. In this context, several strategies are relevant. First, the effect of more careful data curation, e.g. excluding patients with image artefacts and in the case of HNC restrict the auto-segmentation task to the single most frequent cancer site, could be investigated. Second, potential ways to incorporate clinical and contextual information into the supervised deep learning models could possibly lead to improved and more stable performance. Third, strategies such as transfer learning, unsupervised learning, and learning from noisy labels could all potentially reduce the required number of training samples [224, 225]. The latter two approaches are particularly relevant if the number of labeled samples is the limiting factor of the dataset size, which was not the case in this thesis. On another note, multi-centric studies, potentially also including openly available datasets [226], could be performed to increase the number of patients, also allowing for systematically comparing the stability and generalizability of single and multi-center-based models.

Given the risk of variable auto-segmentation performance, full automation of GTV segmentation in clinical practice without any human quality assurance and/or contour revision is not likely at present. In this context, optimal implementation of GTV auto-segmentation tools and documentation of their clinical usefulness are two important fields to consider in future work. With respect to optimal implementation, there is a need for further development of methods that can inform clinicians about uncertainties in the predicted auto-contours and identify new patients that deviate from the training set distribution, also known as out-of-distribution samples [223, 227]. Identification of patients not suited for an existing auto-segmentation tool is important to keep clinicians from wasting time on low-quality auto-contours, as it can be assumed that there is no time-savings associated with auto-segmentation if more than about 40 % of the auto-contour requires manual revision. Clinical usefulness in terms of time-savings and reduced interobserver variability has already been documented for deep learning-based automatic GTV segmentation in several cancer diagnoses, as exemplified in [43, 44]. However, further studies quantifying the clinical usefulness of auto-segmentation tools are likely needed to motivate their implementation into the clinical workflow. As direct quantification of clinical usefulness can be highly resource demanding in itself, it is also relevant to derive and report surrogate performance metrics that correlate with clinical usefulness [172]. A standardized framework with recommendations for the calculation and reporting of quantitative and qualitative perfor-

mance metrics could also be a welcome addition to the field of auto-segmentation for radiotherapy.

Bibliography

[1] Groendahl AR, Knudtsen IS, Huynh BN, Mulstad M, Moe YM, Knuth F, Tomic O, Indahl UG, Torheim T, Dale E, Malinen E, and Futsaether CM. A comparison of methods for fully automatic segmentation of tumors and involved nodes in PET/CT of head and neck cancers. *Physics in Medicine and Biology*, 6:065012, 2021.

[2] Moe YM, Groendahl AR, Tomic O, Dale E, Malinen E, and Futsaether CM. Deep learning-based auto-delineation of gross tumour volumes and involved nodes in PET/CT images of head and neck cancer patients. *European Journal of Nuclear Medicine and Molecular Imaging*, 9:2782–2792, 2021.

[3] Groendahl AR, Moe YM, Kaushal CK, Huynh BN, Rusten E, Tomic O, Hernes E, Hanekamp B, Undseth C, Guren MG, Malinen E, and Futsaether CM. Deep learning-based automatic delineation of anal cancer gross tumour volume: a multimodality comparison of CT, PET and MRI. *Acta Oncologica*, 61:1:89–96, 2021.

[4] Groendahl AR, Huynh BN, Tomic O, Søvik Å, Dale E, Malinen E, Skogmo HK, and Futsaether CM. Automatic gross tumor segmentation of canine head and neck cancer using deep learning and cross-species transfer learning. To be submitted to *Frontiers in Veterinary Science*.

[5] Knuth F, Adde IA, Huynh BN, Groendahl AR, Winter RM, Negård A, Holmedal SH, Meltzer S, Ree AH, Flatmark K, Dueland S, Hole KH, Seierstad T, Redalen KR, and Futsaether CM. MRI-based automatic segmentation of rectal cancer using 2D U-Net on two independent cohorts. *Acta Oncologica*, 61:2:255–263, 2022.

[6] Knuth F, Groendahl AR, Winter RM, Torheim T, Negård A, Holmedal SH, Bakke KM, Meltzer S, Futsaether CM, and Redalen KR. Semi-automatic tumor segmentation of rectal cancer based on functional magnetic resonance imaging. *Physics and imaging in radiation oncology*, 22:77–84, 2022.

[7] Grøndahl AR, Kuttner S, Elschot M, Bathen TF, Sundset R, Knudtsen IS, and Futsæther CM. Maskinlæring og medisinske bilder for bedre diagnostikk

og persontilpasset kreftbehandling. *HMT*, volume 5, 2018. Published in Norwegian.

[8] Ren J, Huynh BN, Groendahl AR, Tomic O, Futsaether CM, and Korreman SS. PET normalizations to improve deep learning auto-segmentation of head and neck tumors in 3D PET/CT. In *Head and Neck Tumor Segmentation and outcome prediction*. HECKTOR 2021. Lecture Notes in Computer Science, vol 13209. Springer, 2022.

[9] Huynh BN, Ren J, Groendahl AR, Tomic O, Korreman SS, and Futsaether CM. Comparing deep learning and conventional machine learning for outcome prediction of head and neck cancer in PET/CT. In *Head and Neck Tumor Segmentation and outcome prediction*. HECKTOR 2021. Lecture Notes in Computer Science, vol 13209. Springer, 2022.

[10] Groendahl AR, Huynh BN, Moe YM, Kaushal CK, Rusten E, Tomic O, Hernes E, Hanekamp B, Undseth C, Guren MG, Malinen E, and Futsaether CM. Deep learning-based automatic delineation of anal cancer gross tumour volume: A multimodality comparison of CT, PET and MRI. Presented at: *BiGART 2021* 2021-10-05–2021-10-06.

[11] Huynh BN, Groendahl AR, Moe YM, Tomic O, Dale E, Malinen E, and Futsaether CM. Deep learning for automatic segmentation of head and neck cancers in PET/CT images: the simpler, the better. Presented at: *BiGART 2021* 2021-10-05–2021-10-06.

[12] Knuth F, Adde IA, Huynh BN, Groendahl AR, Winter RM, Negård A, Holmedal SH, Meltzer S, Ree AH, Flatmark K, Dueland S, Holde KH, Seierstad T, Redalen KR, and Futsaether CM. MRI-based automatic segmentation of rectal cancer using 2D U-Net on two independent cohorts. Presented at: *BiGART 2021* 2021-10-05–2021-10-06.

[13] Groendahl AR, Huynh BN, Moe YM, Kaushal CK, Rusten E, Tomic O, Hernes E, Hanekamp B, Undseth C, Guren MG, Dale E, Malinen E, and Futsaether CM. Deep learning for automatic target volume delineation. Presented at: *NACP 2020/21* 2021-04-11–2021-04-13.

[14] Knuth F, Groendahl AR, Winter RM, Torheim T, Negård A, Holmedal SH, Bakke KM, Meltzer S, Futsaether CM, and Redalen K. Influence of functional MRI sequences on automatic segmentation of rectal cancer. Presented at: *NACP 2020/21* 2021-04-11–2021-04-13.

[15] Groendahl AR, Moe Y, Kaushal CK, Tomic O, Dale E, Guren MG, Malinen E, and CM Futsæther. Machine learning for automatic tumor segmentation. Presented at: *Mini-MedFys 2020* 2020-02-03.

[16] Knuth F, Groendahl AR, Torheim T, Negård A, Holmedal SH, Bakke KM, Meltzer S, Futsaether CM, and Redalen K. Functional MR-based automatic

tumor segmentation of rectal cancer. Presented at: *BiGART 2019* 2019-05-22–2019-05-24.

[17] Knuth F, Groendahl AR, Torheim T, Negård A, Holmedal SH, Bakke KM, Meltzer S, Futsaether CM, and Redalen K. Automatic tumour delineation in rectal cancer using functional MRI and machine learning. Presented at: *ESTRO 2019* 2021-04-26–2021-04-30.

[18] Knuth F, Groendahl AR, Torheim T, Negård A, Holmedal SH, Bakke KM, Meltzer S, Futsaether CM, and Redalen K. Automatic tumor delineation in rectal cancer using functional MRI and machine learning. Presented at: *MedFys 2019* 2019-02-04–2019-02-06.

[19] Groendahl AR, Huynh BN, Moe YM, Kaushal CK, Rusten E, Tomic O, Hernes E, Hanekamp B, Undseth C, Guren MG, Malinen E, and Futsaether CM. Deep learning-based automatic delineation of anal cancer gross tumour volume: A multimodality comparison of CT, PET and MRI. Presented at: *BiGART 2021* 2021-10-06.

[20] Huynh BN, Groendahl AR, Moe YM, Tomic O, Dale E, Malinen E, and Futsaether CM. Tuning deep learning models for automatic segmentation of head and neck cancers in PET/CT images. Presented at: *ESTRO 2021* 2021-08-27–2021-08-31.

[21] Moe YM, Groendahl AR, Mulstad M, Tomic O, Indahl UG Dale E, Malinen E, and Futsaether CM. Deep learning for automatic tumour segmentation in PET/CT images of patients with head and neck cancers. Presented at: *MIDL 2019* 2019-07-08–2019-07-10.

[22] Groendahl AR, Knudtsen IS, Mulstad M, Tomic O, Moe YM, Indahl UG, Torheim T, Dale E, Malinen E, and Futsaether CM. Automatic tumour delineation of head and neck cancers in PET/CT images using thresholding and machine learning methods. Presented at: *BiGART 2019* 2019-05-22–2019-05-24.

[23] Langberg GS, Groendahl AR, Midtfjord AD, Tomic O, Liland KH, Knudtsen IS, Dale E, Malinen E, and Futsaether CM. Establishing a complete radiomics framework for biomarker identification and outcome prediction using PET/CT images of head neck cancers. Presented at: *BiGART 2019* 2019-05-22–2019-05-24.

[24] Groendahl AR, Mulstad M, Moe YM, Knudtsen IS, Torheim T, Tomic O, Indahl UG, Malinen E, Dale E, and Futsaether CM. Comparison of automatic tumour segmentation approaches for head and neck cancers in PET/CT images. Presented at: *ESTRO 2019* 2021-04-26–2021-04-30.

[25] Groendahl AR, Midtfjord AD, Langberg GS, Tomic O, Indahl UG, Knudtsen IS, Malinen E, Dale E, and Futsaether CM. Prediction of treatment outcome

for head and neck cancers using radiomics of PET/CT images. Presented at: *ESTRO 2019* 2021-04-26–2021-04-30.

[26] Kimberly D Miller, Leticia Nogueira, Theresa Devasia, Angela B Mariotto, K Robin Yabroff, Ahmedin Jemal, Joan Kramer, and Rebecca L Siegel. Cancer treatment and survivorship statistics, 2022. *CA: a cancer journal for clinicians*, 72(5):409–436, 2022.

[27] Cancer treatment statistics. Cancer Research UK [Internet]. `https://www.cancerresearchuk.org/health-professional/cancer-statistics/treatment`. Cited 2022 Dec 1.

[28] Radiotherapy board on target 2: updated guidance for image-guided radiotherapy. The Royal College of Radiologists, Society and College of Radiographers, Institute of Physics and Engineering in Medicine. London: The Royal College of Radiologists, 2021.

[29] Vincent Grégoire, Matthias Guckenberger, Karin Haustermans, Jan JW Lagendijk, Cynthia Ménard, Richard Pötter, Ben J Slotman, Kari Tanderup, Daniela Thorwarth, Marcel Van Herk, et al. Image guidance in radiation therapy for better cure of cancer. *Molecular oncology*, 14(7):1470–1491, 2020.

[30] Kyle Wang and Joel E Tepper. Radiation therapy-associated toxicity: Etiology, management, and prevention. *CA: A Cancer Journal for Clinicians*, 71(5):437–454, 2021.

[31] Brian O'Sullivan, RB Rumble, P Warde, IMRT Indications Expert Panel, et al. Intensity-modulated radiotherapy in the treatment of head and neck cancer. *Clinical oncology*, 24(7):474–487, 2012.

[32] Tejpal Gupta, JaiPrakash Agarwal, Sandeep Jain, Reena Phurailatpam, Sadhana Kannan, Sarbani Ghosh-Laskar, Vedang Murthy, Ashwini Budrukkar, Ketayun Dinshaw, Kumar Prabhash, et al. Three-dimensional conformal radiotherapy (3d-crt) versus intensity modulated radiation therapy (imrt) in squamous cell carcinoma of the head and neck: a randomized controlled trial. *Radiotherapy and Oncology*, 104(3):343–348, 2012.

[33] KS Clifford Chao, Navneet Majhail, Chih-jen Huang, Joseph R Simpson, Carlos A Perez, Bruce Haughey, and Gershon Spector. Intensity-modulated radiation therapy reduces late salivary toxicity without compromising tumor control in patients with oropharyngeal carcinoma: a comparison with conventional techniques. *Radiotherapy and oncology*, 61(3):275–280, 2001.

[34] Lisa A Kachnic, Kathryn Winter, Robert J Myerson, Michael D Goodyear, John Willins, Jacqueline Esthappan, Michael G Haddock, Marvin Rotman, Parag J Parikh, Howard Safran, et al. Rtog 0529: a phase 2 evaluation of dose-painted intensity modulated radiation therapy in combination with 5-fluorouracil and mitomycin-c for the reduction of acute morbidity in car-

cinoma of the anal canal. *International Journal of Radiation Oncology* Biology* Physics*, 86(1):27–33, 2013.

[35] N Hodapp. The icru report 83: prescribing, recording and reporting photon-beam intensity-modulated radiation therapy (imrt). *Journal of the ICRU*, 10(1), 2010.

[36] M Kosmin, J Ledsam, B Romera-Paredes, R Mendes, S Moinuddin, D de Souza, L Gunn, C Kelly, CO Hughes, A Karthikesalingam, et al. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. *Radiotherapy and Oncology*, 135:130–140, 2019.

[37] Shalini K Vinod, Michael G Jameson, Myo Min, and Lois C Holloway. Uncertainties in volume delineation in radiation oncology: a systematic review and recommendations for future studies. *Radiotherapy and Oncology*, 121(2):169–179, 2016.

[38] Kenneth Jensen, Jeppe Friborg, Christian Rønn Hansen, Eva Samsøe, Jørgen Johansen, Maria Andersen, Bob Smulders, Elo Andersen, Martin Skovmos Nielsen, Jesper Grau Eriksen, et al. The danish head and neck cancer group (dahanca) 2020 radiotherapy guidelines. *Radiotherapy and Oncology*, 151:149–151, 2020.

[39] Michael Ng, Trevor Leong, Sarat Chander, Julie Chu, Andrew Kneebone, Susan Carroll, Kirsty Wiltshire, Samuel Ngan, and Lisa Kachnic. Australasian gastrointestinal trials group (agitg) contouring atlas and planning guidelines for intensity-modulated radiotherapy in anal cancer. *International Journal of Radiation Oncology* Biology* Physics*, 83(5):1455–1462, 2012.

[40] Barbara Segedin and Primoz Petric. Uncertainties in target volume delineation in radiotherapy–are they relevant and what can we do about them? *Radiology and oncology*, 50(3):254–262, 2016.

[41] Samantha Cox, Anne Cleves, Enrico Clementel, Elizabeth Miles, John Staffurth, and Sarah Gwynne. Impact of deviations in target volume delineation–time for a new rtqa approach? *Radiotherapy and Oncology*, 137:1–8, 2019.

[42] Amy Tien Yee Chang, Li Tee Tan, Simon Duke, and Wai-Tong Ng. Challenges for quality assurance of target volume delineation in clinical trials. *Frontiers in Oncology*, 7:221, 2017.

[43] Li Lin, Qi Dou, Yue-Ming Jin, Guan-Qun Zhou, Yi-Qiang Tang, Wei-Lin Chen, Bao-An Su, Feng Liu, Chang-Juan Tao, Ning Jiang, et al. Deep learning for automated contouring of primary tumor volumes by mri for nasopharyngeal carcinoma. *Radiology*, 291(3):677–686, 2019.

[44] Grzegorz Chlebus, Hans Meine, Smita Thoduka, Nasreddin Abolmaali, Bram Van Ginneken, Horst Karl Hahn, and Andrea Schenk. Reducing inter-

observer variability and interaction time of mr liver volumetry by combining automatic cnn-based liver segmentation and manual corrections. *PloS one*, 14(5):e0217228, 2019.

[45] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[46] Kristy K Brock. Adaptive radiotherapy: moving into the future. In *Seminars in radiation oncology*, volume 29, page 181. NIH Public Access, 2019.

[47] Peter J. Elliott, John M. Knapman, and Wolfgang Schlegel. Interactive image segmentation for radiation treatment planning. *IBM Systems Journal*, 31(4):620–634, 1992.

[48] Daniel J Withey and Zoltan J Koles. Medical image segmentation: Methods and software. In *2007 Joint Meeting of the 6th International Symposium on Noninvasive Functional Source Imaging of the Brain and Heart and the International Conference on Functional Biomedical Imaging*, pages 140–143. IEEE, 2007.

[49] Gregory Sharp, Karl D Fritscher, Vladimir Pekar, Marta Peroni, Nadya Shusharina, Harini Veeraraghavan, and Jinzhong Yang. Vision 20/20: perspectives on automated image segmentation for radiotherapy. *Medical physics*, 41(5):050902, 2014.

[50] Carlos E Cardenas, Jinzhong Yang, Brian M Anderson, Laurence E Court, and Kristy B Brock. Advances in auto-segmentation. In *Seminars in radiation oncology*, volume 29, pages 185–197. Elsevier, 2019.

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

[52] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernadino Romera-Paredes, et al. Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study. *Journal of medical Internet research*, 23(7):e26151, 2021.

[53] Tomaž Vrtovec, Domen Močnik, Primož Strojan, Franjo Pernuš, and Bulat Ibragimov. Auto-segmentation of organs at risk for head and neck radiotherapy planning: from atlas-based to deep learning methods. *Medical physics*, 47(9):e929–e950, 2020.

[54] Deep learning segmentation models. Raysearch Laboratories [Internet]. `https://www.raysearchlabs.com/media/publications/`

deep-learning-segmentation-model-catalogue/. Cited 2022 Dec 15.

[55] Ai-rad companion organs rt. Siemens Healthineers [Internet]. https://www.siemens-healthineers.com/en-th/digital-health-solutions/digital-solutions-overview/clinical-decision-support/ai-rad-companion/organs-rt. Cited 2022 Dec 15.

[56] Challenges. Medical Image Computing and Computer Assisted Intervention [Internet]. https://conferences.miccai.org/2022/en/MICCAI2022-CHALLENGES.html. Cited 2022 Dec 15.

[57] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):1–13, 2022.

[58] GM Cooper and Hausman RE. *The cell: a molecular approach*. Sinauer Associates, 7 edition, 2013.

[59] R Paul Symonds, John A Mills, and Angela Duxbury. *Walter and Miller's Textbook of Radiotherapy: Radiation Physics, Therapy and Oncology*. Elsevier Health Sciences, 8th edition, 2019.

[60] Cancer classification. National Cancer Institute [Internet]. https://training.seer.cancer.gov/disease/categories/classification.html. Cited 2022 Nov 1.

[61] Types of cancer. Cancer Research UK [Internet]. https://www.cancerresearchuk.org/what-is-cancer/how-cancer-starts/types-of-cancer?_gl=1*aq99k5*_ga*ODAxMTg2MDM2LjE2Njg5NDk3ODc.*_ga_58736Z2GNN*MTY2ODk0OTc4Ni4xMi4xLjE2Njg5NDk4NDIuNC4wLjA.&_ga=2.229484316.1888412447.1668949787-801186036.1668949787. Cited 2022 Nov 1.

[62] What is the tnm cancer staging system? The Union for International Cancer Control (UICC) [Internet]. https://www.uicc.org/resources/tnm. Cited 2022 Nov 15.

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

[64] Global health estimates: Leading causes of death. World Health Organization [Internet]. https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death. Cited 2022 Nov 15.

[65] Mayur D Mody, James W Rocco, Sue S Yom, Robert I Haddad, and Nabil F Saba. Head and neck cancer. *The Lancet*, 2021.

[66] ML Hedberg and JR Grandis. *The Molecular Pathogenesis of Head and Neck Cancer, In: The Molecular Basis of Cancer*. W.B. Saunders, 4th edition, 2015.

[67] Matt Lechner, Jacklyn Liu, Liam Masterson, and Tim R Fenton. Hpv-associated oropharyngeal cancer: Epidemiology, molecular biology and clinical management. *Nature Reviews Clinical Oncology*, 19(5):306–327, 2022.

[68] J-P Machiels, C René Leemans, W Golusinski, C Grau, L Licitra, and V Gregoire. Squamous cell carcinoma of the oral cavity, larynx, oropharynx and hypopharynx: Ehns–esmo–estro clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 31(11):1462–1475, 2020.

[69] Farhad Islami, Jacques Ferlay, Joannie Lortet-Tieulent, Freddie Bray, and Ahmedin Jemal. International trends in anal cancer incidence rates. *International journal of epidemiology*, 46(3):924–938, 2017.

[70] S Rao, MG Guren, K Khan, G Brown, Andrew G Renehan, SE Steigen, E Deutsch, E Martinelli, and D Arnold. Anal cancer: Esmo clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 32(9):1087–1100, 2021.

[71] Heather L Gardner, Joelle M Fenger, and Cheryl A London. Dogs as a model for cancer. *Annual review of animal biosciences*, 4:199, 2016.

[72] Comparative oncology program. National Cancer Institute [Internet]. `https://ccr.cancer.gov/comparative-oncology-program`. Cited 2022 Nov 15.

[73] Joshua D Schiffman and Matthew Breen. Comparative oncology: what dogs and other species can teach us about humans with cancer. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1673):20140231, 2015.

[74] Jennie L Rowell, Donna O McCarthy, and Carlos E Alvarez. Dog models of naturally occurring cancer. *Trends in molecular medicine*, 17(7):380–388, 2011.

[75] Michael W Nolan, Michael S Kent, and Mary-Keara Boss. Emerging translational opportunities in comparative oncology with companion canine cancers: radiation oncology. *Frontiers in Oncology*, 9:1291, 2019.

[76] Deli Liu, Huan Xiong, Angela E Ellis, Nicole C Northrup, Kevin K Dobbin, Dong M Shin, and Shaying Zhao. Canine spontaneous head and neck squamous cell carcinomas represent their human counterparts at the molecular level. *PLoS genetics*, 11(6):e1005277, 2015.

[77] Katherine S Hansen and Michael S Kent. Imaging in non-neurologic oncologic treatment planning of the head and neck. *Frontiers in veterinary science*, 6:90, 2019.

[78] Louise B Brønden, Thomas Eriksen, and Annemarie T Kristensen. Oral malignant melanomas and other head and neck neoplasms in danish dogs-data from the danish veterinary cancer registry. *Acta Veterinaria Scandinavica*, 51(1):1–6, 2009.

[79] Sara A Ayres and Julius M Liptak. Head and neck tumors. *Veterinary surgical oncology*, pages 143–181, 2022.

[80] Stella S Bastianello. A survey on neoplasia in domestic species over a 40-year period from 1935 to 1974 in the republic of south africa. VI. Tumours occurring in dogs. *Onderstepoort Journal of Veterinary Research*, 50:199–220, 1983.

[81] K Arnesen, H Gamlem, E Glattre, J Grøndalen, L Moe, and K Nordstoga. The norwegian canine cancer register 1990-1998. report from the project "cancer in the dog". *Eur J Comp Anim Pract*, 11:159–169, 2001.

[82] David M Vail, Douglas H Thamm, and Julias Liptak. *Withrow and MacEwen's Small Animal Clinical Oncology*. Elsevier Health Sciences, 6th edition, 2019.

[83] Donald J Meuten. *Tumors in domestic animals*. John Wiley & Sons, 5th edition, 2020.

[84] John Farrelly and Margaret C McEntee. A survey of veterinary radiation facilities in 2010. *Veterinary Radiology & Ultrasound*, 55(6):638–643, 2014.

[85] Edward C Halperin, David E Wazer, Carlos A Perez, and Luther W Brady. *Perez & Brady's principles and practice of radiation oncology E-book*. Wolters Kluwer, 7th edition, 2018.

[86] Gisele C Pereira, Melanie Traughber, and Raymond F Muzic. The role of imaging in radiation therapy planning: past, present, and future. *BioMed research international*, 2014, 2014.

[87] Stefan Zachow, Michael Zilske, and Hans-Christian Hege. 3d reconstruction of individual anatomy from medical image data: Segmentation and geometry processing. In *Proceedings of the CADFEM Users Meeting*, 2007.

[88] Richard Bibb, Dominic Eggbeer, and Abby Paterson. *Medical modelling: the application of advanced design and rapid prototyping techniques in medicine*. Woodhead Publishing, 2014.

[89] C Glide-Hurst, D Chen, H Zhong, and IJ Chetty. Changes realized from extended bit-depth and metal artifact reduction in ct. *Medical physics*, 40(6Part1):061711, 2013.

[90] A Murphy and J Kube. Windowing (ct). Radiopaedia.org [Internet]. `https://radiopaedia.org/articles/windowing-ct`. Cited 2022 Nov 1.

[91] W Richard Webb, Wiliam E Brant, and Nancy M Major. *Fundamentals of Body CT*. Elsevier Health Sciences, 2019.

[92] Shane Minogue, Charles Gillham, Maeve Kearney, and Laura Mullaney. Intravenous contrast media in radiation therapy planning computed tomography scans–current practice in ireland. *Technical Innovations & Patient Support in Radiation Oncology*, 12:3–15, 2019.

[93] K Williams and Heidi Probst. Use of iv contrast media in radiotherapy planning ct scans: a uk audit. *Radiography*, 22:S28–S32, 2016.

[94] Kyongtae T Bae. Intravenous contrast medium administration and scan timing at ct: considerations and approaches. *Radiology*, 256(1):32–61, 2010.

[95] Simon R Cherry and Magnus Dahlbom. *PET: physics, instrumentation, and scanners*. Springer, 2006.

[96] John Lilley. *Nuclear physics: principles and applications*. John Wiley & Sons, 2013.

[97] Wei Jiang, Yamn Chalich, and M Jamal Deen. Sensors for positron emission tomography applications. *Sensors*, 19(22):5019, 2019.

[98] William W Moses. Fundamental limits of spatial resolution in pet. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 648:S236–S240, 2011.

[99] Thomas Beyer, Luc Bidaut, John Dickson, Marc Kachelriess, Fabian Kiessling, Rainer Leitgeb, Jingfei Ma, Lalith Kumar Shiyam Sundar, Benjamin Theek, and Osama Mawlawi. What scans we will read: imaging instrumentation trends in clinical oncology. *Cancer Imaging*, 20(1):1–38, 2020.

[100] Terry Jones and David W Townsend. History and future technical innovation in positron emission tomography. *Journal of Medical Imaging*, 4(1):011013, 2017.

[101] Magdy M Khalil. *Basic sciences of nuclear medicine*. Springer Nature, 2021.

[102] Paul E Kinahan and James W Fletcher. Positron emission tomography-computed tomography standardized uptake values in clinical practice and assessing response to therapy. In *Seminars in Ultrasound, CT and MRI*, volume 31, pages 496–505. Elsevier, 2010.

[103] Federica Orsini, Alice Lorenzoni, Erinda Puta, and Giuliano Mariani. *Positron-Emitting Radiopharmaceuticals for Diagnostic Applications*, pages 85–98. Springer International Publishing, Cham, 2017.

[104] Ronald Boellaard, Roberto Delgado-Bolton, Wim JG Oyen, Francesco Giammarile, Klaus Tatsch, Wolfgang Eschner, Fred J Verzijlbergen, Sally F Barrington, Lucy C Pike, Wolfgang A Weber, et al. Fdg pet/ct: Eanm

procedure guidelines for tumour imaging: version 2.0. *European journal of nuclear medicine and molecular imaging*, 42(2):328–354, 2015.

[105] ACRACNMSNMMISPR Practice parameter for performing fdg-pet/ct in oncology. ACR Guidelines and Standards Committee [Internet]. `https://www.acr.org/-/media/ACR/Files/Practice-Parameters/fdg-pet-ct.pdf`. Cited 2022 Nov 1.

[106] R Hashemi, C Lisanti, and W Bradley. *MRI: The Basics*. Wolters Kluwer Health, 4th edition, 2017.

[107] C Westbrook and J Talbot. *MRI in Practice*. John Wiley & Sons, 5th edition, 2018.

[108] DW McRobbie, EA Moore, MJ Graves, and MR Prince. *MRI from Picture to Proton*. Cambridge university press, 3rd edition, 2017.

[109] BM Dale, MA Brown, and RC Semelka. *MRI: basic principles and applications*. John Wiley & Sons, 2015.

[110] Pierre Decazes, Pauline Hinault, Ovidiu Veresezan, Sébastien Thureau, Pierrick Gouel, and Pierre Vera. Trimodality pet/ct/mri and radiotherapy: a mini-review. *Frontiers in Oncology*, 10:614008, 2021.

[111] Steven P Sourbron and David L Buckley. Classic models for dynamic contrast-enhanced mri. *NMR in Biomedicine*, 26(8):1004–1027, 2013.

[112] Geetha Soujanya Chilla, Cher Heng Tan, Chenjie Xu, and Chueh Loo Poh. Diffusion weighted magnetic resonance imaging and its recent trend—a survey. *Quantitative imaging in medicine and surgery*, 5(3):407, 2015.

[113] Carmelo Messina, Rodolfo Bignone, Alberto Bruno, Antonio Bruno, Federico Bruno, Marco Calandri, Damiano Caruso, Pietro Coppolino, Riccardo De Robertis, Francesco Gentili, et al. Diffusion-weighted imaging in oncology: an update. *Cancers*, 12(6):1493, 2020.

[114] Martin Georg Zeilinger, Michael Lell, Pascal Andreas Thomas Baltzer, Arnd Dörfler, Michael Uder, and Matthias Dietzel. Impact of post-processing methods on apparent diffusion coefficient values. *European radiology*, 27(3):946–955, 2017.

[115] Moti Freiman, Stephan D Voss, Robert V Mulkern, Jeannette M Perez-Rossello, Michael J Callahan, and Simon K Warfield. In vivo assessment of optimal b-value range for perfusion-insensitive apparent diffusion coefficient imaging. *Medical physics*, 39(8):4832–4839, 2012.

[116] Getting external beam radiotherapy. American Cancer Society [Internet]. `https://www.cancer.org/treatment/treatments-and-side-effects/treatment-types/radiation/external-beam-radiation-therapy.html`. Updated 2022 Apr 10; Cited 2022 Nov 1.

[117] Rajamanickam Baskar, Jiawen Dai, Nei Wenlong, Richard Yeo, and Kheng-Wei Yeoh. Biological response of cancer cells to radiation treatment. *Frontiers in molecular biosciences*, 1:24, 2014.

[118] Little JB. *Principal Cellular and Tissue Effects of Radiation.* In *Holland-Frei Cancer Medicine.* BC Decker, 6th edition, 2003.

[119] Deborah E Citrin. Recent developments in radiotherapy. *New England journal of medicine*, 377(11):1065–1075, 2017.

[120] Laura Beaton, Steve Bandula, Mark N Gaze, and Ricky A Sharma. How rapid advances in imaging are defining the future of precision radiation oncology. *British journal of cancer*, 120(8):779–790, 2019.

[121] Michael Baumann, Mechthild Krause, Jens Overgaard, Jürgen Debus, Søren M Bentzen, Juliane Daartz, Christian Richter, Daniel Zips, and Thomas Bortfeld. Radiation oncology in the era of precision medicine. *Nature Reviews Cancer*, 16(4):234–249, 2016.

[122] Byungchul Cho. Intensity-modulated radiation therapy: a review with a physics perspective. *Radiation oncology journal*, 36(1):1, 2018.

[123] JR Mortier and L Blackwood. Treatment of nasal tumours in dogs: a review. *Journal of Small Animal Practice*, 61(7):404–415, 2020.

[124] Valerie J Poirier, Ethel SY Koh, Johnson Darko, Andre Fleck, Christopher Pinard, and David M Vail. Patterns of local residual disease and local failure after intensity modulated/image guided radiation therapy for sinonasal tumors in dogs. *Journal of Veterinary Internal Medicine*, 35(2):1062–1072, 2021.

[125] Anne T Davis, Antony L Palmer, and Andrew Nisbet. Can ct scan protocols used for radiotherapy treatment planning be adjusted to optimize image quality and patient dose? a systematic review. *The British journal of radiology*, 90(1076):20160406, 2017.

[126] Charles M Washington and Dennis T Leaver. *Principles and Practice of Radiation Therapy.* Elsevier Health Sciences, 2015.

[127] Mohammad Hussein, Ben JM Heijmen, Dirk Verellen, and Andrew Nisbet. Automation in intensity modulated radiotherapy treatment planning—a review of recent innovations. *The British journal of radiology*, 91(1092):20180270, 2018.

[128] Olga L Green, Lauren E Henke, and Geoffrey D Hugo. Practical clinical workflows for online and offline adaptive radiation therapy. In *Seminars in radiation oncology*, volume 29, pages 219–227. Elsevier, 2019.

[129] Neil G Burnet, Simon J Thomas, Kate E Burton, and Sarah J Jefferies. Defining the tumour and target volumes for radiotherapy. *Cancer Imaging*, 4(2):153, 2004.

[130] Douglas Jones. Icru report 50—prescribing, recording and reporting photon beam therapy. *Medical Physics*, 21(6):833–834, 1994.

[131] André Wambersie. Icru report 62, prescribing, recording and reporting photon beam therapy (supplement to icru report 50). *Icru News*, 1999.

[132] Anne Kiil Berthelsen, Jane Dobbs, Elisabeth Kjellén, Torsten Landberg, Torgil R Möller, Per Nilsson, Lena Specht, and André Wambersie. What's new in target volume definition for radiologists in icru report 71? how can the icru volume definitions be integrated in clinical practice? *Cancer Imaging*, 7(1):104, 2007.

[133] Vincent Grégoire, Kian Ang, Wilfried Budach, Cai Grau, Marc Hamoir, Johannes A Langendijk, Anne Lee, Quynh-Thu Le, Philippe Maingon, Chris Nutting, et al. Delineation of the neck node levels for head and neck tumors: a 2013 update. dahanca, eortc, hknpcsg, ncic ctg, ncri, rtog, trog consensus guidelines. *Radiotherapy and Oncology*, 110(1):172–181, 2014.

[134] Vincent Grégoire, Mererid Evans, Quynh-Thu Le, Jean Bourhis, Volker Budach, Amy Chen, Abraham Eisbruch, Mei Feng, Jordi Giralt, Tejpal Gupta, et al. Delineation of the primary tumour clinical target volumes (ctv-p) in laryngeal, hypopharyngeal, oropharyngeal and oral cavity squamous cell carcinoma: Airo, caca, dahanca, eortc, georcc, gortec, hknpcsg, hncig, iagkht, lprhht, ncic ctg, ncri, nrg oncology, phns, sbrt, somera, sro, sshno, trog consensus guidelines. *Radiotherapy and Oncology*, 126(1):3–24, 2018.

[135] R Glynne-Jones, V Goh, A Aggarwal, H Maher, S Dubash, and R Hughes. *Anal Carcinoma.* In *Target Volume Definition in Radiation Oncology.* Springer, 2015.

[136] Dahanca radiotherapy guidelines, 2013. Danish Head and Neck Cancer Group.

[137] Maria Thor, Aditya Apte, Rabia Haq, Aditi Iyer, Eve LoCastro, and Joseph O Deasy. Using auto-segmentation to reduce contouring and dose inconsistency in clinical trials: the simulated impact on rtog 0617. *International Journal of Radiation Oncology* Biology* Physics*, 109(5):1619–1626, 2021.

[138] Vincent Andrearczyk, Valentin Oreiller, Sarah Boughdad, Catherine Cheze Le Rest, Hesham Elhalawani, Mario Jreige, John O Prior, Martin Vallières, Dimitris Visvikis, Mathieu Hatt, et al. Overview of the hecktor challenge at miccai 2021: automatic head and neck tumor segmentation and outcome prediction in pet/ct images. In *Head and neck tumor segmenta-*

tion and outcome prediction, pages 1–37. HECKTOR 2021. Lecture Notes in Computer Science, vol 13209. Springer, 2021.

[139] Jon Magne Moan, Cecilie Delphin Amdal, Eirik Malinen, Jørund Graadal Svestad, Trond Velde Bogsrud, and Einar Dale. The prognostic role of 18f-fluorodeoxyglucose pet in head and neck cancer depends on hpv status. *Radiotherapy and Oncology*, 140:54–61, 2019.

[140] Kathinka S Slørdahl, Dagmar Klotz, Jan-Åge Olsen, Eva Skovlund, Christine Undseth, Heidi Larsen Abildgaard, Morten Brændengen, Arild Nesbakken, Stein Gunnar Larsen, Bettina A Hanekamp, et al. Treatment outcomes and prognostic factors after chemoradiotherapy for anal cancer. *Acta Oncologica*, 60(7):921–930, 2021.

[141] F Chollet. *Deep learning with Python*. Simon and Schuster, 2018.

[142] Rafael C Gonzalez and Richard E Woods. *Digital Image Processing*. Pearson, 4th edition, 2018.

[143] Wilhelm Burger and Mark J Burge. *Digital image processing: an algorithmic introduction*. Springer, 3rd edition, 2022.

[144] Mathieu Hatt, John A Lee, Charles R Schmidtlein, Issam El Naqa, Curtis Caldwell, Elisabetta De Bernardi, Wei Lu, Shiva Das, Xavier Geets, Vincent Gregoire, et al. Classification and evaluation strategies of auto-segmentation approaches for pet: Report of aapm task group no. 211. *Medical physics*, 44(6):e1–e42, 2017.

[145] J Bernard Davis, Beatrice Reiner, Marius Huser, Cyrill Burger, Gábor Székely, and I Frank Ciernik. Assessment of 18f pet signals for automatic target volume definition in radiotherapy treatment planning. *Radiotherapy and Oncology*, 80(1):43–50, 2006.

[146] Turid Torheim, Eirik Malinen, Knut Håkon Hole, Kjersti Vassmo Lund, Ulf G Indahl, Heidi Lyng, Knut Kvaal, and Cecilia M Futsaether. Autodelineation of cervical cancers using multiparametric magnetic resonance imaging and machine learning. *Acta oncologica*, 56(6):806–812, 2017.

[147] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[148] Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14, 2001.

[149] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2nd edition, 2009.

[150] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[151] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[152] S Raschka and V Mirjalili. *Python machine learning: Machine learning and deep learning with python, scikit-Learn, and TensorFlow*. Packt, 2 edition, 2017.

[153] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[154] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc, 2022.

[155] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[156] Yann LeCun. Generalization and network design strategies. *Connectionism in perspective*, 19(143-155):18, 1989.

[157] Yi-Tong Zhou and Rama Chellappa. Computation of optical flow using a neural network. In *ICNN*, pages 71–78, 1988.

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

[159] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[160] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[161] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? *Advances in neural information processing systems*, 31, 2018.

[162] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[163] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[164] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[165] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

[166] Shruti Jadon. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–7. IEEE, 2020.

[167] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[168] Patrice Y Simard, David Steinkraus, John C Platt, et al. Best practices for convolutional neural networks applied to visual document analysis. In *International conference on document analysis and recognition*, volume 3. Edinburgh, 2003.

[169] Modules: Image augmentation. Deoxys-image [Internet]. `https://deoxys-image.readthedocs.io/en/latest/modules.html#module-deoxys_image.augmentation`. Cited 2022 Nov 1.

[170] Davood Karimi, Simon K Warfield, and Ali Gholipour. Transfer learning in medical image segmentation: New insights from analysis of the dynamics of model parameters and learned representations. *Artificial Intelligence in Medicine*, 116:102078, 2021.

[171] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[172] Michael V Sherer, Diana Lin, Sharif Elguindi, Simon Duke, Li-Tee Tan, Jon Cacicedo, Max Dahele, and Erin F Gillespie. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. *Radiotherapy and Oncology*, 160:185–191, 2021.

[173] Thorvald A Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske skrifter*, 5:1–34, 1948.

[174] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[175] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15(1):1–28, 2015.

[176] Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863, 1993.

[177] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701, 1937.

[178] Myles Hollander, Douglas A Wolfe, and Eric Chicken. *Nonparametric statistical methods*. John Wiley & Sons, 2013.

[179] Zhe Guo, Ning Guo, Kuang Gong, Quanzheng Li, et al. Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multimodality network. *Physics in Medicine & Biology*, 64(20):205015, 2019.

[180] Vincent Andrearczyk, Valentin Oreiller, Martin Vallières, Joel Castelli, Hesham Elhalawani, Mario Jreige, Sarah Boughdad, John O Prior, and Adrien Depeursinge. Automatic segmentation of head and neck tumors and nodal metastases in pet-ct scans. In *Medical imaging with deep learning*, pages 33–43. PMLR, 2020.

[181] B Berthon, M Evans, C Marshall, N Palaniappan, N Cole, V Jayaprakasam, T Rackley, and E Spezi. Head and neck target delineation using a novel pet automatic segmentation algorithm. *Radiotherapy and Oncology*, 122(2):242–247, 2017.

[182] Alessandro Stefano, Salvatore Vitabile, Giorgio Russo, Massimo Ippolito, Maria Gabriella Sabini, Daniele Sardina, Orazio Gambino, Roberto Pirrone, Edoardo Ardizzone, and Maria Carla Gilardi. An enhanced random walk algorithm for delineation of head and neck cancers in pet studies. *Medical & biological engineering & computing*, 55(6):897–908, 2017.

[183] Albert Comelli, Alessandro Stefano, Giorgio Russo, Maria Gabriella Sabini, Massimo Ippolito, Samuel Bignardi, Giovanni Petrucci, and Anthony Yezzi. A smart and operator independent system to delineate tumours in positron emission tomography scans. *Computers in Biology and Medicine*, 102:1–15, 2018.

[184] Albert Comelli, Alessandro Stefano, Samuel Bignardi, Giorgio Russo, Maria Gabriella Sabini, Massimo Ippolito, Stefano Barone, and Anthony Yezzi. Active contour algorithm with discriminant analysis for delineating tumors in positron emission tomography. *Artificial Intelligence in Medicine*, 94:67–78, 2019.

[185] Albert Comelli, Alessandro Stefano, Giorgio Russo, Samuel Bignardi, Maria Gabriella Sabini, Giovanni Petrucci, Massimo Ippolito, and Anthony Yezzi. K-nearest neighbor driving active contours to delineate biological tumor volumes. *Engineering Applications of Artificial Intelligence*, 81:133–144, 2019.

[186] Mathieu Hatt, Baptiste Laurent, Anouar Ouahabi, Hadi Fayad, Shan Tan, Laquan Li, Wei Lu, Vincent Jaouen, Clovis Tauber, Jakub Czakon, et al. The first miccai challenge on pet tumor segmentation. *Medical image analysis*, 44:177–195, 2018.

[187] Bin Huang, Zhewei Chen, Po-Man Wu, Yufeng Ye, Shi-Ting Feng, Ching-Yee Oliver Wong, Liyun Zheng, Yong Liu, Tianfu Wang, Qiaoliang Li, et al. Fully automated delineation of gross tumor volume for head and neck cancer on pet-ct using deep learning: a dual-center study. *Contrast media & molecular imaging*, 2018, 2018.

[188] Huan Yu, Curtis Caldwell, Katherine Mah, Ian Poon, Judith Balogh, Robert MacKenzie, Nader Khaouam, and Romeo Tirona. Automated radiation targeting in head-and-neck cancer using region-based texture analysis of pet and ct images. *International Journal of Radiation Oncology* Biology* Physics*, 75(2):618–625, 2009.

[189] David Bird, Andrew F Scarsbrook, Jonathan Sykes, Satiavani Ramasamy, Manil Subesinghe, Brendan Carey, Daniel J Wilson, Neil Roberts, Gary McDermott, Ebru Karakaya, et al. Multimodality imaging with ct, mr and fdg-pet for radiotherapy target volume delineation in oropharyngeal squamous cell carcinoma. *BMC cancer*, 15(1):1–10, 2015.

[190] Shivakumar Gudi, Sarbani Ghosh-Laskar, Jai Prakash Agarwal, Suresh Chaudhari, Venkatesh Rangarajan, Siji Nojin Paul, Rituraj Upreti, Vedang Murthy, Ashwini Budrukkar, and Tejpal Gupta. Interobserver variability in the delineation of gross tumour volume and specified organs-at-risk during imrt for head and neck cancers and the impact of fdg-pet/ct on such variability at the primary site. *Journal of medical imaging and radiation sciences*, 48(2):184–192, 2017.

[191] Espen Rusten, Bernt Louni Rekstad, Christine Undseth, Ghazwan Al-Haidari, Bettina Hanekamp, Eivor Hernes, Taran Paulsen Hellebust, Eirik Malinen, and Marianne Grønlie Guren. Target volume delineation of anal cancer based on magnetic resonance imaging or positron emission tomography. *Radiation Oncology*, 12(1):1–8, 2017.

[192] Beatrice Berthon, Emiliano Spezi, Paulina Galavis, Tony Shepherd, Aditya Apte, Mathieu Hatt, Hadi Fayad, Elisabetta De Bernardi, Chiara D Soffientini, C Ross Schmidtlein, et al. Toward a standard for the evaluation of pet-auto-segmentation methods following the recommendations of aapm task group no. 211: Requirements and implementation. *Medical physics*, 44(8):4098–4111, 2017.

[193] Vincent Andrearczyk, Valentin Oreiller, Mario Jreige, Martin Vallieres, Joel Castelli, Hesham Elhalawani, Sarah Boughdad, John O Prior, and Adrien Depeursinge. Overview of the hecktor challenge at miccai 2020: automatic head and neck tumor segmentation in pet/ct. In *Head and Neck Tumor*

Segmentation, pages 1–21. HECKTOR 2021. Lecture Notes in Computer Science, vol 12603. Springer, 2021.

[194] Jintao Ren, Jesper Grau Eriksen, Jasper Nijkamp, and Stine Sofia Korreman. Comparing different ct, pet and mri multi-modality image combinations for deep learning-based head and neck tumor segmentation. *Acta Oncologica*, 60(11):1399–1406, 2021.

[195] Roque Rodríguez Outeiral, Paula Bos, Abrahim Al-Mamgani, Bas Jasperse, Rita Simões, and Uulke A van der Heide. Oropharyngeal primary tumor segmentation for radiotherapy planning on magnetic resonance imaging using deep learning. *Physics and imaging in radiation oncology*, 19:39–44, 2021.

[196] Kareem A Wahid, Sara Ahmed, Renjie He, Lisanne V van Dijk, Jonas Teuwen, Brigid A McDonald, Vivian Salama, Abdallah SR Mohamed, Travis Salzillo, Cem Dede, et al. Evaluation of deep learning-based multiparametric mri oropharyngeal primary tumor auto-segmentation and investigation of input channel effects: Results from a prospective imaging registry. *Clinical and translational radiation oncology*, 32:6–14, 2022.

[197] Qiaoliang Li, Yuzhen Xu, Zhewei Chen, Dexiang Liu, Shi-Ting Feng, Martin Law, Yufeng Ye, and Bingsheng Huang. Tumor segmentation in contrast-enhanced magnetic resonance imaging for nasopharyngeal carcinoma: deep learning with convolutional neural network. *BioMed Research International*, 9128527, 2018.

[198] Al B Benson, Alan P Venook, Mahmoud M Al-Hawary, Lynette Cederquist, Yi-Jen Chen, Kristen K Ciombor, Stacey Cohen, Harry S Cooper, Dustin Deming, Paul F Engstrom, et al. Anal carcinoma, version 2.2018, nccn clinical practice guidelines in oncology. *Journal of the National Comprehensive Cancer Network*, 16(7):852–871, 2018.

[199] Liyuan Chen, Chenyang Shen, Zhiguo Zhou, Genevieve Maquilan, Kevin Albuquerque, Michael R Folkert, and Jing Wang. Automatic pet cervical tumor segmentation by combining deep learning and anatomic prior. *Physics in Medicine & Biology*, 64(8):085019, 2019.

[200] Joohyung Lee, Ji Eun Oh, Min Ju Kim, Bo Yun Hur, and Dae Kyung Sohn. Reducing the model variance of a rectal cancer segmentation network. *IEEE Access*, 7:182725–182733, 2019.

[201] Jiazhou Wang, Jiayu Lu, Gan Qin, Lijun Shen, Yiqun Sun, Hongmei Ying, Zhen Zhang, and Weigang Hu. A deep learning-based autosegmentation of rectal tumors in mr images. *Medical physics*, 45(6):2560–2564, 2018.

[202] Mengmeng Wang, Peiyi Xie, Zhao Ran, Junming Jian, Rui Zhang, Wei Xia, Tao Yu, Caifeng Ni, Jinhui Gu, Xin Gao, et al. Full convolutional network based multiple side-output fusion architecture for the segmentation of rectal

tumors in magnetic resonance images: a multi-vendor study. *Medical physics*, 46(6):2659–2668, 2019.

[203] Jongin Kim, Ji Eun Oh, Joohyung Lee, Min Ju Kim, Bo Yun Hur, Dae Kyung Sohn, and Boreom Lee. Rectal cancer: Toward fully automatic discrimination of t2 and t3 rectal cancers using deep convolutional neural network. *International Journal of Imaging Systems and Technology*, 29(3):247–259, 2019.

[204] Mumtaz Hussain Soomro, Matteo Coppotelli, Silvia Conforto, Maurizio Schmid, Gaetano Giunta, Lorenzo Del Secco, Emanuele Neri, Damiano Caruso, Marco Rengo, and Andrea Laghi. Automated segmentation of colorectal tumor in 3d mri using 3d multiscale densely connected convolutional neural network. *Journal of healthcare engineering*, 2019, 2019.

[205] Stefano Trebeschi, Joost JM van Griethuysen, Doenja MJ Lambregts, Max J Lahaye, Chintan Parmar, Frans CH Bakers, Nicky HGM Peters, Regina GH Beets-Tan, and Hugo JWL Aerts. Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric mr. *Scientific reports*, 7(1):1–9, 2017.

[206] Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and Ghassan Hamarneh. Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*, 54(1):137–178, 2021.

[207] Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: a review. *Frontiers in Cardiovascular Medicine*, 7:25, 2020.

[208] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[209] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[210] Mathias Perslev, Erik Bjørnager Dam, Akshay Pai, and Christian Igel. One network to segment them all: A general, lightweight system for accurate 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 30–38. Springer, 2019.

[211] Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles RG Guttmann, Frank-Erik de Leeuw, Clare M Tempany, Bram van Ginneken, et al. Transfer learning for domain adaptation in mri: Application in brain lesion segmentation. In

International conference on medical image computing and computer-assisted intervention, pages 516–524. Springer, 2017.

[212] Mazin Abed Mohammed, Mohd Khanapi Abd Ghani, Net al Arunkumar, Salama A Mostafa, Mohamad Khir Abdullah, and MA Burhanuddin. Trainable model for segmenting and identifying nasopharyngeal carcinoma. *Computers & Electrical Engineering*, 71:372–387, 2018.

[213] Kuo Men, Xinyuan Chen, Ye Zhang, Tao Zhang, Jianrong Dai, Junlin Yi, and Yexiong Li. Deep deconvolutional neural network for target segmentation of nasopharyngeal cancer in planning computed tomography images. *Frontiers in oncology*, page 315, 2017.

[214] Oral session 1: Automatic segmentation of primary tumors and lymph nodes, chaired by Karim Wahid. The HECKTOR challenge at MICCAI 2022: automatic head and neck tumor segmentation and outcome prediction in PET/CT images, 2022.

[215] Marta Bogowicz, Stephanie Tanadini-Lang, Matthias Guckenberger, and Oliver Riesterer. Combined ct radiomics of primary tumor and metastatic lymph nodes improves prediction of loco-regional control in head and neck cancer. *Scientific reports*, 9(1):1–7, 2019.

[216] Hecktor 2022 head and neck tumor segmentation and outcome prediction in pet/ct images third edition. Grand challenges [Internet]. `https://hecktor.grand-challenge.org/`. Cited 2022 Dec 20.

[217] Mark Gooding. Oral presentation: Contour similarity metrics and clinical usability. Presented at: *ESTRO 2021*, 2021.

[218] Femke Vaassen, Colien Hazelaar, Ana Vaniqui, Mark Gooding, Brent van der Heyden, Richard Canters, and Wouter van Elmpt. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Physics and Imaging in Radiation Oncology*, 13:1–6, 2020.

[219] Elaine Cha, Sharif Elguindi, Ifeanyirochukwu Onochie, Daniel Gorovets, Joseph O Deasy, Michael Zelefsky, and Erin F Gillespie. Clinical implementation of deep learning contour autosegmentation for prostate radiotherapy. *Radiotherapy and Oncology*, 159:1–7, 2021.

[220] Mark J Gooding, Annamarie J Smith, Maira Tariq, Paul Aljabar, Devis Peressutti, Judith van der Stoep, Bart Reymen, Daisy Emans, Djoya Hattu, Judith van Loon, et al. Comparative evaluation of autocontouring in clinical practice: a practical method using the turing test. *Medical physics*, 45(11):5105–5115, 2018.

[221] Alan M Turing. Computing machinery and intelligence. In *Parsing the turing test*, pages 23–65. Springer, 2009.

[222] Tim Lustberg, Johan van Soest, Mark Gooding, Devis Peressutti, Paul Aljabar, Judith van der Stoep, Wouter van Elmpt, and Andre Dekker. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. *Radiotherapy and Oncology*, 126(2):312–317, 2018.

[223] Davood Karimi and Ali Gholipour. Improving calibration and out-of-distribution detection in medical image segmentation with convolutional neural networks. *arXiv preprint arXiv:2004.06569v3*, 2022.

[224] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54:280–296, 2019.

[225] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, 2020.

[226] The cancer imaging archive. The national cancer institute / The cancer imaging program [Internet]. `https://www.cancerimagingarchive.net/`. Cited 2022 Dec 20.

[227] Cornelis AT van den Berg and Ettore F Meliadò. Uncertainty assessment for deep learning radiotherapy applications. In *Seminars in Radiation Oncology*, volume 32, pages 304–318. Elsevier, 2022.

Appendix A

Paper I

Accepted revised author manuscript
Received: 13 August 2020 | Revised: 22 December 2020 | Accepted for publication: 11 February 2021

Published 4 March 2021
Phys. Med. Biol. 66 (2021) 065012. doi: https://doi.org/10.1088/1361-6560/abe553.

A comparison of methods for fully automatic segmentation of tumors and involved nodes in PET/CT of head and neck cancers

Aurora Rosvoll Groendahl[1], Ingerid Skjei Knudtsen[2], Bao Ngoc Huynh[1], Martine Mulstad[1], Yngve Mardal Moe[1], Franziska Knuth[3], Oliver Tomic[1], Ulf Geir Indahl[1], Turid Torheim[4,5], Einar Dale[6], Eirik Malinen[2,7] and Cecilia Marie Futsaether[1]

[1] Faculty of Science and Technology, Norwegian University of Life Sciences, Ås, Norway
[2] Department of Medical Physics, Oslo University Hospital, Oslo, Norway
[3] Department of Physics, Norwegian University of Science and Technology, Trondheim, Norway
[4] Department of Informatics, University of Oslo, Oslo, Norway
[5] Institute for Cancer Genetics and Informatics, Oslo University Hospital, Oslo, Norway
[6] Department of Oncology, Oslo University Hospital, Oslo, Norway
[7] Department of Physics, University of Oslo, Oslo, Norway

E-mail: cecilia.futsaether@nmbu.no

Abstract

Target volume delineation is a vital but time-consuming and challenging part of radiotherapy, where the goal is to deliver sufficient dose to the target while reducing risks of side effects. For head and neck cancer (HNC) this is complicated by the complex anatomy of the head and neck region and the proximity of target volumes to organs at risk. The purpose of this study was to compare and evaluate conventional PET thresholding methods, six classical machine learning algorithms and a 2D U-Net convolutional neural network (CNN) for automatic gross tumor volume (GTV) segmentation of HNC in PET/CT images. For the latter two approaches the impact of single vs. multimodality input on segmentation quality was also assesed. 197 patients were included in the study. The cohort was split into training and test sets (157 and 40 patients, respectively). Five-fold cross-validation was used on the training set for model comparison and selection. Manual GTV delineations represented the ground truth. Tresholding, classical machine learning and CNN segmentation models were ranked separately according to the cross-validation Sørensen-Dice similarity coefficient ($Dice$). PET thresholding gave a maximum mean $Dice$ of 0.62, whereas classical machine learning resulted in maximum mean $Dice$ scores of 0.24 (CT) and 0.66 (PET; PET/CT). CNN models obtained maximum mean $Dice$ scores of 0.66 (CT), 0.68 (PET) and 0.74 (PET/CT). The difference in cross-validation $Dice$ between multimodality PET/CT and single-modality CNN models was significant ($p \leq 0.0001$). The top-ranked PET/CT-based CNN model outperformed the best-performing thresholding and classical machine

learning models, giving significantly better segmentations in terms of cross-valdiation and test set $Dice$, true positive rate, positive predictive value and surface distance-based metrics ($p \leq 0.0001$). Thus, deep learning based on multimodality PET/CT input resulted in superior target coverage and less inclusion of surrounding normal tissue.

Keywords: Head and neck cancer, PET/CT, gross tumor volume, automatic segmentation, thresholding, machine learning, deep learning, CNN.

1. Introduction

More than 800 000 cases of head and neck cancer (HNC) were diagnosed worldwide in 2018 (Bray *et al* 2018). The majority of HNCs are squamous cell carcinomas (HNSCC) of the oral cavity, oropharynx, hypopharynx and larynx (Argiris *et al* 2008; Haddad and Shin 2008). Most patients are diagnosed with locally advanced, nonmetastatic disease, where standard treatment is concurrent radio-chemotherapy (Halperin *et al* 2013).

For radiotherapy in general, the challenge is to deliver sufficient doses to the target volumes (TVs) whilst keeping the doses to the organs at risk (OARs) at acceptable levels in order to prevent major toxicities. For HNC, this is further complicated by the complex anatomy of the head and neck region, as well as close proximity between TVs and OARs (Grégoire *et al* 2015; O'Sullivan *et al* 2012). Highly conformal dose radiotherapy techniques such as intensity-modulated radiotherapy and volumetric arc therapy reduce radiation-related toxicities (O'Sullivan *et al* 2012) but require precise and accurate volume definitions (Eisbruch and Gregoire 2009; Grégoire *et al* 2015). In clinical practice, the current gold standard for TV delineation is manual contouring in the radiotherapy treatment planning system. However, manual delineation is time-consuming and encumbered with intra- and interobserver variability (Bird *et al* 2015; Gudi *et al* 2017; Kosmin *et al* 2019; Lin *et al* 2019). Automatic segmentation (auto-segmentation) could potentially alleviate intra- and interobserver variations and reduce time spent on

delineations, as shown recently by Lin *et al* (2019), leading to improved TV dose coverage and sparing of OARs.

Radiotherapy planning of HNSCC of the oral cavity, pharynx and larynx is usually performed using CT images of the patient in treatment position. 18F-fluorodeoyxyglucose-PET (FDG-PET) may be used as a supplementary modality as it can provide additional information for TV delineations (Grégoire *et al* 2015). Most previous studies on auto-segmentation of HNC involve segmentation of OARs or nodal/elective TVs in planning CT images (see Kosmin *et al* (2019) for a review) or segmentation of the primary tumor volume in FDG-PET images (Berthon *et al* 2017; Comelli *et al* 2018; Comelli *et al* 2019a; Comelli *et al* 2019b; Stefano *et al* 2017). For the CT-based auto-segmentation studies, methods range from Atlas-based algorithms to deep learning using convolutional neural networks (CNNs) (Kosmin *et al* 2019). The above PET-based studies are based on a limited number of patients (< 30) and apply semi- or fully automatic classical machine learning methods to segment the gross tumor volume (GTV) (Berthon *et al* 2017) or a biologically relevant TV within the GTV (Comelli *et al* 2018; Comelli *et al* 2019a; Comelli *et al* 2019b; Stefano *et al* 2017).

There are, however, a multitude of approaches for PET auto-segmentation available, ranging from fixed thresholding to various machine learning methods (Hatt *et al* 2017). In the most comprehensive comparison of PET auto-segmentation methods to date, including simulated, phantom and clinical FDG-PET image data for patients with HNC, a deep learning approach using CNNs was ranked first, followed by various classical machine learning methods (Hatt *et al* 2018). This comparison study demonstrated that the performance of all methods was dependent on the FDG-PET uptake properties of the tumor and background tissue, leading to substantial variations in patient-wise segmentation quality (Hatt *et al* 2018). As the FDG-PET uptake is not fully cancer

specific, an auto-segmentation method capable of utilizing CT or multimodality PET/CT images is likely to be more robust towards abnormal uptake characteristics, potentially leading to higher quality segmentations for patients with atypical uptake. Both Yu *et al* (2009) and Huang *et al* (2018) have performed auto-segmentation of GTVs in combined PET/CT images but for a limited number of HNC patients (≤ 22). A recent study by Guo *et al* (2019) propose a deep learning framework for GTV segmentation in HNC, obtaining superior auto-segmentations using multimodality PET/CT rather than single modality input for 250 patients. PET/CT also resulted in the highest-quality GTV auto-segmentations when deep learning-based segmentation was applied on images of patients with oropharyngeal cancer ($n = 202$) (Andrearczyk *et al* 2020).

To summarize, most previous studies performing auto-segmentation of GTVs in HNC have been limited by the low number of patients included and/or have not investigated the use of single modality vs. multimodality PET/CT images. Furthermore, most PET-only-based studies have relied on a human operator to locate a region of interest within or around the TV (Comelli *et al* 2018; Comelli *et al* 2019a; Comelli *et al* 2019b; Stefano *et al* 2017), or focus on small pre-selected volumes of interest (VOIs) encompassing the immediate region surrounding each TV (Hatt *et al* 2018). There is, therefore, a need for evaluating different automatic segmentation methods on images with less biased VOIs for larger HNC patient cohorts, comparing the segmentation performance obtained using single and multimodality images. This is the focus of our present work.

The purpose of this study was to evaluate thresholding, classical machine learning and deep learning for automatic GTV segmentation based on PET/CT images of patients with HNSCC. Auto-segmentations were obtained in a cohort of 197 patients, using (i): Conventional PET thresholding methods, (ii): A classical machine learning approach

where different classifiers and advanced feature engineering were explored, and (iii): A deep learning approach using CNNs which inherently derives features automatically. The impact of imaging modality on segmentation quality was evaluated by comparing the results obtained using CT or PET images or the combination of the two modalities as input to the classical machine learning and deep learning approaches.

2. Materials and methods

2.1 Study

2.1.1 Patients and treatment

The present study includes patients with HNSCC of the oral cavity, oropharynx, hypopharynx and larynx, treated with curatively intended radio(chemo)therapy. The patient cohort of 225 patients and the treatment regime have been described previously (Moan *et al* 2019). In the current study, we excluded patients who did not have a contrast enhanced CT along with the PET examination, resulting in 197 patients eligible for analysis. Characteristics of the eligible patients are summarized in table 1. The study was approved by The Regional Ethics Committee (REK) and the Institutional Review Board. Exemption from study-specific informed consent was granted by REK as this is a retrospective study and the patients are de-identified.

2.1.2 Imaging and manual delineations

FDG-PET/CT scans were performed on a Siemens Biograph 16 (Siemens Healthineers GmbH, Erlangen, Germany) with a radiotherapy compatible flat table and radiotherapy fixation mask. The PET/CT protocol consisted of a radiotherapy optimized PET/CT acquisition from the skull base to the mid chest with arms down (5 mins/bed

Table 1. Patient characteristics summarized for all eligible patients and for the patients included in the training and test sets (n = number of patients).

Characteristic[a]	All patients (n = 197)	Training set (n = 157)	Hold-out test set (n = 40)
Age [years]			
Mean (range)	60.3 (39.9–79.1)	60.6 (39.9–79.1)	59.4 (43.0–77.0)
Sex			
Female	49 (24.9%)	38 (24.2 %)	11 (27.5 %)
Male	148 (75.1 %)	119 (75.8 %)	29 (72.5 %)
Tumor stage			
T1/T2	96 (48.7 %)	76 (48.4 %)	20 (50.0 %)
T3/T4	101 (51.3 %)	81 (51.6 %)	20 (50.0 %)
Nodal stage			
N0	47 (23.9 %)	37 (23.6 %)	10 (25.0 %)
N1	23 (11.7 %)	19 (12.1 %)	4 (10.0 %)
N2	120 (60.9 %)	95 (60.5 %)	25 (62.5 %)
N3	7 (3.6 %)	6 (3.8 %)	1 (2.5 %)
Tumor site			
Oral cavity	17 (8.6 %)	14 (8.9 %)	3 (7.5 %)
Oropharynx	143 (72.6 %)	113 (72.0 %)	30 (75.0 %)
Hypopharynx	16 (8.1 %)	15 (9.6 %)	1 (2.5 %)
Larynx	21 (10.7 %)	15 (9.6 %)	6 (15.0 %)
GTV-T[b] [cm^3]			
Mean (range)	25.0 (0.8–285.0)	25.2 (0.8–285.0)	24.3 (1.4–157.6)
GTV-N[c] [cm^3]			
Mean (range)	19.3 (0.5–276.7)	25.6 (0.5–276.7)	19.5 (0.5–76.4)

[a] Percentages may not sum to exactly 100 due to rounding.

[b] Gross primary tumor volume

[c] Involved nodal volume (for patients with nodal stage ≥ N1)

position) followed by a standard whole-body PET/CT acquisition. For the current analysis, only the radiotherapy planning PET and corresponding contrast-enhanced CT images were included. Details on the imaging protocol can be found in appendix A. All TV delineations were done at the time of treatment planning, also described in (Moan *et al* 2019), following the DAHANCA Radiotherapy Guidelines (2013). The gross primary tumor volume (GTV-T) and, if present, the involved nodal volume (GTV-N) were first

contoured manually by an experienced nuclear medicine physician based on FDG-PET findings. The resulting delineations were further refined by one or two oncology residents based on the contrast-enhanced CT and clinical information. For final quality assurance, the delineations were reviewed by a senior oncologist.

2.2 Image pre-processing

The PET/CT image series and DICOM radiotherapy planning structures were exported to an external computer and pre-processed using Interactive Data Language (IDL) v8.5 (Harris Geospatial Solutions, Broomfield, CO, USA). For each patient, the PET, CT and structure series were resampled to an isotropic voxel size of $1 \times 1 \times 1$ mm^3 and registered to a common frame of reference. PET image values were expressed as standardized uptake values (SUV), normalized to body weight. All further pre-processing was performed using MATLAB® 2019a (The Mathworks, Inc. Natick, Massachusetts, USA).

Image regions consisting of high SUV brain tissues were identified and excluded by applying a two-dimensional region-growing function on the maximum intensity projection of the PET images. The images were cropped to a VOI containing the GTV-T and GTV-N. Each VOI was defined by including a 20 mm edge around the manually delineated GTV structures in the axial plane, and one extra slice in the z-direction. Image slices in-between delineated structures were not included in the VOI. The above VOI definition resulted in a reduced data set, where the total number of GTV-T and GTV-N voxels constituted approximately 6 % of all included voxels. This proportion varied moderately between patients, depending on the primary tumor volume and/or the extent of nodal involvement. A typical VOI is shown in figure 1(a), along with the corresponding PET/CT images.

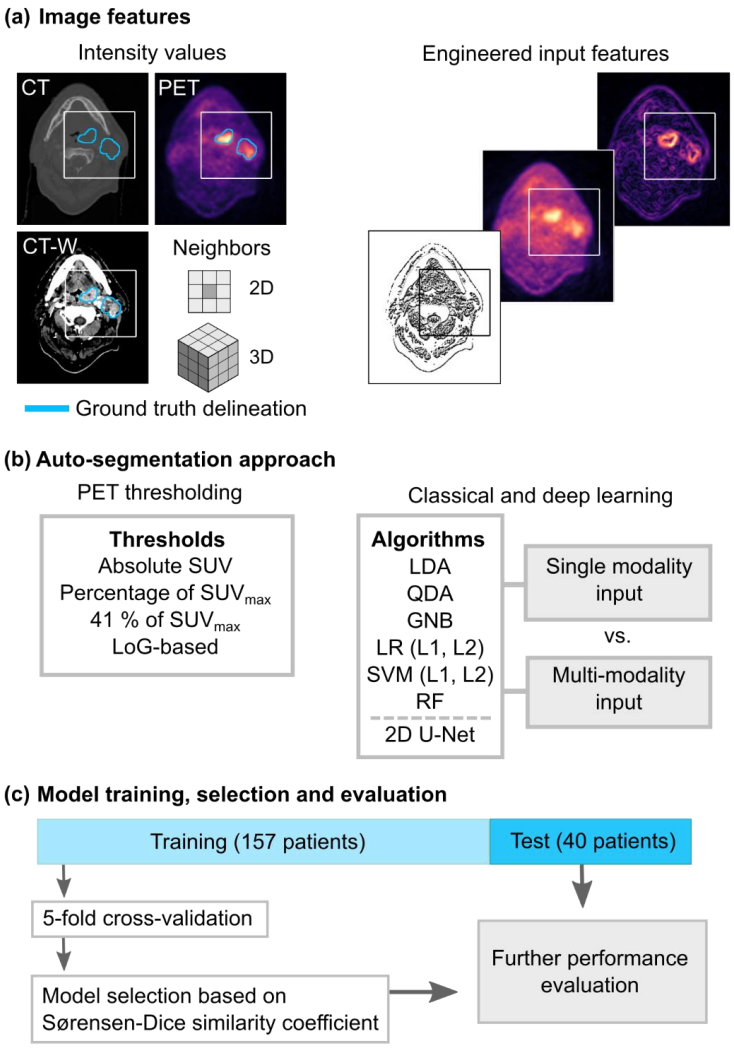


Figure 1. Schematic overview of the experimental set-up. (a) Left: The image data consisted of radiotherapy PET/CT images. CT-W denotes CT with an applied window setting (shown: center 60 HU and width 100 HU). Manual gross tumor volume delineations were used as the ground truth in the experiments. A volume of interest (VOI; white square) was defined prior to auto-segmentation. Each image voxel within the VOI was represented by its intensity value. For the classical machine learning approach, intensities within 2D or 3D voxel neighborhoods were used as features. Right: Various 1D and 2D image transformations were further evaluated as features for the classical learners (examples from left: local binary patterns of CT-W, natural logarithm and gradient magnitude of PET). (b) Auto-segmentations were obtained using PET thresholding, six classical machine learning classifiers (linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), Gaussian naïve Bayes (GNB), logistic regression (LR), support vector machines (SVM) and random forest (RF)) and one CNN architecture (2D U-Net). (c) The cohort of 197 patients was split into training and test sets. Superior models were selected for further evaluation, based on the Sørensen-Dice similarity coefficient between auto-segmentations and the ground truth.

2.3 Model training and validation

Auto-segmentations were obtained using four PET thresholding methods, six classical machine learning classification algorithms and one CNN architecture. The thresholding, classical machine learning and CNN methodologies are described in detail in sections 2.5, 2.6 and 2.7, respectively. An overview of the analysis is shown in figure 1.

The manual GTV-T and GTV-N delineations were used as the ground truth when training and evaluating auto-segmentation models. The auto-segmentation task was considered as a two-class classification problem: tumor and involved lymph node tissues belonging to GTV-T or GTV-N (class 1), or unaffected tissues (class 0).

Patients were divided into a training set (157 patients; 80 %) and an internal hold-out test set (40 patients; 20 %), stratified with respect to tumor stage (cf. table 1). A random sampling five-folded cross-validation procedure, stratified by tumor stage, was used with the training set for hyper-parameter tuning and model comparison. This ensured comparable tumor stage distributions across the training and hold-out test sets, as well as the cross-validation folds. The generalization performance of the superior thresholding, classical machine learning and CNN auto-segmentation models was evaluated on the hold-out test set. Prior to final test set evaluation, the superior thresholding and classical machine learning models were retrained on the full training set.

Thresholding and classical machine learning were performed using MATLAB®. The CNN models were trained using Python and TensorFlow.

2.4 Performance evaluation

The Sørensen-Dice similarity coefficient ($Dice$) (Dice 1945; Sørensen 1948) was used to assess the cross-validation performance of each auto-segmentation model. $Dice$

may be defined using either set notation or the number of true positive (TP), false positive (FP) and false negative (FN) voxels:

$$Dice = \frac{2|P \cap G|}{|P| + |G|} = \frac{2TP}{2TP + FP + FN}. \quad (1)$$

In equation (1) G and P denotes the set of voxels included in the ground truth (G) delineation and the predicted (P) segmentation mask, respectively. Hence, $Dice$ measures the degree of spatial overlap between the ground truth and the predicted segmentation mask given as output by an auto-segmentation model. $Dice$ scores of 0.70 or higher may be considered as good overlap between volumetric segmentations (Zijdenbos *et al* 1994). Based on the $Dice$ performance, we selected the superior thresholding, classical machine learning and CNN models, respectively, for further assessment and comparison, using the methodology described in section 2.8 below.

The superior models were further evaluated in terms of true positive rate (TPR), positive predictive value (PPV), the 95th percentile Hausdorff distance (HD_{95}) (Huttenlocher *et al* 1993) and mean surface distance (MSD) (Taha and Hanbury 2015). These metrics provide complementary information on the quality of the predictions.

TPR, also called sensitivity, is the fraction of the ground truth delineation that overlaps with the predicted segmentation mask, and is defined in terms of TP and FN voxels:

$$TPR = \frac{TP}{TP + FN}. \quad (2)$$

On the other hand, PPV, also called precision, is the fraction of the predicted segmentation mask that overlaps with the ground truth delineation, expressed as:

$$PPV = \frac{TP}{TP + FP}. \quad (3)$$

The Hausdorff distance (HD) is defined as the maximum distance between the surface voxels of the ground truth set G and the predicted mask set P, expressed as (Taha and Hanbury 2015):

$$HD(G,P) = \max(h(G,P), h(P,G)), \tag{4}$$

where $h(G,P)$ is the directed HD defined as:

$$h(G,P) = \max_{g \in G} \min_{p \in P} \| g - p \|. \tag{5}$$

$\| g - p \|$ in equation (5) denotes the Euclidian norm between surface points g and p. As HD is known to be sensitive to outliers, it is not recommended to use this metric directly (Zhang and Lu 2004). Therefore, it was replaced by the 95th percentile HD (HD_{95}) which excludes the most extreme observations. Furthermore, the MSD is defined as the mean distance between the surface voxels of the ground truth set G and the predicted set P, given by:

$$MSD(G,P) = \max(d(G,P), d(P,G)), \tag{6}$$

where $d(G,P)$ is the directed average HD defined as:

$$d(G,P) = \frac{1}{|G|} \sum_{g \in G} \min_{p \in P} \| g - p \|. \tag{7}$$

The MSD metric provides the typical distance between the ground truth delineation and the predicted segmentation, whereas HD_{95} reflects the longest distance and thus the most severe mismatch between the surface voxels of the two sets. The above distance metrics can provide clinically relevant information about the differences in edges between segmentations not captured by $Dice$ due to its intrinsic volume dependency. Furthermore, $Dice$ does not separate between type 1 errors (FP) and type II errors (FN). However,

reporting both TPR and PPV conveys distinct information about both error types (Hatt *et al* 2017). All our performance metrics were calculated on a per patient basis. HD_{95} and MSD were calculated using an in-house-developed Python library available at https://github.com/yngvem/mask_stats.

2.5 Auto-segmentation using PET thresholding

PET thresholding was performed using either an absolute SUV threshold, a percentage of the maximum SUV (SUV_{max}) threshold, or a method based on Laplacian of Gaussian (LoG) filtering, with basis in the procedure outlined in (Gonzalez and Woods 2010), hereinafter referred to as LoG-based thresholding.

LoG-based thresholding consisted of several sequential operations, where the overall objective was to use edges, as indicated by the LoG filter, to improve segmentations obtained by thresholding. Initially, the original images (PET) were transformed with a 3D LoG filter with standard deviations (SDs) of {1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5} mm in the convolution kernel. For each SD the corresponding kernel size was n^3 with $n \geq 6\ SD$. To avoid unwanted edge effects, LoG filtering was performed prior to exclusion of brain tissues and image cropping. The resulting LoG filtered images (fLoG) were converted to absolute values and percentile thresholding was used to create a binary mask. This mask was applied on the product of fLoG and PET to exclude the least relevant background voxels. Otsu's method (Otsu 1979) was then applied on the masked fLoG × PET to segment tumor/lymph nodes from the background.

The above thresholding models were optimized with respect to $Dice$, by maximizing the mean $Dice$ per patient ($mDice$) for each training fold in the cross-validation procedure. The absolute SUV threshold was varied from 0 to 8 with an incremental change of 0.25, while the percentage of SUV_{max} threshold was varied from 0

to 100 % using increments of 1 %. For LoG-based thresholding, the percentile value was varied from 50 to 95 with increments of 5.

For comparison, we also used a fixed percentage threshold equal to 41 % of the SUV_{max}, which has been recommended for PET thresholding (Boellaard *et al* 2015; Davis *et al* 2006).

2.6 Auto-segmentation using classical machine learning

2.6.1 Feature extraction

Segmentation based on classical machine learning was performed using different combinations of image features derived from the original PET and CT images as model input. The effect of imaging modality was assessed using either CT features, PET features or combinations of PET and CT features as input to the classifier. In addition, we evaluated the effect of reducing the dynamic range of the CT images (windowing) by replacing the original CT features with features based solely on windowed CT images (CT-W).

The original voxel intensity values constituted the simplest features, while the original intensities in 2D and 3D neighborhoods surrounding each voxel were used as spatial features. As illustrated in figure 1(a), the 2D neighbors were defined as the 8-neighborhood within a given axial image slice (previously described in (Torheim *et al* 2017)), whereas the 3D neighbors were defined as the 26-neighborhood also including voxels from adjacent image slices.

For CT, CT-W and PET, we also evaluated several 1D and 2D image transformations for inclusion as features by assessing their point-biserial correlation with the ground truth, using training data only. The 1D transformations included the natural logarithm, exponential, square and square root of intensity values as defined in (van

Griethuysen *et al* 2017). Evaluated 2D transformations included the gradient magnitude and direction using the Sobel operator (Gonzalez and Woods 2010), the LoG filter (Gonzalez and Woods 2010), local binary patterns (LBP) (Ojala *et al* 2002) and the 1st level Haar (Gonzalez and Woods 2010) and Coiflet (Daubechies 1993) discrete stationary wavelets. LoG filtering was performed with SDs of {1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5} mm, whereas LBP used sample radii of {1, 2, 4, 6} mm and a constant sample size of 8 voxels. As for LoG-based thresholding, the spatially dependent (2D) transformations were performed prior to exclusion of high SUV brain voxels and VOI definition.

Additionally, we evaluated 10 different CT window settings with window centers of {60, 70} HU, corresponding to typical intra-tumor intensity values, and window widths of {100, 200, 350, 500, 1000} HU. The various window settings were considered both for inclusion as features in the CT-based classification and for determining the settings for CT-W.

All transformations with an absolute point-biserial correlation with the ground truth equal to or larger than that of the corresponding non-transformed images were included as features. Similarly, the window setting having the highest point-biserial correlation with the ground truth was selected for creating CT-W, namely (centre: 60 HU; width: 100 HU).

2.6.2 Standardization, image unfolding and class-imbalance

For each patient, the original and transformed 3D image stacks were standardized separately to zero mean and a standard deviation of one. Prior to performing voxel-wise classification, the standardized 3D image stacks were unfolded into 2D data matrices X where each row consisted of the given input feature(s) for one unique voxel, as described in detail in (Torheim *et al* 2017). The delineated structures were unfolded to a

corresponding response vector Y, containing the class membership of each voxel (class 0 or 1 according to the ground truth). In total, 31 different X matrices (see figure 4) were used as input to the classification algorithms. For the X matrices consisting of single modality 2D or 3D neighborhoods, we also assessed the effect of sorting each voxel and its neighborhood in descending order according to intensity value. These sorted X matrices gave an overall representation of changes in neighborhood voxel intensities, rather than focusing on the intensities with respect to voxel location.

To alleviate the severe class-imbalance between affected (class 1) and unaffected tissues (class 0), the majority class (i.e. the unaffected tissues class) was randomly under-sampled to obtain 50-50 class-balance per patient for each training fold in the cross-validation scheme. Random under-sampling is a naïve approach for handling imbalanced data sets but has been shown to improve classification accuracy for the minority class (Batuwita and Palade 2010; Chawla *et al* 2002; Zhang and Mani 2003).

2.6.3 Classification algorithms

Machine learning-based auto-segmentation was performed using six classical machine learning algorithms, namely linear discriminant analysis (LDA) (Fisher 1936), quadratic discriminant analysis (QDA) (Hastie 2001), Gaussian naïve Bayes (GNB) (Hastie 2001), logistic regression (LR) (Hastie 2001), linear support vector machines (SVM) (Cortes and Vapnik 1995) and random forest (RF) (Breiman 2001). Both LR and SVM were trained for a range of logarithmically spaced regularization parameter values λ, using either LASSO-type (least absolute shrinkage and selection operator) (Tibshirani 1996) or Ridge-type (Hastie 2001) regularization, also referred to as L1 and L2 regularization, respectively. The λ value was varied until a peak in $mDice$ was observed for the cross-validation procedure (LR: $\lambda \in [10^{-6}, 10^{4}]$; SVM: $\lambda \in [10^{-6}, 10^{2}]$). RF

was trained with fixed parameters (number of predictors (P) to select at random for each split: $\sqrt{P}$; minimum number of observations per tree leaf: 1; bootstrap sample size equal to the number of training set observations), apart from the number of trees which was varied from 2 and up to 128 until convergence of the cross-validation $mDice$.

2.7 Auto-segmentation using CNNs

A 2D U-Net CNN architecture (Ronneberger *et al* 2015) with the Dice loss function (Milletari *et al* 2016) was trained to perform auto-segmentation in axial image slices, based on single- (PET, CT, CT-W) or multimodality (PET/CT, PET/CT-W) image input (without standardization of the image stacks). The settings used for CT-W was the same as for classical learning (centre: 60 HU; width: 100 HU). Due to varying VOI size between patients, the image slices were padded with zeros to obtain a common matrix dimension of 176 × 176 mm^2. CNN models were trained using the Adam optimizer (Kingma and Ba 2014) with a learning rate of 10^{-5}. Details on the CNN architecture are given in appendix B.

For CNN model selection, the superior input modality was determined first based on the $mDice$ from five-fold cross-validation (cf. section 2.8 below). Next, the one out of the five cross-validation models with the highest $mDice$ in its associated validation fold was selected for final test set evaluation.

2.8 Statistical analysis

Thresholding, classical machine learning and CNNs were first evaluated separately based on the $Dice$ scores from five-fold cross validation, using a non-parametric Friedman test (Friedman 1937) for repeated measures one-way analysis of variance on ranks. If the Friedman test detected significant differences, Nemenyi's many-

to-one test (Hollander *et al* 2014; Pohlert 2020) was used to compare the treatment effect of the considered models or algorithms on the $Dice$ or $mDice$, respectively. The model or algorithm with the highest Friedman rank sum was used as the control, testing the null hypothesis stating no difference in rank sums, against the one-sided alternative hypothesis stating that the control had a significantly higher rank sum than the model or algorithm it was compared against. Thus, a rejection of the null hypothesis indicated that the control model or algorithm obtained superior segmentation quality in terms of $Dice$ or $mDice$.

Friedman test with Nemenyi pairwise comparisons (Hollander *et al* 2014; Pohlert 2020) was further used to compare the per patient segmentation performance of the selected thresholding, classical machine learning and CNN models.

The statistical analysis was conducted in R (R Development Core Team 2019), using the PMCMRplus package (Pohlert 2020). All tests were conducted with a significance level of 0.05.

The patient-wise performance metrics and associated summary statistics are shown in combined box- and violin plots, where the violin part visualizes the data distribution using kernel density estimation to obtain the probability density function. Box plots include median value and interquartile range (white box), mean value (red dot), whiskers for the 5^{th}–95^{th} percentile and outliers (black dots). Violin plots were created using a Gaussian smoothing kernel and distribution tails were trimmed to only include the observed data range.

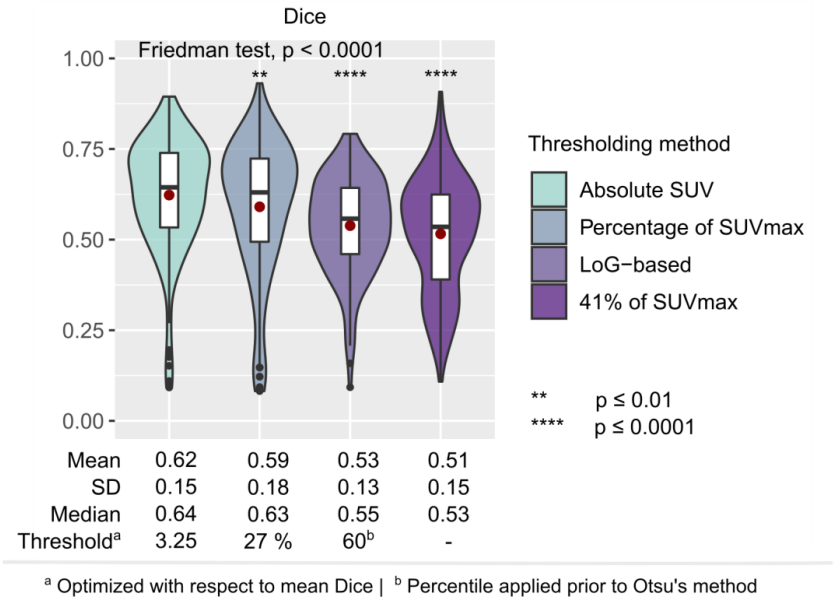


Figure 2. Combined box- and violin plots of the per patient Sørensen-Dice similarity coefficient ($Dice$) between manual delineations and auto-segmentations obtained using PET thresholding on the training set ($n = 157$ patients). Threshold optimization and subsequent auto-segmentation was performed using five-fold cross-validation. Reported thresholds are averaged over the five cross-validation training folds. Results of Friedman test (evaluating the difference in per patient $Dice$ between models) and subsequent many-to-one comparisons with absolute SUV as control model, are indicated in the figure (significance level $\alpha = 0.05$, one-sided Nemenyi many-to-one test).

3. Results

3.1 Auto-segmentations obtained by PET thresholding

The per patient $Dice$ of the threshold-based segmentation models are shown in figure 2, along with corresponding summary statistics and optimal thresholding parameters. Absolute SUV thresholding (T_{abs}) and the optimized percentage of SUV_{max} threshold resulted in $mDice$ scores of 0.62 and 0.59, respectively. In comparison, the reference 41 % of SUV_{max} threshold and LoG-based thresholding obtained $mDice$ scores of 0.51 and 0.53, respectively.

The absolute SUV thresholding model (T_{abs}) ranked highest and displayed significantly higher per patient $Dice$ than the remaining models (figure 2). Thus, T_{abs} was selected for further evaluation.

3.2 Auto-segmentations obtained by classical machine learning

3.2.1 Comparisons across imaging modalities

An overview of the cross-validation $mDice$ for the various classification models, based on data from either CT, CT-W, PET, PET/CT or PET/CT-W is shown in figure 3, and the exact numeric $mDice$ values for each algorithm and input combination are given in figure 4.

Auto-segmentation models based solely on PET resulted in $mDice$ in the same range as PET/CT and PET/CT-W-based models ($mDice$: 0.39–0.66; figure 4). Thus, there was no added gain in $mDice$ segmentation performance when combining PET with the CT or CT-W data (figure 3). Auto-segmentation based solely on CT or CT-W performed poorly with $mDice$ ranging from 0.12 to 0.24 (figure 4), indicating inferior agreement with the ground truth delineations. For CT and CT-W, there was only minor variation in $mDice$ across classification algorithms, but RF resulted in the highest $mDice$ for most inputs (figure 4).

Replacing the original CT data with CT-W generally led to improvements in $mDice$ (figure 4). However, the highest ranked CT and CT-W models had an identical $mDice$ of 0.24. Both models resulted in patchy and imprecise segmentations.

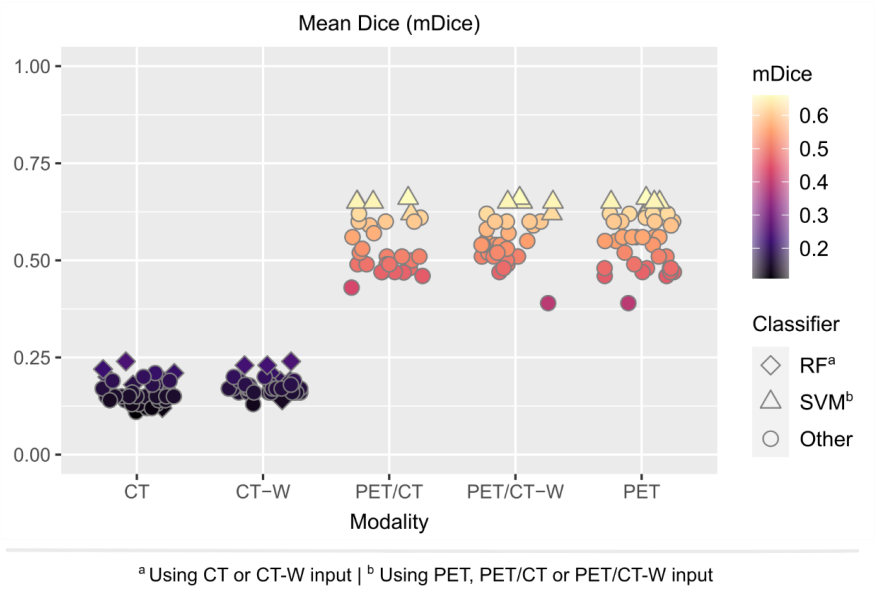


Figure 3. Overview of the mean per patient Sørensen-Dice similarity coefficient ($mDice$) between manual delineations and auto-segmentations obtained by six machine learning algorithms, using image input based on either CT, CT-W (CT with windowing), PET, PET/CT or PET/CT-W. Each point corresponds to a unique combination of image input and classifier. Results were obtained from five-fold cross-validation on the training set (n = 157 patients). Results for random forest (RF) using CT or CT-W input and support vector machines (SVM) using PET, PET/CT or PET/CT-W are highlighted.

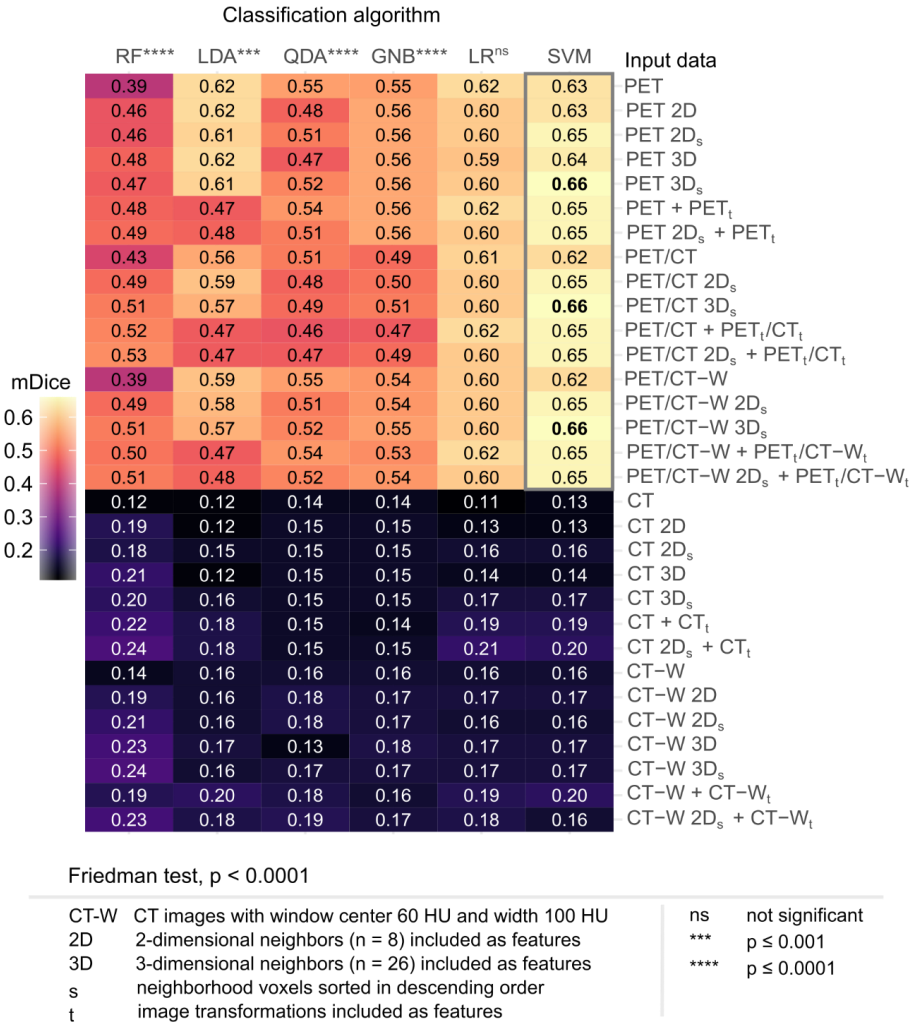


Figure 4. Heat map of the mean per patient Sørensen-Dice similarity coefficient ($mDice$) between manual delineations and auto-segmentations obtained using the machine learning classifiers random forest (RF), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), Gaussian naïve Bayes (GNB), logistic regression (LR) and support vector machines (SVM), with different image input data based on CT, CT-W, PET or combinations of these. The models achieving highest $mDice$ are shown in bold. Results were obtained from five-fold cross-validation on the training set (n = 157 patients). Friedman test results (evaluating the difference in $mDice$ between algorithms for PET, PET/CT and PET/CT-W-based input) and subsequent many-to-one comparisons with SVM as control algorithm, are indicated in the figure (significance level $\alpha = 0.05$, one-sided Nemenyi many-to-one test).

3.2.1 Algorithm and model selection

Due to the poor overlap with ground truth delineations, models based solely on CT or CT-W were not included in further model selection. For the remaining machine learning models, the Friedman test indicated a significant difference in $mDice$ depending on which classification algorithm was used (figure 4). The highest ranked algorithm was the SVM classifier, obtaining consistently higher $mDice$ than the other algorithms. For the SVM models based on PET, PET/CT or PET/CT-W, $mDice$ ranged from 0.63 to 0.66.

The Friedman test also indicated a significant difference in per patient $Dice$ segmentation performance across the included SVM models ($p < 0.0001$). The three SVM models with the highest $mDice$ (figure 4) were also ranked highest based on the per patient $Dice$, having identical Friedman rank sums. These models used L1 regularization and PET, PET/CT or PET/CT-W intensity values with 3D neighbors sorted in descending order as input features, respectively. Due to the feature selection property of the L1 regularization, the highest ranked PET/CT and PET/CT-W models were essentially the same, giving approximately identical per patient $Dice$ cross-validation results.

Each of the three top-ranked SVM models achieved significantly higher per patient *Dice* segmentation performance than the lower ranked SVM models (Nemenyi many-to-one test, all $p \leq 0.05$). Although the top-ranked models were inseparable based on rank sums, the SVM model that only considered PET intensity values (PET $3D_s$ in figure 4, henceforth referred to as SVM_{PET}) was singled out for further evaluation, based on its superior SD and median $Dice$ performance (SD: 0.13 vs. 0.17; median: 0.68 vs. 0.65).

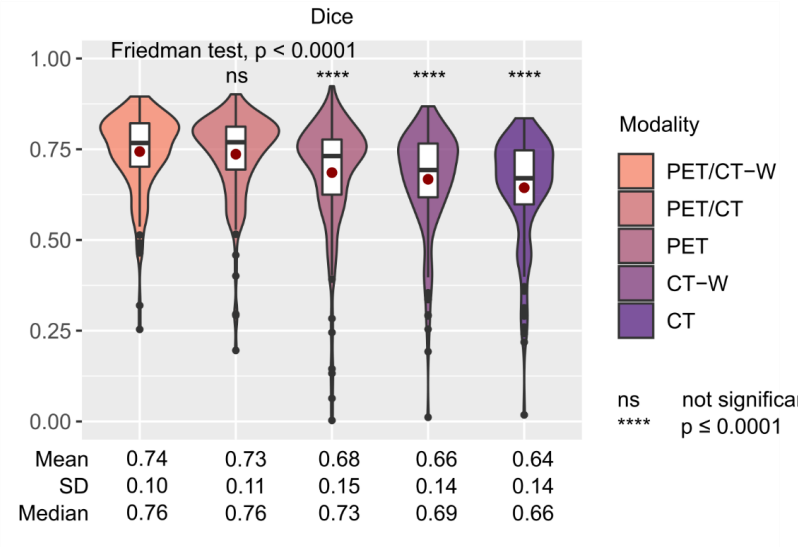


Figure 5. Combined box- and violin plots of the per patient Sørensen-Dice similarity coefficient (*Dice*) between manual delineations and auto-segmentations obtained using a 2D U-Net CNN architecture with the following image input (from left): PET/CT-W (CT-W: CT with windowing), PET/CT, PET, CT-W or CT. Results were obtained from five-fold cross-validation on the training set (n = 157 patients). Results of Friedman test (evaluating the difference in per patient *Dice* between CNN models) and subsequent many-to-one comparisons with the PET/CT-W-based model as control, are indicated in the figure (significance level $\alpha = 0.05$, one-sided Nemenyi many-to-one test).

3.3 Auto-segmentations obtained by CNNs

Dice scores obtained using the 2D CNN models with different input modalities are shown in figure 5. The CNN approach resulted in adequate to high overall *Dice* performance for all input modalities, including CT and CT-W, and there was a substantial increase in performance for multimodality vs. single modality input (*mDice*: 0.74 (PET/CT-W); 0.73 (PET/CT): 0.68 (PET); 0.66 (CT-W); 0.64 (CT)).

The CNN model based on PET/CT-W images (henceforth referred to as U-NET$_{\text{PET/CT-W}}$) was ranked highest according to the Friedman rank sum and obtained significantly higher per patient *Dice* than all single-modality models (figure 5). Based on its superior ranking, U-NET$_{\text{PET/CT-W}}$ was selected for hold-out test set evaluation.

3.4 Performance of the superior models

The combined cross-validation and hold-out test set segmentation performance of the superior thresholding (T_{abs}), classical machine learning (SVM_{PET}) and CNN (U-$NET_{PET/CT\text{-}W}$) models are shown in figure 6. Separate summary statistics for the cross-validation and hold-out test sets are given in table 2. For all three models, the segmentation performance on the hold-out test set was comparable to its cross-validation performance (table 2).

As shown in figure 6, U-$NET_{PET/CT\text{-}W}$ obtained significantly higher quality segmentations than the thresholding and classical machine learning models for all metrics ($p \leq 0.0001$; figure 6 (a-e)). SVM_{PET} further achieved significantly better $Dice$ than T_{abs} ($p \leq 0.0001$; Figure 6 (a)), and somewhat higher mean and median TPR and PPV. SVM_{PET} and T_{abs} obtained comparable mean, SD and median MSD, whereas T_{abs} thresholding resulted in the lowest mean, SD and median HD_{95} of the two.

Representative auto-segmentations obtained on the hold-out test set are shown in figure 7, where the segmentation contours predicted by the superior models are compared to the ground truth. All models resulted in relatively high-quality segmentations for low background FDG-PET signal, combined with high and homogeneous tracer uptake within the GTV-T and/or GTV-N. Neither of the PET-only models (SVM_{PET} and T_{abs}) were capable of segmenting low FDG-PET uptake regions within ground truth delineations (figure 7 (a)). These two models were also prone to include FP voxels with moderate to high SUV. This tendency was, however, more pronounced for thresholding than for classical machine learning (figure 7 (b) and (c)). Superior HD_{95} of T_{abs} relative to SVM_{PET} (cf. figure 6; table 2) was in many cases related to the thresholding model's inclusion of more voxels in the segmentation mask, thereby coinciding with or being in the proximity of ground truth edges, and in particular the GTV-N boundaries (figure 7 (b)). Auto-

segmentations obtained by SVM_{PET} were generally more refined than thresholding (figure 7 (b) and (c)). The multimodality $U\text{-}NET_{PET/CT\text{-}W}$ model was as expected more robust towards atypical FDG-PET uptake characteristics than the PET-only models and was, therefore, to a greater extent capable of segmenting low-uptake regions (figure 7 (a)) and GTV-N edges (figure 7 (b)), as well as avoiding inclusion of FP voxels.

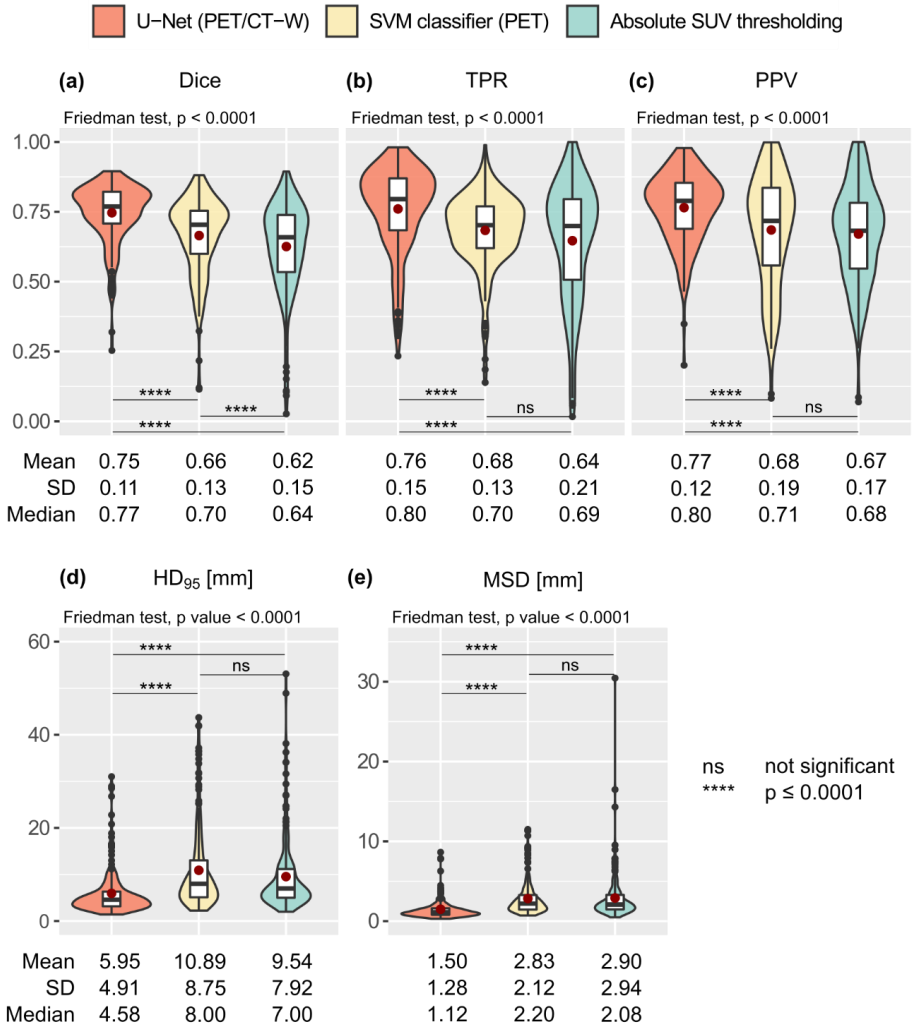


Figure 6. Combined box- and violin plots of the per patient segmentation performance of the superior CNN (U-Net with PET/CT-W input), classical machine learning (SVM classifier using 3D neighborhood PET information) and PET thresholding (absolute SUV) models. The data shown are combined results from cross-validation on the training set (n= 157 patients) and evaluation on the hold-out test set (n = 40 patients). (a) Sørensen-Dice similarity coefficient ($Dice$), (b) true positive rate (TPR), (c) positive predictive value (PPV), (d) 95th percentile Hausdorff distance (HD_{95}), (e) mean surface distance (MSD). For each performance metric (a-e), the results of Friedman test (evaluating the difference in per patient performance between models) and subsequent pairwise comparisons are indicated (significance level $\alpha = 0.05$, two-sided Nemenyi pairwise comparisons).

Table 2. Median, mean and standard deviation (SD) of per patient segmentation performance metrics ($Dice$, TPR, PPV, HD_{95}, MSD) for the superior PET thresholding (T_{abs}), classical machine learning (SVM_{PET}) and CNN (U-$NET_{PET/CT\text{-}W}$) models. Results were obtained using five-fold cross-validation (CV) on the training set (n = 157 patients) and evaluation on the hold-out test set (n = 40 patients).

		T_{abs} (thresholding)		**SVM_{PET}** (classical learning)		**U-$NET_{PET/CT\text{-}W}$** (deep learning)	
Metric		CV	Test set	CV	Test set	CV	Test set
Dice	*Mean*	0.62	0.63	0.66	0.68	0.74	0.75
	SD	0.15	0.16	0.13	0.14	0.10	0.09
	Median	0.64	0.69	0.68	0.72	0.76	0.78
TPR	*Mean*	0.64	0.66	0.68	0.68	0.75	0.76
	SD	0.21	0.20	0.13	0.13	0.15	0.15
	Median	0.69	0.72	0.70	0.68	0.79	0.81
PPV	*Mean*	0.67	0.66	0.68	0.69	0.76	0.78
	SD	0.17	0.19	0.20	0.18	0.12	0.10
	Median	0.68	0.68	0.71	0.74	0.78	0.79
HD_{95} [mm]	*Mean*	9.45	9.88	10.96	10.62	5.98	5.79
	SD	7.99	7.53	9.02	7.48	4.96	4.60
	Median	6.92	7.31	7.81	8.80	4.47	4.74
MSD [mm]	*Mean*	2.94	2.73	2.88	2.64	1.53	1.36
	SD	3.16	1.74	2.24	1.50	1.37	0.79
	Median	2.08	2.17	2.17	2.43	1.11	1.15

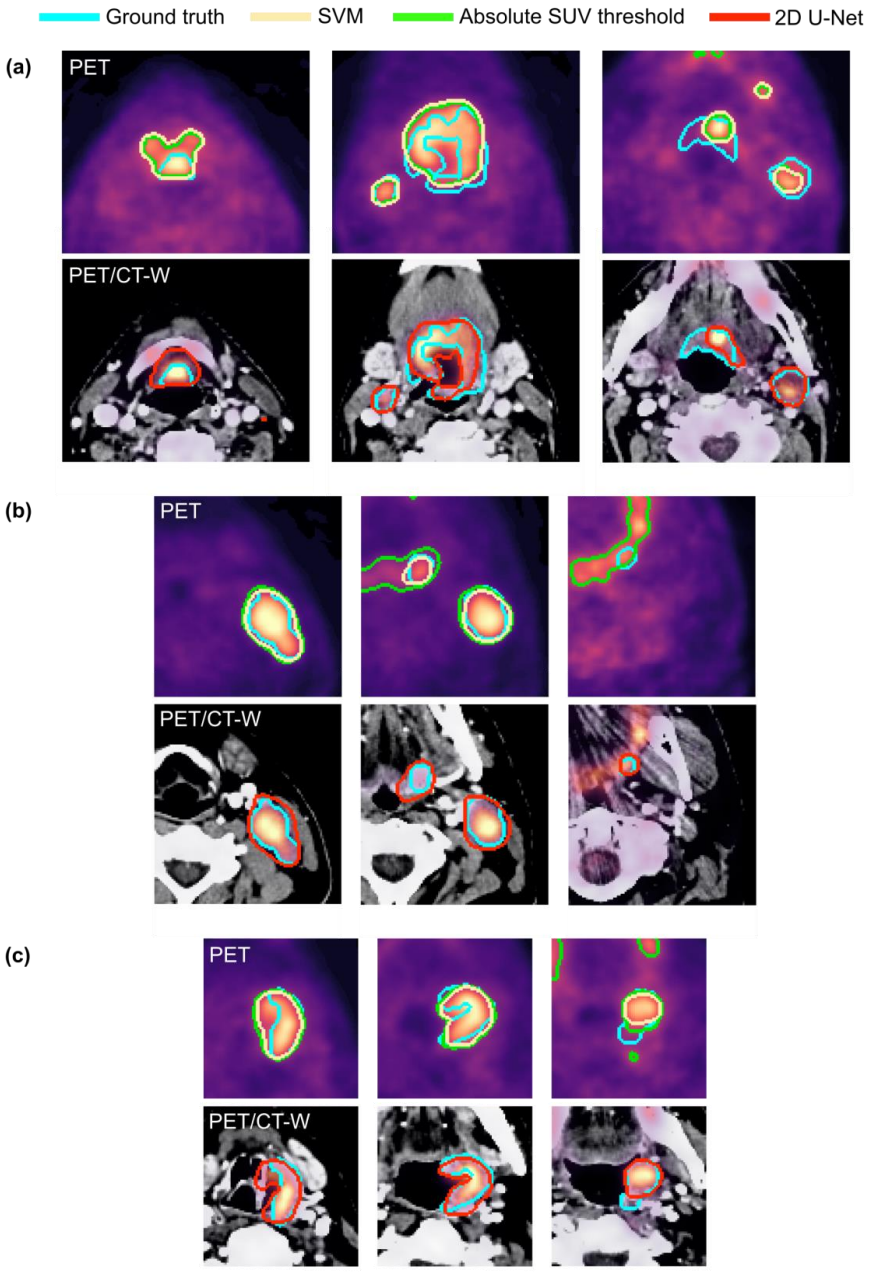


Figure 7. Example auto-segmentation contours in superior, central and inferior image slices (columns) from three different patients (a-c). The shown auto-segmentations were obtained by absolute SUV PET thresholding (green), classical machine learning using the support vector machines (SVM) classifier with 3D neighborhood PET information (yellow) and a 2D U-Net CNN based on combined PET/CT-W image input (red). The per patient Sørensen-Dice similarity coefficient for thresholding, SVM and U-Net models, respectively, were: (a) 0.68, 0.70, 0.74; (b) 0.62, 0.79, 0.76; (c) 0.71, 0.73, 0.79.

4. Discussion

In this study, we evaluated PET thresholding methods, classical machine learning classifiers and a U-Net CNN based on either CT, CT-W, PET, PET/CT or PET/CT-W input for fully automatic segmentation of the primary tumor and involved nodes in a cohort of 197 patients with HNSCC. The CNN approach outperformed both thresholding and classical machine learning, providing the highest overlap with the ground truth for all input modalities. The top-ranked CNN model used combined PET/CT-W information ($\text{U-NET}_{\text{PET/CT-W}}$) and resulted in significantly better segmentations in terms of $Dice$, TPR, PPV, HD_{95} and MSD, compared to the best-performing classical learning (SVM_{PET}) and thresholding (T_{abs}) models. SVM_{PET} further achieved significantly better $Dice$ performance than the top-ranked thresholding model. The higher TPR and PPV of the CNN model indicates better TV coverage as well as less inclusion of normal tissue, respectively, than the other two model approaches. The significantly smaller distance metrics of the selected CNN model further indicate more accurate and precise delineations, which could translate into less need for manual revision of the segmentations in a clinical setting.

Fixed PET thresholding has limitations but is still in widespread use due to its simplicity. The observed difference in segmentation performance between optimized and non-optimized percentage thresholds in our study emphasizes the advantages of threshold optimization, both with respect to application and patient cohort. As an example, our optimized percentage threshold led to an increase in $mDice$ of 0.08 (16 %), relative to the fixed 41 % threshold. The optimization also resulted in a considerably lower thresholding percentage (27 %) than the established 41 % of SUV_{max} (Boellaard *et al* 2015; Davis *et al* 2006). This could in part be attributed to the fact that the ground truth

delineations in our study were based on CT, in addition to PET. Therefore, the delineations typically encompassed larger volumes than the hypermetabolic regions.

Of the evaluated classical machine learning algorithms, SVM provided the highest-quality auto-segmentations in terms of $mDice$, whereas the LR algorithm ranked second. Contrary to most of the other evaluated algorithms, linear SVM and LR do not require the assumption of a specific probability distribution, only that the classes are linearly separable (Hastie *et al* 2001). The two classifiers will generally give similar results, but smaller variations may occur due to their different construction. LR is developed from a probabilistic perspective, assigning optimal probabilities for each voxel belonging to the tumor class. In contrast, linear SVM is developed from a geometric perspective, finding the linear subspace that maximises class separation of the voxels. The superior SVM model, SVM_{PET}, used original PET image intensities with 3D neighbors sorted in descending order according to intensity. Thus, classification was based on the extent of high SUV in the voxel neighborhood, rather than the exact position of such voxels.

For PET-only-based segmentation, the superior thresholding and classical machine learning models performed comparably to the CNN model, achieving a cross-validation $mDice$ of about 0.65, indicating that information required for PET segmentation lay in the voxel intensities rather than subtle and complex spatial patterns. A simple thresholding approach may thus be sufficient for preparatory PET segmentation, as an assistance to the human expert.

Previous studies on PET-based auto-segmentation have achieved $mDice$ in the range 0.77–0.87 for classical learners (Berthon *et al* 2017; Comelli *et al* 2018; Comelli *et al* 2019a; Comelli *et al* 2019b; Hatt *et al* 2018; Stefano *et al* 2017). In the comparison study by Hatt *et al.* (2018), fixed thresholding based on 40 % of the SUV_{max} resulted in

an $mDice$ of 0.70, whereas an $mDice$ of 0.80 was reported for the CNN model. However, differences in image data (Hatt *et al* 2018), the basis of ground truth delineations (Berthon *et al* 2017), TV definition (Comelli *et al* 2018; Comelli *et al* 2019a; Comelli *et al* 2019b; Hatt *et al* 2018; Stefano *et al* 2017), prior VOI definition (Hatt *et al* 2018), and/or the level of automation (Comelli *et al* 2018; Comelli *et al* 2019a; Comelli *et al* 2019b; Stefano *et al* 2017) make direct comparisons between the above studies and our present work challenging. Methods in (Comelli *et al* 2018; Comelli *et al* 2019a; Comelli *et al* 2019b; Stefano *et al* 2017) depended on the user manually drawing a line within or a contour around the cancer-region excluding healthy tissues with high tracer uptake and thus limiting the number of FP voxels. Moreover, involved nodes were not included in any of these studies. In (Hatt *et al* 2018) the methods were applied on a combination of simulated, phantom and clinical data, where the number of clinical images was limited, and the pre-defined VOI was restricted to encompass only the immediate background of each primary tumor. Therefore, the auto-segmentation task of these studies may be considered less challenging than performing fully automatic segmentation of both the GTV-T and GTV-N within larger VOIs, as in our present work.

The most recent studies performing PET-based auto-segmentation of HNSCC (Guo *et al* 2019; Andrearczyk *et al* 2020) evaluate deep learning for single- and multimodality PET/CT input, including both the nodal and primary tumor GTV, in addition to larger image VOIs. Both these studies were based on multi-center patient cohorts, which may be more challenging than our present single-center task. When basing the segmentation solely on PET images, the DenseNet (Guo *et al* 2019) and 3D V-Net (Andrearczyk *et al* 2020) obtained $mDice$ of 0.64 and 0.58, respectively. The former is comparable to out SVM_{PET} and PET-based CNN models, which obtained cross-validation $mDice$ of 0.66 and 0.68.

Regardless of segmentation approach, our PET-only-based models performed relatively poorly on patients with considerable false positive and/or false negative FDG-PET uptake regions, indicating the disadvantages of performing auto-segmentation based solely on molecular imaging. None of the classical machine learning algorithms considered in this study provided satisfactory auto-segmentations in CT or CT-W images, and there was no gain in segmentation performance when PET images were combined with the anatomical CT or CT-W information. In general, segmentation performance of the classical CT or CT-W models benefitted from the inclusion of image transformations, suggesting a non-linear relationship between the CT signal and the ground truth. The success of CNNs, which automatically find complex as well as subtle patterns within images, for CT-based segmentation of a range of diagnoses and applications (see Cardenas *et al* 2019), as well as the fair performance of our CT-based CNN models, support this hypothesis. Guo *et al* (2019) and Andrearczyk *et al* (2020) achieved $mDice$ scores of 0.31 and 0.49 for solely CT-based GTV segmentation using CNNs. This is considerably lower than our CT-based CNN models ($mDice$: 0.64–0.66) but still constitutes a substantial increase in $mDice$ compared to our classical machine learning approach ($mDice$: 0.24). Thus, information required for segmentation was present in CT images, but manually engineering the relevant features with a classical machine learning approach is difficult. Using CNNs, bypassing the feature engineering step, should therefore be the preferred approach for CT-based segmentation.

In contrast to the classical machine learning algorithms, our 2D U-Net was further able to improve segmentation performance significantly given multimodality PET/CT input, resulting in $mDice$ scores of 0.73–0.75. This is in line with both Guo *et al* (2019) and Andrearczyk *et al* (2020), where PET/CT also gave superior results with $mDice$

scores of 0.71 and 0.60, respectively. Thus, it appears that simpler classification algorithms such as our proposed SVM model are comparable to CNNs when segmentation is based solely on PET images, but that CNNs can take advantage of the complementary information contained within PET/CT images to improve segmentation.

Several previous studies have evaluated inter-observer variability in manual GTV delineations for HNSCC (Bird *et al* 2015; Gudi *et al* 2017; Kajitani *et al* 2013; Murakami *et al* 2008; Riegel *et al* 2006). However, only Gudi *et al.* (2017) report the inter-observer agreement for PET/CT using $Dice$, allowing for direct comparison to our present work. Gudi *et al.* (2017) investigated variations in GTV-T and OAR delineations between three experienced radiation oncologists, each with more than 10 years' experience in contemporary HNC radiotherapy. Manual delineations were made for 10 different HNSCC cases. The agreement between observers using either (contrast-enhanced) CT or FDG-PET/CT to perform GTV-T delineations corresponded to an overall $Dice$ performance of ($mDice \pm SD$) 0.57 ± 0.12 and 0.69 ± 0.08, respectively. Thus, the reported inter-observer agreements were comparable to $Dice$ performances of all our CNN models ($mDice$: 0.64–0.75; SDs: 0.10–0.15) and our classical SVM_{PET} model (0.66 ± 0.13).

5. Conclusions

In this study, we conducted an extensive evaluation of the applicability of several PET thresholding methods, classical machine learning classifiers and a 2D U-Net CNN architecture using single or multimodality PET/CT input, for fully automatic segmentation of the primary and nodal GTV in 197 patients with HNSCC. Such auto-segmentation methods have not previously been evaluated and compared in a large HNSCC patient cohort. All models using only PET-based input resulted in fair overall *Dice* segmentation performance. Classical machine learning classifiers were unable to provide satisfactory auto-segmentations based solely on input derived from CT images, nor could they utilize the combined anatomical and molecular information in multimodality PET/CT to improve segmentation quality over PET-only models. This was not the case for the CNN models, which outperformed classical learners for auto-segmentation based on CT-only or combined PET/CT input. The superior model was the U-Net based on PET and windowed CT images, resulting in higher-quality auto-segmentations than the best-performing PET thresholding and classical learning methods.

Acknowledgements

This work is supported by the Norwegian Cancer Society (Grant Number 160907-2014 and 182672-2016). We thank Dr. Jens Petter Wold for valuable suggestions. The authors declare no conflicts of interest.

Ethical Statement

The study was approved by The Regional Ethics Committee (REK) and the Institutional Review Board. Exemption from study-specific informed consent was granted by REK as this is a retrospective study and the patients are de-identified.

Appendix A. Protocol for PET/CT image acquisition

Table A1. Image acquisition and reconstruction parameters for the radiotherapy PET/CT (n = number of patients).

CT	
Scan mode	Helical (rotation time 0.5 s, pitch 0.75)
Peak tube voltage	120 kV
Automatic exposure control	CareDose with quality reference mAs 300
Reconstructed slice thickness	2.00 mm
Reconstruction filter:	B30f/B30s
Matrix size	512 x 512
Pixel size	0.98 × 0.98 mm^2 (n = 161)
	1.37 × 1. 37 mm^2 (n = 30)
	0.89 × 0.89 mm^2 (n = 2)
	0.96 × 0.96 mm^2 (n = 1)
	0.92 × 0.92 mm^2 (n = 1)
	0.88 ×0.88 mm^2 (n = 1)
	0.82 × 0.82 mm^2 (n = 1)
PET	
Reconstruction algorithm	Ordered Subset Expectations maximization (OSEM),
	4 iterations, 8 subsets
Bed position overlap	25 %
Post reconstruction filter	Gaussian, full width at half maximum 3.5 mm
	Gaussian, full width at half maximum 2.0 mm (n = 3)
	Gaussian, full width at half maximum 5.0 mm (n = 1)
Matrix size:	256 × 256
Voxel size ($x - y - z$)	2.66 × 2.66 × 2.00 mm^3 (n = 143)
	2.66 × 2.66 × 5.00 mm^3 (n = 21)
	1.77 × 1.77 × 2.00 mm^3 (n = 20)
	2.66 × 2.66 × 1.00 mm^3 (n = 5)
	1.33 × 1.33 × 2.00 mm^3 (n = 4)
	4.06 × 4.06 × 2.00 mm^3 (n = 2)
	4.06 × 4.06 × 1.00 mm^3 (n = 2)

According to the hospital's procedures, FDG was administered intravenously to the patient after at least six hours of fasting. Between injection and imaging, the patient rested in a quiet, dimly lit room. Median time from injection to imaging in this cohort was 89 mins (SD: 21.5; range: 60-270). Median administered dose was 378 MBq (range: 328–422). The radiotherapy planning PET/CT (contrast enhanced CT) was performed prior to

a standard whole-body PET/CT on a radiotherapy compatible flat table with head support in a radiotherapy fixation mask. The radiotherapy planning CT was optimized for the head and neck region; using the contrast agent Visipaque 320 mg I/mL 100 mL with flow 3.5 mL/s, and CT acquisition performed after a delay of about 30 s. This CT scan was used for attenuation correction and image fusion for image interpretation. Only the radiotherapy PET/CT data were included in our analysis, and the image acquisition and reconstruction parameters for these series are found in table A1.

Appendix B. CNN architecture

Table B1. CNN model architecture.

Name	Inputs	Output shape
Conv1	Input	176 × 176 × 64
Conv2	Conv1	176 × 176 × 64
MaxPool1	Conv2	88 × 88 × 64
Conv3	MaxPool1	88 × 88 × 128
Conv4	Conv3	88 × 88 × 128
MaxPool2	Conv4	44 × 44 × 128
Conv5	MaxPool2	44 × 44 × 256
Conv6	Conv5	44 × 44 × 256
MaxPool1	Conv6	22 × 22 × 256
Conv7	MaxPool3	22 × 22 × 512
Conv8	Conv7	22 × 22 × 512
MaxPool1	Conv8	11 × 11 × 512
Conv9	MaxPool4	11 × 11 × 1024
Conv10	Conv9	11 × 11 × 1024
ConvTranspose1	Conv10	22 × 22 × 512
Conv11	ConvTranspose1 & Conv8	22 × 22 × 512
Conv12	Conv11	22 × 22 × 512
ConvTranspose2	Conv12	44 × 44 × 256
Conv13	ConvTranspose2 & Conv6	44 × 44 × 256
Conv14	Conv11	44 × 44 × 256
ConvTranspose3	Conv10	88 × 88 × 128
Conv15	ConvTranspose3 & Conv4	88 × 88 × 128
Conv16	Conv11	88 × 88 × 128
ConvTranspose4	Conv10	176 × 176 × 64
Conv17	ConvTranspose4 & Conv2	176 × 176 × 64
Conv18	Conv11	176 × 176 × 64
FinalConv	Conv18	176 × 176 × 1

The CNN architecture is given in table B1. The convolutional (Conv) and transposed convolutional (ConvTranspose) layers used a 3 × 3 convolution kernel and were followed by the ReLU activation function. Batch normalization was used after each Conv layer, and a bilinear interpolation was included after each ConvTranspose layer. Relevant code for the experiments is available on https://github.com/huynhngoc/PMB-2020.

ORCID iDs

Aurora Rosvoll Groendahl: https://orcid.org/0000-0003-1327-3844

Ingerid Skjei Knudtsen: https://orcid.org/0000-0001-9313-2878

Bao Ngoc Huynh: https://orcid.org/0000-0001-5210-132X

Yngve Mardal Moe: https://orcid.org/0000-0002-5159-9012

Oliver Tomic: https://orcid.org/0000-0003-1595-9962

Ulf Geir Indahl: https://orcid.org/0000-0002-3236-463X

Turid Torheim: https://orcid.org/0000-0001-6191-2036

Einar Dale: https://orcid.org/0000-0001-9483-2788

Eirik Malinen: https://orcid.org/0000-0002-1308-9871

Cecilia Marie Futsaether: https://orcid.org/0000-0001-7944-0719

References

Andrearczyk V, Oreiller V, Vallières M, Castelli J, Elhalawani, H, Boughdad S, Jreige M, Prior, J O and Depeursinge, A 2020 Automatic Segmentation of Head and Neck Tumors and Nodal Metastases in PET-CT scans *Proceedings of the Third Conf. on Medical Imaging with Deep Learning* **121** 33–43

Argiris A, Karamouzis M V, Raben D and Ferris R L 2008 Head and neck cancer *Lancet* **371**1695–709

Batuwita R and Palade V Efficient resampling methods for training support vector machines with imbalanced datasets 2010 *The 2010 International Joint Conf. on Neural Networks* 1–8

Berthon B, Evans M, Marshall C, Palaniappan N, Cole N, Jayaprakasam V, Rackley T and Spezi E 2017 Head and neck target delineation using a novel PET automatic segmentation algorithm *Radiother. Oncol.* **122** 242–7

Bird D *et al.* 2015 Multimodality imaging with CT, MR and FDG-PET for radiotherapy target volume delineation in oropharyngeal squamous cell carcinoma *BMC Cancer* **15** 844

Boellaard R *et al.* 2015 FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0 *Eur. J. Nucl. Med. Mol. Imaging* **42** 328–54

Bray F, Ferlay J, Soerjomataram I, Siegel R L, Torre L A and Jemal A 2018 Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries *CA: Cancer J. Clin.* **68** 394–424

Breiman L 2001 Random Forests *Mach. Learn.* **45** 5–32

Cardenas C E, Yang J, Anderson B M, Court L E and Brock K B 2019 Advances in Auto-Segmentation *Semin. Radiat. Oncol.* **29** 185–97

Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 SMOTE: synthetic minority over-sampling technique *J. Artif. Intell. Res.* **16** 321–57

Comelli A, Stefano A, Bignardi S, Russo G, Sabini M G, Ippolito M, Barone S and Yezzi A 2019a Active contour algorithm with discriminant analysis for delineating tumors in positron emission tomography *Artif. Intell. Med.* **94** 67–78

Comelli A, Stefano A, Russo G, Bignardi S, Sabini M G, Petrucci G, Ippolito M and Yezzi A 2019b K-nearest neighbor driving active contours to delineate biological tumor volumes *Eng. Appl. Artif. Intell.* **81** 133–44

Comelli A, Stefano A, Russo G, Sabini M G, Ippolito M, Bignardi S, Petrucci G and Yezzi A 2018 A smart and operator independent system to delineate tumours in Positron Emission Tomography scans *Comput. Biol. Med.* **102** 1–15

Cortes C and Vapnik V 1995 Support-vector networks *Mach. Learn.* **20** 273–97

Daubechies I 1993 Orthonormal Bases of Compactly Supported Wavelets II. Variations on a Theme *SIAM J. Math. Anal.* **24** 499–519

Davis J B, Reiner B, Huser M, Burger C, Székely G and Ciernik I F 2006 Assessment of 18F PET signals for automatic target volume definition in radiotherapy treatment planning *Radiother. Oncol.* **80** 43–50

Dice L R 1945 Measures of the Amount of Ecologic Association Between Species *Ecology* **26** 297–302

Eisbruch A and Gregoire V 2009 Balancing risk and reward in target delineation for highly conformal radiotherapy in head and neck *Semin. Radiat. Oncol.* **19** 43–52

Fisher R A 1936 The use of multiple measurements in taxonomic problems *Ann. Eugen.* **7** 179–88

Friedman M 1937 The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance *J. Am. Stat. Assoc.* **32** 675–701

Gonzalez R C and Woods R E 2010 *Digital Image Processing* 3rd edn (Upper Saddle River: Prentice-Hall)

Grégoire V, Langendijk J A and Nuyts S 2015 Advances in Radiotherapy for Head and Neck Cancer *J. Clin. Oncol.* **33** 3277–84

Gudi S, Ghosh-Laskar S, Agarwal J P, Chaudhari S, Rangarajan V, Nojin Paul S, Upreti R, Murthy V, Budrukkar A and Gupta T 2017 Interobserver Variability in the Delineation of Gross Tumour Volume and Specified Organs-at-risk During IMRT for Head and Neck Cancers and the Impact of FDG-PET/CT on Such Variability at the Primary Site *J. Med. Imaging Radiat. Sci.* **48** 184–92

Guo Z, Guo N, Gong K, Zhong S a and Li Q 2019 Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multi-modality network *Phys. Med. Biol.* **64** 205015
Haddad R I and Shin D M 2008 Recent Advances in Head and Neck Cancer *N. Engl. J. Med.* **359** 1143–54
Halperin E C, Brady L W and Perez C A 2013 *Perez & Brady's Principles and Practice of Radiation Oncology* (Philadelphia: Wolters Kluwer Health)
Hastie T J, Tibshirani T and Friedman J 2001 *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (New York: Springer)
Hatt M *et al* 2018 The first MICCAI challenge on PET tumor segmentation *Med. Image Anal.* **44** 177–95
Hatt M *et al* 2017 Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM task group No. 211 *Med. Phys.* **44** e1–e42
Hollander M, Wolfe D A and Chicken E 2014 *Nonparametric Statistical Methods* (Hoboken, New Jersey: John Wiley & Sons, Inc)
Huang B *et al* 2018 Fully Automated Delineation of Gross Tumor Volume for Head and Neck Cancer on PET-CT Using Deep Learning: A Dual-Center Study *Contrast Media Mol. Imaging* **2018** 8923028
Huttenlocher D P, Klanderman G A and Rucklidge W J 1993 Comparing images using the Hausdorff distance *IEEE Trans. Pattern Anal. Mach. Intell.* **15** 850–63
Kajitani C, Asakawa I, Uto F, Katayama E, Inoue K, Tamamoto T, Shirone N, Okamoto H, Kirita T and Hasegawa M 2013 Efficacy of FDG-PET for defining gross tumor volume of head and neck cancer *J. Radiat. Res.* pp 671–8
Kingma D P and Ba J 2014 Adam: A Method for Stochastic Optimization (arXiv: 1412.6980)
Kosmin M *et al* 2019 Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer *Radiother. Oncol.* **135** 130–40
Lin L *et al* 2019 Deep Learning for Automated Contouring of Primary Tumor Volumes by MRI for Nasopharyngeal Carcinoma *Radiology* **291** 677–86
Milletari F, Navab N and Ahmadi S-A 2016 V-net: Fully convolutional neural networks for volumetric medical image segmentation *Fourth International Conf. on 3D Vision* 565–571
Moan J M, Amdal C D, Malinen E, Svestad J G, Bogsrud T V and Dale E 2019 The prognostic role of 18F-fluorodeoxyglucose PET in head and neck cancer depends on HPV status *Radiother. Oncol*. **140** 54–61
Murakami R *et al.* 2008 Impact of FDG-PET/CT fused imaging on tumor volume assessment of head-and-neck squamous cell carcinoma: intermethod and interobserver variations *Acta radiol.* **49** 693–9
O'Sullivan B, Rumble R B and Warde P 2012 Intensity-modulated Radiotherapy in the Treatment of Head and Neck Cancer *J. Clin. Oncol.* **24** 474–87
Ojala T, Pietikainen M and Maenpaa T 2002 Multiresolution gray-scale and rotation invariant texture classification with local binary patterns *IEEE Trans. Pattern Anal. Mach. Intell.* **24** 971–87
Otsu N 1979 A Threshold Selection Method from Gray-Level Histograms *IEEE Trans. Syst., Man, Cybern.* **9** 62–6
Pohlert T 2020 Pairwise Multiple Comparisons of Mean Rank Sums Extended: R package version 1.4.4.
R Development Core Team 2019 R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing)
Riegel A C, Berson A M, Destian S, Ng T, Tena L B, Mitnick R J and Wong P S 2006 Variability of gross tumor volume delineation in head-and-neck cancer using CT and PET/CT fusion *Int. J. Radia. Oncol. Biol. Phys.* **65** 726–32
Ronneberger O, Fischer P and Brox T 2015 U-net: Convolutional networks for biomedical image segmentation *International Conf. Medical Image Computing & Computer Assisted Intervention* 234–241
Stefano A, Vitabile S, Russo G, Ippolito M, Sabini M G, Sardina D, Gambino O, Pirrone R, Ardizzone E and Gilardi M C 2017 An enhanced random walk algorithm for delineation of head and neck cancers in PET studies *Med. Biol. Eng.Comput.* **55** 897–908

Sørensen T 1948 A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons *Kongelige Danske Videnskabernes Selskab* **5** 1–34

Taha A A and Hanbury A 2015 Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool *BMC Medical Imaging* **15** 29

Tibshirani R 1996 Regression Shrinkage and Selection via the Lasso *J. Royal Stat. Soc. Series B (Methodological)* **58** 267–88

Torheim T, Malinen E, Hole K H, Lund K V, Indahl U G, Lyng H, Kvaal K and Futsaether C M 2017 Autodelineation of cervical cancers using multiparametric magnetic resonance imaging and machine learning *Acta Oncol.* **56** 806–12

van Griethuysen J J M, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan R G H, Fillion-Robin J-C, Pieper S and Aerts H J W L 2017 Computational Radiomics System to Decode the Radiographic Phenotype *Cancer Res.* **77** e104–e7

Yu H, Caldwell C, Mah K, Poon I, Balogh J, MacKenzie R, Khaouam N and Tirona R 2009 Automated radiation targeting in head-and-neck cancer using region-based texture analysis of PET and CT images *Int. J. Radiat. Oncol. Biolol. Phys*.**75** 618–25

Zhang D and Lu G 2004 Review of shape representation and description techniques *Pattern Recognit.* **37** 1–19

Zhang J P and Mani I 2003 KNN Approach to Unbalanced Data Distributions; A Case Study Involving Infromation Extraction *Proceeding of International Conf. on Machine Learning*

Zijdenbos A P, Dawant B M, Margolin R A and Palmer A C 1994 Morphometric analysis of white matter lesions in MR images: method and validation *IEEE Trans. Med. Imaging* **13** 716–24

Appendix B

Paper II

European Journal of Nuclear Medicine and Molecular Imaging (2021) 48:2782–2792
https://doi.org/10.1007/s00259-020-05125-x

ORIGINAL ARTICLE

Deep learning-based auto-delineation of gross tumour volumes and involved nodes in PET/CT images of head and neck cancer patients

Yngve Mardal Moe[1] · Aurora Rosvoll Groendahl[1] · Oliver Tomic[1] · Einar Dale[2] · Eirik Malinen[3,4] · Cecilia Marie Futsaether[1]

Received: 10 August 2020 / Accepted: 15 November 2020 / Published online: 9 February 2021

Abstract

Purpose Identification and delineation of the gross tumour and malignant nodal volume (GTV) in medical images are vital in radiotherapy. We assessed the applicability of convolutional neural networks (CNNs) for fully automatic delineation of the GTV from FDG-PET/CT images of patients with head and neck cancer (HNC). CNN models were compared to manual GTV delineations made by experienced specialists. New structure-based performance metrics were introduced to enable in-depth assessment of auto-delineation of multiple malignant structures in individual patients.

Methods U-Net CNN models were trained and evaluated on images and manual GTV delineations from 197 HNC patients. The dataset was split into training, validation and test cohorts ($n = 142$, $n = 15$ and $n = 40$, respectively). The Dice score, surface distance metrics and the new structure-based metrics were used for model evaluation. Additionally, auto-delineations were manually assessed by an oncologist for 15 randomly selected patients in the test cohort.

Results The mean Dice scores of the auto-delineations were 55%, 69% and 71% for the CT-based, PET-based and PET/CT-based CNN models, respectively. The PET signal was essential for delineating all structures. Models based on PET/CT images identified 86% of the true GTV structures, whereas models built solely on CT images identified only 55% of the true structures. The oncologist reported very high-quality auto-delineations for 14 out of the 15 randomly selected patients.

Conclusions CNNs provided high-quality auto-delineations for HNC using multimodality PET/CT. The introduced structure-wise evaluation metrics provided valuable information on CNN model strengths and weaknesses for multi-structure auto-delineation.

Keywords Deep learning · Delineation · Head and neck cancer · Automatic delineation

Introduction

Radiotherapy (RT) with concurrent chemotherapy is the preferred curative treatment option for inoperable head and neck cancer (HNC) [1]. An essential part of RT is tumour delineation, where the tumour and involved lymph nodes are carefully outlined in medical images. This task is vital to ensure that all malignant tissues are included in the RT treatment volume.

Positron emission tomography/X-ray computed tomography (PET/CT) is a highly useful modality for imaging and subsequent delineation of HNC for RT [2]. In most cases, CT is performed with an iodinated contrast agent [3]. Tumours and involved nodes may be detected on PET images, as these regions normally have higher metabolic activity than surrounding healthy tissue. However, PET is limited by low spatial resolution, and combining PET images with high-resolution CT images may improve delineation quality. Several studies have found a significant reduction in interobserver variability for manual gross tumour volume (GTV) delineations in HNC when

This article is part of the Topical Collection on Advanced Image Analyses (Radiomics and Artificial Intelligence)

✉ Cecilia Marie Futsaether
cecilia.futsaether@nmbu.no

1 Faculty of Science and Technology, Norwegian University of Life Sciences, Ås, Norway

2 Department of Oncology, Oslo University Hospital, Oslo, Norway

3 Department of Medical Physics, Oslo University Hospital, Oslo, Norway

4 Department of Physics, University of Oslo, Oslo, Norway

using combined PET/CT instead of CT [4–7]. Despite this, considerable interobserver variations still occur. In a recent HNC interobserver study, the average overlap between PET/CT-based GTV delineations made by expert radiation oncologists was 69% (as measured by the Dice score) [7]. Moreover, the manual delineation process is time consuming and can be a bottleneck in RT planning. Finding methods to improve delineation quality and reduce the workload is therefore highly warranted.

Automatic tumour delineation using deep convolutional neural networks (CNNs) can potentially provide delineation consistency and time-efficiency. Recent studies show a high degree of overlap between expert's tumour delineations and those proposed by CNNs [8–10]. There has, to date, been few studies on CNN auto-delineation of HNC lesions using multimodality images. In [8], Lin et al. used a 3D CNN to successfully auto-delineate the GTV of nasopharyngeal cancers from PET/MRI images. These delineations were evaluated both quantitatively and qualitatively, the latter by expert oncologists. Likewise, Huang et al. [9] used a 2D CNN to delineate the GTV of HNC lesions in PET/CT images. Although very promising, these studies did not consider involved lymph nodes, which, according to current practice, should be prescribed the same RT dose as the GTV. A typical patient with HNC may have multiple involved neck nodes and delineating these is essential for adequate RT [2]. This issue was addressed by Guo et al. [10] who used 3D CNNs to delineate both the GTV and involved nodes in CT, PET and combined PET/CT images. In the latter study, the quality of the network delineations was evaluated quantitatively on a patient-wise basis, regardless of the number of structures delineated.

The scoring of involved lymph nodes does, however, constitute a challenge in evaluating the quality of auto-delineations. This is apparent for occult or small lesions that have been judged as malignant by the expert but not by the auto-delineation method (false negatives). For these situations, the overall performance of the automatic method may be interpreted as poor when assessed using, for example, distance-based metrics, despite a high agreement for the tumour and larger nodes. The same problem also arises for false positive predictions, where the auto-delineation program may incorrectly delineate an hypermetabolic region as part of the GTV. In this case, the distance between this falsely delineated structure and the true GTV may be very large, even though there is high agreement between all other predicted structures and the ground truth. Thus, there is a need for standardised methods to estimate the performance of multi-structure auto-delineation, when the expert delineations include both primary tumour and involved nodes.

The aim of the current study was threefold. First, we evaluated 2D CNN models for fully automatic delineation of both the gross (primary) tumour volume (GTV-T) and the malignant nodal volume (GTV-N) in patients with HNC. Secondly, as all patients underwent a combined PET/CT examination prior to treatment, network performance was assessed using single-modality (CT or PET) as well as multimodality (PET/CT) image input, to determine which modality or modality combination provided the most accurate auto-delineations. Thirdly, we introduce a new framework for structure-wise performance evaluation of multi-structure auto-delineations, as a supplement to already well-established performance metrics. This framework provides additional metrics to quantify the similarity between the expert's ground truth and the network predictions when more than one contoured structure is present in the ground truth, thereby enabling thorough evaluation of the strengths and weaknesses of auto-delineation approaches. Finally, auto-delineations were qualitatively assessed by an expert oncologist.

Material and methods

Imaging and contouring

HNC patients referred to curative chemoradiotherapy at Oslo University Hospital from January 2007 to December 2013 were retrospectively included, as described in [11]. Briefly, inclusion criteria were squamous cell carcinoma of the oral cavity, oropharynx, hypopharynx and larynx treated with curatively intended radio(chemo)therapy and available radiotherapy plans based on FDG PET/CT. Nasopharyngeal cancers were excluded, as were patients with known distal metastases and post-operative radiotherapy without residual tumour. In addition, patients without a contrast-enhanced planning CT were excluded, resulting in 197 patients included in the current analysis. Patient characteristics are provided in Supplementary Table A1 (Online Resource 1). The study was approved by the Regional Ethics Committee (REK) and the Institutional Review Board. Exemption from study-specific informed consent was granted by REK.

All patients had an RT optimised PET/CT scan (CT with contrast enhancement) taken on a Siemens Biograph 16 scanner (Siemens Healthineers Gmbh, Erlangen, Germany). After $\geq$ 6 h of fasting 370$\pm$20 MBq FDG was injected, and the patient rested for about one hour until imaging. Image acquisition was performed on an RT-compatible flat table with head support in an RT fixation mask. PET acquisition time was 5 min/bed position with 25% overlap between positions. The PET coincidence data were reconstructed using the OSEM4,8 algorithm with a Gaussian post-reconstruction filter with full width at half maximum equal to 3.5 mm for 193 patients, 2 mm for 3 patients and 5 mm for 1 patient. PET pixel size varied between 1.33 and 4.06 mm

(mode 2.66 mm for 143 patients) in a 256×256 matrix with a slice thickness of 1.0–5.00 mm (mode 2.00 mm for 169 patients). CT images were obtained with a peak tube voltage of 120 kV, giving a reconstructed matrix of 512×512, a pixel size of around 1.0 mm and a slice thickness of 2.0 mm. The Visipaque contrast agent was used, and the CT acquisition was performed after a delay of about 30 s post-injection. All PET and CT image series were resampled to a common isotropic $1\times1\times1$ mm^3 reference frame. The resulting image slices were cropped to a 191-by-265 mm^2 axial region of interest, keeping the patient in the centre of the full image stack.

The primary tumour (GTV-T) and, if present, malignant lymph nodes (GTV-N) were manually delineated by an experienced nuclear medicine specialist, based on the FDG uptake. These delineations were further refined by one to two (of many) oncology residents based on the contrast-enhanced CT and clinical information such as the endoscopy report. The delineations were finally approved by one of several senior oncologists. All delineations were performed at the time of initial RT (i.e. the patients received RT based on these delineations). The union of the manual GTV-T and GTV-N delineations were defined as the ground truth and used for training and evaluation of the CNN models. An overview of the number of manually delineated structures per patient and their volumes is given in Supplementary Tables A2 and A3 (Online Resource 1).

Model architecture and training

A U-Net architecture following the setup described in [12] was trained to delineate GTV-T and GTV-N in the PET/CT image slices. There was one addition to the original U-Net architecture, namely that batch normalisation [13] was applied after each ReLU non-linearity. Model details are provided in Supplementary Table A4 (Online Resource 1).

Four different loss functions were compared as follows: (1) the cross-entropy loss, (2) the Dice loss [14] and (3) the f_β loss with $\beta \in \{2, 4\}$ [15]. For each loss function, the models were trained using CT images only, PET images only and both PET and CT images. Additionally, the impact of CT windowing on model performance was assessed, using a narrow soft-tissue window of width 200 HU and a centre of 70 HU (range: [−30, 170] HU). The window centre of 70 HU corresponded to the median HU value within the GTV-T and GTV-N in the training set. In total, 20 models were run (i.e. 4 loss functions × (3 image input combinations without windowing + 2 input combinations with windowing)).

To assess model performance, we split the patients into three cohorts, stratifying by the primary tumour (T) stage of the TNM staging system to ensure similar patient characteristics across cohorts: A training cohort (142 patients), a validation cohort (15 patients) and a test cohort (40 patients). Patient characteristics of these cohorts are given in Supplementary Table A1 (Online Resource 1). To compare models, the patient-wise Dice score (1) was evaluated on patients in the validation cohort. Then, for each modality, the model achieving the highest Dice score was used to delineate in images from the test cohort. These test cohort auto-delineations were evaluated in depth, using the qualitative and quantitative methods described below.

To train the model, we used the Adam optimiser [16] with the β-values[1] recommended in [16] and a learning rate of 10^{-4}. The model was trained for 20 epochs, and the network coefficients were saved to disc (checkpointed) every second epoch. After training a model, we compared the average Dice score per image slice of each coefficient checkpoint. The coefficient checkpoint with the highest slice-wise Dice was used for subsequent performance analysis.

No post-processing was applied on the model output, such that the raw delineations provided by the CNN models were assessed without modifications.

Quantitative performance evaluation

Patient-wise metrics

Similarity and surface-distance metrics were used to assess the quality of the predicted delineations generated by the CNN models. Firstly, we measured overall delineation accuracy by the patient-wise (i.e. per patient) Dice score. The Dice score is given by:

$$\mathrm{Dice}(X, \hat{X}) = \frac{\left|X \cap \hat{X}\right|}{\frac{1}{2}\left|X\right| + \frac{1}{2}\left|\hat{X}\right|}, \tag{1}$$

where $|X|$ and $|\hat{X}|$ are the number of voxels in the ground truth, and the predicted delineations, respectively, and $|X \cap \hat{X}|$ denotes the number of voxels in the intersection between the ground truth and predicted delineations.

Next, we computed three surface-distance-based metrics for each patient (i.e. patient-wise): (1) the 95th percentile Hausdorff distance (HD_{95}), (2) the average surface distance (ASD) and (3) the median surface distance (MSD), all three of which were calculated from the same set of boundary distances. For a boundary voxel i in the predicted delineation, we computed its smallest distance D_i to the ground truth boundary $\tilde{X}$, given by:

$$D_i = \min_{\tilde{\boldsymbol{x}}_j \in \tilde{X}} \mathrm{dist}(\hat{\tilde{\boldsymbol{x}}}_i, \tilde{\boldsymbol{x}}_j), \tag{2}$$

where $\mathrm{dist}(\hat{\tilde{\boldsymbol{x}}}_i, \tilde{\boldsymbol{x}}_j)$ is the (Euclidean) distance between the predicted boundary voxel i with coordinates $\hat{\tilde{\boldsymbol{x}}}_i$ and the true

[1] These are different β values than those for the f_β loss.

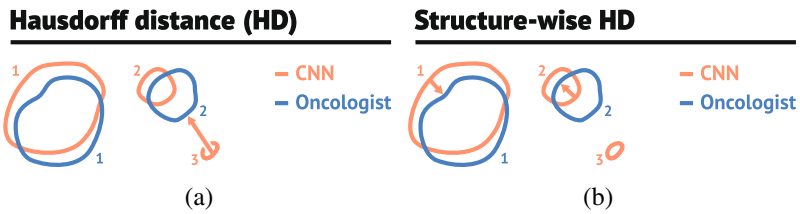


Fig. 1 The illustration in **a** demonstrates how the surface-distance based metrics (HD $= \max_i D_i$) can be non-informative in the presence of falsely predicted structures. As CNN structure 3 (*red*, CNN) does not overlap with any of the true manually delineated structures (*blue*, Oncologist), it is defined as a falsely predicted structure. Computing distance metrics between the false structure 3 and true structures will increase the metrics. The illustration in **b** shows how this problem can be alleviated by only computing the surface-distance-based metrics for predicted structures (*red*, labelled 1 and 2) that have sufficient overlap with the manually delineated ground truth (*blue*, labelled 1 and 2)

boundary voxel j with coordinates $\tilde{\boldsymbol{x}}_j$. From the set of all such distances, we computed HD_{95} as its 95% quantile, ASD as its average and MSD as its median. Thus, the HD_{95} measures how severe the largest delineation error is, and the ASD and MSD measure the overall delineation error. These surface-distance metrics should be as small as possible.

Structure-wise metrics

The distance-based metrics can be skewed if the CNN model misses a true structure or falsely predicts an additional structure not included in the ground truth, as illustrated in Fig. 1a. If distance-based metrics are to capture the delineation quality of the structures that are actually detected, they should be computed for true and predicted structures that overlap. Thus, we computed the degree of overlap between true and predicted structures, giving the *coverage fraction* (CFrac):

$$\mathrm{CFrac}(\hat{X}_k, X) = \frac{\left|\hat{X}_k \cap X\right|}{\left|\hat{X}_k\right|}. \tag{3}$$

Here $\hat{X}_k$ is the set of voxels in the kth structure of the predicted mask. An illustration of the CFrac is given in Fig. 2. If the CFrac was greater than 0.5, the predicted structure was defined as correctly identified by the CNN model. Thereafter, HD_{95}, MSD and ASD were computed separately for all structures in the auto-delineation with CFrac $\geq$ 0.5, as shown in Fig. 1b, giving structure-wise distance metrics not skewed by falsely predicted structures.

To further assess the performance of the CNN model, we defined a structure-wise sensitivity and positive predictive value. The number of true negative structures cannot be defined, and the number of true positive structures varies according to perspective (ground truth vs auto-delineation). As illustrated by the example in Fig. 2, there are two structures (1 and 2, red) in the auto-delineated mask that obtain a CFrac above 0.5 with the ground truth, and one that does not (red structure 3). The auto-delineation, therefore, has two true positive structures and one false positive structure. Thus, we define the structure-wise positive predictive value with respect to the CNN model (PPV_{CNN}) as:

$$\mathrm{PPV}_{\mathrm{CNN}} = \frac{\mathrm{TP}_{\mathrm{CNN}}}{\mathrm{TP}_{\mathrm{CNN}} + \mathrm{FP}}, \tag{4}$$

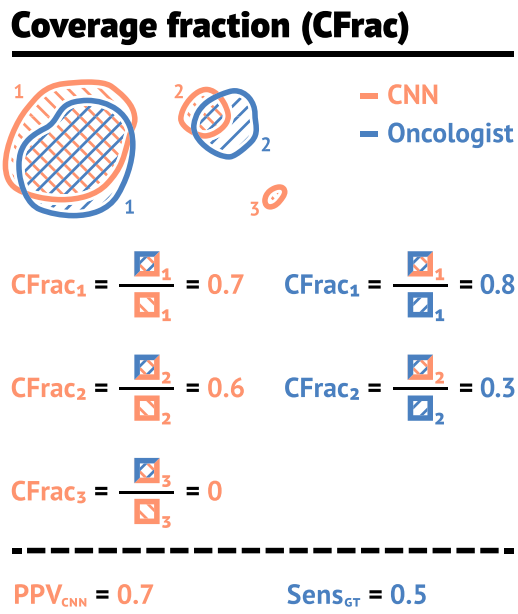


Fig. 2 Illustration of the coverage fraction metric CFrac. The left column gives the CFrac (3) for the overlap between the predicted (*red*, CNN) and true (*blue*, Oncologist) structures relative to the CNN structure (*red*). In this case, the PPV_{CNN} is 0.7, as two of the structures (*red*, 1 and 2) proposed by the CNN-model were delineated by the oncologist, and one (*red*, structure 3) was not. The right column gives CFrac for the overlap between the CNN (*red*) and true (*blue*) structures relative to the true structure (*blue*). The $Sens_{GT}$ is equal to 0.5 since one true structure (*blue*, 1) was identified by the CNN-model and one was not (*blue*, 2), as its coverage fraction was less than 0.5

where TP_{CNN} is the number of true positive structures in the auto-delineation mask and FP is the number of false positive structures. For Fig. 2, $TP_{CNN} = 2$ and FP= 1, giving a PPV_{CNN}=0.7 (rounded to one significant digit).

Likewise, for Fig. 2, there is one structure (1, blue) in the ground truth that obtain a CFrac above 0.5 with the auto-delineation mask, meaning that there is only one true positive with respect to the ground truth. Consequently, we define structure-wise sensitivity with respect to the ground truth ($Sens_{GT}$):

$$Sens_{GT} = \frac{TP_{GT}}{TP_{GT} + FN}, \quad (5)$$

where TP_{GT} is the number of true positive structures with respect to the ground truth and FN is the number of structures in the ground truth not delineated by the CNN model (false negatives). In Fig. 2, $TP_{GT} = 1$ and FN= 1, giving a $Sens_{GT} = 0.5$.

Finally, to further assess errors made by the CNN model, we calculated (1) the volume of structures in the auto-delineation that obtained a CFrac > 0.5 with the ground truth ($Volume_{true}$), and (2) the volume of structures in the auto-delineation that obtained a CFrac ≤ 0.5 with the ground truth ($Volume_{false}$). For the delineations in Fig. 2, the $Volume_{true}$ is the mean volume of CNN (red) structures 1 and 2 and $Volume_{false}$ is the volume of CNN structure 3.

Qualitative evaluation

The CNN model with superior mean Dice performance was qualitatively evaluated by an expert oncologist with more than 7-year experience in HNC target volume delineation. The expert was presented with the ground truth and the delineations made by the CNN-model for 15 patients randomly selected from the test cohort. The expert did not know which contour was CNN-generated and which was human-generated. For each of these patients, the oncologist was asked to identify (if possible) which delineation was generated by the CNN model. The oncologist scored the quality of the selected auto-delineation masks using a score from one to ten. A score of one represented a delineation with little to no clinical value and a score of ten represented a delineation where the oncologist was unable to identify whether the mask was generated by the CNN model or human specialists, implying high clinical value.

Code

Models were trained using Python and TensorFlow. Code for running the experiments is provided at https://github.com/yngvem/EJNMMI-20. Performance metrics were computed using an in-house developed Python library provided at: https://github.com/yngvem/mask_stats.

Results

Comparison of models

The average model performance on the validation cohort is summarised in Table 1. All models had an average Dice between 0.40 and 0.65. Note that standard deviations of

Table 1 The performance (mean ± one standard deviation) of CNN models trained using different modalities

			Modality				
			PET	CT		PET/CT	
			–	CTW	CT	CTW	CT
Patient-wise	Dice	(%)	61 ± 2	55 ± 2	48 ± 5	63 ± 1	62 ± 1
	ASD	(mm)	8.1 ± 2.6	11 ± 3	13 ± 7	7.0 ± 0.8	8.0 ± 3.1
	MSD	(mm)	4.5 ± 1.9	5.6 ± 0.8	7.8 ± 2.5	4.6 ± 1.0	4.6 ± 2.6
	HD_{95}	(mm)	31 ± 14	38 ± 17	50 ± 44	24 ± 2	32 ± 17
	$Sens_{GT}$	(%)	75 ± 5	60 ± 9	53 ± 11	75 ± 4	78 ± 7
	PPV_{CNN}	(%)	25 ± 5	22 ± 6	21 ± 11	26 ± 4	28 ± 11
Structure-wise	ASD	(mm)	1.6 ± 0.2	2.1 ± 0.4	2.4 ± 0.6	1.6 ± 0.3	1.4 ± 0.2
	MSD	(mm)	1.1 ± 0.2	1.7 ± 0.4	1.9 ± 0.6	1.1 ± 0.3	0.92 ± 0.14
	HD_{95}	(mm)	4.9 ± 0.7	5.6 ± 1.0	6.0 ± 1.0	4.5 ± 0.5	4.4 ± 0.6
	$Volume_{true}$	(cm^3)	17 ± 5	9.7 ± 3.7	11 ± 3	15 ± 4	16 ± 3
	$Volume_{false}$	(cm^3)	1.1 ± 1.3	0.41 ± 0.17	1.6 ± 2.6	0.58 ± 0.54	0.58 ± 0.41

Choice of loss function had little effect on performance, and averaging therefore was done over models trained with different loss functions. CTW and CT columns represent models trained with and without CT windowing, respectively. All models were evaluated on the validation cohort

Table 2 Performance on the test cohort for the CNN models with the highest Dice score on the validation set, using different input modalities

			Modality					
			PET		CT		PET/CT	
			Mean	Std.	Mean	Std.	Mean	Std.
Patient-wise	Dice	(%)	69	17	56	21	71	16
	ASD	(mm)	4.2	4.3	6.1	5.1	4.7	4.8
	MSD	(mm)	1.9	4.0	3.1	5.4	1.8	4.0
	HD_{95}	(mm)	18	16.5	22.2	11.7	21.2	17.1
	$Sens_{GT}$	(%)	77	31	53	41	86	27
	PPV_{CNN}	(%)	45	29	28	17	33	23
Structure-wise	ASD	(mm)	1.1	0.6	1.4	1.4	1.0	0.6
	MSD	(mm)	0.61	0.54	0.96	1.45	0.56	0.67
	HD_{95}	(mm)	4.0	2.6	4.1	2.5	3.3	1.8
	$Volume_{true}$	(cm^3)	17	22	7.2	12	15	24
	$Volume_{false}$	(cm^3)	0.45	1.1	0.56	1.7	0.54	3.0

The CT images were pre-processed using windowing

the Dice score and structure-wise performance metrics were relatively small, indicating that the model performance was stable between models trained with the same modality and windowing option, but with different loss functions. Thus, loss function choice had little effect on performance.

Imaging modality and Hounsfield windowing of CT images, however, had a clear effect on performance. Models trained on both PET and CT images had the highest patient-wise Dice performance and the lowest surface distances, indicating a high degree of overlap between the model prediction and the ground truth. Models trained solely on PET images had lower Dice and larger surface distances than PET/CT models, but outperformed models based on CT images on all performance metrics. Note that patient-wise surface distances were both larger and more varied than structure-wise due to measurements between false positive structures and the ground truth.

From Table 1, it is also apparent that models, on average, identified more than 50% of the manually delineated structures in the validation cohort ($Sens_{GT}$). Particularly models built using PET images had high detection rates, identifying more than 75% of true structures. However, the models generated many false positive structures, which is apparent from the low PPV_{CNN}. On average, less than a third of all delineated structures in the CNN masks (for all modalities) were also present in the ground truth. Despite this, the Dice was high, indicating that the false positive structures were small in volume (see $Volume_{false}$ in Table 1).

Performance on the test cohort

For each input-modality, the model that achieved the highest average Dice score on the validation cohort was selected for further evaluation on the test cohort. The best PET-based model was trained with the Dice loss function. The best CT-based and PET/CT-based models were trained using CT-windowing and the f_2 loss function. Test cohort performance metrics are shown in Table 2 and Fig. 3. The CT model had the lowest patient-wise Dice score (56%) as well as the largest patient-wise distance metrics, indicating poorer overlap between the predicted delineation and the ground truth relative to PET and PET/CT models. PET and PET/CT models achieved high Dice performance (69% and 71%, respectively) and structure-wise sensitivity ($Sens_{GT}$) (77% and 86%, respectively), indicating that these models had high overlap with the ground truth and detected the majority of the manually delineated structures (i.e. few false negative structures, Eq. 5).

Even though the CT model on average identified 53% of the manually delineated structures ($Sens_{GT}$), it was unable to identify even a single structure for 10 patients in the test cohort (data not shown). In contrast, the PET and PET/CT models failed to identify a single structure for only two patients. Moreover, from the boxplots overlayed on the violin plots in Fig. 3a, we see that the 25th percentile $Sens_{GT}$ was 15% for the CT-based model, while it was 50% for the PET-based model and 93% for the PET/CT-based model, again highlighting the PET/CT model's high rate of structure identification.

The structure-wise metrics illustrate (see Table 2 and Fig. 3) that using both the PET and CT signal simultaneously was beneficial compared to only using one modality. All structure-wise surface distance metrics were smaller for the PET/CT model and spanned a narrower range with fewer large outliers compared to the models that used only a single modality. Thus, PET/CT-based auto-delineations were more

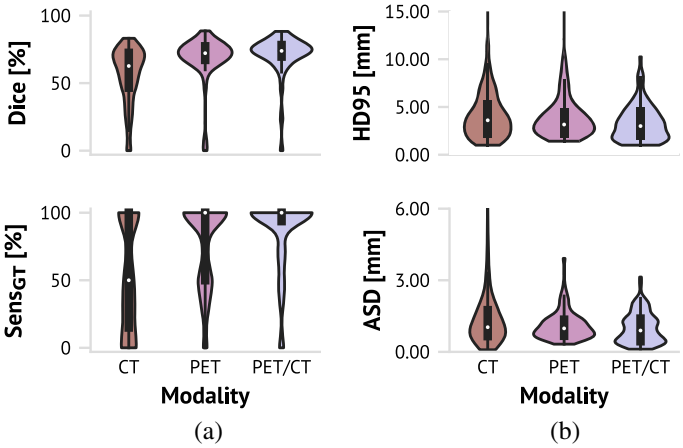


Fig. 3 Violin plots with boxplots overlayed (*dark gray box* within) for test cohort performance metrics of models achieving the highest Dice based on each imaging modality. **a** Patient-wise Dice score and structure-wise sensitivity ($Sens_{GT}$). **b** Structure-wise HD_{95} and ASD distance metrics. Note that the axis for ASD is cut off at 6 mm to improve visualisation for the PET and PET/CT-based models. There was one structure generated by the CT-based model which had an ASD outside this range (13 mm). Refer to Table 2 for details

accurate with fewer large deviations between the predicted and true structure boundaries.

Note also that the structure-wise distance metrics (ASD, MSD, HD_{95}) were considerably smaller than the corresponding patient-wise distance metrics (Table 2). This indicates that the patient-wise distance metrics were influenced by measurements between falsely delineated structures (false positives) and the ground truth (see Fig. 1a). This is further supported by the metric PPV_{CNN}, which was below 50%, indicating that the CNN models tended to delineate several false positive structures (i.e. large FP, Eq. 4). However, the volumes of the erroneously predicted structures ($Volume_{false}$) were small compared to the true structure volumes (see Supplementary Table A2, Online Resource 1). This is also reflected by the high Dice score of all the models. Furthermore, the average volume of the erroneous structures ($Volume_{false}$) for the PET/CT model was 0.54 cm^3. There were only five true structures in the entire data set (< 5%) smaller than or equal to this size (data not shown).

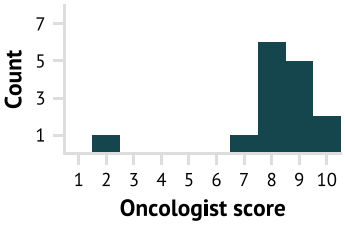


Fig. 4 Qualitative evaluation by an experienced oncologist of 15 PET/CT-based CNN delineations, randomly selected from the test cohort. A score of 1 corresponds to a CNN delineation requiring extensive revision whereas a score of 10 corresponds to a CNN delineation that was indistinguishable from a manual delineation

Qualitative performance evaluation

The score distribution of the oncologist's evaluation of the PET/CT-based CNN delineations for 15 patients in the test cohort is shown in Fig. 4, with performance details given in Table 3. The majority of the CNN delineations were of high quality. Thirteen delineations were scored 8 or higher, indicating that the CNN delineations only required minor modifications by an oncologist. For the two cases receiving a score of 10, the oncologist was unable to decide which delineation was generated by the CNN model and human specialists. Only one case was assessed to a score < 7. This auto-delineation received a score of 2, indicating that major revision was required.

Figure 5 shows a representative image slice for three patients whose PET/CT delineation was qualitatively assessed by an oncologist. Animations of these delineations are provided in Online Resources 2, 3 and 4 and performance metric details are highlighted in italics in Table 3. The upper row shows a patient for whom the oncologist was unable to differentiate between the PET/CT-based auto-delineation and the ground truth (Online Resource 2). For this patient, both the PET- and the PET/CT-based models performed well, whereas the CT-based model missed the primary tumour in the larynx. The middle row shows a patient for whom the PET/CT-based auto-delineation obtained a qualitative score of 8 (Online Resource 3). The CT-based auto-delineation only identified one structure and contained two false positive structures. PET-based auto-delineation, however, correctly identified both structures. Likewise, the PET/CT-based auto-delineation correctly identified both structures, but included one false positive structure. This false positive structure resulted in a high patient-wise HD_{95} of 50 mm. However, the correctly identified structures had an average

Table 3 Performance metrics for 15 randomly selected patients in the test cohort, whose auto-delineated contour (generated by the superior PET/CT model) was evaluated by an experienced oncologist

				HD_{95} (mm)		ASD (mm)		MSD (mm)	
Score	Dice (%)	$Sens_{GT}$ (%)	PPV_{CNN} (%)	SW	PW	SW	PW	SW	PW
2	*24*	*0*	*45*	*3.4*	*45*	*1.8*	*13*	*1.8*	*4.9*
7	73	100	17	4.6	30	1.5	4.8	1.1	1.0
8	74	100	17	4.9	33	1.5	4.7	1.0	1.4
8	74	40	29	4.1	17	1.1	2.6	0.57	1.0
8	*74*	*100*	*33*	*3.4*	*50*	*1.1*	*5.4*	*0.63*	*1.4*
8	77	100	25	4.6	30	1.4	5.4	0.72	1.4
8	73	67	14	3.9	23	1.1	4.8	0.59	1.4
8	88	100	25	4.9	6.8	1.5	1.8	0.88	0
9	75	100	33	3.4	9.7	1.0	2.1	0.56	1.0
9	85	100	22	4.3	8.1	1.2	1.9	0.50	1.0
9	74	50	56	3.1	3.7	1.1	1.1	0.75	0
9	85	100	38	3.0	5.0	0.73	1.1	0.14	0
9	78	100	33	3.0	42	0.90	5.8	0.40	1.0
10	77	75	75	3.0	4.1	0.86	1.2	0.40	1.0
10	*73*	*100*	*33*	*3.6*	*8.8*	*1.2*	*2.5*	*0.74*	*1.0*

Dice, $Sens_{GT}$ and PPV_{CNN} are given patient-wise. For split metrics, PW represents the patient-wise metric and SW represents the structure-wise metric. Representative auto-delineations for the italicised rows (patients) are shown in Fig. 5

structure-wise HD_{95} of 3.4 mm (Table 3), indicating that the CNN model delineated these structures adequately, more in line with the oncologist's evaluation. The bottom row shows the patient for whom the PET/CT-based auto-delineation obtained a qualitative score of 2 (Online Resource 4). Here, all models, regardless of input-modality, failed at delineating the true structures, likely caused by the strong beam hardening artefacts and low PET-signal.

Discussion

Comparison to previous work

To the best of our knowledge, only three previous studies have evaluated the use of CNNs for auto-delineation of the GTV in HNC using multimodal images [8–10]. The PET/MRI-based 3D CNN of Lin et al. [8] and the PET/CT-based 2D CNN described in Huang et al. [9] obtained a median Dice score of 79% and a mean Dice score of 74%, respectively, for auto-delineation of the primary tumour volume. As in the present study, Guo et al. [10] achieved superior auto-delineation performance of the GTV-T and GTV-N for combined PET/CT input, compared to using single modality CT or PET input. Their PET/CT-based 3D network (Dense Net) resulted in a mean Dice of 71%. Similarly, our 2D U-Net obtained a mean Dice of 71% for PET/CT-based GTV-T and GTV-N auto-delineation. One notable difference between the present study and the results reported by Guo et al. is the quality of auto-delineations obtained using solely CT images. By only including CT intensities in the range [−30, 170] HU, we obtained a mean Dice of 56%, whereas Guo et al. reported a considerably lower mean Dice of 31% using a wider CT window in the range [−200, 200] HU [10].

Despite differences across imaging modalities in the above studies, the median or mean agreement between CNNs using multimodality input and the expert's ground truth is above 70%. Previous studies conclude that there are considerable interobserver variations in manual HNC target volume delineations [4–7, 17]. In Bird et al. [17], the Dice agreement between five clinicians (three radiation oncologists and two radiologists) was only 56% when delineating the GTV in CT images. Similarly, Gudi et al. [7] found that the Dice agreement between three radiation oncologists was 57% when the GTV was delineated using CT images and 69% when delineated using PET/CT-images. Thus, the interobserver variability of clinicians is similar to the performance of the present CNN-model, which, when evaluated on the test cohort, had an average Dice score of 56% and 71% for the CT and PET/CT-based models, respectively. Furthermore, Lin et al. [8] found that the interobserver and intraobserver variability between oncologists, as well as the contouring time, decreased

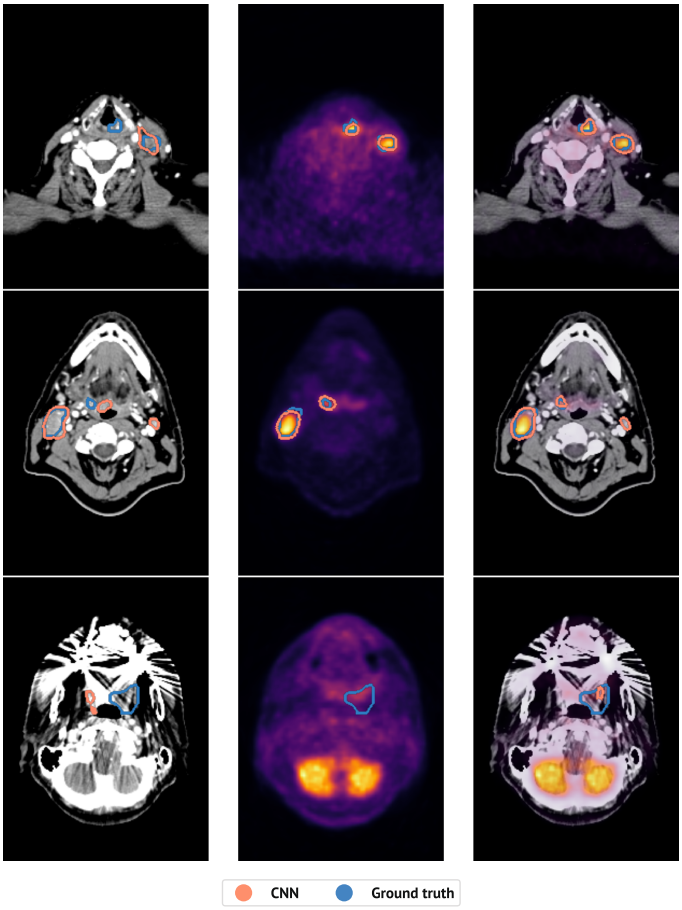


Fig. 5 The predicted (*red*) and true (*blue*) delineations for one representative image slice from three different patients in the test cohort. From left to right: CT-based predictions; PET-based predictions; PET/CT-based predictions. From top to bottom, different subjects for whom an experienced oncologist gave the PET/CT-based predictions a qualitative score of 10, 8 and 2, respectively. Performance metrics for the shown patients are marked with italicised text in Table 3. Animations are given in Online Resources 2 (top patient), 3 (middle patient) and 4 (bottom patient)

significantly when CNN-based auto-delineations were used to assist manual delineations, highlighting the possible clinical value of auto-delineation tools.

Clinical usefulness of CNN-based auto-delineation

Both the quantitative performance metrics and the qualitative oncologist's evaluation illustrate that despite the moderate amount of training samples and the simple CNN architecture, the models produced delineations of high quality. We observed that the auto-delineations could be useful in RT with just minor to moderate refinements required, such as removing false positive structure, delineating a missing structure, or refining the delineation boundary. We infer this conclusion both from the qualitative scores provided by an expert oncologist as well as the quantitative surface-distance metrics. The average structure-wisesurface distances between true and predicted delineated structures were on the same order of magnitude as the CT resolution ($\sim$ 1 mm). Furthermore, structure-wise HD_{95} were on the same order of magnitude as the PET resolution ($\sim$ 3 mm). We can, in other words, conclude that the CNN model generated highly accurate auto-delineations, with few exceptions.

The CNN model does, however, exhibit some weaknesses. Some CNN delineated structures are false positives. Moreover, not all ground truth structures are detected. The false positive structures are of minor concern, as most of them are smaller than the resolution of the PET-images. A simple post-processing procedure can easily remove such small structures. In contrast, the lack of sensitivity is more problematic, since all malignant structures should be treated. However, the average $Sens_{GT}$ was 86% for the PET/CT model, and all structures were identified for 75% of the patients in the test cohort. It was further noted that many of the patients with low CNN performance had beam

hardening artefacts on the CT image, leading to slices with little-to-no information from the CT-signal, as can be seen in the bottom row of Fig. 5. This implies that the model worked well for a large portion of the patients, but a small number of patients would still require considerable manual refinements before RT. Lastly, our models were trained and evaluated on images acquired at one single centre. An important next step is an assessment of our models' generalisability to images stemming from other centres.

The effect of imaging modality

The CT signal indicates the mass density of tissue. However, we are interested in the properties of soft-tissue tumours and involved lymph nodes, which are only represented in a small section of the CT range. As such, analysing the entire CT range is unnecessary and could even make it harder to find relevant features. This motivated the reduction in dynamic range of the CT images, utilising a soft-tissue window ranging from −30 to 170 HU.

From a deep learning perspective, such an a priori decrease in dynamic range is not expected to affect model performance to a great extent, as the same transformation can be learned by a two-layer neural network with ReLU activation functions. Nevertheless, our experiments strongly suggest that decreasing the dynamic range of the CT images can have a considerable positive effect on model performance. This increase in performance will likely be less prominent as the dataset size grows, because then, it may be easier for the model to learn the windowing-operation. Further discussion of imaging modalities will, therefore, only consider CT and PET/CT models where the CT images were pre-processed using the given Hounsfield window settings.

When we compare the performance of the models based on their input modality, we notice that the PET signal was essential for discovering the involved lymph nodes correctly. Without the PET-signal, the models, on average, only discovered 60% ($Sens_{GT}$) of the manually delineated structures (GTV-T and GTV-N) in the validation cohort. For patients in the test cohort, the CNN model performed worse. The highest performing CT-based CNN-model only managed to identify 53% of the manually delineated structures. Conversely, models trained using only PET images and models trained using both PET and CT images, delineated on average 75% of the malignant structures for the validation cohort. On the test cohort, the highest performing PET and PET/CT models discovered 77% and 86%, respectively. Thus, we conclude that the PET signal was crucial for obtaining auto-delineation models with sufficient sensitivity.

A benefit of CT, compared to PET, is its higher spatial resolution. In our experiments, the surface distances between the detected structures and their corresponding ground truth boundaries were smaller for the models that incorporated the CT signal as compared to those without CT input. Hence, the high resolution of the CT was essential to identify the small details and provide an accurate boundary of the structures. Finally, combining the PET signal with the CT signal improved all quantitative performance metrics except for the PPV_{CNN}, for most patients. We therefore recommend using a fused PET/CT approach for auto-delineation of head and neck tumours and involved nodes.

The performance metrics

By including structure-wise performance metrics, as opposed to only voxel- and patient-wise performance metrics, we were able to quantitatively analyse the results in a more in-depth fashion. These structure-wise metrics are meant as a supplement providing additional information on the quality of the auto-delineation, not as a replacement of the well-established and commonly reported metrics, which must be reported to enable cross-study comparisons. Thus, it is the joint information provided by the different types of metrics that we consider useful.

The added information content of the structure-wise metrics is demonstrated in Tables 2 and 3. We see that the patient-wise surface distances are relatively large—especially the HD_{95} metric. However, the corresponding structure-wise distance metrics are much smaller. As these metrics were only calculated for auto-delineated structures that overlapped by more than 50% with the ground truth, falsely predicted structures that would otherwise skew the distance-metrics were avoided. Thus, the discrepancy between the patient-wise and the structure-wise distance metrics indicates that the models predict false positive structures. Furthermore, the small structure-wise distance metrics demonstrate that the models performed well for the true structures they actually detected and delineated, thereby indicating how much manual modification is required by the clinician. Likewise, the structure-wise sensitivity, PPV and volumes were very informative as to the types of errors the models make, such as how many true structures they miss, how many false structures they predict and how large these are.

Conclusions

In summary, we show that CNNs can be used for accurate and precise GTV delineations of HNC using multimodality PET/CT. Furthermore, our proposed structure-wise performance metrics enabled in-depth assessment of CNN predictions and errors, which may facilitate the use of such auto-delineation tools in RT planning.

Acknowledgements We thank Marie Roald for making the illustrations for the quantitative performance metrics, Inês Das Neves for valuable discussions on writing and Kristian Hovde Liland for valuable feedback.

Funding The study was funded by the Norwegian Cancer Society (Grants 160907-2014 and 182672-2016).

Data availability Data access requires approval by the Regional Ethics Committee.

Code availability Code used for model training and evaluation is available at https://github.com/yngvem/EJNMMI-20 and https://github.com/yngvem/mask_stats, respectively.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

Ethics approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Consent to participate Exemption from study-specific informed consent was granted by the Regional Ethics Committee.

Consent for publication Exemption from study-specific informed consent was granted by the Regional Ethics Committee.

Supplementary Information The online version contains supplementary material available at (https://doi.org/10.1007/s00259-020-05125-x).

References

1. Alterio D, Marvaso G, Ferrari A, Volpe S, Orecchia R, Jereczek-Fossa BA. Modern radiotherapy for head and neck cancer. Semin Oncol. 2019;46(3):233–45.
2. van den Bosch S, Doornaert PAH, Dijkema T, Zwijnenburg EM, Verhoef LCG, Hoeben BAW, Kasperts N, Smid EJ, Terhaard CHJ, Kaanders JHAM. 18F-FDG-PET/CT-based treatment planning for definitive (chemo)radiotherapy in patients with head and neck squamous cell carcinoma improves regional control and survival. Radiother Oncol. 2020;142:107–114. ISSN 0167-8140.
3. Grégoire V, Thorwarth D, Lee JA. Molecular imaging-guided radiotherapy for the treatment of head-and-neck squamous cell carcinoma: does it fulfill the promises? Semin Radiat Oncol. 2018;28(1):35–45.
4. Ashamalla H, Guirgius A, Bieniek E, Rafla S, Evola A, Goswami G, Oldroyd R, Mokhtar B, Parikh K. The impact of positron emission tomography/computed tomography in edge delineation of gross tumor volume for head and neck cancers. Int J Radiat Oncol. 2007;68(2):388–95. ISSN 0360-3016.
5. Murakami R, Uozumi H, Hirai T, Nishimura R, Katsuragawa S, Shiraishi S, Toya R, Tashiro K, Kawanaka K, Oya N, Tomiguchi S, Yamashita Y. Impact of FDG-PET/CT fused imaging on tumor volume assessment of head-and-neck squamous cell carcinoma: intermethod and interobserver variations. Acta Radiol. 2008;49(6):693–9.
6. Kajitani C, Asakawa I, Uto F, Katayama E, Inoue K, Tamamoto T, Shirone N, Okamoto H, Kirita T, Hasegawa M. Efficacy of FDG-PET for defining gross tumor volume of head and neck cancer. J Radiat Res. 2013;01(4):671–8. ISSN 0449-3060.
7. Gudi S, Ghosh-Laskar S, Agarwal JP, Chaudhari S, Rangarajan V, Paul SN, Upreti R, Murthy V, Budrukkar A, Gupta T. Interobserver variability in the delineation of gross tumour volume and specified organs-at-risk during IMRT for head and neck cancers and the impact of FDG-PET/CT on such variability at the primary site. J Med Imaging Radiat Sci. 2017;48(2):184–92. ISSN 1939-8654.
8. Lin L, Dou Q, Jin Y-M, Zhou G-Q, Tang Y-Q, Chen W-L, Su B-A, Liu F, Tao C-J, Jiang N, et al. Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma. Radiology. 2019;291(3):677–86. PMID: 30912722.
9. Huang B, Chen Z, Wu P-M, Ye Y, Feng S-T, Wong C-YO, Zheng L, Liu Y, Wang T, Li Q, et al. Fully automated delineation of gross tumor volume for head and neck cancer on PET-CT using deep learning: a dual-center study. Contrast Media Mol Imaging 2018. 2018.
10. Guo Z, Guo N, Gong K, Li Q. Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multi-modality network. Phys Med Biol. 2019;64(20):205015.
11. Moan JM, Amdal CD, Malinen E, Svestad JG, Bogsrud TV, Dale E. The prognostic role of 18F-fluorodeoxyglucose PET in head and neck cancer depends on HPV status. Radiother Oncol. 2019;140:54–61. ISSN 0167-8140.
12. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: International conference med image comput comp assist interv. Springer; 2015. p. 234–41.
13. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International conference mach learn; 2015. p. 448–56.
14. Milletari F, Navab N, Ahmadi S-A. V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 4th international conference on 3d vision. IEEE; 2016. p. 565–71.
15. Hashemi SR, Salehi SSM, Erdogmus D, Prabhu SP, Warfield SK, Gholipour A. Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: application to multiple sclerosis lesion detection. IEEE Access. 2019;7:1721–35. ISSN 2169-3536.
16. Kingma DP, Ba JL. ADAM: a method for stochastic optimization. In: International conference learn represent; 2014.
17. Bird D, Scarsbrook AF, Sykes J, Ramasamy S, Subesinghe M, Carey B, Wilson DJ, Roberts G, McDermott N, Karakaya E, Bayman E, Sen M, Speight R, Prestwich RJD. Multimodality imaging with CT, MR and FDG-PET for radiotherapy target volume delineation in oropharyngeal squamous cell carcinoma. BMC Cancer. 2015;15(1):1–10.

Deep learning based auto-delineation of gross tumour volumes and involved nodes in PET/CT images of head and neck cancer patients

Yngve Mardal Moe · Aurora Rosvoll Groendahl · Oliver Tomic · Einar Dale · Eirik Malinen · Cecilia Marie Futsaether

A Supplementary material

A.1 Dataset statistics

Table A1 shows patient characteristics for the entire patient cohort and the patients included in the training, validation and test cohorts. In addition, Table A2 shows summary statistics for the distribution of structure sizes in the ground truth for the three cohorts. Table A3 shows summary statistics for the number of structures in the ground truth for each patient in the cohorts.

A.2 Architecture

Table A4 shows the architecture of the neural networks. All convolutional layers except the last and all up-convolutional (transposed strided convolution) layers consisted of a (3×3)-(up-)convolution, followed by a ReLU activation function and finally a batch normalisation layer. Up-convolutional layers also included a bilinear interpolation layer after batch normalisation to match the layer before downsampling. The final convolutional layer consisted of a (1×1)-convolution followed by a sigmoidal activation function. All convolutional and up-convolutional layers included a bias-term for each output channel and all convolution-weights were initialised using the normally distributed He-scheme. The code for the experiments are available on GitHub: `https://github.com/yngvem/EJNMMI-2020`.

C.M. Futsaether
Norwegian University of Life Sciences
E-mail: cecilia.futsaether@nmbu.no

Table A1 Patient characteristics.

Characteristic[a]	All patients ($n = 197$)	Train ($n = 142$)	Validation ($n = 15$)	Test ($n = 40$)
Age [years]				
Mean	60.3	60.7	58.8	59.4
Range	39.9–79.1	39.9–79.1	43.2–73.7	43.0–77.0
Sex				
Female	24.9 %	25.4 %	13.3 %	27.5 %
Male	75.1 %	74.7 %	86.7 %	72.5 %
TNM[b]				
T1	9.1 %	9.2 %	6.7 %	10.0 %
T2	39.6 %	39.4 %	40.0 %	40.0 %
T3	23.4 %	23.9 %	20.0 %	22.5 %
T4	27.9 %	27.5 %	33.3 %	27.5 %
N0	23.9 %	25.4 %	6.7 %	25.0 %
N1	11.7 %	12.0 %	13.3 %	10.0 %
N2	60.9 %	58.5 %	80 %	62.5 %
N3	3.6 %	4.2 %	0 %	2.5 %
AJCC/UICC[b] stage				
I	1.0 %	1.4 %	0 %	0 %
II	8.6 %	9.2 %	0 %	10.0 %
III	19.8 %	19.7 %	20.0 %	20.0 %
IV	70.1 %	69.0 %	80.0 %	70.0 %
Tumour site				
Oral cavity	8.6 %	7.0 %	26.7 %	7.5 %
Oropharynx	72.6 %	73.2 %	60.0 %	75.0 %
Hypopharynx	8.1 %	9.2 %	13.3 %	2.5 %
Larynx	10.7 %	10.1 %	0 %	15.0 %
GTV-T[c] [cm^3]				
Mean	25.0	23.9	37.3	24.3
Range	0.8–285.0	0.8–285.0	2.6–247.2	1.4–157.6
GTV-N[d] [cm^3]				
Mean	19.3	26.6	37.4	19.5
Range	0.5–276.7	0.5–276.7	2.6–247.2	0.5–76.4

[a] Percentages may not sum to exactly 100 due to rounding.
[b] 7$^{\text{th}}$ edition
[c] Gross primary tumour volume
[d] Involved nodal volume (for patients with nodal stage $\geq$ N1)

Table A2 Summary statistics for structure sizes.

		Cohort		
		Train	Validation	Test
Mean	[cm^3]	17	22	16
Standard deviation	[cm^3]	31	43	24
25% percentile	[cm^3]	2.2	2.3	2.3
Median	[cm^3]	7.2	6.3	6.2
75% percentile	[cm^3]	18	27	21
Min	[cm^3]	0.10	0.35	0.31
Max	[cm^3]	28	25	17

Table A3 Summary statistics for the number of structures per patient.

	Cohort		
	Train	Validation	Test
Mean	2.5	2.3	2.3
Standard deviation	1.4	0.86	1.2
25% percentile	2.0	2.0	1.0
Median	2.0	2.0	2.0
75% percentile	3.0	3.0	3.0
Min	1.0	1.0	1.0
Max	10	4.0	6.0

Table A4 Model architecture.

Name	Inputs	Output shape
Conv1	Dataset	$191 \times 265 \times 64$
Conv2	Conv1	$191 \times 265 \times 64$
MaxPool1	Conv2	$95 \times 132 \times 64$
Conv3	MaxPool1	$95 \times 132 \times 128$
Conv4	Conv3	$95 \times 132 \times 128$
MaxPool2	Conv4	$47 \times 66 \times 128$
Conv5	MaxPool2	$47 \times 66 \times 256$
Conv6	Conv5	$47 \times 66 \times 256$
MaxPool3	Conv6	$23 \times 33 \times 256$
Conv7	MaxPool3	$23 \times 33 \times 512$
Conv8	Conv7	$23 \times 33 \times 512$
MaxPool4	Conv8	$11 \times 16 \times 512$
Conv9	MaxPool4	$11 \times 16 \times 1024$
Conv10	Conv9	$11 \times 16 \times 1024$
UpConv1	Conv10	$23 \times 33 \times 512$
Conv11	UpConv1 & Conv8	$23 \times 33 \times 512$
Conv12	Conv11	$23 \times 33 \times 512$
UpConv2	Conv12	$47 \times 66 \times 256$
Conv13	UpConv2 & Conv6	$47 \times 66 \times 256$
Conv14	Conv13	$47 \times 66 \times 256$
UpConv3	Conv14	$95 \times 132 \times 128$
Conv15	UpConv3 & Conv4	$95 \times 132 \times 128$
Conv16	Conv15	$95 \times 132 \times 128$
UpConv4	Conv16	$191 \times 265 \times 64$
Conv17	UpConv4 & Conv2	$191 \times 265 \times 64$
Conv18	Conv17	$191 \times 265 \times 64$
FinalConv	Conv18	$191 \times 265 \times 1$

Appendix C

Paper III

Accepted revised author manuscript

Received: 20 June 2021 | Revised: 8 October 2021 | Accepted for publication: 13 October 2021

Published Online 16 November 2021
Acta Oncol. 61 (2021) 1:89–96. doi: https://doi.org/10.1080/0284186X.2021.1994645.

Deep learning-based automatic delineation of anal cancer gross tumour volume: A multimodality comparison of CT, PET and MRI

Aurora Rosvoll Groendahl[a], Yngve Mardal Moe[a], Christine Kiran Kaushal[a], Bao Ngoc Huynh[a], Espen Rusten[b], Oliver Tomic[a], Eivor Hernes[c], Bettina Hanekamp[c], Christine Undseth[d], Marianne Grønlie Guren[d,e], Eirik Malinen[b,f], Cecilia Marie Futsaether[a]*.

[a]Faculty of Science and Technology, Norwegian University of Life Sciences, Ås, Norway; [b]Department of Medical Physics, Oslo University Hospital, Oslo, Norway; [c]Department of Radiology and Nuclear Medicine, Oslo University Hospital, Oslo, Norway; [d]Department of Oncology, Oslo University Hospital, Oslo, Norway; [e]Division of Cancer Medicine, Institute of Clinical Medicine, University of Oslo, Oslo, Norway; [f]Department of Physics, University of Oslo, Oslo, Norway.

*Corresponding author: Cecilia Marie Futsaether
Faculty of Science and Technology
Norwegian University of Life Sciences
P. O. box 5003 NMBU
NO-1432 Ås
e-mail: cecilia.futsaether@nmbu.no

Word count: 3 498

Deep learning-based automatic delineation of anal cancer gross tumour volume: A multimodality comparison of CT, PET and MRI

Abstract

Background: Accurate target volume delineation is a prerequisite for high-precision radiotherapy. However, manual delineation is resource-demanding and prone to interobserver variation. An automatic delineation approach could potentially save time and increase delineation consistency. In this study, the applicability of deep learning for fully automatic delineation of the gross tumour volume (GTV) in patients with anal squamous cell carcinoma (ASCC) was evaluated for the first time. An extensive comparison of the effects single modality and multimodality combinations of computed tomography (CT), positron-emission tomography (PET) and magnetic resonance imaging (MRI) have on automatic delineation quality was conducted.

Materials and methods: 18F-fluorodeoxyglucose PET/CT and contrast-enhanced CT (ceCT) images were collected for 86 patients with ASCC. A subset of 36 patients also underwent a study-specific 3T MRI examination including T2- and diffusion-weighted imaging. The resulting two datasets were analysed separately. A two-dimensional U-Net convolutional neural network (CNN) was trained to delineate the GTV in axial image slices based on single or multimodality image input. Manual GTV delineations constituted the ground truth for CNN model training and evaluation. Models were evaluated using the Dice similarity index (Dice) and surface distance metrics computed from five-fold cross-validation.

Results: CNN-generated automatic delineations demonstrated good agreement with the ground truth, resulting in mean Dice scores of 0.65–0.76 and 0.74–0.83 for the 86 and 36-patient datasets, respectively. For both datasets, the highest mean Dice scores were obtained using a multimodal combination of PET and ceCT (0.76–0.83). However, models based on single modality ceCT performed comparably well (0.74–0.81). T2W-only models performed acceptably but were somewhat inferior to the PET/ceCT and ceCT-based models.

Conclusion: CNNs provided high-quality automatic GTV delineations for both single and multimodality image input, indicating that deep learning may prove a versatile tool for target volume delineation in future patients with ASCC.

Keywords: anal cancer; radiotherapy; gross tumour volume; automatic delineation; deep learning.

Background

Anal squamous cell carcinoma (ASCC) is a rare malignancy with increasing global incidence [1], for which chemoradiotherapy is the standard curative treatment [2]. Modern radiotherapy (RT) techniques such as intensity-modulated RT (IMRT) delivers highly conformal dose distributions to the target volume (TV), ensuring high tumour control and reduced normal tissue doses with subsequent decrease in acute and late toxicities [2]. IMRT has been shown to be superior to three-dimensional (3D) conformal RT in reducing acute toxicity in patients with ASCC [3]. However, the reduced TV margins and steep dose gradients resulting from high precision RT require consistent and precise TV delineations [4, 5].

Manual TV delineation is a highly resource-demanding task and has been recognised as a main source of uncertainty within the RT workflow [4]. Inadequate TV definition can potentially lead to under-dosing of the tumour, thereby increasing relapse risk, or result in increased morbidity [5–7]. Guidelines, delineation atlases and quality assurance protocols decrease delineation variability [4–6, 8], yet interobserver variations remain an issue, with considerable delineation discrepancies occurring in approximately 10 % of RT plans [5].

The optimal imaging modalities for TV delineation remain unclear for many diagnoses [4, 5, 7], including ASCC [9]. Patients with ASCC may undergo several imaging procedures in conjunction with staging and RT planning, including computed tomography (CT), magnetic resonance imaging (MRI) and/or positron-emission tomography (PET) using 18F-fluorodeoxyglucose (FDG) [2, 10]. Pelvic MRI is

considered state-of-the-art imaging for ASCC and is used routinely for staging at many European centres, as it provides superior soft-tissue contrast resulting in detailed information on local and regional tumour extent [11]. PET has high sensitivity for detecting primary, regional and metastatic disease [12], and should be considered for RT planning according to U.S. guidelines [10]. TV delineation for ASCC is commonly performed in RT planning CT images. Ideally, the CT images should be co-registered with MR and/or PET images [13], but this practice varies. According to current guidelines [8] all available clinical and image information should be used when contouring gross tumour volumes (GTVs) for IMRT.

The more extensive use of multimodality image interpretation accompanying modern RT techniques as well as the requirement of highly accurate delineations has increased TV contouring time substantially [5, 7]. It is therefore of vital importance to develop and evaluate methods to improve the efficacy and accuracy of TV delineation. Deep convolutional neural networks (CNNs) have successfully been used for automatic tumour delineation in a range of diagnoses [14]. Using CNNs to support manual GTV delineations can reduce contouring time and delineation variability [15]. CNNs can further be used for systematic studies of the impact of imaging modality on automatic delineation quality by introducing single or multiple input channels to the network [16, 17].

Automated TV delineation for ASCC using deep learning has to our knowledge not yet been explored. However, the widespread use of IMRT and multimodality imaging, as well as the proximity of TVs and critical organs such as the small bowel and

bladder [2], suggest that auto-delineation could be advantageous for this group of patients. In this study, we investigated the use of a deep two-dimensional (2D) CNN for fully automatic GTV delineation of ASCC. As ASCC patients may undergo various imaging procedures, we conducted a comprehensive comparison of the effects single modality and multimodality combinations of CT, PET and MRI sequences have on the quality of CNN-generated auto-delineations.

Materials and methods

Patient cohort and imaging

ASCC patients scheduled for curative chemoradiotherapy at Oslo University Hospital between 2013–2016 and enrolled in the prospective ANCARAD observational study (NCT01937780) [18] were included. The study was approved by the Regional Ethics Committee, and all patients gave written informed consent. Patients with biopsy-proven ASCC and visible tumour on baseline PET, previously included for analysis in [19], were eligible (n = 93). Staging was performed according to the 7th edition AJCC tumour–node–metastasis (TNM) system [20]. RT was delivered with volumetric modulated arc therapy/IMRT (n = 62) or 3D conformal RT (n = 31) to doses of 54 or 58 Gy. Concomitant chemotherapy with one or two cycles of mitomycin C and 5-fluorouracil was given to most patients.

Patients without a RT planning contrast-enhanced CT (ceCT) scan acquired with Iomeron® contrast agent were excluded from the present study, resulting in 86 patients with a complete set of pre-treatment PET and planning ceCT images. PET was

performed with a non-enhanced low-dose CT (ldCT) scan. A subset of 36 patients had consented to a study-specific 3T MRI examination before treatment, with a dedicated protocol including T2- and diffusion-weighted sequences (T2W and DW, respectively). See Supplementary Table A1 for image acquisition details.

To properly assess the effect of MRI sequences on auto-delineation quality, both in combination and comparison with other modalities, the included patients were analysed as two separate datasets consisting of: (i) PET, ldCT and ceCT images (n = 86) and (ii) PET, ldCT, ceCT, T2W and DW images (n =36), hereinafter referred to as DS-86 and DS-36 respectively. Patient characteristics are given in Table 1.

Manual contouring

Contouring was performed in the Eclipse Treatment Planning System (Varian Medical Systems - Paolo Alto, CA, USA) by a trained oncologist (C.U.) supported when necessary by an experienced MR radiologist (B.H.) and/or a nuclear medicine specialist (E.H.). The GTV was manually delineated to include the visible tumour and the circumference of the anal canal and/or rectum when involved, as seen on axial images, according to current practice [8, 21]. GTV contours were defined in the planning ceCT image basis, supported by PET and MR T2W images, where T2W images were acquired with a standard (i.e. not study-specific) MRI protocol. Involved lymph nodes were not evaluated in the present study.

Image pre-processing

Images were imported to the MICE toolkit (NONPI Medical AB, Umeå, Sweden) and co-registered using rigid transformation and mutual information criteria [22]. The image matrices were linearly interpolated to the reference frame and resolution of the planning ceCT. A representative volume of interest (VOI), encompassing approximately the same pelvic region in the image series was selected for all patients (median in-plane matrix dimensions: 188×188; median number of image slices per patient: 41) during co-registration. This reduced memory consumption and computational burden throughout the analysis. The VOI was selected to be as large as possible, keeping the patient in the centre of axial image slices, but at the same time excluding irrelevant areas outside the patient.

The DW series was condensed into an apparent diffusion coefficient (ADC) map by regression analysis [23], using b-values of 200, 400, 600, 800 mm^2/s, to produce images reflecting water diffusion. The condensation into one single image series also made the impact of this MRI sequence comparable to the other single modality image series.

The study-specific DW sequence did not cover the entire tumour for all patients. This was also the case for some T2W image series. The affected image slices, mostly containing peripheral tumour regions located cranially or caudally, were also removed from the other image series in DS-36, reducing the number of slices in this dataset by about 14 %.

Both CT series were pre-processed using a narrow soft-tissue window, with centres and widths defined as the mode and two times the standard deviation, respectively, of the intensity values within manual GTV delineations. This resulted in ldCT and ceCT window settings ({centre, width}) of {32, 220} HU and {70, 300} HU, respectively.

After VOI-definition and elimination of image slices with insufficient MRI field-of-view from DS-36, approximately 50 % of the slices in the two datasets did not contain manual GTV delineations. To focus model training on delineation of the GTV, 80 % of the slices without manual delineations were randomly removed from the datasets. The resulting number of image slices in DS-86 and DS-36 are given in Supplementary Table A2.

CNN architecture and training

A 2D U-Net CNN architecture [24] with the Dice loss function [25] was implemented for GTV delineation in axial image slices using Python and TensorFlow (see Supplementary Appendix B for architecture details). The U-Net was preferred over more recent architectures as it is a mature network with documented strong performance, often surpassing specialised deep learning pipelines for a wide range of medical applications and diagnoses [26]. Due to the limited number of training samples and large number of input modalities and combinations thereof, as well as the FOV limitations of the MRI sequences, we used a 2D rather than a 3D architecture, thus reducing the number of trainable parameters and GPU-memory requirements.

Five-fold cross-validation was used to acquire a robust estimate of CNN performance. Each dataset was divided into five cross-validation folds (Supplementary Table A2) using stratified random sampling to ensure comparable GTV distributions across folds. The manual GTV delineations constituted the ground truth used for training and evaluation. Models were trained using the Adam optimiser with standard parameters and a learning rate of 10^{-4} [27]. The effect of imaging modality and modality combinations was assessed for both datasets by training separate single and multimodality models based on the available image input. In the case of multimodal models, each image modality was treated as a separate network input channel. All evaluated model inputs are specified in Figures 1 (DS-86) and 2 (DS-36).

Expanding the number of training images by adding modified versions of the existing image data (image augmentation) has been shown to increase CNN model performance for TV delineation [28]. To assess the impact of such increased input variance, all our experiments were run without and with image augmentation, consisting of randomised, excessive elastic image deformation [29] applied on a random subset of 50 % of the images during training.

Performance evaluation

CNN performance was first evaluated using the Dice similarity coefficient (Dice) [30], describing the spatial overlap between voxels belonging to the ground truth and the auto-delineations. Perfect and no overlap correspond to Dice scores of 1 and 0, respectively. Models were further assessed using the 95th percentile Hausdorff distance

(HD_{95}) [31] and the average and median Hausdorff surface distances (ASD and MSD) between the ground truth and automatic GTV contours. Performance metrics were calculated on a per patient basis, as defined in Supplementary Appendix C.

Statistical analysis

Effects of imaging modality and image augmentation on validation Dice were evaluated using the non-parametric Friedman test [32] followed by Nemenyi multiple pairwise comparisons [33] (PMCMRplus R package [34]) and the paired Wilcoxon signed-rank test in R, respectively. Statistical tests were two-sided with a significance level of 0.05.

Results

Cross-validation Dice scores of models based on different image inputs are shown in Figures 1 (DS-86) and 2 (DS-36). Several single and multimodality models had good overlap with the ground truth (mean Dice $\gtrsim$ 0.75–80). The highest mean Dice was obtained for PET/ceCT or ceCT (DS-86; Figure 1) and PET/ceCT, PET/ceCT/T2W or ceCT (DS-36; Figure 2). PET, ldCT and ADC resulted in the poorest overall Dice scores.

Overall validation Dice performances were somewhat poorer for DS-86 (Figure 1) than for DS-36 (Figure 2). This was explained by a higher number of difficult delineation cases in DS-86. The most problematic DS-86 patients had a substantial number of image slices without ground truth delineations, where the CNN falsely delineated a contour. The two patients where the PET, ldCT or PET/ldCT models failed

completely (Dice ~ 0–0.2) also had atypical PET uptake or an atypical shape of the ground truth GTV. Such difficult patients were not present in the DS-36 dataset, which contained fewer challenging delineation cases. This was in part attributed to the limited FOV of the DS-36 MRI series (see Materials and methods: *Image pre-processing*). The model rankings according to cross-validation Dice (Figures 1 and 2) were, however, consistent between DS-86 and DS-36, indicating agreement on the effect of the imaging modalities common for both sets.

Randomised image augmentation gave moderate increases in mean and median Dice for most models (Figures 1 and 2; Wilcoxon signed-rank test; DS-86: not significant; DS-36: $p < 0.001$). Furthermore, the DS-86 models based on PET/ceCT and ceCT obtained significantly better Dice than PET/ldCT, PET and ldCT-based models (Supplementary Table D1), whereas no significant difference was found between the PET/ceCT and ceCT-based models. Similar trends were detected for the smaller DS-36 set, where the PET/ceCT model obtained significantly better Dice than all other models except those based on PET/ceCT/T2W or ceCT (models with image augmentation; Supplementary Table D2). Furthermore, the ceCT model (with augmentation) obtained significantly better Dice than all other single-modality models, including T2W.

HD_{95}, ASD and MSD performances of selected single and multimodality models are given in Table 2 (see Supplementary Table E1 for all model results). The two models obtaining the highest mean and median Dice scores (PET/ceCT and ceCT) also gave the lowest distance-based metrics. For both datasets, the PET/ceCT and ceCT models resulted in median MSD smaller than or equal to the slice thickness of the

planning ceCT (2.50 mm). As for the Dice metric, the larger distance metrics of DS-86 can be explained by the higher incidence of difficult delineation cases. On a per patient basis, there was a significant negative correlation between Dice and each distance-based metric for the top-ranked DS-36 and DS-86 PET/ceCT models (Spearman's rank correlation; Supplementary Figure E1). For the DS-86 a small sub-group of patients with smaller and less distinct tumours (stage T1-T2) showed poor performance. No other tumour stage groupings were seen.

Example auto-delineations are shown in Figure 3 for models based on PET/ceCT, ceCT and T2W images, respectively (see also Supplementary Videos E1 and E2). All three models gave high-quality GTV auto-delineations for most image slices (Figure 3 (a)), but the T2W model generally had somewhat lower overlap with the ground truth (Figure 3 (b)). The models could result in poor auto-delineations for atypical image slices. Particularly very large ground truth GTVs, often including heterogeneous regions with up to several substantial air-filled zones were challenging (Figure 3 (c)). All models were also prone to delineating false positive regions encompassing the rectum or anal canal circumference in slices not containing a ground truth delineation (see Supplementary Videos E1 and E2 for examples). Though the PET/ceCT and ceCT models resulted in very similar auto-delineations, the PET/ceCT model generally gave the most refined auto-delineations for somewhat atypical GTVs (Figure 3 (d)).

Discussion

In this study, deep learning for automatic GTV contouring in patients with ASCC was evaluated for the first time. The 2D U-Net approach produced high-quality GTV delineations for a range of single or multimodality CT, PET and/or MR image inputs (mean Dice $\gtrsim$ 0.75–0.80). Our results suggest that deep learning tools could be useful for GTV contouring in ASCC patients, regardless of the exact imaging regimens.

Multimodality PET/ceCT images provided the highest mean agreement with the ground truth, but single modality ceCT models performed comparably well. Though the T2W-only models were somewhat inferior to ceCT-only models, the T2W-based model with image augmentation provided an acceptable mean Dice of 0.77. Combining all three modalities (PET/ceCT/T2W) did not increase DS-36 Dice performance, compared to PET/ceCT or ceCT-based models.

Using either PET/ceCT, ceCT or MRI for TV delineation could simplify the RT workflow, compared to using both MRI and ceCT or all three modalities. Both PET/ceCT and MRI-based auto-delineation is highly clinically relevant, as these modalities represent state-of-the-art practice in the U.S. and Europe. Single modality imaging with ceCT, on the other hand, is not considered state-of-the-art for ASCC diagnosis and staging [2, 10], but TV delineation based solely on RT planning ceCT images could be relevant in some clinical settings due to cost and efficiency benefits. MRI is the imaging modality of choice for visualisation of tumour and normal tissue anatomy in the pelvic region, and MRI-only RT planning could therefore be relevant for patients with ASCC. The feasibility of MRI-only RT planning using synthetic-CT (sCT)

for dose calculations was recently studied for anorectal cancer patients, where results indicate that conditional generative adversarial networks can produce clinically acceptable sCT images [37]. An MRI-only RT approach has certain potential advantages compared to using CT as primary and MRI as secondary RT planning modality, the most important being avoidance of geometric uncertainties introduced by registration of CT and MRI and the possible cost and efficiency benefits of avoiding the CT scan [38]. The latter could be particularly relevant within an adaptive RT workflow based on MR-only imaging [38], where auto-delineation may play a key role in making the required replanning feasible. Further exploration of MRI-based auto-delineation of both TVs and OARs in patients with ASCC is, therefore, warranted.

The lowest mean Dice was obtained for models based solely on PET or ADC. As the manual GTV delineation in this study, following clinical practice, included the anal canal and/or rectum circumference, the ground truth may include areas of non-affected tissues and air volumes. This influences the CNN learning process, particularly for models based solely on functional imaging (PET, ADC) which is generally more cancer specific than anatomical imaging with clearer boundaries between visible tumour tissue and surrounding regions included in the GTV. For example, FDG PET images have higher correlation with tumour specimens than CT and MR (T2W) images [35]. Moreover, manual tumour delineations based on ADC maps are significantly smaller than T2W-based delineations [36]. Hence, our inferior results for PET and ADC models could be related to the GTV definition, and our ranking of modalities may not translate to auto-delineation of visible tumour tissue only.

The limited number of training samples available in our datasets may have affected the generalisability and robustness of our models. Though the evaluated CNN generally performed well, the network failed on certain images. There are two main modes of failure, both likely related to the few representative training samples. First, the network delineates a false positive region encompassing the rectum and/or anal canal in many image slices without ground truth delineations. Second, the network struggles to correctly delineate atypical GTVs in general and particularly very large GTV contours, encompassing heterogeneous regions with multiple substantial air-filled areas, typical for patients with an anal abscess and/or fistula. Cross-institutional studies involving more patients, potentially using distributed learning for cross-centre training [39], or using transfer learning [40] based on other tumour sites could potentially alleviate some of these difficulties, thereby increasing model robustness and generalisability. A two-step procedure where image slices with tumour tissue are pre-selected, either manually or by a separate deep learning classification [41], could further limit the former failure mode. For ASCC patients presenting with abscess, fistula or other conditions where the tumour is not fully visible, special attention to manual revision of the auto-delineated GTV contours would still be recommended.

For the more frequently occurring pelvic cancers, cervical squamous cell carcinoma and rectal adenocarcinoma, CNNs for TV delineation have been explored but only for single modality PET [41], T2W [28, 43–47] or combined T2W/DW images [47]. Reported mean Dice scores for T2W-based CNN auto-delineation of the visible rectal or cervical tumour volume (0.72–0.84) [28, 43–47] are comparable to our T2W-

based results (0.75–0.77). Combining T2W and DW images for auto-delineation of the visible rectal tumour did not improve mean Dice (0.70 [48]), relative to the above T2W-only studies [28, 43–47], which is consistent with our results for T2W and ADC. U-Net delineation of cervical tumours in PET images [42] gave a mean Dice of 0.80, which is higher than our PET-only models (mean Dice: 0.65–0.76). This discrepancy is likely related to differences in TV definition. For auto-delineation of the rectum, previous studies obtained mean Dice values of 0.90–0.94 [28, 45] for T2W and 0.79 [49] for ceCT. The latter is comparable to our result for ceCT (0.74–0.81). Note that in [28] and [45] a very limited number of representative image slices (1–3) were selected from each patient, potentially affecting performance.

Few studies have evaluated the impact of imaging modality on interobserver variability for manual tumour delineations in ASCC [9, 50, 51], where only [9] included PET and ceCT images in addition to MRI sequences. Rusten et al. [9] investigated the interobserver and inter-modality variability for PET/ceCT vs. MRI/ceCT-based delineations of the GTV and the visible tumour tissue in a subset of the patients (n=19) included in our current study. The reported interobserver agreement was 0.80 and 0.74 (median Dice) for GTV delineations based on PET/ceCT and MRI/ceCT (including T2W and DW), respectively, which is somewhat lower than the median Dice obtained by our CNN approach for DS-36 (PET/ceCT: 0.82–0.85; T2W/ceCT: 0.80–0.82). Furthermore, inter-modality agreement between PET/ceCT and MRI/ceCT corresponded to a median Dice of 0.75 for the GTV [9], indicating good agreement between PET/ceCT and MRI/ceCT-based manual delineations. The latter is in

accordance with our current results, where PET/ceCT and T2W/ceCT models provided similar auto-delineations.

In conclusion, our proposed CNN approach provided high-quality fully automatic GTV delineations in ASCC patients, based on either single or multimodality images. The overlap between CNN-generated auto-delineations and the ground truth was comparable to interobserver agreement between experts performing manual GTV delineation in ASCC. This demonstrates the potential of CNN as a versatile tool for automatic TV delineation.

Funding

Eirik Malinen was supported by a grant from the Norwegian Cancer Society (grant no. 182672-2016) during the conduct of this study.

Disclosure statement

The authors have no conflicts of interest to declare.

Acknowledgements

We thank PhD Per-Ivar Lønne and MD Kathinka Schmidt Slørdahl for assistance with the patient characteristics data

References

[1] Islami F, Ferlay J, Lortet-Tieulent J, et al. International trends in anal cancer incidence rates. Int J Epidemiol. 2016;46(3):924–938. https://doi.org/10.1093/ije/dyw276.

[2] Rao S, Guren MG, Khan K, et al. Anal cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up☆. Ann Oncol. 2021;32(9):1087–1100. https://doi.org/10.1016/j.annonc.2021.06.015

[3] Kachnic LA, Winter K, Myerson R. RTOG 0529: a phase 2 evaluation of dose-painted intensity modulated radiation therapy in combination with 5-fluorouracil and mitomycin-C for the reduction of acute morbidity in carcinoma of the anal canal. Int J Radiat Oncol Biol Phys. 2013;86(1):27–33. https://doi.org/10.1016/j.ijrobp.2012.09.023.

[4] Segedin B, Petric P. Uncertainties in target volume delineation in radiotherapy - are they relevant and what can we do about them? Radiol Oncol. 2016;50(3):254–262. https://doi.org/10.1515/raon-2016-0023.

[5] Cox S, Cleves A, Clementel E, et al. Impact of deviations in target volume delineation - Time for a new RTQA approach? Radiother Oncol. 2019;137:1–8. https://doi.org/10.1016/j.radonc.2019.04.012.

[6] Chang ATY, Tan LT, Duke S, et al. Challenges for Quality Assurance of Target Volume Delineation in Clinical Trials. Front Oncol. 2017;7:221. https://doi.org/10.3389/fonc.2017.00221.

[7] Vinod SK, Jameson MG, Min M, et al. Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies. Radiother Oncol. 2016;121(2):169–179. https://doi.org/10.1016/j.radonc.2016.09.009.

[8] Ng M, Leong T, Chander S, et al. Australasian Gastrointestinal Trials Group (AGITG) Contouring Atlas and Planning Guidelines for Intensity-Modulated Radiotherapy in Anal Cancer. Int J Radiat Oncol Biol Phys. 2012;83(5):1455–1462. https://doi.org/10.1016/j.ijrobp.2011.12.058.

[9] Rusten E, Rekstad BL, Undseth C, et al. Target volume delineation of anal cancer based on magnetic resonance imaging or positron emission tomography. Radiat Oncol. 2017;12(1):147. https://doi.org/10.1186/s13014-017-0883-z.

[10] Benson AB, Venoo AP, Al-Hawary MM, et al. Anal Carcinoma, Version 2.2018, NCCN Clinical Practice Guidelines in Oncology. J Natl Compr Canc Netw. 2018; 16(7):852–871. https://doi.org/10.6004/jnccn.2018.0060.

[11] Goh V, Gollub FK, Liaw J, et al. Magnetic Resonance Imaging Assessment of Squamous Cell Carcinoma of the Anal Canal Before and After Chemoradiation:

Can MRI Predict for Eventual Clinical Outcome? Int J Radiat Oncol Biol Phys. 2010;78(3):715–721. https://doi.org/10.1016/j.ijrobp.2009.08.055.

[12] Jones M, Hruby G, Solomon M, et al. The Role of FDG-PET in the Initial Staging and Response Assessment of Anal Cancer: A Systematic Review and Meta-analysis. Ann Surg Oncol. 2015;22:3574–3581. https://doi.org/10.1245/s10434-015-4391-9.

[13] Glynne-Jones R, Tan D, Hughes R, et al. Squamous-cell carcinoma of the anus: progress in radiotherapy treatment. Nat Rev Clin Oncol. 2016;13:447–459. https://doi.org/10.1038/nrclinonc.2015.218.

[14] Cardenas CE, Yang J, Anderson BM, et al. Advances in Auto-Segmentation. Semin Radiat Oncol. 2019;29(3):185–197. https://doi.org/10.1016/j.semradonc.2019.02.001.

[15] Lin L, Dou Q, Jin YM, et al. Deep Learning for Automated Contouring of Primary Tumor Volumes by MRI for Nasopharyngeal Carcinoma. Radiology. 2019;291(3):677–686. https://doi.org/10.1148/radiol.2019182012.

[16] Guo Z, Li X, Huang H, et al. Deep Learning-Based Image Segmentation on Multimodal Medical Imaging. IEEE Trans Radiat Plasma Med Sci. 2019;3(2):162–169. https://doi.org/10.1109/TRPMS.2018.2890359.

[17] Guo Z, Guo N, Gong K, et al. Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multi-modality network. Phys Med Biol. 2019;64(20):205015. https://dx.doi.org/10.1088%2F1361-6560%2Fab440d.

[18] Slørdahl KS, Klotz D, Olsen JÅ, et al. Treatment outcomes and prognostic factors after chemoradiotherapy for anal cancer. Acta Oncol. 2021. https://doi.org/10.1080/0284186X.2021.1918763.

[19] Rusten E, Rekstad BL, Undseth C, et al. Anal cancer chemoradiotherapy outcome prediction using ^{18}F-fluorodeoxyglucose positron emission tomography and clinicopathological factors. Br J Radiol. 2019;92(1097):20181006. https://doi.org/10.1259/bjr.20181006.

[20] Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. Ann Surg Oncol. 2010;17(6):1471–1474. https://doi.org/10.1245/s10434-010-0985-4.

[21] Glynne-Jones R, Goh V, Aggarwal A, et al. Anal Carcinoma. In: Grosu AL, Nieder C, editors. Target Volume Definition in Radiation Oncology. Berlin: Springer; 2015. p. 193–218. https://doi.org/10.1007/978-3-662-45934-8.

[22] Maes F, Collignon A, Vandermeulen D, et al. Multimodality image registration by maximization of mutual information. IEEE Trans Med Imaging. 1997; 16(2):187–198. https://doi.org/10.1109/42.563664.

[23] Freiman M, Voss SD, Mulkern RV, et al. In vivo assessment of optimal b-value range for perfusion-insensitive apparent diffusion coefficient imaging. Med Phys. 2012;39(8):4832–4839. https://doi.org/10.1118/1.4736516.

[24] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells W, et al., editors. Medical Image Computing and Computer-Assisted Intervention – MICCA 2015. MICCAI 2015; 2015 Oct 5–9; Munich. Springer; 2015. p. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

[25] Milletari F, Mavab N, Ahmadi S. V-net: Fully convolutional neural networks for volumetric medical image segmentation. 2016 Fourth International Conference on 3D vision (3DV); 2016 Oct 25–28; Stanford, CA. IEEE; 2016. p. 565–571. https://doi.org/10.1109/3DV.2016.79.

[26] Isensee F, Jaeger PF, Kohl SAA, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18:203–211. https://doi.org/10.1038/s41592-020-01008-z.

[27] Kingma DP, Ba JL. ADAM: A Method for Stochastic Optimization. 3rd International Conference on Learning Representations (ICLR 2015); 2015 May 7–9; San Diego, CA.

[28] Lee J, Oh JE, Kim MJ, et al. Reducing the Model Variance of a Rectal Cancer Segmentation Network. IEEE Access. 2019;7:182725–182733. https://doi.org/10.1109/ACCESS.2019.2960371.

[29] Simard PY, Steinkraus D, Platt JC. Best practices for convolutional neural networks applied to visual document analysis. Proceedings of the Seventh International Conference on Document Analysis and Recognition; 2003 Aug 6; Edinburgh. IEEE; 2003. p. 958–63. https://doi.org/10.1109/ICDAR.2003.1227801.

[30] Dice, LR. Measures of the Amount of Ecologic Association Between Species. Ecology. 1945;26(3):297–302. https://doi.org/10.2307/1932409.

[31] Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing images using the Hausdorff distance. IEEE Trans Pattern Anal Mach Intell. 1993;15(9):850–863.

[32] Friedman M. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. J Am Stat Assoc. 1937;32:675–701. https://doi.org/10.1080/01621459.1937.10503522.

[33] Hollander M, Wolfe DA, Chicken E. Nonparametric Statistical Methods. 3rd ed. Hoboken (NJ): Wiley; 2014. https://doi.org/10.1002/9781119196037.

[34] Pohlert T. PMCMRplus: Calculate Pairwise Multiple Comparisons of Mean Rank Sums Extended [software]. R package version 1.9.0. 2021 [cited 2021 June 9]. Available from: https://CRAN.R-project.org/package=PMCMRplus.

[35] Buijsen J, van den Bogaard J, Janssen MHM, et al. FDG-PET provides the best correlation with the tumor specimen compared to MRI and CT in rectal cancer. Radiother Oncol. 2011;98(2):270–276. https://doi.org/10.1016/j.radonc.2010.11.018.

[36] Rosa C, Delli Pizzi D, Augurio A, et al. Volume Delineation in Cervical Cancer With T2 and Diffusion-weighted MRI: Agreement on Volumes Between Observers. In Vivo. 2020;34(4):1981–1986. https://dx.doi.org/10.21873%2Finvivo.11995.

[37] Bird D, Nix MG, McCallum H, et al. Multicentre, deep learning, synthetic-CT generation for ano-rectal MR-only radiotherapy treatment planning. Radiother Oncol. 2021;156:23–28. https://doi.org/10.1016/j.radonc.2020.11.027.

[38] Jonsson J, Nyholm T, Söderkvist K. The rationale for MR-only treatment planning for external radiotherapy. Clin Transl Radiat Oncol. 2019;18:60–65. https://doi.org/10.1016/j.ctro.2019.03.005.

[39] Choudhury A, Theophanous S, Lønne PI, et al. Predicting outcomes in anal cancer patients using multi-centre data and distributed learning - A proof-of-

concept study. Radiother Oncol. 2021;159:183–189. https://doi.org/10.1016/j.radonc.2021.03.013.

[40] Karimi D, Warfield SK, Gholipour A. Critical Assessment of Transfer Learning for Medical Image Segmentation with Fully Convolutional Neural Networks. arXiv:2006.00356v1 [Preprint]. 2020 [cited 2021 June 9]: [11 p.]. Available from: https://arxiv.org/abs/2006.00356v1.

[41] Zhu S, Dai Z, Wen N. Two-Stage Approach for Segmenting Gross Tumor Volume in Head and Neck Cancer with CT and PET Imaging. In: Andrearczyk V, Oreiller V, Depeursinge A, editors. Head and Neck Tumor Segmentation. HECKTOR 2020. Lecture Notes in Computer Science, vol 12603. Springer. Springer; 2021. p. 22–27. https://doi.org/10.1007/978-3-030-67194-5_2.

[42] Chen L, Shen C, Zhou Z. Automatic PET cervical tumor segmentation by combining deep learning and anatomic prior. Phys Med Biol. 2019;64(8):085019. https://doi.org/10.1088/1361-6560/ab0b64.

[43] Wang J, Lu J, Qin G, et al. Technical Note: A deep learning-based autosegmentation of rectal tumors in MR images. Med Phys. 2018;45(6):2560–2564. https://doi.org/10.1002/mp.12918.

[44] Wang M, Xie P, Ran Z, et al. Full convolutional network based multiple side-output fusion architecture for the segmentation of rectal tumors in magnetic resonance images: A multi-vendor study. Med Phys. 2019;46(6):2659–2668. https://doi.org/10.1002/mp.13541.

[45] Kim J, Oh JE, Lee J, et al. Rectal cancer: Toward fully automatic discrimination of T2 and T3 rectal cancers using deep convolutional neural network. Int J Imaging Syst Technol. 2019;29:247–259. https://doi.org/10.1002/ima.22311.

[46] Soomro MH, Coppotelli M, Conforto S, et al. Automated Segmentation of Colorectal Tumor in 3D MRI Using 3D Multiscale Densely Connected Convolutional Neural Network. J Healthc Eng. 2019;1075434. https://doi.org/10.1155/2019/1075434.

[47] Bnouni N, Islem R, Rhim MS, et al. Dynamic Multi-scale CNN Forest Learning for Automatic Cervical Cancer Segmentation. In: Shi Y, Suk HI, Liu M, editors. Machine Learning in Medical Imaging. MLMI 2018; 2018 Sept 28; Granada. Springer; 2018. p. 19–27. https://doi.org/10.1007/978-3-030-00919-9_3.

[48] Trebeschi S, van Griethuysen JJM, Lambregts DMJ, et al. Deep Learning for Fully-Automated Localization and Segmentation of Rectal Cancer on Multiparametric MR. Sci Rep. 2017; 7(5301). https://doi.org/10.1038/s41598-017-05728-9.

[49] Liu Z, Liu X, Xiao B, et al. Segmentation of organs-at-risk in cervical cancer CT images with a convolutional neural network. Phys Med. 2020;69:184–191. https://doi.org/10.1016/j.ejmp.2019.12.008.

[50] Prezzi D, Mandegaran R, Gourtsoyianni S, et al. The impact of MRI sequence on tumour staging and gross tumour volume delineation in squamous cell carcinoma of the anal canal. Eur Radiol. 2018;28:1512–1519. https://dx.doi.org/10.1007%2Fs00330-017-5133-0.

[51] Min LA, Vacher YJL, Dewit L, et al. Gross tumour volume delineation in anal cancer on T2-weighted and diffusion-weighted MRI – Reproducibility between radiologists and radiation oncologists and impact of reader experience level and DWI image quality. Radiother Oncol. 2020;150:81–88. https://doi.org/10.1016/j.radonc.2020.06.012.

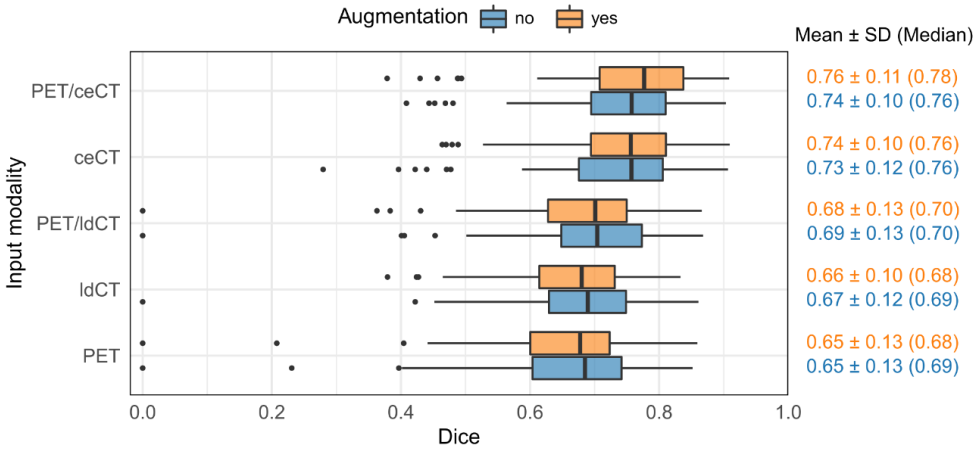


Figure 1. Box plots of the cross-validation Dice scores (DS-86 dataset). CNN models were trained on different modalities (y-axis) with or without image augmentation (orange vs. blue). Dice: Dice similarity coefficient; SD: standard deviation; PET: positron-emission tomography; ceCT: contrast-enhanced computed tomography; ldCT: low-dose computed tomography.

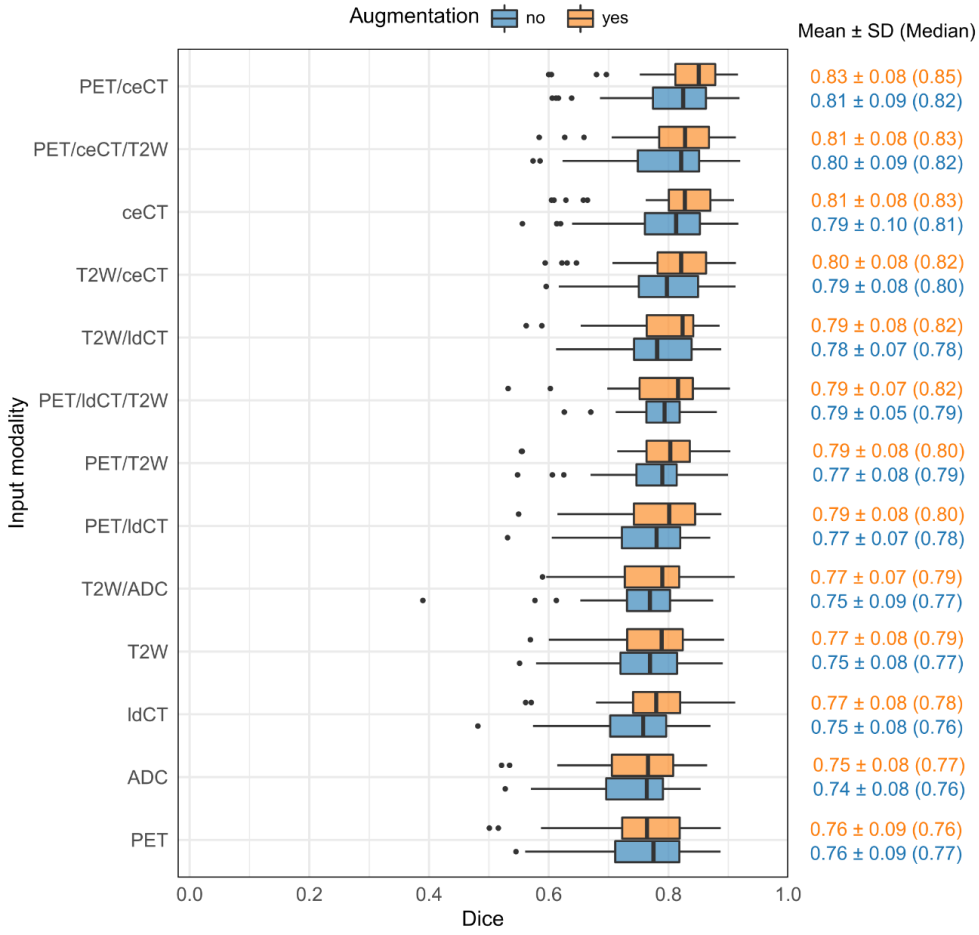


Figure 2. Box plots of the cross-validation Dice scores (DS-36 dataset). CNN models were trained on different modalities (y-axis) with or without image augmentation (orange vs. blue). Dice: Dice similarity coefficient; SD: standard deviation; PET: positron-emission tomography; ceCT: contrast-enhanced computed tomography; T2W: T2-weighted; ldCT: low-dose computed tomography; ADC: apparent diffusion coefficient.

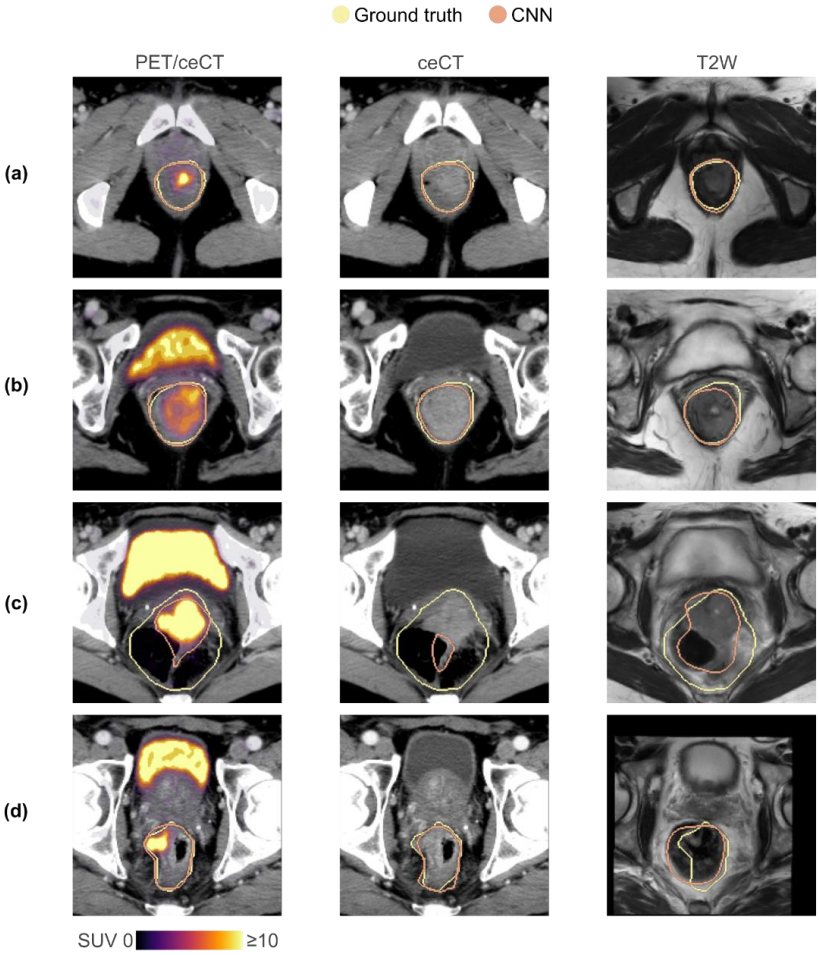


Figure 3. Manual ground truth (yellow) and CNN-generated (orange) GTV contours in representative image slices (a)–(d) for the DS-36 dataset. CNN-generated auto-delineations were based on (from left) PET/ceCT, ceCT and T2W images, respectively. These DS-36 auto-delineations were representative of the corresponding DS-86 results. CNN: convolutional neural network; GTV: gross tumour volume; PET: positron-emission tomography; ceCT: contrast-enhanced computed tomography; T2W: T2-weighted; SUV: standardised uptake value.

Table 1. Patient and tumour characteristics for the DS-86 and DS-36 datasets.

Characteristic[a]	**DS-86**[b] (n = 86)	**DS-36**[c] (n = 36)
Age [years]		
Mean (range)	62.0 (40.8–88.8)	60.9 (40.8–88.8)
Sex		
Female	66 (76.7 %)	29 (80.6 %)
Male	20 (23.3 %)	7 (19.4 %)
Tumour stage[d]		
T1	4 (4.7 %)	0 (0 %)
T2	41 (47.7 %)	20 (55.6 %)
T3	16 (18.6 %)	6 (16.7 %)
T4	25 (29.1 %)	10 (27.8 %)
Nodal stage[d]		
N0	40 (46.5 %)	15 (41.7 %)
N1	11 (12.8 %)	4 (11.1 %)
N2	18 (20.9 %)	10 (27.8 %)
N3	17 (19.8 %)	7 (19.4 %)
HPV status		
Positive	74 (86.1 %)	33 (91.7 %)
Negative	11 (12.8 %)	3 (8.3 %)
Unknown	1 (1.2 %)	0 (0 %)
GTV [cm^3]		
Mean (range)	59.2 (9.9–311.8)	49.7 (11.4–136.8)

[a]Percentages may not sum to exactly 100 due to rounding. [b]All 86 included patients having positron-emission tomography and contrast-enhanced computed tomography images available. [c]Subset of 36 patients that agreed to a study-specific 3T magnetic resonance imaging examination. [d]According to the 7th edition AJCC TNM system [20]. GTV: gross tumour volume; HPV: human papillomavirus.

Table 2. Distance-based performance metrics for selected single and multimodality models (with image augmentation).

Dataset	Input modality	HD_{95} [mm]		ASD [mm]		MSD [mm]	
		Mean ± SD	Median	Mean ± SD	Median	Mean ± SD	Median
DS-86	PET/ceCT	16.06 ± 8.88	13.73	3.31 ± 1.64	3.09	2.85 ± 2.34	2.50
	ceCT	19.15 ± 13.98	16.40	3.66 ± 1.75	3.27	3.14 ± 2.22	2.50
DS-36	PET/ceCT	7.07 ± 4.43	5.43	1.76 ± 1.04	1.58	1.30 ± 1.16	0.90
	PET/ceCT/T2W	8.34 ± 6.42	6.44	1.98 ± 1.13	1.75	1.54 ± 1.01	1.27
	ceCT	8.07 ± 4.43	7.50	1.89 ± 1.08	1.68	1.41 ± 1.24	1.27
	T2W	12.13 ± 15.59	8.30	2.57 ± 1.16	2.43	2.05 ± 1.09	1.80

PET: positron-emission tomography; ceCT: contrast-enhanced computed tomography; T2W: T2-weighted; HD_{95}: 95^{th} percentile Hausdorff distance; ASD: average surface distance; MSD: median surface distance; SD: standard deviation.

Supplemental material

Deep learning-based automatic delineation of anal cancer gross tumour volume: A multimodality comparison of CT, PET and MRI

Aurora Rosvoll Groendahl[a], Yngve Mardal Moe[a], Christine Kiran Kaushal[a], Bao Ngoc Huynh[a], Espen Rusten[b], Oliver Tomic[a], Eivor Hernes[c], Bettina Hanekamp[c], Christine Undseth[d], Marianne Grønlie Guren[d,e], Eirik Malinen[b,f], Cecilia Marie Futsaether[a].

[a]Faculty of Science and Technology, Norwegian University of Life Sciences, Ås, Norway; [b]Department of Medical Physics, Oslo University Hospital, Oslo, Norway; [c]Department of Radiology and Nuclear Medicine, Oslo University Hospital, Oslo, Norway; [d]Department of Oncology, Oslo University Hospital, Oslo, Norway; [e]Division of Cancer Medicine, Institute of Clinical Medicine, University of Oslo, Oslo, Norway; [f]Department of Physics, University of Oslo, Oslo, Norway.

Appendix A. Image data

Supplementary Table A1. Image acquisition details.

MRI	
Scanner	Philips Ingenia 3.0 T, Philips, Amsterdam, Netherlands
T2W	
Pulse sequence	TSE
Flip angle	90°
TR/TE	3712/80 ms
Matrix size	627 × 627
Voxel size (x – y – z)	0.36 × 0.36 × 5 mm^3
DW	
Pulse sequence	SE-EP single-shot
Flip angle	90°
TR/TE	6843/65 ms
B-values	0, 10, 20, 40, 80, 160, 200, 400, 800, 1 000, 1 200, 1 500 mm^2/s
Matrix size	320 × 320
Voxel size (x – y – z)	1.25 × 1.25 × 4 mm^3
PET/ldCT[a]	
Scanner	Biograph mCT 40, Siemens Medical Solutions, Erlangen, Germany
Patient positioning	Flat tabletop
PET	
Administered FDG activity (standard IV dose)	3 MBq/kg, ~ 60 min prior to imaging
Reconstruction algorithm	OSEM, 2 iterations, 21 subsets, including PSF-TOF
Matrix size	400 × 400
Voxel size (x – y – z)	2 × 2 × 3 mm^3
ldCT	
Reconstructed slice thickness	3 mm
Pixel size (range)	0.43 × 0.43 mm^2 – 0.98 × 0.98 mm^2
ceCT[b]	
Scanner	General Electric LightSpeed Pro 16 scanner GE Healthcare, Chicago, Illinois, USA
Contrast medium	Iomeron®
Matrix size	512 × 512
Voxel size (x – y – z)	0.9 × 0.9 × 2.5 mm^3

[a]PET with a non-enhanced ldCT scan used for attenuation correction and localisation. [b]ceCT for radiotherapy planning purposes. MRI: magnetic resonance imaging; T2W: T2-weighted; TSE: turbo spin-echo; TR: repetition time; TE: echo time; DW: diffusion-weighted; SE-EP: spin-echo echo-planar; PET: positron-emission tomography; FDG: 18F-fluorodeoxyglucose; IV: intravenous; OSEM: ordered subset expectations maximization; PSF-TOF: point spread function-time of flight; ldCT: low-dose computed tomography; ceCT: contrast-enhanced computed tomography.

Supplementary Table A2. Number of patients and image slices in the cross-validation folds of datasets DS-86 and DS-36.

	DS-86		DS-36	
	Patients	Image slices	Patients	Image slices
Fold 1	17	441	7	104
Fold 2	18	380	8	140
Fold 3	17	430	7	133
Fold 4	17	422	7	139
Fold 5	17	418	7	128
Total	86	2 091	36	644

Appendix B. CNN architecture and experiments

All experiments were run using deoxys (https://deoxys.readthedocs.io/en/latest/), our in-house-developed publicly available Python framework for running deep-learning experiments, with emphasis on target volume auto-delineation.

The CNN architecture used in the present study is outlined in Supplementary Table B1. As the volume of interest varied between patients, the image slices were padded with zeros to get a common in-plane matrix dimension of 236 × 236 before inputting them to the network. The convolutional (Conv) and transposed convolutional (ConvTransp) layers used a 3×3 convolution kernel and were followed by the ReLU activation function. Batch normalization was used after each Conv layer, and a bilinear interpolation was included after each ConvTransp layer. Dropout was applied to the Conv10 layer (end of the encoder path) with a keep probability of 0.5. Relevant code for the experiments, including the customised image augmentation scheme, is available from https://github.com/argrondahl/Acta-Oncologica-2021.

Supplementary Table B1. 2D U-Net convolutional neural network architecture.

Name	Inputs	Output shape	Activation Function	Batch Normalisation[a]
Conv1	Input	236 × 236 × 64	ReLU	Yes
Conv2	Conv1	236 × 236 × 64	ReLU	Yes
MaxPool1	Conv2	118 × 118 × 64	-	No
Conv3	MaxPool1	118 × 118 × 128	ReLU	Yes
Conv4	Conv3	118 × 118 × 128	ReLU	Yes
MaxPool2	Conv4	59 × 59 × 128	-	No
Conv5	MaxPool2	59 × 59 × 256	ReLU	Yes
Conv6	Conv5	59 × 59 × 256	ReLU	Yes
MaxPool3	Conv6	29 × 29 × 256	-	No
Conv7	MaxPool3	29 × 29 × 512	ReLU	Yes
Conv8	Conv7	29 × 29 × 512	ReLU	Yes
MaxPool4	Conv8	14 × 14 × 512	-	No
Conv9	MaxPool4	14 × 14 × 1024	ReLU	Yes
Conv10	Conv9	14 × 14 × 1024	ReLU	Yes
Dropout[b]	Conv10	14 × 14 × 1024	-	No
ConvTransp1	Dropout	29 × 29 × 512	ReLU	Yes
Conv11	ConvTransp1 & Conv8	29 × 29 × 512	ReLU	Yes
Conv12	Conv11	29 × 29 × 512	ReLU	Yes
ConvTransp2	Conv12	59 × 59 × 256	ReLU	Yes
Conv13	ConvTransp2 & Conv6	59 × 59 × 256	ReLU	Yes
Conv14	Conv11	59 × 59 × 256	ReLU	Yes
ConvTransp3	Conv10	118 × 118 × 128	ReLU	Yes
Conv15	ConvTransp3 & Conv4	118 × 118 × 128	ReLU	Yes
Conv16	Conv11	118 × 118 × 128	ReLU	Yes
ConvTransp4	Conv10	236 × 236 × 64	ReLU	Yes
Conv17	ConvTransp4 & Conv2	236 × 236 × 64	ReLU	Yes
Conv18	Conv11	236 × 236 × 64	ReLU	Yes
FinalConv	Conv18	236 × 236 × 1	Sigmoid	No

[a]Batch size of 16. [b]Keep probability of 0.5. Conv: convolutional layer; MaxPool: max pooling layer; ConvTransp: transposed convolutional layer; ReLU: rectified linear unit.

Appendix C. Performance metric calculations

The reported Dice similarity index was calculated per patient based on the true positive (TP), false positive (FP) and false negative (FN) voxels, respectively, summarised over all included image slices of the given patient:

$$\text{Dice} = \frac{2TP}{2TP + FP + FN} = \frac{2|P \cap G|}{|P| + |G|}, \quad \text{(A1)}$$

where $|P|$ and $|G|$ denotes the number of voxels in the patient's predicted auto-delineation and ground truth delineation, respectively, and $|P \cap G|$ is the number of voxels in the intersection between the prediction and ground truth.

Similarly, the distance-based metrics (95th percentile Hausdorff distance, average surface distance, median surface distance) were calculated over all the included image slices (constituting a 3D image stack) of each patient, using the deepmind Python library (https://github.com/deepmind/surface-distance).

Appendix D. Statistical analysis

Supplementary Table D1. *p* values from Nemenyi post-hoc pairwise comparisons of per patient Dice similarity coefficient for models based on different image modalities with (without) the inclusion of image augmentation for DS-86.

	ceCT	ldCT	PET	PET/ceCT
ldCT	**<0.0001**			
	(<0.0001)			
PET	**<0.0001**	0.999		
	(<0.0001)	(0.899)		
PET/ceCT	0.689	**<0.0001**	**<0.0001**	
	(0.277)	**(<0.0001)**	**(<0.0001)**	
PET/ldCT	**<0.0001**	0.053	**0.047**	**<0.0001**
	(0.034)	**(0.047)**	**(0.002)**	**(<0.0001)**

Statistically significant *p* values in bold.
Friedman test for models with image augmentation: $p = 3.46\times10^{-30}$.
Friedman test for models without image augmentation: $p = 1.12\times10^{-21}$.
PET: positron-emission tomography; ceCT: contrast-enhanced computed tomography; ldCT: low-dose computed tomography.

Supplementary Table D2. *p* values from Nemenyi post-hoc pairwise comparisons of per patient Dice similarity coefficient for models based on different image modalities with (without) the inclusion of image augmentation for DS-36.

	ADC	ceCT	ldCT	PET	PET/ceCT	PET/ ceCT/ T2W	PET/ ldCT/	PET/ ldCT/ T2W	PET/ T2W	T2W	T2W/ ADC	T2W/ ceCT
ceCT	**<0.0001**											
	(0.021)											
ldCT	0.999	**0.003**										
	(1.00)	**(0.029)**										
PET	0.999	**0.001**	1.00									
	(0.919)	(0.773)	(0.946)									
PET/ceCT	**<0.0001**	0.971	**<0.0001**	**<0.0001**								
	(<0.0001)	(0.960)	**(<0.0001)**	**(0.035)**								
PET/ceCT/T2W	**<0.001**	0.999	**0.008**	**0.003**	0.908							
	(0.004)	(0.999)	**(0.006)**	(0.472)	(0.997)							
PET/ldCT	0.584	0.224	0.986	0.938	**0.002**	0.366						
	(0.857)	(0.857)	(0.897)	(1.00)	(0.057)	(0.584)						
PET/ldCT/T2W	0.209	0.606	0.809	0.629	**0.019**	0.773	0.999					
	(0.089)	(0.999)	(0.114)	(0.960)	(0.773)	(0.999)	(0.983)					
PET/T2W	0.366	0.407	0.929	0.809	**0.007**	0.584	1.00	1.00				
	(0.792)	(0.908)	(0.842)	(1.00)	(0.082)	(0.672)	(1.00)	(0.992)				
T2W	0.999	**0.001**	1.00	1.00	**<0.0001**	**0.003**	0.946	0.651	0.826			
	(0.998)	(0.346)	(0.999)	(0.999)	**(0.003)**	(0.134)	(0.999)	(0.672)	(0.999)			
T2W/ADC	0.999	**<0.001**	1.00	1.00	**<0.0001**	**0.002**	0.929	0.606	0.792	1.00		
	(0.998)	(0.346)	(0.999)	(0.999)	**(0.003)**	(0.134)	(0.999)	(0.672)	(0.999)	(1.00)		
T2W/ceCT	**<0.001**	0.999	**0.032**	**0.012**	0.714	0.999	0.629	0.938	0.826	**0.013**	**0.011**	
	(0.029)	(1.00)	**(0.039)**	(0.826)	(0.938)	(0.999)	(0.897)	(1.00)	(0.938)	(0.407)	(0.407)	
T2W/ldCT	0.256	0.539	0.857	0.693	**0.013**	0.714	0.999	1.00	1.00	0.714	0.672	0.908
	(0.714)	(0.946)	(0.773)	(0.999)	(0.114)	(0.754)	(1.00)	(0.997)	(1.00)	(0.998)	(0.998)	(0.965)

Statistically significant *p* values in bold.

Friedman test for models with image augmentation: $p = 1.50\times10^{-15}$.

Friedman test for models without image augmentation: $p = 8.87\times10^{-8}$.

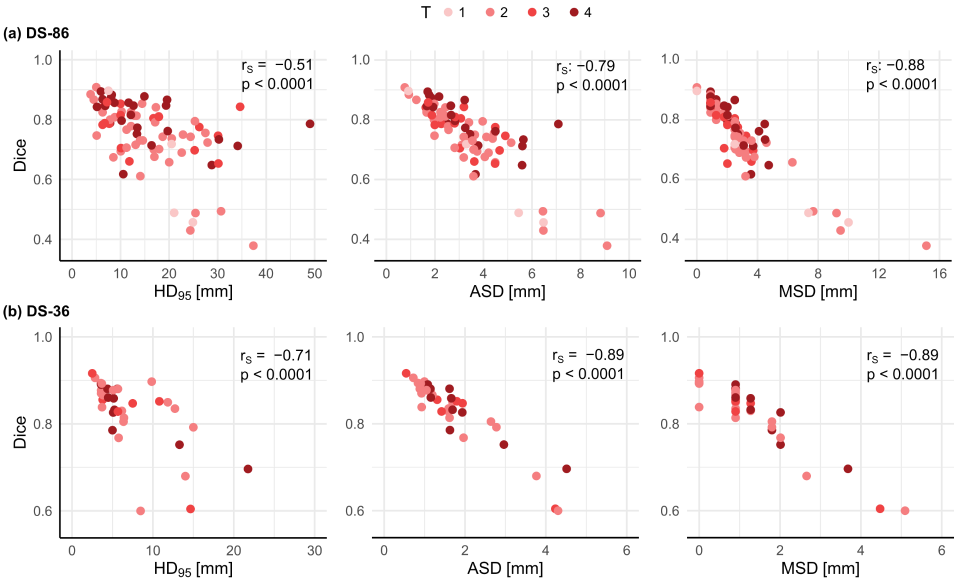
PET: positron-emission tomography; ceCT: contrast-enhanced computed tomography; ldCT: low-dose computed tomography; T2W: T2-weighted; ADC: apparent diffusion coefficient.

Appendix E. Auto-delineation performance

Supplementary Table E1. Distance-based performance metrics for all DS-86 and DS-36 models with image augmentation (in bold: the two models of each dataset displaying the best distance metrics).

Dataset	Input modality	HD_{95} [mm]		ASD [mm]		MSD [mm]	
		Mean ± SD	Median	Mean ± SD	Median	Mean ± SD	Median
DS-86	**PET/ceCT**	**16.06 ± 8.88**	**13.73**	**3.31 ± 1.64**	**3.09**	**2.85 ± 2.34**	**2.50**
	ceCT	**19.15 ± 13.98**	**16.40**	**3.66 ± 1.75**	**3.27**	**3.14 ± 2.22**	**2.50**
	PET/ldCT	19.23 ± 14.52	15.66	4.30 ± 2.07	3.76	4.06 ± 2.59	3.23
	ldCT	20.01 ± 13.70	17.23	4.54 ± 1.50	4.41	4.46 ± 2.22	3.75
	PET	24.49 ± 13.27	23.08	4.98 ± 2.15	4.51	4.77 ± 2.84	4.02
DS-36	**PET/ceCT**	**7.07 ± 4.43**	**5.43**	**1.76 ± 1.04**	**1.58**	**1.30 ± 1.16**	**0.90**
	PET/ceCT/T2W	8.34 ± 6.42	6.44	1.98 ± 1.13	1.75	1.54 ± 1.01	1.27
	ceCT	**8.07 ± 4.43**	**7.50**	**1.89 ± 1.08**	**1.68**	**1.41 ± 1.24**	**1.27**
	T2W/ceCT	9.42 ± 7.27	5.44	2.12 ± 1.19	1.71	1.68 ± 1.07	1.27
	T2W/ldCT	9.26 ± 8.13	7.01	2.19 ± 0.99	1.92	1.80 ± 1.13	1.80
	PET/ldCT/T2W	9.91 ± 12.21	6.87	2.19 ± 1.06	1.85	1.82 ± 1.12	1.80
	PET/T2W	8.80 ± 5.49	7.23	2.22 ± 0.98	1.92	1.84 ± 1.15	1.80
	PET/ldCT	7.93 ± 3.27	7.86	2.19 ± 0.91	2.05	1.81 ± 1.05	1.80
	T2W/ADC	8.87 ± 3.74	8.58	2.39 ± 0.87	2.34	1.99 ± 1.03	1.80
	T2W	12.13 ± 15.59	8.30	2.57 ± 1.16	2.43	2.05 ± 1.09	1.80
	ldCT	11.79 ± 9.86	9.87	2.60 ± 1.07	2.54	2.10 ± 1.18	2.50
	ADC	12.07 ± 12.03	9.76	2.92 ± 2.00	2.48	2.48 ± 1.95	1.91
	PET	10.73 ± 8.36	7.55	2.55 ± 1.44	2.22	2.23 ± 1.13	2.01

Two patients with no predicted gross tumour volumes were excluded from the affected calculations (pertain to DS-86 PET/ldCT, ldCT and PET-based models).
HD95: 95th percentile Hausdorff distance; ASD: average surface distance; MSD: median surface distance; SD: standard deviation. PET: positron-emission tomography; ceCT: contrast-enhanced computed tomography; ldCT: low-dose computed tomography; T2W: T2-weighted; ADC: apparent diffusion coefficient.



Supplementary Figure E1. Per patient Dice plotted against the different distance-based performance metrics for the best DS-86 (a) and DS-36 (b) models (both based on PET/ceCT). HD_{95}: 95th percentile Hausdorff distance; ASD: average surface distance; MSD: median surface distance; r_s: Spearman rank correlation; T: tumour stage; PET: positron-emission tomography; ceCT: contrast-enhanced computed tomography.

Appendix D

Paper IV

Automatic gross tumor segmentation of canine head and neck cancer using deep learning and cross-species transfer learning

Aurora Rosvoll Groendahl[1], Bao Ngoc Huynh[1], Oliver Tomic[1], Åste Søvik[2a], Einar Dale[3], Eirik Malinen[4,5], Hege Kippenes Skogmo[2], Cecilia Marie Futsaether[1*]

[1]Faculty of Science and Technology, Department of Physics, Norwegian University of Life Sciences, Ås, Norway

[2]Faculty of Veterinary Science, Department of Companion Animal Clinical Sciences, Norwegian University of Life Sciences, Ås, Norway.

[3]Department of Oncology, Oslo University Hospital, Oslo, Norway

[4]Department of Physics, University of Oslo, Oslo, Norway

[5]Department of Medical Physics, Oslo University Hospital, Oslo, Norway

*** Correspondence:**
Cecilia Marie Futsaether
cecilia.futsaether@nmbu.no

Keywords: canine head and neck cancer, dogs, radiotherapy, automatic target volume segmentation, gross tumor volume, artificial intelligence, deep learning, cross-species transfer learning.

Abstract

Background: Radiotherapy (RT) is increasingly being used on dogs with spontaneous head and neck cancer (HNC), which account for a large percentage of veterinary patients treated with RT. Accurate definition of the gross tumor volume (GTV) is a vital part of RT planning, ensuring adequate dose coverage of the tumor while limiting the radiation dose to surrounding tissues. Currently the GTV is contoured manually in medical images, which is a time-consuming and challenging task.

Purpose: The purpose of this study was to evaluate the applicability of deep learning-based automatic segmentation of the GTV in canine patients with HNC.

Materials and methods: Contrast-enhanced computed tomography (CT) images and corresponding manual GTV contours of 36 canine HNC patients and 197 human HNC patients were included. A 3D U-Net convolutional neural network (CNN) was trained to automatically segment the GTV in canine

[a] Current affiliation: Under Pelsen AS, Ås, Norway.

patients using two main approaches**:** (i) training models from scratch based solely on canine CT images, and (ii) using cross-species transfer learning where models were pretrained on CT images of human patients and then fine-tuned on CT images of canine patients. For the canine patients, automatic segmentations were assessed using the Dice similarity coefficient (*Dice*), the positive predictive value, the true positive rate, and surface distance metrics, calculated from a four-fold cross-validation strategy where each fold was used as a validation set and test set once in independent model runs.

Results: CNN models trained from scratch on canine data or by using transfer learning obtained mean test set *Dice* scores of 0.55 and 0.52, respectively, indicating acceptable auto-segmentations, similar to the mean *Dice* performances reported for CT-based automatic segmentation in human HNC studies. Automatic segmentation of nasal cavity tumors appeared particularly promising, resulting in mean test set *Dice* scores of 0.69 for both approaches.

Conclusion: In conclusion, deep learning-based automatic segmentation of the GTV using CNN models based on canine data only or a cross-species transfer learning approach shows promise for future application in RT of canine HNC patients.

1 Introduction

Head and neck cancer (HNC) is a heterogeneous group of malignant neoplasms originating from the different anatomical sites of the upper aerodigestive tract [1] and is relatively frequent in both humans and dogs. For humans, HNC is the seventh leading cancer by incidence worldwide [2], of which 90 % are squamous cell carcinomas (SCCs) of the oral cavity, oropharynx, hypopharynx, and larynx [3]. The incidence rate of HNC in dogs is similar to that of humans, but canine HNC patients present a greater variety of cancer subtypes and SCCs are less predominant than in humans [4-6]. For the same cancer subtypes, however, dogs with spontaneous tumors have been used as a comparative species in cancer research, taking advantage of the relative similarity of tumor biology and anatomic size between human and canine patients [7-9].

In humans, the main curative treatment modalities for HNC are surgery, radiotherapy (RT), chemotherapy, or a combination of these. Treatment decisions are typically based on primary tumor site and stage. However, most human HNC patients receive RT as an integral part of treatment [1]. At present, the most frequently used RT technique for HNC in humans is intensity-modulated RT (IMRT) [1]. IMRT is a high-precision technique, offering highly conformal radiation doses to the target and improved sparing of surrounding critical normal tissue structures, known as organs at risk (OARs), compared to conventional and 3-dimensional (3D) conformal RT [10-12]. These advantages are highly relevant for the treatment of HNC due to the complex anatomy of the head and neck region with immediate proximity between irradiated target volumes (TVs) and OARs.

In dogs, surgery is the primary treatment for most HNCs, but RT is indicated as the primary treatment for sinonasal tumors where full surgical resection is challenging [4, 13, 14]. Multimodal treatment with surgery, RT and chemotherapy may also be considered for canine HNC patients, particularly for cancers with significant risk of metastatic spread [13]. Though veterinary RT facilities are small in size and number compared to human facilities, RT has increasingly become available for veterinary patients [15]. Tumors of the head and neck in dogs and cats account for a large percentage of the neoplasms treated with RT in veterinary patients [15, 16]. Recently, more precise RT techniques such as image guided RT and IMRT have also been used for many patients in veterinary medicine [14, 17].

Accurate definition of TVs and OARs is required for successful high precision RT [18], regardless of species. Tumor and/or organ volume contours can also be required for extraction of quantitative image-based features used in radiomics studies [19], where the primary aim is to identify imaging biomarkers. In clinical practice, TV and OAR definition is typically performed manually by clinical experts who contour the given structures on axial anatomical images, usually RT planning computed tomography (CT) images, using functional images as support if available. Manual contouring is, however, inherently subject to intra- and interobserver variability, introducing

significant geometric uncertainties in RT planning and delivery [18]. Inaccurate contour definitions can severely affect treatment outcome, potentially leading to underdosing of TVs and associated increased risk of locoregional failure or too high dose to normal tissues and subsequent increased RT toxicity [20-22]. Furthermore, manual contouring is time and labor-intensive, particularly for HNC where the complexity and number of structures are considerable [23].

Recognizing the limitations of manual contouring, various automatic segmentation (auto-segmentation) methods and their potential application in the RT planning workflow have received significant attention. Over the past decade, deep learning methods have rapidly gained a central position within medical image analysis, particularly for semantic segmentation tasks such as contouring of RT structures. Many studies have shown that deep learning with convolutional neural networks (CNNs) can provide highly accurate auto-segmentations in human subjects, surpassing alternative segmentation methods [24-30]. Moreover, the use of CNNs to guide manual contouring can decrease both contouring time and inter-observer variability [31, 32]. Several studies have evaluated the use of CNNs for segmentation of the gross tumor volume (GTV) or OARs in human HNC subjects, achieving high-quality segmentations based on RT planning CT, positron emission tomography (PET) and/or magnetic resonance (MR) images [29-31, 33-41]. Even though there is increased focus on various deep learning applications in veterinary medicine, as exemplified by [42-46], few studies have evaluated the use of CNNs for semantic segmentation tasks in veterinary patients [47-49]. Only two studies [48, 49] have focused on RT structures. Park et al. [48] used CNNs to contour various OARs in canine HNC patients (n = 90) based on CT images, obtaining similar segmentation performance as reported for humans, whereas Schmid et al. [49] applied CNNs to contour the medial retropharyngeal lymph nodes in CT images of canine HNC patients (n = 40) obtaining acceptable performance. Auto-segmentation of the GTV or any other TV has, to the best of our knowledge, not previously been explored for veterinary patients including dogs. Given the increased use of RT for canine HNC patients, it is highly warranted to investigate the applicability of automatic GTV segmentation in this group of patients.

One challenge for machine learning (in general) and deep learning (in particular) in the medical domain is that the number of available samples is often limited. Supervised CNN algorithms generally require large, labelled training sets. As the contouring process is laborious and must be done by a clinical expert to ensure satisfactory contour quality, it might not be feasible to label numerous images if this is not done prospectively at the time of treatment. Moreover, in the case of relatively rare diseases the number of available subjects will be low. Transfer learning has been proposed as a strategy to tackle limited training data [50].

The essence of transfer learning is to apply knowledge gained from solving one problem, referred to as the source problem, to solving a novel, separate problem, referred to as the target

problem [50, 51]. This approach has also been applied to deep learning-based medical segmentation tasks (for a summary, see [52]). In veterinary science, transfer learning has been used successfully to segment acutely injured lungs in a limited number of CT images of dogs, pigs and sheep using a CNN model pretrained on a larger number of CT images of humans [47]. These findings suggest that cross-species transfer learning from humans to dogs could potentially be used to increase the performance of other segmentation tasks such as GTV segmentation, particularly when the number of canine subjects is low [50].

The objective of the present study was to evaluate the applicability of CNNs for fully automatic segmentation of the GTV in canine HNC based on CT images. In addition, the impact of transfer learning from a larger cohort of human HNC patients on auto-segmentation performance was investigated. Two main approaches to model training were assessed: (i) training CNN models from scratch based solely on CT images of canine patients (n = 36), and (ii) using a transfer learning approach where CNN models were pretrained on CT images of human HNC patients (n = 197) and subsequently fine-tuned on CT images of canine patients. These two approaches were compared to a reference approach (iii) where CNN models trained solely on human data were applied directly to canine data, without transfer learning.

2 Materials and Methods

In the present work, two different datasets consisting of contrast-enhanced CT images of canine and human patients, referred to as the canine and human datasets, respectively, were used to train CNN auto-segmentation models. Characteristics of the patients in the canine and human datasets can be found in Table 1 and Table 2, respectively. CT imaging and reconstruction parameters are summarized in Table 3.

2.1 Patients and imaging

Canine dataset

The canine data was collected retrospectively by reviewing the imaging database and the patient record system of the University Animal Hospital at the Norwegian University of Life Sciences (NMBU). Potential patients were identified by searching the imaging database over the years 2004–2019, resulting in 1 304 small animal cases that were reviewed using the following inclusion criteria: canine patients with confirmed malignant neoplasia of the head or cervical region with a complete imaging examination including contrast-enhanced CT. A total of 36 canine cases met these criteria and were included in the canine HNC dataset. As these data were generated as part of routine patient workup, approval from the animal welfare committee was not required. Baseline CT imaging was performed pre and 1 min post intravenous contrast administration, using a GE BrightSpeed S CT scanner (GE Healthcare, Chicago, Illinois, USA). The animals were scanned in sternal recumbency under general anesthesia.

Human dataset

The human data used in this study was obtained from a retrospective study of HNC patients with SCC of the oral cavity, oropharynx, hypopharynx, and larynx, treated with curative radio(chemo)therapy at Oslo University Hospital between 2007 and 2013 [53]. The study was approved by The Regional Ethics Committee and the Institutional Review Board. 18F-fluorodeoxyglucose (FDG) PET/CT imaging was performed at baseline on a Siemens Biograph 16 (Siemens Healthineers GmbH, Erlangen, Germany) with a RT compatible flat table and RT fixation mask. Only the RT planning contrast-enhanced CT images were included in our present work and patients who did not receive contrast agent were excluded from the analysis, resulting in a dataset of 197 patients. This set of patients has previously been described and analyzed in two separate auto-segmentation studies [29, 34]. Further details on the FDG PET/CT imaging protocol can be found in [29].

Table 1. Patient characteristics of the canine dataset.

Characteristic[a]	All patients (*n* = 36)	Fold 1 (*n* = 9)	Fold 2 (*n* = 9)	Fold 3 (*n* = 9)	Fold 4 (*n* = 9)
Age [years]					
Mean (range)	7.7 (1.1–13.6)	8.3 (4.6–13.6)	7.9 (1.1–11.1)	8.2 (4.4–10.1)	6.2 (2.0–10.0)
Sex					
Female	13 (36 %)	2 (22 %)	4 (44 %)	4 (44 %)	3 (33 %)
Male	23 (64 %)	7 (78 %)	5 (56 %)	5 (55 %)	6 (67 %)
Weight [kg]					
Mean (range)	32.1 (8.2–74.5)	26.9 (13.0–38.9)	35.3 (8.8–74.5)	30.9 (8.2–45.4)	35.4 (14.8–52.0)
Tumor site					
Oral cavity	5 (14 %)	3 (33 %)	1 (11 %)	0 (0 %)	1 (11 %)
Nasal cavity	14 (39 %)	2 (22 %)	5 (56 %)	5 (56 %)	2 (22 %)
Nasopharynx	1 (3 %)	0 (0 %)	0 (0 %)	0 (0 %)	1 (11 %)
Other	16 (44 %)	4 (44 %)	3 (33 %)	4 (44 %)	5 (56 %)
Nodal status					
Node involvement	4 (11 %)	2 (22 %)	1 (11 %)	1 (11 %)	0 (0 %)
GTV-T [cm^3]					
Mean (range)	69.7 (4.5–358.7)	50.0 (8.8–123.9)	57.0 (4.5–195.0)	48.3 (8.5–91.5)	123.4 (12.4–358.7)
GTV-N [cm^3]					
Mean (range)	9.8 (0.002–38.1)	0.5 (0.03–1.1)	38.1 (NA)	0.002 (NA)	NA (NA)

[a] Percentages may not sum to exactly 100 due to rounding.

GTV-T: gross primary tumor volume; GTV-N: involved nodal volume (for patients with node involvement); NA: not applicable.

Table 2. Patient characteristics of the human dataset.

Characteristic[a]	All patients (*n* = 197)	Training set (*n* = 126)	Validation set (*n* = 31)	Test set (*n* = 40)
Age [years]				
Mean (range)	60.3 (39.9–79.1)	60.5 (39.9–78.9)	60.7 (48.4–79.1)	59.4 (43.0–77.0)
Sex				
Female	49 (25 %)	28 (22 %)	10 (32 %)	11 (28 %)
Male	148 (75 %)	98 (78 %)	21 (68 %)	29 (72 %)
Tumor stage[b]				
T1/T2	96 (49 %)	61 (48 %)	15 (48 %)	20 (50 %)
T3/T4	101 (51 %)	65 (52 %)	16 (52 %)	20 (50 %)
Nodal stage[b]				
N0	47 (24 %)	29 (23 %)	8 (26 %)	10 (25 %)
N1	23 (12 %)	15 (12 %)	4 (13 %)	4 (10 %)
N2	120 (61 %)	78 (62 %)	17 (55 %)	25 (62 %)
N3	7 (4 %)	4 (3 %)	2 (6 %)	1 (3 %)
Tumor site				
Oral cavity	17 (9 %)	10 (8 %)	4 (13 %)	3 (7 %)
Oropharynx	143 (73 %)	91 (72 %)	22 (71 %)	30 (75 %)
Hypopharynx	16 (8 %)	12 (10 %)	3 (10 %)	1 (3 %)
Larynx	21 (11 %)	13 (10 %)	2 (6 %)	6 (15 %)
GTV-T [cm^3]				
Mean (range)	25.0 (0.8–285.0)	26.0 (0.8–285.0)	21.8 (0.8–78.2)	24.3 (1.4–157.6)
GTV-N [cm^3]				
Mean (range)	24.3 (0.5–276.7)	27.5 (0.5–276.7)	17.7 (0.9–77.8)	19.5 (0.5–76.4)

[a] Percentages may not sum to exactly 100 due to rounding.

[b] Staging according to the 7th edition AJCC/UICC tumor-node-metastasis system.

GTV-T: gross primary tumor volume; GTV-N: involved nodal volume (for patients with nodal stage ≥ N1)

Table 3. CT imaging and reconstruction parameters.

Human dataset (n = 197)	
Scanner	Siemens Biograph 16, Siemens Healthineers GmbH, Erlangen, Germany
Scan mode	Helical (rotation time 0.5 s, pitch 0.75)
Peak tube voltage	120 kV
Reconstructed slice thickness	2.00 mm
Reconstruction kernel	B30f/B30s
Matrix size	512 × 512
Pixel size	0.98 × 0.98 mm^2 (n = 161)
	1.37 × 1. 37 mm^2 (n = 30)
	0.89 × 0.89 mm^2 (n = 2)
	0.96 × 0.96 mm^2 (n = 1)
	0.92 × 0.92 mm^2 (n = 1)
	0.88 × 0.88 mm^2 (n = 1)
	0.82 × 0.82 mm^2 (n = 1)
Contrast agent	Visipaque 320 mg iodine/mL
Canine dataset (n = 36)	
Scanner	GE BrightSpeed S, GE Healthcare, Chicago, Illinois, USA
Scan mode	Helical (rotation time 1.0 s, pitch 0.75)
Peak tube voltage	120 kV
Reconstructed slice thickness	1.25 mm (n = 3)
	2.00 mm (n = 3)
	2.50 mm (n = 24)
	3.00 mm (n = 4)
	3.75 mm (n = 2)
Reconstruction kernel	Standard
Matrix size	512 × 512
Pixel size (range)	0.22 × 0.22 mm^2 – 0.49 × 0.49 mm^2
Contrast agent	Omnipaque 300 mg iodine/mL

2.2 Manual GTV contouring

Manual GTV contours were used as the ground truth for training and evaluation of auto-segmentation models. For both datasets, manual contouring was performed in axial image slices and the GTV was defined to encompass the gross primary tumor volume (GTV-T) and any involved nodal volume (GTV-N) if present.

For the human patients, the GTV was contoured prospectively in the treatment planning system at the time of initial RT planning and in accordance with the previous DAHANCA Radiotherapy Guidelines [54]. Contouring was based on FDG PET and contrast-enhanced CT images. First, the GTV was contoured by an experienced nuclear medicine physician based on FDG PET findings. Next, one or two oncology residents refined the delineations based on contrast-enhanced CT images and clinical information. Finally, the delineations were quality assured by a senior oncologist.

Contouring of the canine GTVs was performed retrospectively by a board-certified veterinary radiologist (H.K.S.) with radiation oncology residency training. Contours were defined based on contrast-enhanced CT images using the 3D Slicer software (https://www.slicer.org) [55]. The resulting delineations were smoothed in 3D Slicer using an in-plane median filter (5×5 kernel) before further image pre-processing. This was done to minimize the differences between the canine GTVs and the human GTVs, as the latter were smoothed by default in the hospital treatment planning system.

2.3 Image pre-processing

All CT images and corresponding manual GTV delineations were resampled to an isotropic voxel size of $1.0 \times 1.0 \times 1.0$ mm^3 to achieve a consistent voxel size and retain the actual anatomical size ratio between patients/species. Details regarding the resampling of the human HNC dataset can be found in [29]. All other image pre-processing was performed using Python and SimpleITK [56].

The images of the human dataset were first cropped to a volume of interest (VOI) of size $191 \times 265 \times 173$ mm^3, defined to encompass the head and neck region. Subsequently, the canine images were cropped and/or padded symmetrically about each axis to obtain the same image dimensions as the above VOI while keeping the patient in the center of each 3D image stack. If padding was applied, added voxels were given a value corresponding to background/air.

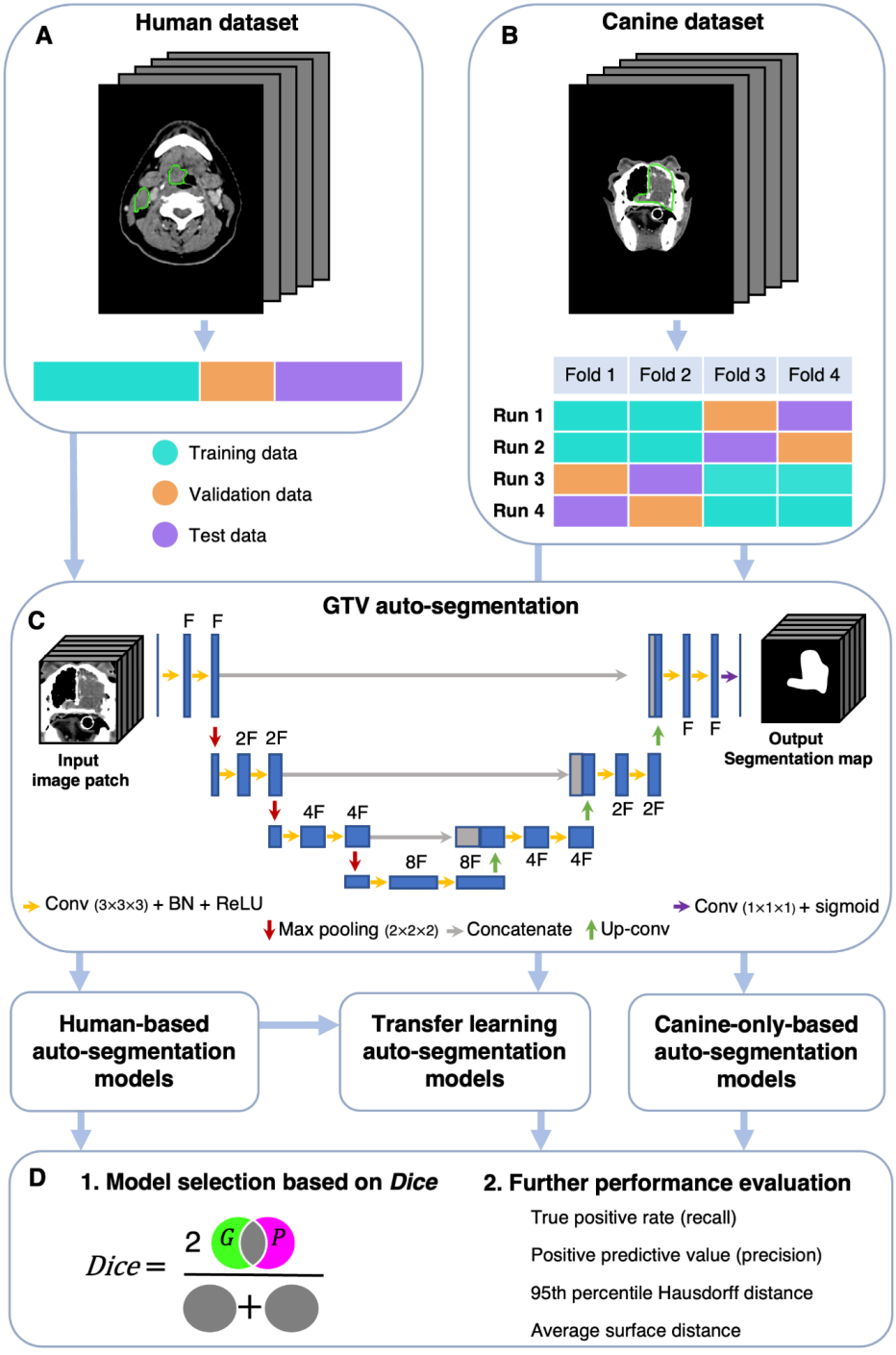

A
Human dataset
B
Canine dataset
Fold 1
Fold 2
Fold 3
Fold 4
Run 1
Run 2
Run 3
Run 4
Training data
Validation data
Test data
C
GTV auto-segmentation
Input image patch
Output Segmentation map
F
2F
4F
8F
Conv (3×3×3) + BN + ReLU
Conv (1×1×1) + sigmoid
Max pooling (2×2×2)
Concatenate
Up-conv
Human-based auto-segmentation models
Transfer learning auto-segmentation models
Canine-only-based auto-segmentation models
D
1. Model selection based on Dice
Dice = 2 G P / G + P
2. Further performance evaluation
True positive rate (recall)
Positive predictive value (precision)
95th percentile Hausdorff distance
Average surface distance

Figure 1. Schematic overview of the analysis. The human (A) and canine (B) datasets consisted of CT images and corresponding manual GTV delineations (green contours) cropped to a volume of interest of 191 × 265 × 173 mm^3. The human dataset was divided into a training, validation, and test set, whereas the canine dataset was divided into four folds used for model training and evaluation. (C) GTV auto-segmentations were generated using a 3D U-Net architecture with input image patches of size 112 × 112 × 112 mm^3 (shown U-Net: depth of 3 (3 max pooling operations) and F filters in the first convolutional layer). Auto-segmentation models were trained on either human or canine data, where the model trained on human data was further used for transfer learning (fine-tuning with canine data). (D) Model performances were first assessed using the Dice similarity coefficient (*Dice*; cf. equation (1), Section 2.5), measuring the overlap between manual ground truth delineations (G) and predicted auto-segmentations (P). The models with superior *Dice* characteristics were selected for further performance evaluation. CT: computed tomography; GTV: gross tumor volume; Conv: Convolution; BN: Batch Normalization; ReLU: Rectified Linear Unit; Up-conv: Up-convolution.

2.4 Deep learning architecture and model training

Canine auto-segmentations were obtained using two main approaches, namely (i) by training CNN models from scratch based on the canine dataset only, and (ii) a transfer learning approach where CNN models were pretrained on the human dataset and subsequently fine-tuned on the canine dataset. As a comparison to the above approaches, the CNN models trained on the human dataset only were evaluated directly on the canine dataset (i.e., without transfer learning). A schematic overview of the analysis is given in Figure 1.

A 3D U-Net CNN architecture [57] with the Dice loss function [58] was used throughout this study. All models were trained using the Adam optimizer with an initial learning rate of 10^{-4} [59]. Further details about the CNN architecture are outlined in Figure 1C. Experiments were run on the Orion High Performance Computing resource at NMBU using deoxys, our in-house developed Python framework for running deep learning experiments with emphasis on TV auto-segmentation (https://deoxys.readthedocs.io/en/latest/).

We assessed the impact of varying the following: (1) the complexity of the U-Net architecture, (2) the CT window settings of the input images, and (3) the training set image augmentation configurations. First, for the models trained from scratch on canine data, different U-Net complexities were assessed using network depths of 3, 4 and 5 with a corresponding number of filters in the first network layer of 32, 64 and 64. Second, we explored using CT window settings with a window center equal to the median Hounsfield unit (HU) value within the ground truth GTV voxels of the relevant training data (human training set: 65 HU; canine training sets: 93 HU (folds 1 and 2) and 96 HU (folds 3 and 4)) and a window width of either 200 HU or 400 HU. CNN models were trained using either one single input channel with windowed CT images, two separate input channels consisting of CT images with and without windowing, or three input channels where two were with different

window settings according to the canine and human training data, and the third channel consisted of CT images without windowing. Third, the following image augmentation configurations were evaluated: no image augmentation, image augmentation in the form of 3D rotation, zooming, and flipping, or 3D elastic deformations. Code for running the experiments, including the above image augmentation schemes, is available at https://github.com[b].

To train models and evaluate model performance, the datasets were divided as follows: Patients in the human dataset were split into a training (n = 126), validation (n = 31) and test (n = 40) set (Figure 1A) using randomly stratified sampling to obtain similar primary tumor stage distributions in each set (cf. Table 2; staging according to the 7th edition AJCC/UICC tumor-node-metastasis system). Patients in the canine dataset were randomly divided into four equally sized folds (n = 9). Following the cross-validation and test set evaluation strategy outlined in Figure 1B, each of these folds was used twice for model training (cyan), once as a validation set (orange) and once as a test set (purple). With this strategy, each canine model configuration was trained four times, and each patient was twice in the training set, once in the validation set and once in the test set. Thus, the validation and test set performances could be calculated for each of the 36 patients. This procedure was chosen to acquire a robust estimate of the auto-segmentation performance despite a limited number of canine patients, taking individual differences across patients into account and making the validation and test set performances less dependent on how the data was split.

Most models were trained for 100 epochs, saving model weights (checkpointing) to disc every epoch. However, for the pretraining of models on human data, early stopping with patience 30 (i.e., stop training if validation loss does not improve for 30 consecutive epochs) was used to avoid overfitting to the source domain. For continued training (fine-tuning) of pretrained models, we compared initializing the Adam optimizer with an initial epoch set to 50 vs. 100. After training of one model, the optimal epoch was identified as the epoch maximizing the mean per patient Sørensen-Dice similarity coefficient [60, 61] (*Dice*; cf. Section 2.5 below) on validation data.

2.5 Performance evaluation

The quality of the CNN-generated auto-segmentations were first assessed using *Dice* [60, 61], which is a volumetric overlap metric quantifying the degree of spatial overlap between the set of voxels in the ground truth G and the predicted auto-segmentation P (Figure 1D). *Dice* is defined as:

$$Dice = \frac{2\,|P \cap G|}{|P| + |G|} = \frac{2TP}{2TP + FP + FN}, \quad (1)$$

[b] A permanent link to the GitHub repository will be provided upon acceptance.

where *TP*, *FP* and *FN* refer to the true positive, false positive, and false negative voxels, respectively. *Dice* ranges from 0 to 1, where 0 corresponds to no overlap and 1 corresponds to perfect overlap between the sets. Based on the *Dice* performances on validation and test data, we selected one model trained from scratch on canine data and one model trained with transfer learning for more in-depth performance evaluation and comparison.

As *Dice* does not separate between *FP* and *FN* voxels and is known to be volume-dependent, the auto-segmentation performance of the two selected models were further assessed using the positive predictive value (*PPV*), the true positive rate (*TPR*), the 95th percentile Hausdorff distance (HD_{95}) [62] and the average surface distance (*ASD*) [63].

PPV and *TPR*, commonly also referred to as precision and recall, are defined as:

$$PPV = \frac{TP}{TP + FP}, \tag{2}$$

and

$$TPR = \frac{TP}{TP + FN}. \tag{3}$$

As seen from equations (2) and (3), *PPV* is the fraction of the predicted auto-segmentation *P* that overlaps with *G,* while *TPR* is the fraction of the ground truth *G* that overlaps with *P*. In the context of TVs used for RT, *PPV* measures the degree of avoiding inclusion of normal tissue voxels in the auto-segmentation, while *TPR* measures the degree of target coverage.

The distance metrics were calculated from the two sets of directed Euclidian distances between the surface voxels of *P* and *G* (set 1: all distances from *P* to *G*; set 2: all distances from *G* to *P*). The HD_{95} and *ASD* were then defined as the maximum value of the 95th percentiles and averages, respectively, of the above two sets of surface distances. HD_{95} reflects the largest mismatch between the surfaces of *P* and *G*, whereas *ASD* is used to quantify the typical displacement between the two surfaces. These metrics should both be as small as possible.

The above performance metrics were calculated per patient, based on all voxels in the pre-defined 3D VOI (cf. Section 2.3). The Python deepmind library was used for calculation of surface-distance-based metrics (https://github.com/deepmind/surface-distance).

3 Results

The validation and test set *Dice* performances of canine models trained with varying network complexity, CT window settings, number of input channels, and image augmentation schemes are summarized in Figure 2. Models trained from scratch (Figure 2A–2B) resulted in mean validation and test set *Dice* scores in the range 0.45–0.62 and 0.39–0.55, whereas models trained with transfer learning (Figure 2C–2D) resulted in validation and test set *Dice* scores ranging from 0.52–0.57 and 0.46–0.52. In comparison, when evaluated on human data the pretrained human-based models resulted in mean validation and test set *Dice* scores of 0.46–0.55 and 0.48–0.54. Models trained on human data only and evaluated directly on canine data resulted in unacceptably low mean *Dice* test scores of 0.02–0.08, even though some models achieved relatively high *Dice* scores for some patients (range of maximum *Dice* per model: 0.16–0.67) (data not shown).

For models trained from scratch on canine data, the highest mean validation *Dice* score (0.62) was observed for models S4 and S8 (Figure 2A), which both used one input channel with a narrow CT window width (200 HU), standard image augmentation (flipping, rotation, zooming) and a high model complexity (depth of 4 and 5, respectively, and 64 filters in the first layer). On the other hand, the less complex model S9 (depth of 3 and 32 filters in the first layer), which was otherwise identical to models S4 and S8, showed comparable mean validation *Dice* performance (0.60) and the highest median *Dice* (0.72). Moreover, models S8 and S9 resulted in similar overall test set *Dice* performances, whereas model S4 had poorer performance on test data (Figure 2B). As model S9 was the least complex and, therefore, the least resource-demanding to train, while at the same time providing competitive *Dice* performance, it was selected for further performance evaluation (Figure 3) and the given complexity and CT window width was used for the transfer learning experiments.

For the transfer learning models, the highest mean validation *Dice* score (0.57) was observed for model T4 (Figure 2C; depth of 3 with 32 filters in the first layer, CT window width of 200 HU, 2 input channels with window center derived from (1) human and (2) canine training data, standard image augmentation and initial epoch set to 50). However, model T2, which included an additional CT channel with no windowing, but was otherwise the same as model T4, displayed the highest test set mean *Dice* (0.52) and a favorable test set *Dice* interquartile range (Figure 2D), indicating a moderately better ability to generalize to previously unseen data. Thus, among the transfer learning models, model T2 was selected for computation of additional performance metrics (Figure 3).

The two selected models (S9 and T2, Figure 2) generally showed similar auto-segmentation performances on test data, as indicated by the plots and summary statistics of Figure 3. For both models, there was substantial inter-patient variation in the resulting auto-segmentation quality. The model trained from scratch on canine data resulted in the best mean performances for all included metrics. In general, the canine-only model had larger tumor coverage (higher mean and median *TPR*)

but tended to include more normal tissue (lower median *PPV*) than the transfer learning model. The transfer learning model did, however, achieve the highest per patient overlap with ground truth contours (maximum *Dice*: 0.89) and the lowest per patient *ASD* (minimum *ASD*: 1.3 mm). In addition, the transfer learning model resulted in a higher number of very high-quality auto-segmentations (*Dice* ≥ 0.85; $n = 5$) than the model trained from scratch ($n = 2$). However, the transfer learning model tended to perform the poorest on more difficult-to-segment canine patients, as reflected by the poorer first quartile *Dice*, *TPR*, *PPV* and HD_{95} values.

Example auto-segmentations are shown in Figures 4–6. In general, the two selected models achieved the highest quality auto-segmentations for patients with nasal cavity tumors, which was the most frequently occurring tumor site in the canine dataset. The canine-only and transfer learning models both achieved a mean test set *Dice* of 0.69 for nasal cavity tumors, compared to the corresponding *Dice* scores of 0.55 and 0.52 for all tumor sites. As exemplified in Figure 4, tumor regions with relatively homogeneous HU values within the ground truth were generally easier to segment correctly. High quality auto-segmentations were also seen for other tumor sites where the tumor was distinct from the surrounding normal tissues and clearly affected the anatomical shape/boundary of the animal (Figure 5). Peripheral parts of the GTV were often more difficult to segment than central parts (Figure 5; bottom row). In some cases, the auto-segmentations included substantial normal tissue regions due to over-estimation of the GTV boundaries or prediction of separate smaller false positive structures. False positive structures and inclusion of particularly brain and eye tissues in the predicted auto-segmentation were more pronounced for the model trained only on canine data (S9). An example is shown in Figure 6.

Both models resulted in poor auto-segmentations for patients with atypical tumor sites, atypical GTV shapes and/or a substantial number of image slices with atypical/very heterogeneous HU values inside the ground truth GTV. Neither of the models was able to successfully segment the smaller canine GTV-N structures. The above patterns indicate that the auto-segmentation performance was dependent on the number of representative canine training samples, regardless of model training approach.

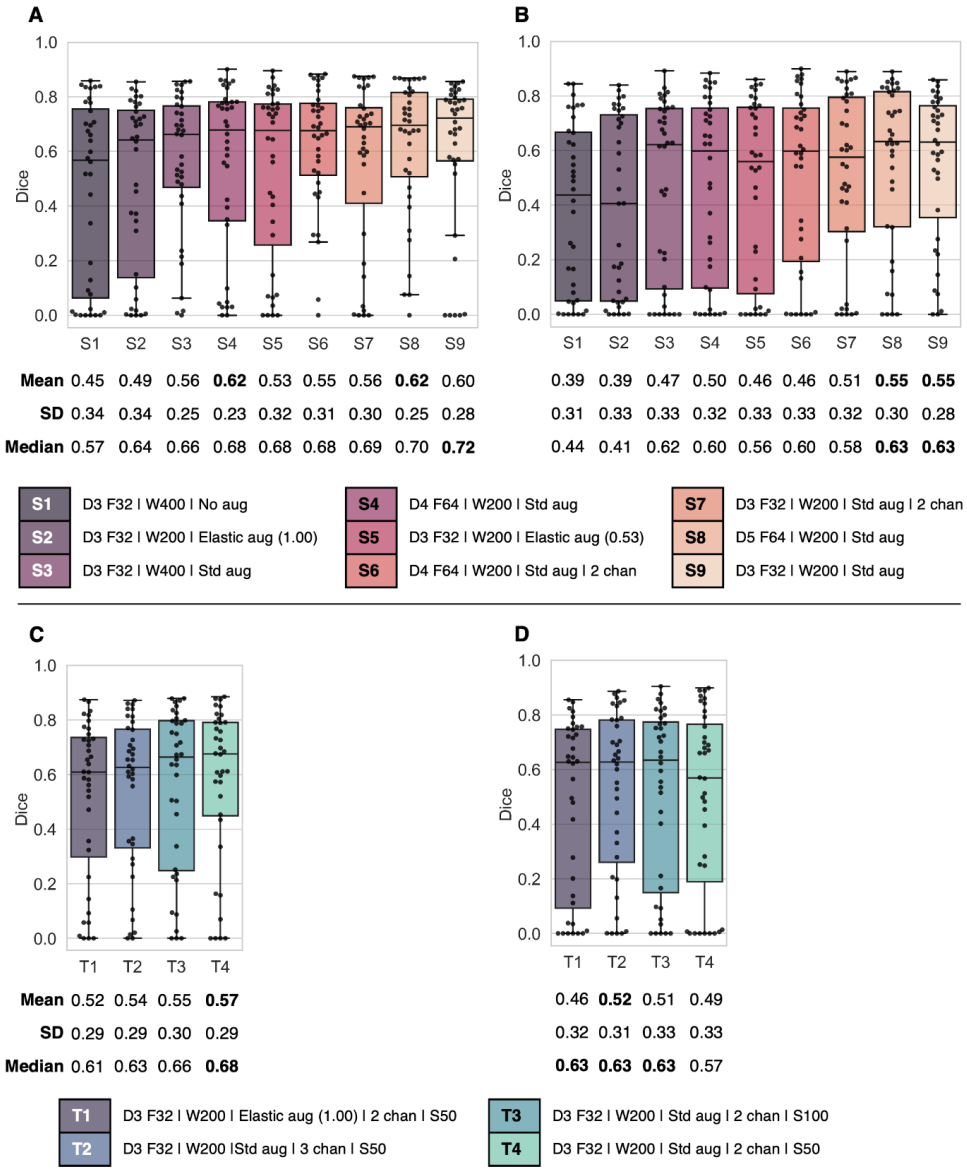
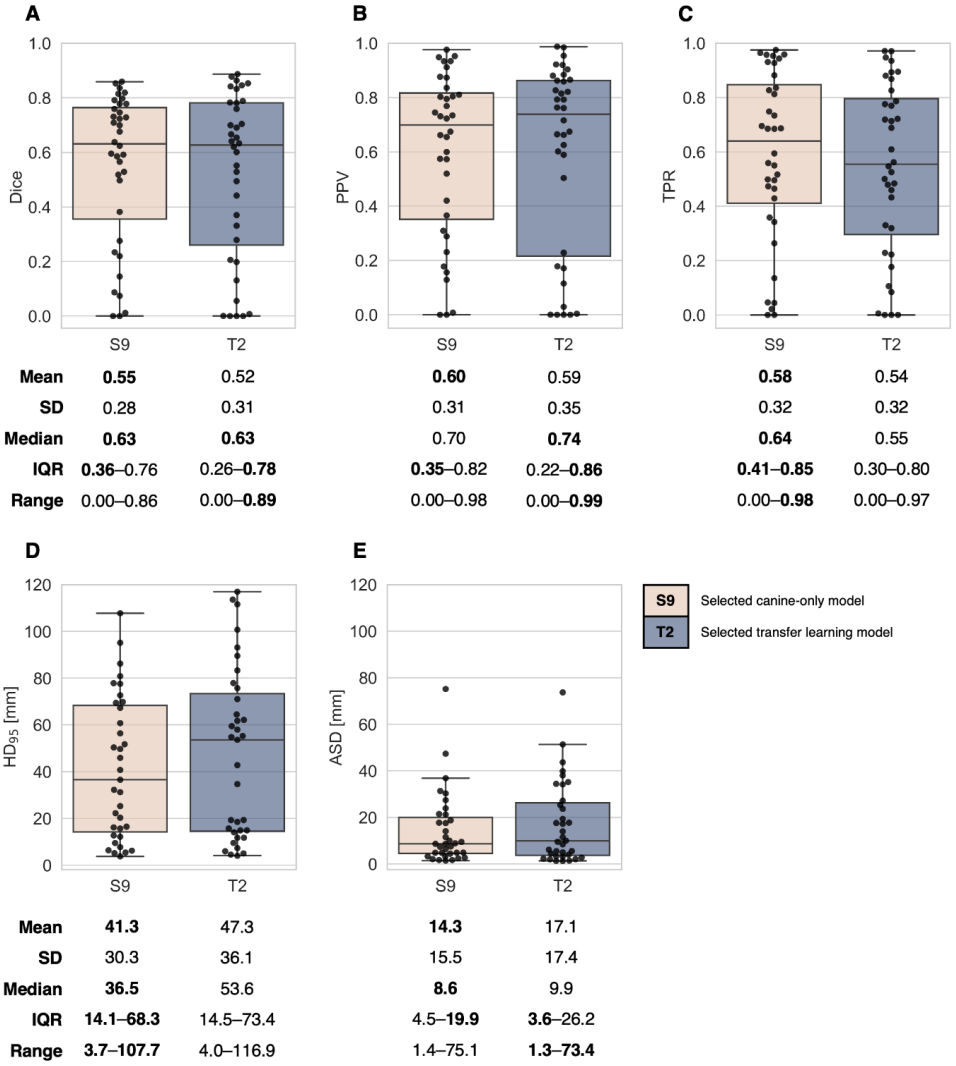


Figure 2. Combined box and swarm plots of per patient *Dice* scores for different model configurations, showing each patient as a separate data point (black). Top: (A) Validation and (B) test results for models S1–S9 trained from scratch on canine data. Bottom: (C) Validation and (D) test result for models T1–T4 trained using the transfer learning approach. Model configurations (S1-S9 and T1-T4) are as follows: Model complexity given by U-Net depth D and number of filters F in the first layer; CT window setting with window width W in HU; Image augmentation settings (Std aug: zooming, rotation and flipping; Elastic aug: elastic deformation on a proportion (0.53 or 1.00) of the training set images); Number of input channels (chan), default 1 channel unless otherwise stated; Initial epoch setting S, either 50 or 100 epochs (transfer learning only). SD: standard deviation.



	A: Dice S9	A: Dice T2	B: PPV S9	B: PPV T2	C: TPR S9	C: TPR T2
Mean	**0.55**	0.52	**0.60**	0.59	**0.58**	0.54
SD	0.28	0.31	0.31	0.35	0.32	0.32
Median	**0.63**	**0.63**	0.70	**0.74**	**0.64**	0.55
IQR	**0.36**–0.76	0.26–**0.78**	**0.35**–0.82	0.22–**0.86**	**0.41**–**0.85**	0.30–0.80
Range	0.00–0.86	0.00–**0.89**	0.00–0.98	0.00–**0.99**	0.00–**0.98**	0.00–0.97

	D: HD_{95} [mm] S9	D: HD_{95} [mm] T2	E: ASD [mm] S9	E: ASD [mm] T2
Mean	**41.3**	47.3	**14.3**	17.1
SD	30.3	36.1	15.5	17.4
Median	**36.5**	53.6	**8.6**	9.9
IQR	**14.1**–**68.3**	14.5–73.4	4.5–**19.9**	**3.6**–26.2
Range	**3.7**–**107.7**	4.0–116.9	1.4–75.1	**1.3**–**73.4**

Figure 3. Combined box and swarm plots of per patient auto-segmentation performance metrics for the two selected models trained from scratch on canine data (S9, Figure 2A, B) and trained using the transfer learning approach (T2, Figure 2C, D). Each patient is shown as a separate data point (black). Performance metrics: (A) Dice similarity coefficient (*Dice)*, (B) positive predictive value (*PPV*), (C) true positive rate (*TPR*), (D) 95th percentile Hausdorff distance (HD_{95}), (E) average surface distance (*ASD*). The exact positioning of individual data points in (A) may differ from the respective plots in Figure 2B and D, due to randomness in the swarm plots. SD: standard deviation; IQR: interquartile range. One patient without any predicted auto-segmentation was excluded from calculations of HD_{95} (D) and *ASD* (E).

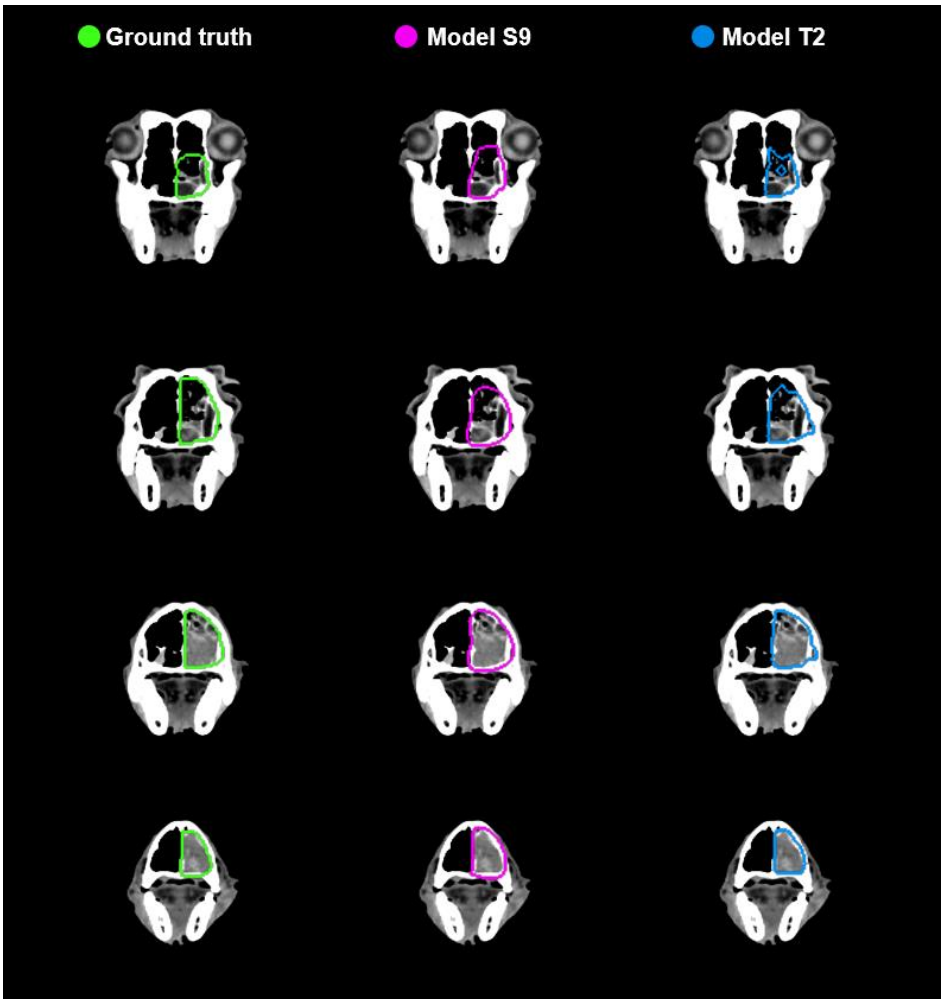


Figure 4. Manual ground truth and automatic deep learning generated gross tumor volume contours in four CT image slices from one canine test set patient (nasal cavity tumor). Left column: Manual ground truth contours (green). Middle column: Auto-segmentation generated by model S9 (magenta; model trained from scratch on canine data only). Right column: Auto-segmentation generated by model T2 (blue; model trained using transfer learning). The two models resulted in Sørensen-Dice similarity coefficients of 0.85 (model S9) and 0.89 (model T2) for the given patient (calculated over all 173 image slices).

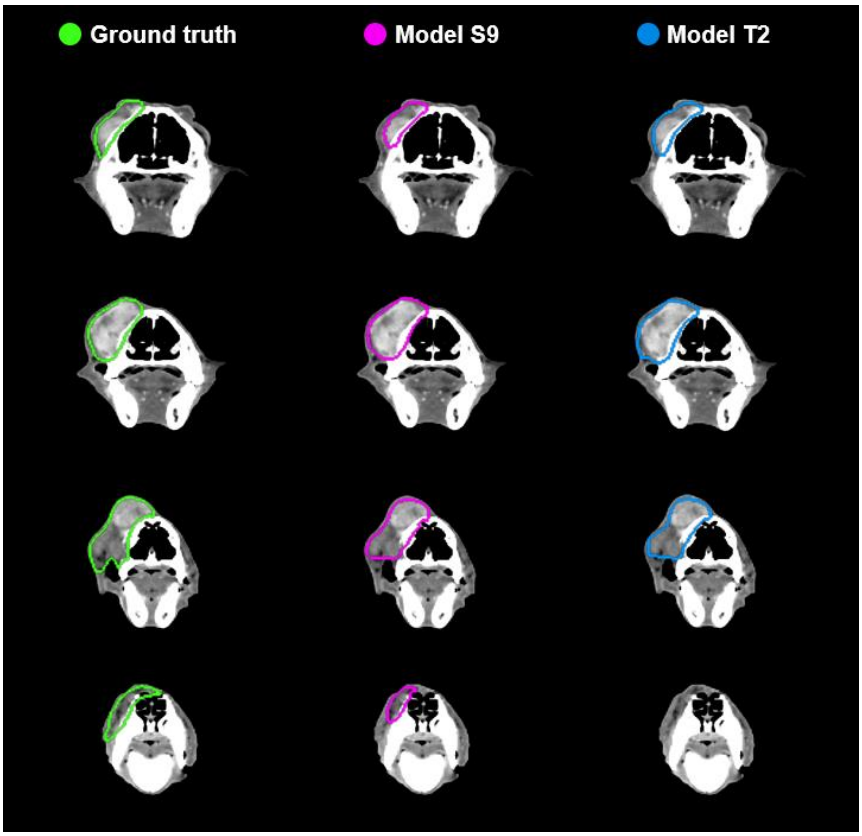


Figure 5. Manual ground truth and automatic deep learning generated gross tumor volume contours in four CT image slices from one canine test set patient (sarcoma). Left column: Manual ground truth contours (green). Middle column: Auto-segmentation generated by model S9 (magenta; model trained from scratch on canine data only). Right column: Auto-segmentation generated by model T2 (blue; model trained using transfer learning). The two models resulted in Sørensen-Dice similarity coefficients of 0.86 (model S9) and 0.84 (model T2) for the given patient (calculated over all 173 image slices).

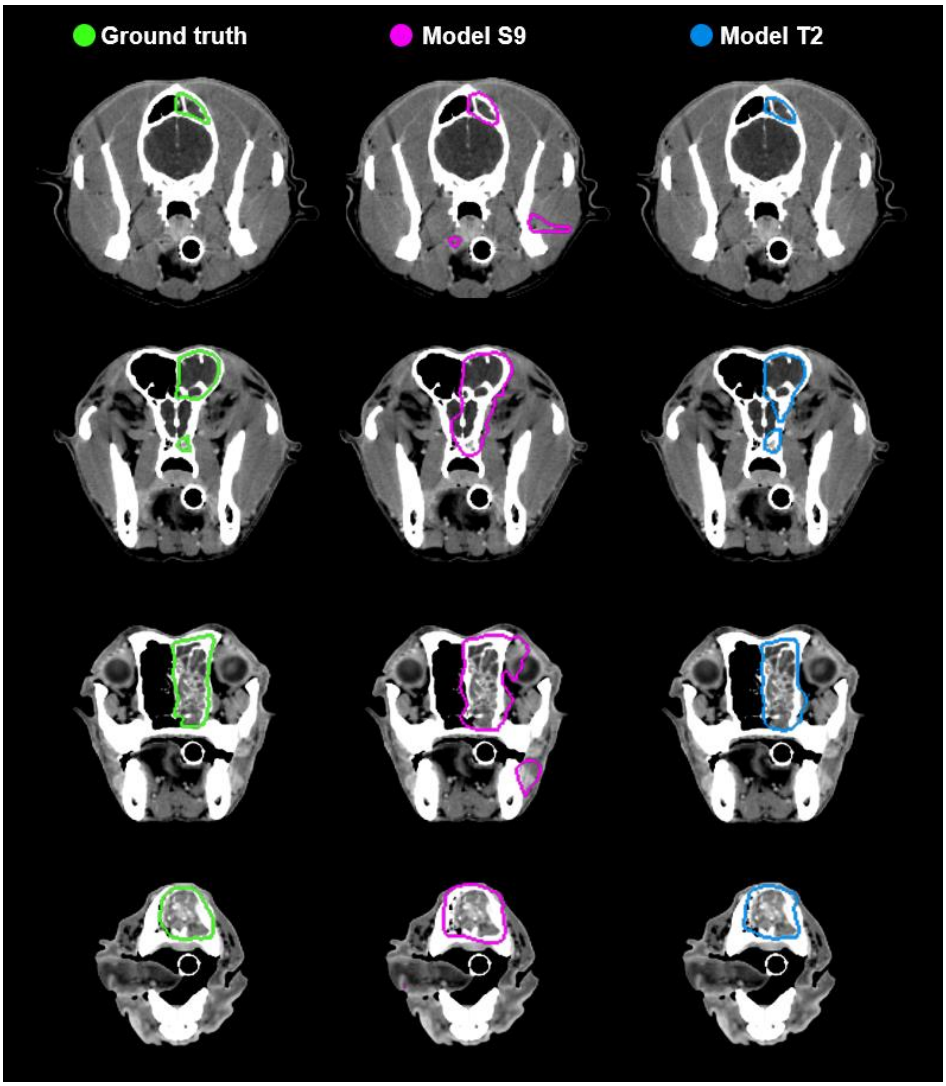


Figure 6. Manual ground truth and automatic deep learning generated gross tumor volume contours in four CT image slices from one canine test set patient (nasal cavity tumor). Left column: Manual ground truth contours (green). Middle column: Auto-segmentation generated by model S9 (magenta; model trained from scratch on canine data only). Right column: Auto-segmentation generated by model T2 (blue; model trained using transfer learning). The two models resulted in Sørensen-Dice similarity coefficients of 0.72 (model S9) and 0.85 (model T2) for the given patient (calculated over all 173 image slices).

4 Discussion

This is the first study to evaluate deep learning-based auto-segmentation of TVs for RT in veterinary patients. Although dogs display breed-related variation in the head and neck anatomy and size, which could potentially complicate the auto-segmentation task, our results show that CNNs can provide high-quality GTV auto-segmentations for this group of patients, despite a limited number of training samples. Our two main approaches, namely (i) CNN models trained from scratch on canine data or (ii) CNN models pretrained on human HNC patients and fine-tuned using canine patients (transfer learning), generally gave similar results. In both cases the mean overlap with the expert ground truth contours was similar to what is obtained for human HNC patients.

Previous studies on human HNC subjects report mean validation and/or test set *Dice* scores in the range of 0.31–0.66 for CNN-generated auto-segmentations of the GTV based on CT images [29, 33, 34, 64]. The relatively large variation in reported performances is likely related to differences in image pre-processing, such as CT window settings and VOI dimensions, the composition of the datasets and/or CNN architecture. Of the above studies, the highest mean *Dice* (0.66; cross-validation result [29]) was obtained using a 2-dimensional (2D) U-Net architecture and a considerably smaller pre-defined VOI than in our present work. Moe et al. [34] obtained a mean test set *Dice* of 0.56 using the same 2D U-Net architecture on larger image VOIs encompassing the entire head and neck region but excluding image slices without any ground truth delineation. Both [29] and [34] used the same single-center HNC patients as in our present study. The lowest mean *Dice* scores (0.31 [33] and 0.49 [64]) were reported for auto-segmentation in multi-center patient cohorts, which is generally more challenging than single-center segmentation, using wider CT window widths. Both latter studies used similarly sized image VOIs and 3D architectures, which are generally superior to their 2D counterparts, as in or present work. In comparison to the above human studies, our best-performing canine models trained from scratch or with transfer learning, both using a 3D U-Net architecture and a narrow CT window, resulted in similar or higher mean validation (test) *Dice* scores of 0.62 (0.55) and 0.57 (0.52), respectively, compared to the above studies. The *Dice* performances of our CNN models were also comparable to the reported *Dice* agreement (0.56–0.57) between clinical experts performing manual GTV contouring in human HNC patients based solely on CT images [65, 66].

Human cancer patients normally undergo several imaging procedures as part of diagnosis and treatment planning. It is also common to base contouring of the GTV on multimodal image information. Thus, most of the recent studies on GTV segmentation in human HNC patients investigate using multimodality images as input to the network for increased performance. As PET/CT imaging is becoming more common in veterinary medicine [67], it is worth noting that all the above human HNC studies reported significant increases (range: 12–129 %; median: 25 %) in mean *Dice* scores when using both FDG PET and CT images as CNN model input. Similar improvements

are likely possible for canine patients, provided that the lesions are comparably FDG PET avid. PET imaging is, however, not likely to become widely available for veterinary patients in the near future. A more realistic approach at present would be to investigate the potential added benefit of including both pre and post contrast CT images as input to CNN models trained from scratch on canine data. In the present work, however, we chose to focus solely on post contrast CT images as these images were also available for the human patients.

As HNC is a heterogeneous group of cancers, many studies on human HNC subjects focus only on one anatomical primary tumor site. Specifically oropharyngeal cancer which is one of the most frequently occurring HNC sites in humans worldwide [2], or nasopharyngeal cancer which display very distinctive properties, are commonly analyzed separately [31, 36, 41, 64, 68-72]. A similar approach could be beneficial for further analyses of auto-segmentation of the GTV in canine HNC subjects. In our present work, the highest quality auto-segmentations were generally obtained in patients with nasal cavity tumors. This is likely influenced by the tumor site distribution in our dataset, where this was the most frequent site. However, nasal cavity tumors display distinctive characteristics in terms of shape and location and generally have high contrast between tumor tissue and normal tissues/background, all of which could aid auto-segmentation. GTV segmentation is also particularly relevant for this group of canine HNC patients, as RT is indicated as the primary treatment [4, 13, 14].

Even though our results show that deep learning can provide high-quality GTV auto-segmentations in canine HNC patients, there are currently several limitations to this approach that must be resolved to increase its potential clinical usefulness. First, regardless of model training approach, the auto-segmentation quality was variable between patients. The poorest performance was seen for patients with rare tumor sites and GTVs with atypical shapes or heterogeneous HU intensity values. This could be alleviated by having a larger training set where all tumor sites are represented to a greater extent. Another possibility, as outlined in the previous paragraph, is to focus on each tumor site separately. Furthermore, inclusion of both pre and post contrast CT images as model input may mitigate the issue of heterogeneous tumor intensities in some cases, as it can be relevant whether the heterogeneity is due to inherent anatomical factors or heterogeneous contrast enhancement. However, GTVs with very heterogeneous HU intensity values including relatively large proportions of bone and/or air might still be difficult to automatically segment correctly and would likely require intervention by a human expert. Secondly, the auto-segmentations could encompass false positive regions including OARs such as the eye and brain. To limit the need for human revision, smaller false positive structures could be removed in a post-processing step, using for example morphologic operations, whereas inclusion of OAR regions due to over-estimation of GTV boundaries could be reduced by combining OAR and TV segmentation. Segmentation of normal tissue structures such as OARs typically achieve higher *Dice* scores than TV segmentation, as organ shapes, locations and

intensities generally are less variable between patients than tumors, though some OARs are more difficult to segment than others due to, e.g., poor CT contrast. Reported mean *Dice* scores of CT-based OAR segmentation using deep learning are 0.78–0.87 [30, 38-40] and 0.83 [48] for human and canine HNC patients, respectively, when averaged over various organ structures. Deep learning-based OAR segmentation may be considered clinically applicable for several OARs [37, 73] and is currently commercially available for RT in humans.

Transfer learning provided high performance but did not improve the mean performance metrics compared to training canine models from scratch. There are several potential factors that can contribute to why transfer learning did not outperform training from scratch, specifically related to the differences between the human and canine datasets. First, there are obvious anatomical differences between the human and canine head and neck region that might not be overcome by the use of image augmentation and fine-tuning of the pretrained human model. Second, the presence and degree of nodal involvement was significantly higher for the human patients. The majority of the human patients (76 %) had known nodal involvement and the mean GTV-N size was similar to the mean GTV-T size, whereas few canine patients (11 %) had known nodal involvement and the GTV-N structures were all small in size compared to the GTV-T. Third, the anatomical tumor site and cancer subtype distributions were not comparable between the two species. Fourth, the ground truth GTV contours were delineated under different conditions. Fifth, the CT imaging was conducted using different scanners with different imaging and reconstruction parameters. Regardless of the above differences between source and target domains, the transfer learning approach resulted in the highest per patient *Dice* score and to a greater extent avoidance of OARs. Thus, there is reason to assume that some features learned in the source domain were useful in the target domain, but that the usefulness was variable among the canine subjects.

A recent thorough investigation of transfer learning for different deep learning-based medical image segmentation tasks in humans, conducted by Karimi et al. [50], shows that transfer learning in general primarily decreased the training time for the target task and that improvements in auto-segmentation performance often was marginal and largely relied on the data and task. According to their results, statistically significant effects of transfer learning only occurred when the number of target training samples was low (~ 3–15 subjects). In other cases, models trained from scratch and transfer learning models were comparable in terms of auto-segmentation quality. Cross-species transfer learning was not evaluated in Karimi et al. [50] but our results are in line with their findings for transfer learning between human domains and tasks. Gerard et al. [47] applied transfer learning to segment acutely injured lungs in CT images of dogs, pigs, and sheep, obtaining median Jaccard index scores ≥ 0.90, which corresponds to *Dice* scores ≥ 0.95, using a multi-resolution CNN model pretrained on CT images of humans without acutely injured lungs. Their proposed transfer learning method was, however, not compared to training models from scratch on the target domain. Thus, the

effect of transfer learning was not quantified, and the high performance could be related to the task or influenced by the CNN configuration rather than the transfer learning approach.

To summarize, segmentation of the GTV in canine and human HNC patients is an inherently challenging task. In this study, CNN models for auto-segmentation of the GTV in canine HNC patients, trained either from scratch on canine data or by using a cross-species transfer learning approach, provided promising results with high performance metrics comparable to results achieved in human HNC auto-segmentation studies. Our results show that transfer learning has the potential to increase segmentation performance for some patients, but differences between source and target domains as well as the heterogeneity of the disease within species can complicate the modelling. Therefore, care must be taken when transferring auto-segmentation models between species.

5 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

6 Author Contributions

ARG: conceptualization, methodology, software, data curation, formal analysis, visualization, writing – original draft preparation, writing – review and editing. BNH: conceptualization, methodology, software, writing – review and editing. OT and ÅS: conceptualization, methodology, writing – review and editing. ED: data curation, funding acquisition, writing – review and editing. EM: conceptualization, methodology, data curation, funding acquisition, writing – review and editing. HKS: conceptualization, methodology, data curation, funding acquisition, writing – review and editing. CMF: conceptualization, methodology, funding acquisition, project administration, supervision, writing – review and editing. All authors contributed to the article and approved the submitted version.

7 Funding

The collection of the human dataset used in the present study was conducted as part of work supported by the Norwegian Cancer Society (Grant Number 160907-2014 and 182672-2016).

8 Acknowledgments

We thank former veterinary students at NMBU, Andrea Knudsen Bye, Jenny Amundsen Dahle, and Elisabet Rønneberg Nilsen, for their help in collecting the canine dataset. The authors acknowledge the Orion High Performance Computing Center (OHPCC) at NMBU for providing computational resources.

9 Data Availability Statement

Access to the human dataset requires approval by the Regional Ethics Committee. The canine raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

References

1. Mody MD, Rocco JW, Yom SS, Haddad RI, Saba NF. Head and Neck Cancer. Lancet (2021) 398(10318):2289-99. doi: 10.1016/s0140-6736(21)01550-6.

2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* (2021) 71(3):209-49. doi: 10.3322/caac.21660.

3. Hedberg ML, Grandis JR. The Molecular Pathogenesis of Head and Neck Cancer. In: Mendelsohn J, Gray JW, Howley PM, Israel MA, Thompson CB, editors. The Molecular Basis of Cancer (Fourth Edition). Philadelphia: W.B. Saunders (2015). p. 491-8.

4. Culp TN. Tumors of the Respiratory System. In: Vail DM, Thamm DH, Liptak J, editors. Withrow and Macewen's Small Animal Clinical Oncology (Sixth Edition). St. Louis: Elsevier Health Sciences (2019). p. 492-523.

5. Wilson DW. Tumors of the Respiratory Tract. In: Meuten DJ, editor. Tumors in Domestic Animals (Fifth Edition). Hoboken: Wiley-Blackwell (2016). p. 467-98.

6. Munday JS, Löhr CV, Kiupel M. Tumors of the Alimentary Tract. In: Meuten DJ, editor. Tumors in Domestic Animals (Fifth Edition). Hoboken: Wiley-Blackwell (2016). p. 499-601.

7. Schiffman JD, Breen M. Comparative Oncology: What Dogs and Other Species Can Teach Us About Humans with Cancer. *Philos Trans R Soc Lond, B, Biol Sci* (2015) 370(1673):20140231. doi: 10.1098/rstb.2014.0231.

8. Rowell JL, McCarthy DO, Alvarez CE. Dog Models of Naturally Occurring Cancer. *Trends Mol Med* (2011) 17(7):380-8. doi: 10.1016/j.molmed.2011.02.004.

9. Liu D, Xiong H, Ellis AE, Northrup NC, Dobbin KK, Shin DM, et al. Canine Spontaneous Head and Neck Squamous Cell Carcinomas Represent Their Human Counterparts at the Molecular Level. *PLoS Genet* (2015) 11(6):e1005277. doi: 10.1371/journal.pgen.1005277.

10. O'Sullivan B, Rumble RB, Warde P. Intensity-Modulated Radiotherapy in the Treatment of Head and Neck Cancer. *Clin Oncol* (2012) 24(7):474-87. doi: 10.1016/j.clon.2012.05.006.

11. Gupta T, Agarwal J, Jain S, Phurailatpam R, Kannan S, Ghosh-Laskar S, et al. Three-Dimensional Conformal Radiotherapy (3D-CRT) Versus Intensity Modulated Radiation Therapy (IMRT) in Squamous Cell Carcinoma of the Head and Neck: A Randomized Controlled Trial. *Radiother Oncol* (2012) 104(3):343-8. doi: 10.1016/j.radonc.2012.07.001.

12. Chao KSC, Majhail N, Huang C-j, Simpson JR, Perez CA, Haughey B, et al. Intensity-Modulated Radiation Therapy Reduces Late Salivary Toxicity without Compromising Tumor Control in Patients with Oropharyngeal Carcinoma: A Comparison with Conventional Techniques. *Radiother Oncol* (2001) 61(3):275-80. doi: 10.1016/S0167-8140(01)00449-2.

13. Hansen KS, Kent MS. Imaging in Non-Neurologic Oncologic Treatment Planning of the Head and Neck. *Front Veterinary Sci* (2019) 6:90. doi: 10.3389/fvets.2019.00090.

14. Mortier J, Blackwood L. Treatment of Nasal Tumours in Dogs: A Review. *J Small Anim Pract* (2020) 61(7):404-15. doi: 10.1111/jsap.13173.

15. Farrelly J, McEntee MC. A Survey of Veterinary Radiation Facilities in 2010. *Vet Radiol Ultrasound* (2014) 55(6):638-43. doi: 10.1111/vru.12161.

16. McEntee MC. A Survey of Veterinary Radiation Facilities in the United States During 2001. *Vet Radiol Ultrasound* Ultrasound (2004) 45(5):476-9. doi: 10.1111/j.1740-8261.2004.04082.x.

17. Poirier VJ, Koh ES, Darko J, Fleck A, Pinard C, Vail DM. Patterns of Local Residual Disease and Local Failure after Intensity Modulated/Image Guided Radiation Therapy for Sinonasal Tumors in Dogs. *J Vet Intern Med* (2021) 35(2):1062-72. doi: 10.1111/jvim.16076.

18. Segedin B, Petric P. Uncertainties in Target Volume Delineation in Radiotherapy–Are They Relevant and What Can We Do About Them? *Radiol Oncol* (2016) 50(3):254-62. doi: 10.1515/raon-2016-0023.

19. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More Than Pictures, They Are Data. *Radiology* (2016) 278(2):563. doi: 10.1148/radiol.2015151169.

20. Cox S, Cleves A, Clementel E, Miles E, Staffurth J, Gwynne S. Impact of Deviations in Target Volume Delineation–Time for a New RTQA Approach? *Radiother Oncol* (2019) 137:1-8. doi: 10.1016/j.radonc.2019.04.012.

21. Chang ATY, Tan LT, Duke S, Ng W-T. Challenges for Quality Assurance of Target Volume Delineation in Clinical Trials. *Front Oncol* (2017) 7:221. doi: 10.3389/fonc.2017.00221.

22. Vinod SK, Jameson MG, Min M, Holloway LC. Uncertainties in Volume Delineation in Radiation Oncology: A Systematic Review and Recommendations for Future Studies. *Radiother Oncol* (2016) 121(2):169-79. doi: 10.1016/j.radonc.2016.09.009.

23. Kosmin M, Ledsam J, Romera-Paredes B, Mendes R, Moinuddin S, de Souza D, et al. Rapid Advances in Auto-Segmentation of Organs at Risk and Target Volumes in Head and Neck Cancer. *Radiother Oncol* (2019) 135:130-40. doi: 10.1016/j.radonc.2019.03.004.

24. Hatt M, Laurent B, Ouahabi A, Fayad H, Tan S, Li L, et al. The First MICCAI Challenge on PET Tumor Segmentation. *Med Image Anal* (2018) 44:177-95. doi: 10.1016/j.media.2017.12.007.

25. Bernard O, Lalande A, Zotti C, Cervenansky F, Yang X, Heng P-A, et al. Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? *IEEE Trans Med Imaging* (2018) 37(11):2514-25. doi: 10.1109/TMI.2018.2837502.

26. Kamnitsas K, Ledig C, Newcombe VF, Simpson JP, Kane AD, Menon DK, et al. Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation. *Med Image Anal* (2017) 36:61-78. doi: 10.1016/j.media.2016.10.004.

27. Ghaffari M, Sowmya A, Oliver R. Automated Brain Tumor Segmentation Using Multimodal Brain Scans: A Survey Based on Models Submitted to the BraTS 2012–2018 Challenges. *IEEE Rev Biomed Eng* (2019) 13:156-68. doi: 10.1109/RBME.2019.2946868.

28. Choi MS, Choi BS, Chung SY, Kim N, Chun J, Kim YB, et al. Clinical Evaluation of Atlas- and Deep Learning-Based Automatic Segmentation of Multiple Organs and Clinical Target Volumes for Breast Cancer. *Radiother Oncol* (2020) 153:139-45. doi: 10.1016/j.radonc.2020.09.045.

29. Groendahl AR, Knudtsen IS, Huynh BN, Mulstad M, Moe YM, Knuth F, et al. A Comparison of Methods for Fully Automatic Segmentation of Tumors and Involved Nodes in PET/CT of Head and Neck Cancers. *Phys Med Biol* (2021) 66(6):065012. doi: 10.1088/1361-6560/abe553.

30. Zhu W, Huang Y, Zeng L, Chen X, Liu Y, Qian Z, et al. AnatomyNet: Deep Learning for Fast and Fully Automated Whole-Volume Segmentation of Head and Neck Anatomy. *Med Phys* (2019) 46(2):576-89. doi: https://doi.org/10.1002/mp.13300.

31. Lin L, Dou Q, Jin Y-M, Zhou G-Q, Tang Y-Q, Chen W-L, et al. Deep Learning for Automated Contouring of Primary Tumor Volumes by MRI for Nasopharyngeal Carcinoma. *Radiology* (2019) 291(3):677-86. doi: 10.1148/radiol.2019182012.

32. Chlebus G, Meine H, Thoduka S, Abolmaali N, van Ginneken B, Hahn HK, et al. Reducing Inter-Observer Variability and Interaction Time of MR Liver Volumetry by Combining Automatic CNN-Based Liver Segmentation and Manual Corrections. *PLoS One* (2019) 14(5):e0217228. doi: 10.1371/journal.pone.0217228.

33. Guo Z, Guo N, Gong K, Zhong Sa, Li Q. Gross Tumor Volume Segmentation for Head and Neck Cancer Radiotherapy Using Deep Dense Multi-Modality Network. *Phys Med Biol* (2019) 64(20):205015. doi: 10.1088/1361-6560/ab440d.

34. Moe YM, Groendahl AR, Tomic O, Dale E, Malinen E, Futsaether CM. Deep Learning-Based Auto-Delineation of Gross Tumour Volumes and Involved Nodes in PET/CT Images of Head and Neck Cancer Patients. *Eur JNucl Med Mol Imaging* (2021) 48(9):2782-92. doi: 10.1007/s00259-020-05125-x.

35. Ren J, Eriksen JG, Nijkamp J, Korreman SS. Comparing Different CT, PET and MRI Multi-Modality Image Combinations for Deep Learning-Based Head and Neck Tumor Segmentation. *Acta Oncol* (2021) 60(11):1399-406. doi: 10.1080/0284186X.2021.1949034.

36. Andrearczyk V, Oreiller V, Jreige M, Vallieres M, Castelli J, Elhalawani H, et al. Overview of the HECKTOR Challenge at MICCAI 2020: Automatic Head and Neck Tumor Segmentation in PET/CT. In: Andrearczyk V, Oreiller V, Depeursinge A, editors. Head and Neck Tumor Segmentation. HECKTOR 2020. Lecture Notes in Computer Science vol 12603. Cham: Springer (2021). doi: 10.1007/978-3-030-67194-5_1.

37. Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, De Fauw J, et al. Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study. *J Medical Internet Res* (2021) 23(7):e26151. doi: 10.2196/26151.

38. Tang H, Chen X, Liu Y, Lu Z, You J, Yang M, et al. Clinically Applicable Deep Learning Framework for Organs at Risk Delineation in CT Images. *Nat Mach Intell* (2019) 1(10):480-91. doi: 10.1038/s42256-019-0099-z.

39. Van der Veen J, Willems S, Deschuymer S, Robben D, Crijns W, Maes F, et al. Benefits of Deep Learning for Delineation of Organs at Risk in Head and Neck Cancer. *Radiother Oncol* (2019) 138:68-74. doi: 10.1016/j.radonc.2019.05.010.

40. Wang W, Wang Q, Jia M, Wang Z, Yang C, Zhang D, et al. Deep Learning-Augmented Head and Neck Organs at Risk Segmentation from CT Volumes. *Front Phys* (2021):612. doi: 10.3389/fphy.2021.743190. 41.

41. Andrearczyk V, Oreiller V, Boughdad S, Rest CCL, Elhalawani H, Jreige M, et al. Overview of the HECKTOR Challenge at MICCAI 2021: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT Images. In: Andrearczyk, V, Oreiller V, Hatt M, Depeursinge A, editors. Head and Neck Tumor Segmentation and Outcome

Prediction. HECKTOR 2021. Lecture Notes in Computer Science vol 13209. Cham: Springer (2022). https://doi.org/10.1007/978-3-030-98253-9_1.

42. Nejedly P, Kremen V, Sladky V, Nasseri M, Guragain H, Klimes P, et al. Deep-Learning for Seizure Forecasting in Canines with Epilepsy. *J Neural Eng* (2019) 16(3):036031. doi: 10.1088/1741-2552/ab172d.

43. Chambers RD, Yoder NC, Carson AB, Junge C, Allen DE, Prescott LM, et al. Deep Learning Classification of Canine Behavior Using a Single Collar-Mounted Accelerometer: Real-World Validation. *Animals* (2021) 11(6):1549. doi: 10.3390/ani11061549.

44. Salvi M, Molinari F, Iussich S, Muscatello LV, Pazzini L, Benali S, et al. Histopathological Classification of Canine Cutaneous Round Cell Tumors Using Deep Learning: A Multi-Center Study. *Front Vet Sci* (2021) 8. doi: 10.3389/fvets.2021.640944.

45. Banzato T, Wodzinski M, Burti S, Osti VL, Rossoni V, Atzori M, et al. Automatic Classification of Canine Thoracic Radiographs Using Deep Learning. *Sci Rep* (2021) 11(1):1-8. doi: 10.1038/s41598-021-83515-3.

46. Florkow MC, Zijlstra F, Willemsen K, Maspero M, van den Berg CA, Kerkmeijer LG, et al. Deep Learning–Based MR-to-CT Synthesis: The Influence of Varying Gradient Echo–Based MR Images as Input Channels. *Magn Reson Med* (2020) 83(4):1429-41. doi: 10.1002/mrm.28008.

47. Gerard SE, Herrmann J, Kaczka DW, Musch G, Fernandez-Bustamante A, Reinhardt JM. Multi-Resolution Convolutional Neural Networks for Fully Automated Segmentation of Acutely Injured Lungs in Multiple Species. *Med Image Anal* (2020) 60:101592. doi: 10.1016/j.media.2019.101592.

48. Park J, Choi B, Ko J, Chun J, Park I, Lee J, et al. Deep-Learning-Based Automatic Segmentation of Head and Neck Organs for Radiation Therapy in Dogs. *Front Vet Sci* (2021):1006. doi: 10.3389/fvets.2021.711612.

49. Schmid D, Scholz VB, Kircher PR, Lautenschlaeger IE. Employing Deep Convolutional Neural Networks for Segmenting the Medial Retropharyngeal Lymph Nodes in CT Studies of Dogs. *Vet Radiol Ultrasound* (2022). doi: 10.1111/vru.13132.

50. Karimi D, Warfield SK, Gholipour A. Transfer Learning in Medical Image Segmentation: New Insights from Analysis of the Dynamics of Model Parameters and Learned Representations. *Artif Intell Med* (2021) 116:102078. doi: 10.1016/j.artmed.2021.102078.

51. Pan SJ, Yang Q. A Survey on Transfer Learning. *IEEE Trans Knowl Data Eng* (2010) 22(10):1345-59. doi: 10.1109/TKDE.2009.191.

52. Cheplygina V, de Bruijne M, Pluim JPW. Not-So-Supervised: A Survey of Semi-Supervised, Multi-Instance, and Transfer Learning in Medical Image Analysis. *Med Image Anal* (2019) 54:280-96. doi: 10.1016/j.media.2019.03.009.

53. Moan JM, Amdal CD, Malinen E, Svestad JG, Bogsrud TV, Dale E. The Prognostic Role of 18f-Fluorodeoxyglucose Pet in Head and Neck Cancer Depends on Hpv Status. *Radiother Oncol* (2019) 140:54-61. doi: 10.1016/j.radonc.2019.05.019.

54. DAHANCA Radiotherapy Guidelines 2013. Danish Head and Neck Cancer Group (DAHANCA) (2013).

55. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, et al. 3D Slicer as an Image Computing Platform for the Quantitative Imaging Network. *Magn Reson Imaging* (2012) 30(9):1323-41. doi: 10.1016/j.mri.2012.05.001.

56. Yaniv Z, Lowekamp BC, Johnson HJ, Beare R. SimpleITK Image-Analysis Notebooks: A Collaborative Environment for Education and Reproducible Research. *J Digit Imaging* (2018) 31(3):290-303. doi: 10.1007/s10278-017-0037-8.

57. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O, editors. 3d U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In: Ourselin S, Joskowicz L, Sabuncu M, Unal G, Wells W, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. MICCAI 2016. Lecture Notes in Computer Science vol 9901. Cham: Springer (2016). Doi: 10.1007/978-3-319-46723-8_49.

58. Milletari F, Navab N, Ahmadi S-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. 2016 Fourth International Conference on 3D Vision (3DV) (2016):565-71.

59. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv [Preprint] (2014). Available at: https://arxiv.org/abs/1412.6980 (Accessed December 15, 2022).

60. Sørensen T. A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species and Its Application to Analyses of the Vegetation on Danish Commons. *Kongelige Danske Videnskabernes Selskab* (1948) 5(4):1–34.

61. Dice LR. Measures of the Amount of Ecologic Association between Species. *Ecology* (1945) 26(3):297-302. doi: 10.2307/1932409.

62. Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing Images Using the Hausdorff Distance. *IEEE Trans Pattern Anal Mach Intell* (1993) 15(9):850-63. doi: 10.1109/34.232073.

63. Taha AA, Hanbury A. Metrics for Evaluating 3D Medical Image Segmentation: Analysis, Selection, and Tool. *BMC Med Imaging* (2015) 15:29. doi: 10.1186/s12880-015-0068-x.

64. Andrearczyk V, Oreiller V, Vallières M, Castelli J, Elhalawani HM, Jreige M, et al. Automatic Segmentation of Head and Neck Tumors and Nodal Metastases in PET-CT Scans. Proceedings of 3rd International Conference on Medical Imaging with Deep Learning (2020) 121:33-43.

65. Bird D, Scarsbrook AF, Sykes J, Ramasamy S, Subesinghe M, Carey B, et al. Multimodality Imaging with CT, MR and FDG-PET for Radiotherapy Target Volume Delineation in Oropharyngeal Squamous Cell Carcinoma. *BMC Cancer* (2015) 15(1):844. doi: 10.1186/s12885-015-1867-8.

66. Gudi S, Ghosh-Laskar S, Agarwal JP, Chaudhari S, Rangarajan V, Nojin Paul S, et al. Interobserver Variability in the Delineation of Gross Tumour Volume and Specified Organs-at-Risk During IMRT for Head and Neck Cancers and the Impact of FDG-PET/CT on Such Variability at the Primary Site. *J Med Imaging Radiat Sci* (2017) 48(2):184-92. doi: 10.1016/j.jmir.2016.11.003.

67. Randall EK. PET-Computed Tomography in Veterinary Medicine. *Vet Clin North Am Small Anim Pract* (2016) 46(3):515-33. doi: 10.1016/j.cvsm.2015.12.008.

68. Outeiral RR, Bos P, Al-Mamgani A, Jasperse B, Simões R, van der Heide UA. Oropharyngeal Primary Tumor Segmentation for Radiotherapy Planning on Magnetic Resonance Imaging

Using Deep Learning. *Phys Imaging Radiat Oncol* (2021) 19:39-44. doi: 10.1016/j.phro.2021.06.005.

69. Wahid KA, Ahmed S, He R, van Dijk LV, Teuwen J, McDonald BA, et al. Evaluation of Deep Learning-Based Multiparametric MRI Oropharyngeal Primary Tumor Auto-Segmentation and Investigation of Input Channel Effects: Results from a Prospective Imaging Registry. *Clinical Transl Radiat Oncol* (2022) 32:6-14. doi: 10.1016/j.ctro.2021.10.003.

70. Men K, Chen X, Zhang Y, Zhang T, Dai J, Yi J, et al. Deep Deconvolutional Neural Network for Target Segmentation of Nasopharyngeal Cancer in Planning Computed Tomography Images. *Front Oncol* (2017) 7:315. doi: 10.3389/fonc.2017.00315.

71. Li Q, Xu Y, Chen Z, Liu D, Feng S-T, Law M, et al. Tumor Segmentation in Contrast-Enhanced Magnetic Resonance Imaging for Nasopharyngeal Carcinoma: Deep Learning with Convolutional Neural Network. *BioMed Res Int* (2018) 2018:9128527. doi: 10.1155/2018/9128527.

72. Mohammed MA, Abd Ghani MK, Arunkumar N, Mostafa SA, Abdullah MK, Burhanuddin MA. Trainable Model for Segmenting and Identifying Nasopharyngeal Carcinoma. *Comp Electr Eng* (2018) 71:372-87. doi: 10.1016/j.compeleceng.2018.07.044.

73. Vrtovec T, Močnik D, Strojan P, Pernuš F, Ibragimov B. Auto-Segmentation of Organs at Risk for Head and Neck Radiotherapy Planning: From Atlas-Based to Deep Learning Methods. *Med Phys* (2020) 47(9):e929-e50. doi: 10.1002/mp.14320.

ISBN: 978-82-575-2041-0
ISSN: 1894-6402

Norwegian University
of Life Sciences

Postboks 5003
NO-1432 Ås, Norway
+47 67 23 00 00
www.nmbu.no

116005 / ANDVORDGRAFISK.NO