Fully Automated Gross Tumor Volume Delineation From PET in Head and Neck Cancer Using Deep Learning Algorithms

Isaac Shiri, MSc, Hossein Arabi, PhD,* Amirhossein Sanaat, MSc,* Elnaz Jenabi, MD,† Minerva Becker, MD,‡ and Habib Zaidi, PhD*§||¶*

Purpose: The availability of automated, accurate, and robust gross tumor volume (GTV) segmentation algorithms is critical for the management of head and neck cancer (HNC) patients. In this work, we evaluated 3 state-of-the-art deep learning algorithms combined with 8 different loss functions for PET image segmentation using a comprehensive training set and evaluated its performance on an external validation set of HNC patients.

Patients and Methods: ^{18}F-FDG PET/CT images of 470 patients presenting with HNC on which manually defined GTVs serving as standard of reference were used for training (340 patients), evaluation (30 patients), and testing (100 patients from different centers) of these algorithms. PET image intensity was converted to SUVs and normalized in the range (0–1) using the SUV_{max} of the whole data set. PET images were cropped to $12 \times 12 \times 12$ cm^3 subvolumes using isotropic voxel spacing of $3 \times 3 \times 3$ mm^3 containing the whole tumor and neighboring background including lymph nodes. We used different approaches for data augmentation, including rotation (−15 degrees, +15 degrees), scaling (−20%, 20%), random flipping (3 axes), and elastic deformation (sigma = 1 and proportion to deform = 0.7) to increase the number of training sets. Three state-of-the-art networks, including Dense-VNet, NN-UNet, and Res-Net, with 8 different loss functions, including Dice, generalized Wasserstein Dice loss, Dice plus XEnt loss, generalized Dice loss, cross-entropy, sensitivity-specificity, and Tversky, were used. Overall, 28 different networks were built. Standard image segmentation metrics, including Dice similarity, image-derived PET metrics, first-order, and shape radiomic features, were used for performance assessment of these algorithms.

Results: The best results in terms of Dice coefficient (mean ± SD) were achieved by cross-entropy for Res-Net (0.86 ± 0.05; 95% confidence interval [CI], 0.85–0.87), Dense-VNet (0.85 ± 0.058; 95% CI, 0.84–0.86), and Dice plus XEnt for NN-UNet (0.87 ± 0.05; 95% CI, 0.86–0.88). The difference between the 3 networks was not statistically significant ($P > 0.05$). The percent relative error (RE%) of SUV_{max} quantification was less than 5% in networks with a Dice coefficient more than 0.84, whereas a lower RE% (0.41%) was achieved by Res-Net with cross-entropy loss. For maximum 3-dimensional diameter and sphericity shape features, all networks achieved a RE ≤ 5% and ≤10%, respectively, reflecting a small variability.

Conclusions: Deep learning algorithms exhibited promising performance for automated GTV delineation on HNC PET images. Different loss functions performed competitively when using different networks and cross-entropy for Res-Net, Dense-VNet, and Dice plus XEnt for NN-UNet emerged as reliable networks for GTV delineation. Caution should be exercised for clinical deployment owing to the occurrence of outliers in deep learning–based algorithms.

Key Words: PET, segmentation, head and neck, quantification, deep learning

(*Clin Nucl Med* 2021;46: 872–883)

PET using various molecular imaging probles is commonly used in clinical setting for various tasks in clinical oncology, including diagnosis and malignant lesion detection, staging and restaring, and monitoring of response to treatment.[1] Various semiquantitative and quantitative image-derived PET metrics are used in clinical and research settings to complement visual interpretation. This includes simple indices, such as the SUV and advanced quantitative metrics extracted from PET images. Quantification of metabolic and physiological processes in vivo provides valuable information for clinical diagnosis/prognosis of disease.[2] The delineation of gross tumor volume (GTV) to calculate the metabolic tumor volume (MTV) and hence total lesion glycolysis (TLG) or to plan external beam radiation therapy is highly demanded in clinical setting.[3] Manual delineation of the GTV is time-consuming and prone to interobserver/intraobserver variability and depend on physician experience.[4,5] In addition, accurate delineation is challenging owing to the noisy nature, poor spatial resolution, and resulting partial volume effect in PET images.[6–10]

Conventional algorithms, including adaptive iterative thresholding,[11–14] active contours,[15,16] region-growing,[8] k-mean iterative clustering,[17] fuzzy c-mean iterative clustering,[18] Gaussian mixture model,[19] random walk,[20] watershed transform,[21,22] graph-based,[23,24] and Markov random field (MRF)–based[25–27] techniques, have been developed for PET image segmentation. However, deployment of these algorithms in clinical setting faces various challenges as they commonly require user input to define background or foreground of tumors (seed or volume of interest), setting parameters specific to each patient, and prior knowledge regarding the clinical indication, scanner performance, and clinical acquisition and processing protocols.[10] Various strategies have been developed for GTV delineation from PET images; however, conventional algorithms commonly fail to achieve good outcome owing to the heterogeneous anatomical nature of the head and neck region, and the presence neighboring metabolically active regions, such as lymph nodes.[7,28] Providing an automatic, accurate, and robust GTV segmentation algorithm is highly demanded for effective head and neck cancer (HNC) patient management.

Received for publication April 2, 2021; revision accepted May 14, 2021.
From the *Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, Geneva, Switzerland; †Research Centre for Nuclear Medicine, Shariati Hospital, Tehran University of Medical Sciences, Tehran, Iran; ‡Division of Radiology, Geneva University Hospital; §Geneva University Neurocenter, University of Geneva, Geneva, Switzerland; ||Department of Nuclear Medicine and Molecular Imaging, University of Groningen, Groningen, the Netherlands; and ¶Department of Nuclear Medicine, University of Southern Denmark, Odense, Denmark.
Conflicts of interest and sources of funding: This work was supported by the Swiss National Science Foundation under grants SNSF 320030_176052 and 320030_173091/1. None declared to all authors.
Correspondence to: Habib Zaidi, PhD, Division of Nuclear Medicine and Molecular Imaging, Department of Medical Imaging, Geneva University Hospital, 4 Rue Gabrielle-Perret-Gentil, CH-1211 Geneva, Switzerland. E-mail: habib.zaidi@hcuge.ch.
Supplemental digital content is available for this article. Direct URL citation appears in the printed text and is provided in the HTML and PDF versions of this article on the journal's Web site (www.nuclearmed.com).

ISSN: 0363-9762/21/4611–0872
DOI: 10.1097/RLU.0000000000003789

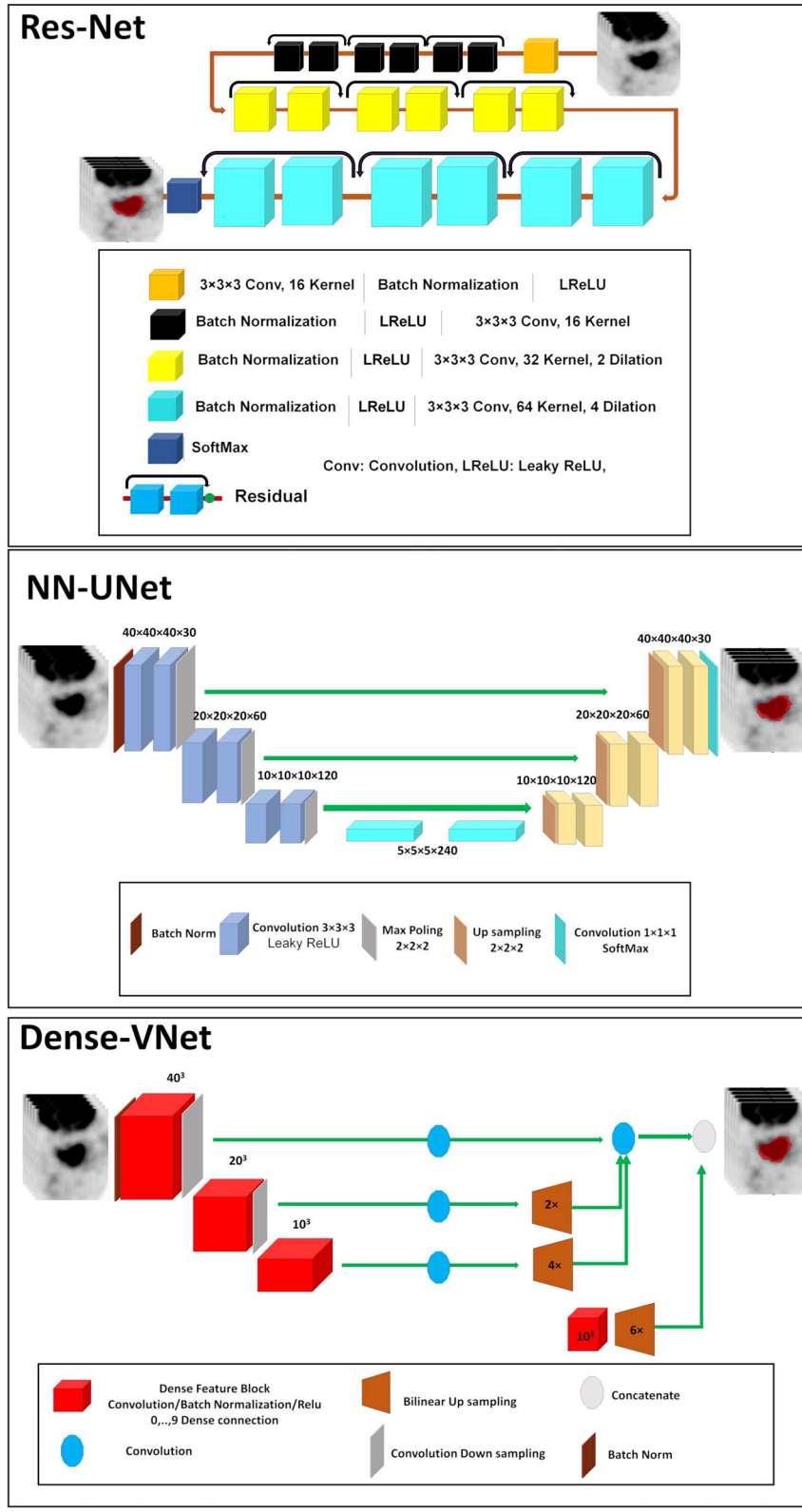

FIGURE 1. Flowcharts of the 3 neural network architectures evaluated in this work, including Res-Net, Dense-VNet, and NN-UNet.

Deep learning is a subfield of machine learning algorithms capable of feature extraction, selection, and classification in 1 step.[29] Machine/deep learning algorithms have been applied in PET for various applications,[30] including image reconstruction,[31] attenuation and scatter correction,[32] image denoising,[33] reduction of acquisition time,[34] and voxel-based dosimetry.[35] Most previous studies demonstrated that deep learning–based algorithms provide equivalent or more reliable results compared with conventional algorithms. A number of machine learning algorithms have been developed for segmentation of medical images.[36–39] In the context of PET image segmentation, K-nearest neighbor,[40] decision tree,[41] support vector machine,[42] and random forest[43] machine learning algorithms have been thoroughly investigated. However, these techniques require handcrafted feature extraction and feature selection, hence limiting their accuracy and robustness.

There is a growing body of literature reporting on the use of deep learning algorithms for PET image segmentation. Leung et al[44] applied a 2-dimensional (2D) U-Net architecture to simulated lung PET images to fine tune the algorithm for patients' images. This approach resulted in better accuracy for small tumors segmentation and more reliable and generalizable results for multiscanner data sets. Fu et al[45] proposed a multimodal attention module using a U-Net network backbone to segment lung tumors through exploiting both physiological and anatomical information. Another study performed by Zhong et al[46] for simultaneous segmentation of non–small cell lung carcinoma tumors from PET and CT images reported that their proposed 3-dimensional (3D) deep learning algorithms outperformed conventional algorithms. Likewise, other work focusing on cervical cancer quantification from PET/CT images using a combination of anatomical prior and deep learning algorithm reported improvement in segmentation accuracy.[47]

Among the studies focusing mainly on PET images segmentation of HNC malignant lesions, Guo et al[48] developed a 3D deep convolutional network for multimodal (PET and CT) image segmentation targeting potential applications in radiotherapy planning. Jin et al[49] set up a fully automatic segmentation for esophagus GTV delineation using 2 streams chained deep learning fusion of PET and CT images. Andrearczyk et al[50] investigated automatic HNC tumors and metastatic lesion segmentation using 2D and 3D V-Net from PET and CT images separately and also in a multichannel manner of early and late fusion. They reported that multichannel input improved segmentation accuracy in HNC patients. Afshari et al[51] proposed a weakly supervised convolutional neural network with significantly improved segmentation performance. Huang

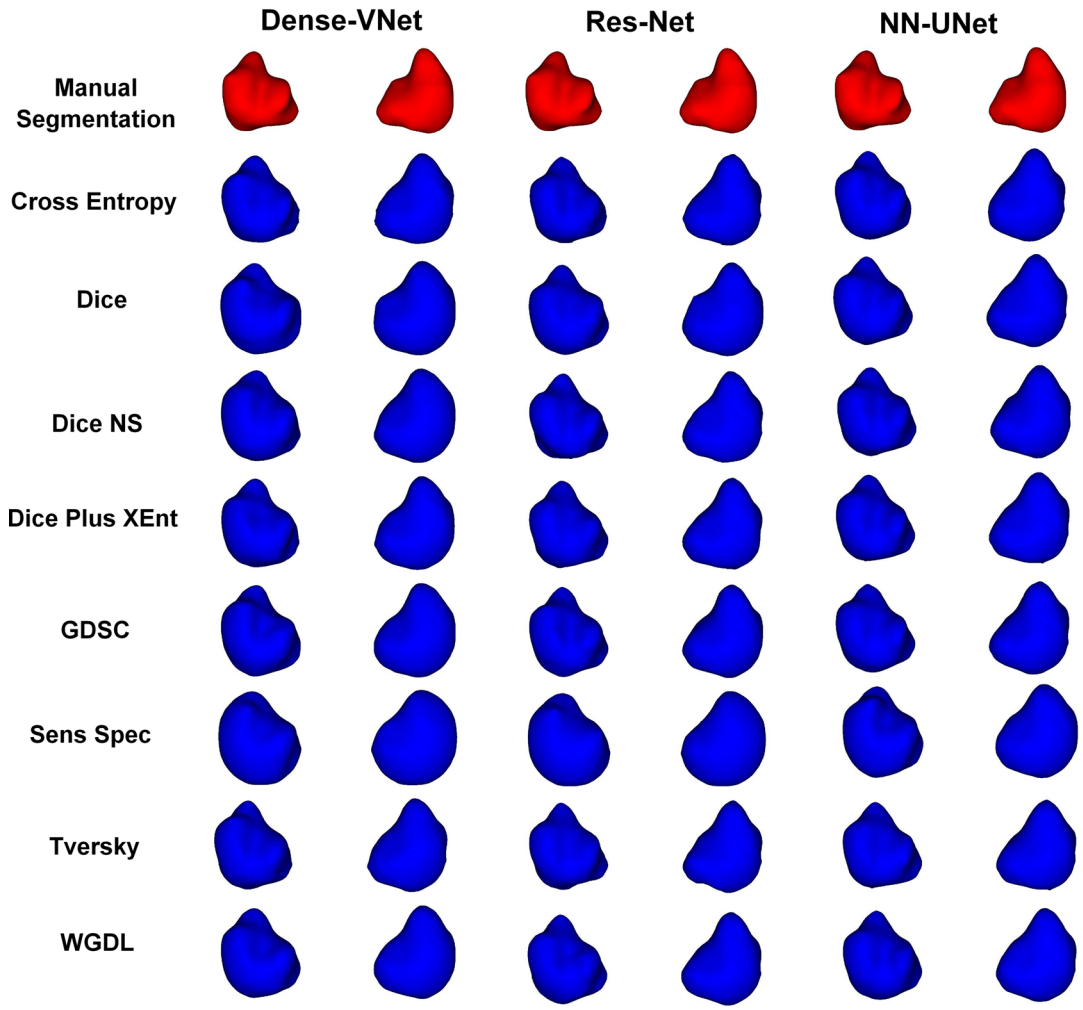

FIGURE 2. Two different 3D views of manual and automated segmentations of malignant lesions using different networks and loss functions for a clinical study from the HUG database.

et al[52] performed a dual-center study for segmentation from PET/CT images using a deep learning algorithm, emphasizing efficiency and generalizability of the segmentation process.

A number of studies, task groups,[9,53] and international medical imaging contests[54] have been undertaken to address the challenges of PET image segmentation. In this work, we evaluated 3 state-of-the-art deep learning PET image segmentation algorithms combined with 8 different loss functions using a large training set and evaluated their performance on external validation sets of HNC patients.

PATIENTS AND METHODS

PET/CT Data Acquisition and Description

This study was conducted on PET images of HNC patients obtained from the open-source database (430 patients) of The Cancer Imaging Archive (TCIA)[55–59] and data gathered at Geneva University Hospital (HUG) (40 patients). Some patients were excluded from the TCIA data set owing to some technical issues, such as image noise, artifacts, absence of images, and image misregistration. The acquisition and reconstruction protocols adopted for the TCIA data sets are given in Prior et al,[55] Gevaert et al,[56] Clark et al,[57] Bakr et al,[58] Vallières et al.[59] The following protocol was used for the acquisition of the HUG data set.

The injected ^{18}F-FDG activity was in the range (119–276 MBq; mean, 201 MBq), whereas the time between injection and data acquisition was in the range (37–117 minutes; mean, 88 minutes). The ordered subset expectation maximization iterative algorithm with time-of-flight and point spread function modeling was used for PET image reconstruction. CT-based attenuation and Compton scatter correction was performed for all PET studies.

Manual Image Segmentation and Preprocessing

All GTVs were manually delineated on PET images by an experienced nuclear medicine physician using the OSIRIX software.[60] PET image intensities were converted to SUV and normalized between 0 and 1 using the maximum value of the used data sets. To produce a rotationally invariant data set, preprocessing of PET images was performed through interpolation to isotropic voxel spacing of $3 \times 3 \times 3$ mm^3. To create uniform data sets in terms of matrix size and voxel size and to handle computational barriers, we cropped the PET images to $12 \times 12 \times 12$ cm subvolumes containing the whole tumor and background, including the lymph nodes.

Neural Networks

We implemented 3 state-of-the-art deep learning–based segmentation algorithms, including Res-Net,[61] Dense-VNet,[62] and

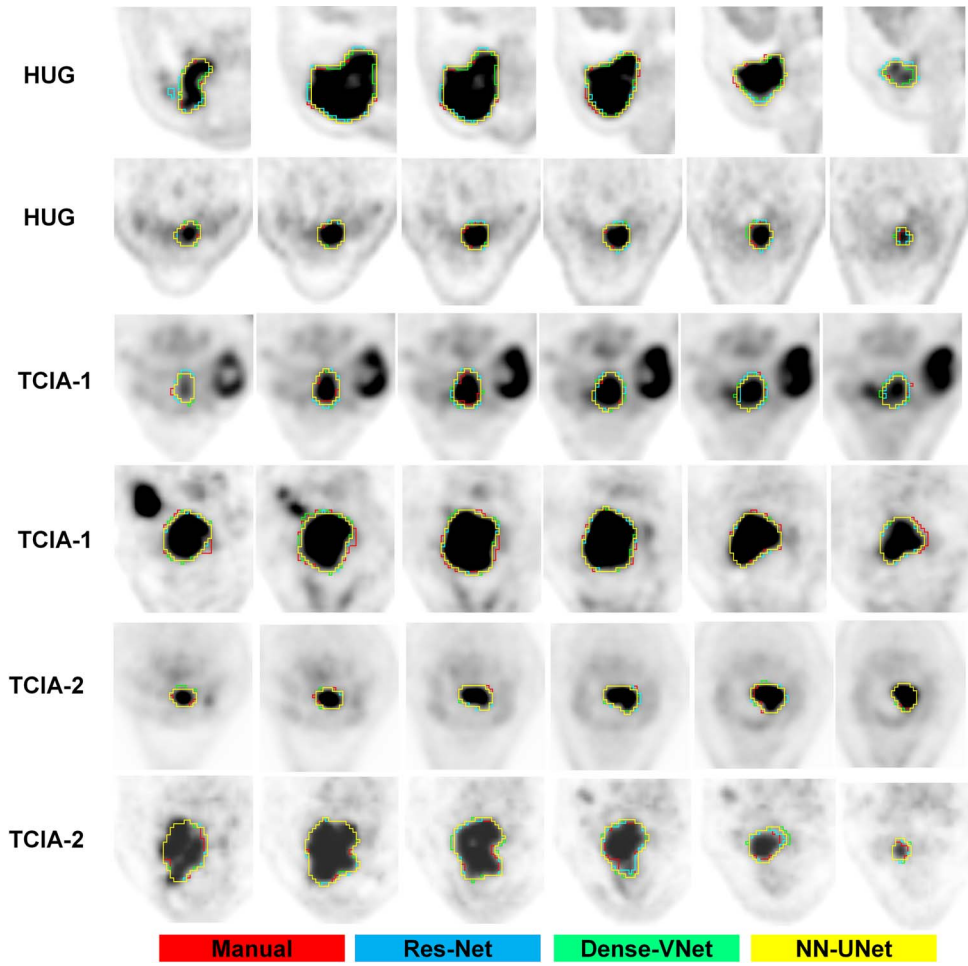

FIGURE 3. Two-dimensional views of manual delineation of lesions and automated segmentation output of the different networks Res-Net (cross entropy), NN-UNET (Dice plus XEnt), and Dense-VNet (cross entropy) for different clinical studies from TCIA database and our institution.

NN-UNet.[37] The architecture of each network is presented in Figure 1. Res-Net is composed of 20 layers, including three 6-level layers with different dilation factors of 0, 2, and 4 to extract the different image features, where every 2 layers are connected by a residual connection.[61] Dense-VNet is a fully convolutional network with 3 dense feature stack blocks (for information flow and decreasing the network parameters), downsampled to the next one and having skip connections with upsample blocks.[62] NN-UNet or no-new-U-Net is a modified standard encoder-decoder network (U-Net architecture-based[63]) with skip connections.[37]

Loss Functions

As loss functions in deep learning algorithms instigate networks for the training process, the choice of the loss function is vital for deep semantic segmentation, which determines the performance of image segmentation.[64] We implemented 8 different well-known loss functions for each network, including Dice, Dice no-square (Dice NS), Dice plus XEnt, Cross-Entropy, generalized Dice loss (GDSC), sensitivity-specificity, Tversky, and generalized Wasserstein Dice loss (WGDL). Further details about loss functions are provided in the Supplemental Material, http://links.lww.com/CNM/A336.

Training

Altogether, 28 different networks were built by combining 3 networks and 8 loss functions. ^{18}F-FDG PET images of 470 patients presenting with HNC, on which manually defined (reference) GTVs were used as training (340 patients from TCIA), evaluation (30 patients from TCIA), and external validation sets (100 patients from different centers) for these algorithms. PET images in SUV units were fed as input to the networks to generate the corresponding binary masks of GTVs.

Data Augmentation

To increase the number of training sets to avoid overfitting and increase the generalizability of networks, we used different data augmentation approaches, including rotation (−15 degrees, +15 degrees), scaling (−20%, 20%), random flipping (3 axes), and elastic deformation (sigma = 1 and proportion to deform = 0.7).

Quantitative Evaluation

All evaluations were performed using 100 patients from different centers, referred to as TCIA 1 (30 patients), TCIA 2 (30 patients), and HUG (40 patients). The evaluation was performed with respect to the manual delineations considered as standard of

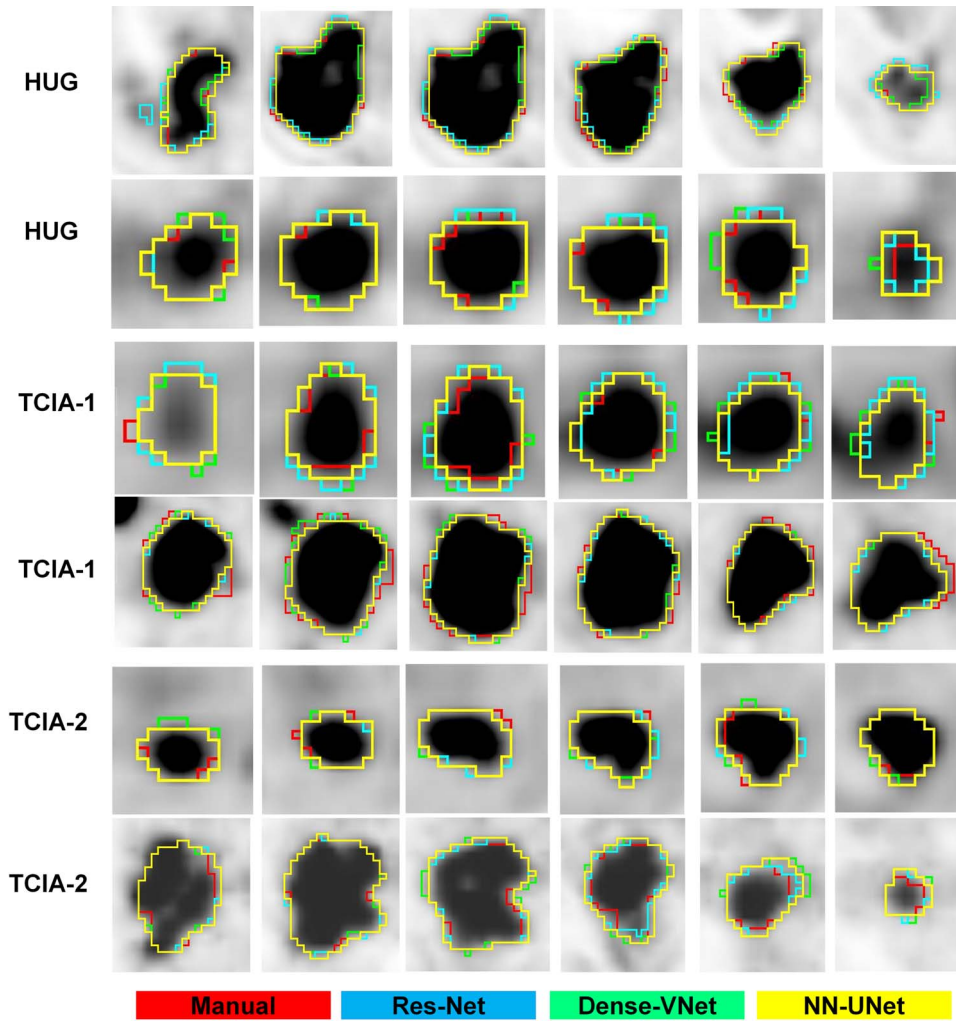

FIGURE 4. Magnified 2D views of manual delineation of lesions and segmentation output of the different networks Res-Net (cross entropy), N-N-UNET (Dice plus XEnt), and Dense VNet (cross entropy) for different clinical studies shown in Figure 3.

TABLE 1. Summary of Quantitative Image Segmentation Performance Metrics (Mean ± SD) for Different Networks and Loss Functions

Network	Loss Function	Dice	Jaccard	False-Negative	False-Positive	Volume Similarity	Mean Surface Distance, mm	SD Surface Distance, mm	Max Surface Distance, mm
Dense-VNET	Cross entropy	0.85 ± 0.058	0.75 ± 0.08	0.12 ± 0.10	0.16 ± 0.10	0.02 ± 0.11	0.17 ± 0.07	0.39 ± 0.09	2.0 ± 0.76
	Dice	0.78 ± 0.091	0.64 ± 0.12	0.05 ± 0.09	0.32 ± 0.15	0.17 ± 0.15	0.26 ± 0.12	0.49 ± 0.12	2.5 ± 0.65
	Dice NS	0.79 ± 0.083	0.66 ± 0.11	0.05 ± 0.09	0.30 ± 0.14	0.16 ± 0.13	0.25 ± 0.1	0.47 ± 0.11	2.3 ± 0.63
	Dice plus XEnt	0.82 ± 0.068	0.71 ± 0.09	0.07 ± 0.10	0.24 ± 0.12	0.10 ± 0.12	0.2 ± 0.087	0.43 ± 0.12	2.2 ± 0.82
	GDSC	0.78 ± 0.09	0.64 ± 0.12	0.06 ± 0.11	0.31 ± 0.15	0.16 ± 0.16	0.27 ± 0.12	0.50 ± 0.12	2.6 ± 0.81
	Sens-spec	0.71 ± 0.091	0.55 ± 0.11	0.03 ± 0.11	0.42 ± 0.14	0.25 ± 0.16	0.37 ± 0.12	0.61 ± 0.11	2.8 ± 0.59
	Tversky	0.85 ± 0.065	0.75 ± 0.09	0.12 ± 0.12	0.15 ± 0.10	0.01 ± 0.12	0.17 ± 0.085	0.39 ± 0.13	2.0 ± 0.85
	WGDL	0.76 ± 0.09	0.62 ± 0.12	0.03 ± 0.08	0.35 ± 0.14	0.20 ± 0.14	0.29 ± 0.11	0.51 ± 0.12	2.5 ± 0.56
Res-Net	Cross entropy	0.86 ± 0.05	0.76 ± 0.08	0.11 ± 0.10	0.14 ± 0.09	0.01 ± 0.11	0.15 ± 0.071	0.38 ± 0.12	1.9 ± 0.92
	Dice	0.85 ± 0.06	0.75 ± 0.08	0.07 ± 0.09	0.19 ± 0.11	0.07 ± 0.11	0.16 ± 0.074	0.39 ± 0.10	2.0 ± 0.70
	Dice NS	0.86 ± 0.06	0.75 ± 0.09	0.08 ± 0.10	0.17 ± 0.11	0.05 ± 0.12	0.16 ± 0.076	0.38 ± 0.10	2.0 ± 0.80
	Dice plus XEnt	0.86 ± 0.05	0.76 ± 0.07	0.08 ± 0.08	0.18 ± 0.10	0.05 ± 0.1	0.15 ± 0.059	0.38 ± 0.08	2.0 ± 0.84
	GDSC	0.85 ± 0.05	0.74 ± 0.08	0.08 ± 0.09	0.19 ± 0.11	0.06 ± 0.11	0.17 ± 0.068	0.4 ± 0.098	2.1 ± 0.91
	Sens-spec	0.72 ± 0.09	0.57 ± 0.11	0.03 ± 0.10	0.41 ± 0.14	0.24 ± 0.16	0.35 ± 0.12	0.58 ± 0.12	2.7 ± 0.68
	Tversky	0.86 ± 0.05	0.76 ± 0.08	0.11 ± 0.11	0.15 ± 0.09	0.02 ± 0.11	0.15 ± 0.069	0.38 ± 0.09	2.0 ± 0.79
	WGDL	0.86 ± 0.05	0.75 ± 0.08	0.07 ± 0.08	0.18 ± 0.11	0.06 ± 0.11	0.16 ± 0.067	0.38 ± 0.08	1.9 ± 0.76
NN-UNet	Cross entropy	0.84 ± 0.06	0.73 ± 0.09	0.07 ± 0.09	0.21 ± 0.13	0.07 ± 0.12	0.18 ± 0.08	0.42 ± 0.16	2.2 ± 1.3
	Dice	0.85 ± 0.06	0.74 ± 0.09	0.07 ± 0.09	0.19 ± 0.12	0.07 ± 0.12	0.17 ± 0.07	0.40 ± 0.12	2.0 ± 0.99
	Dice NS	0.85 ± 0.06	0.74 ± 0.09	0.06 ± 0.08	0.20 ± 0.12	0.08 ± 0.11	0.17 ± 0.07	0.39 ± 0.10	1.9 ± 0.83
	Dice plus XEnt	0.87 ± 0.05	0.77 ± 0.07	0.09 ± 0.08	0.15 ± 0.10	0.03 ± 0.1	0.15 ± 0.06	0.38 ± 0.11	2.0 ± 1.1
	GDSC	0.85 ± 0.06	0.75 ± 0.09	0.07 ± 0.09	0.19 ± 0.11	0.06 ± 0.11	0.17 ± 0.07	0.39 ± 0.12	2.0 ± 0.93
	Sens-spec	0.72 ± 0.09	0.58 ± 0.12	0.03 ± 0.11	0.39 ± 0.15	0.23 ± 0.16	0.34 ± 0.13	0.57 ± 0.13	2.8 ± 0.79
	Tversky	0.86 ± 0.06	0.76 ± 0.08	0.07 ± 0.08	0.18 ± 0.11	0.05 ± 0.11	0.16 ± 0.07	0.38 ± 0.11	2.0 ± 1.0
	WGDL	0.85 ± 0.06	0.75 ± 0.09	0.06 ± 0.08	0.20 ± 0.12	0.08 ± 0.11	0.16 ± 0.07	0.39 ± 0.11	2.0 ± 0.87

Sens-spec, sensitivity-specificity.

TABLE 2. Lower and Upper Bound of 95% CI of Quantitative Image Segmentation Performance Metrics for Different Networks and Loss Functions

Method	Loss Function	Dice	Jaccard	False-Negative	False-Positive	Volume Similarity	Mean Surface Distance, mm	SD Surface Distance, mm	Max Surface Distance, mm
Dense-VNET	Cross entropy	0.84–0.86	0.73–0.76	0.09–0.14	0.14–0.18	0.01–0.04	0.15–0.18	0.37–0.41	1.90–2.20
	Dice	0.76–0.80	0.62–0.67	0.03–0.07	0.29–0.35	0.14–0.2	0.24–0.29	0.47–0.51	2.40–2.60
	Dice NS	0.77–0.81	0.64–0.68	0.03–0.06	0.28–0.33	0.13–0.19	0.23–0.27	0.45–0.49	2.20–2.50
	Dice plus XEnt	0.81–0.84	0.69–0.73	0.05–0.09	0.21–0.26	0.07–0.12	0.18–0.22	0.41–0.46	2.10–2.40
	GDSC	0.76–0.79	0.62–0.67	0.04–0.08	0.28–0.34	0.13–0.19	0.25–0.29	0.48–0.53	2.40–2.80
	Sens-spec	0.69–0.72	0.53–0.57	0.01–0.05	0.39–0.45	0.22–0.28	0.34–0.39	0.58–0.63	2.70–2.90
	Tversky	0.84–0.86	0.73–0.77	0.09–0.14	0.13–0.17	0.01–0.04	0.15–0.18	0.37–0.42	1.80–2.20
	WGDL	0.74–0.78	0.59–0.64	0.02–0.05	0.33–0.38	0.17–0.23	0.27–0.31	0.49–0.54	2.40–2.60
Res-Net	Cross entropy	0.85–0.87	0.75–0.78	0.09–0.13	0.12–0.16	0.01–0.03	0.14–0.17	0.35–0.40	1.80–2.10
	Dice	0.84–0.86	0.73–0.76	0.05–0.09	0.17–0.21	0.04–0.09	0.15–0.18	0.37–0.41	1.80–2.10
	Dice NS	0.84–0.87	0.73–0.77	0.06–0.11	0.15–0.20	0.02–0.07	0.14–0.17	0.36–0.40	1.90–2.20
	Dice plus XEnt	0.85–0.87	0.74–0.77	0.06–0.10	0.16–0.20	0.03–0.07	0.14–0.17	0.36–0.40	1.80–2.20
	GDSC	0.84–0.86	0.72–0.76	0.06–0.10	0.17–0.22	0.04–0.08	0.16–0.18	0.38–0.42	1.90–2.30
	Sens-spec	0.70–0.74	0.54–0.59	0.01–0.05	0.38–0.43	0.21–0.27	0.33–0.37	0.56–0.60	2.60–2.80
	Tversky	0.85–0.87	0.74–0.78	0.08–0.13	0.13–0.17	0.00–0.04	0.14–0.17	0.36–0.39	1.80–2.10
	WGDL	0.85–0.87	0.74–0.77	0.06–0.09	0.16–0.20	0.04–0.08	0.14–0.17	0.36–0.39	1.80–2.10
NN-UNet	Cross entropy	0.83–0.85	0.71–0.75	0.05–0.09	0.18–0.23	0.05–0.10	0.16–0.20	0.39–0.45	1.90–2.40
	Dice	0.84–0.86	0.73–0.76	0.05–0.09	0.17–0.22	0.04–0.09	0.15–0.18	0.37–0.42	1.90–2.20
	Dice NS	0.84–0.86	0.73–0.76	0.04–0.08	0.18–0.23	0.06–0.10	0.15–0.18	0.37–0.41	1.80–2.10
	Dice plus XEnt	0.86–0.88	0.75–0.78	0.07–0.11	0.13–0.17	0.01–0.05	0.13–0.16	0.36–0.40	1.80–2.20
	GDSC	0.84–0.86	0.73–0.77	0.05–0.09	0.17–0.21	0.04–0.09	0.15–0.18	0.37–0.41	1.80–2.20
	Sens-spec	0.70–0.74	0.55–0.60	0.01–0.06	0.36–0.42	0.20–0.26	0.32–0.37	0.55–0.60	2.60–2.90
	Tversky	0.85–0.87	0.74–0.78	0.06–0.09	0.15–0.20	0.03–0.08	0.14–0.17	0.36–0.40	1.80–2.20
	WGDL	0.84–0.86	0.73–0.76	0.04–0.08	0.18–0.22	0.05–0.10	0.15–0.18	0.37–0.41	1.80–2.10

Sens-spec, sensitivity-specificity.

reference. Standard segmentation metrics, including Dice coefficient, Jaccard, false-negative rate, false-positive rate, volume similarity, mean, and standard deviation of surface distance, were calculated (see Supplemental Material for equations, http://links.lww.com/CNM/A336).

Metabolic Activity Intensity and Shape Analysis

We calculated conventional clinically relevant image-derived PET metrics, including SUV_{max}, SUV_{mean}, and SUV_{median}. In addition to conventional clinical PET quantification parameters, we extracted first-order radiomic features, including 10 and 90 percentile, energy, interquartile range, kurtosis, mean absolute deviation, rang, robust mean absolute deviation, root mean squared, total energy, and variance. The shape radiomic features include elongation, flatness, least axis length, major axis length, maximum 2D diameter column, maximum 2D diameter row, maximum 2D diameter slice, maximum 3D diameter, minor axis length, sphericity, surface area, and surface volume ratio (see Supplemental Material, http://links.lww.com/CNM/A336). All feature extractions were performed according to the image biomarker standardization initiative.[65,66] We calculated the mean relative error with respect to manual segmentation using the following formula:

$$\text{Relative Error (\%)} = (\text{Predicted Segmentation} - \text{Manual Segmentation}) / \text{Manual Segmentation} \times 100\%$$

Statistical Analysis

We compared the different networks using Student *t* test statistical analysis and reported the mean ± SD and 95% confidence interval (CI) for the different metrics. All statistical analyses were performed using the *R* software.

RESULTS

Figure 2 presents representative examples of 2 different views of 3D rendered volumes of GTVs for different networks and loss functions along with the manual GTV segmentation of each tumor for a clinical study. Supplemental Figures 1–3, http://

FIGURE 5. Comparison of different models (*P* values) in terms of Dice coefficient. Manual segmentation was used as the criterion standard.

links.lww.com/CNM/A336, present additional examples of 3D GTVs of different patients. Figure 3 illustrates 2D axial views of different patients from the external validation set. Figure 4 depicts a zoomed version of the GTVs shown in Figure 3. As shown in both Figures, the segmentations generated by the different networks are in good agreement with manual segmentations defined on GTVs of malignant lesions presenting with different size, texture, and contrast. Supplemental Figures 4–30, http://links.lww.com/CNM/A336, illustrated the outcome of segmentations achieved by the various networks for different patients from the external validation set.

Tables 1 and 2 summarize PET image segmentation performance metrics (mean ± SD and 95% CI) for different networks and loss functions. A comparison between the various networks in terms of Dice coefficient is illustrated in Figure 5. It can be seen from the results reported in Tables 1 and 2 that the cross-entropy loss function yielded the highest Dice coefficient (mean ± SD) (0.85 ± 0.05; 95% CI, 0.84–0.86) and Jaccard index (0.75 ± 0.08; 95% CI, 0.73–0.76) and lowest surface distances. Tversky loss provided almost the same results with Dice coefficient of (0.85 ± 0.06; 95% CI, 0.84–0.86) and Jaccard index (0.75 ± 0.09; 95% CI, 0.73–0.77). There is no proof of statistically significant difference between cross-entropy and Tversky ($P = 0.70$), and these 2 loss functions are significantly outperformed by others for the Dens-VNet network ($P < 0.05$).

For Res-Net network, cross-entropy loss resulted in the highest performance in term of Dice coefficient (0.86 ± 0.05; 95% CI, 0.85–0.87) and Jaccard index (0.76 ± 0.08; 95% CI, 0.75–0.78). In Res-Net, all loss functions had almost similar performance with no proof of statistically significant difference between them, except the sensitivity-specificity loss function, which showed the most unsatisfactory results (0.72 ± 0.09; 95% CI, 0.70–0.74). It can be seen from Tables 1 and 2 that, for NN-UNet, the Dice plus XEnt loss function yielded the highest Dice coefficient (0.87 ± 0.05; 95% CI, 0.86–0.88) and Jaccard index (0.77 ± 0.07; 95% CI, 0.75–0.78) followed by Tversky with a Dice coefficient of 0.86 ± 0.06 (95% CI, 0.85–0.87) and Jaccard index of 0.76 ± 0.08 (95% CI, 0.74–0.78). There is no proof of statistically significant difference between the 2 loss functions ($P = 0.48$). Except the sensitivity-specificity loss function, the remaining loss functions achieved the same level of accuracy.

It was observed that the sensitivity-specificity loss function resulted in the lowest performance for the different networks.[67]

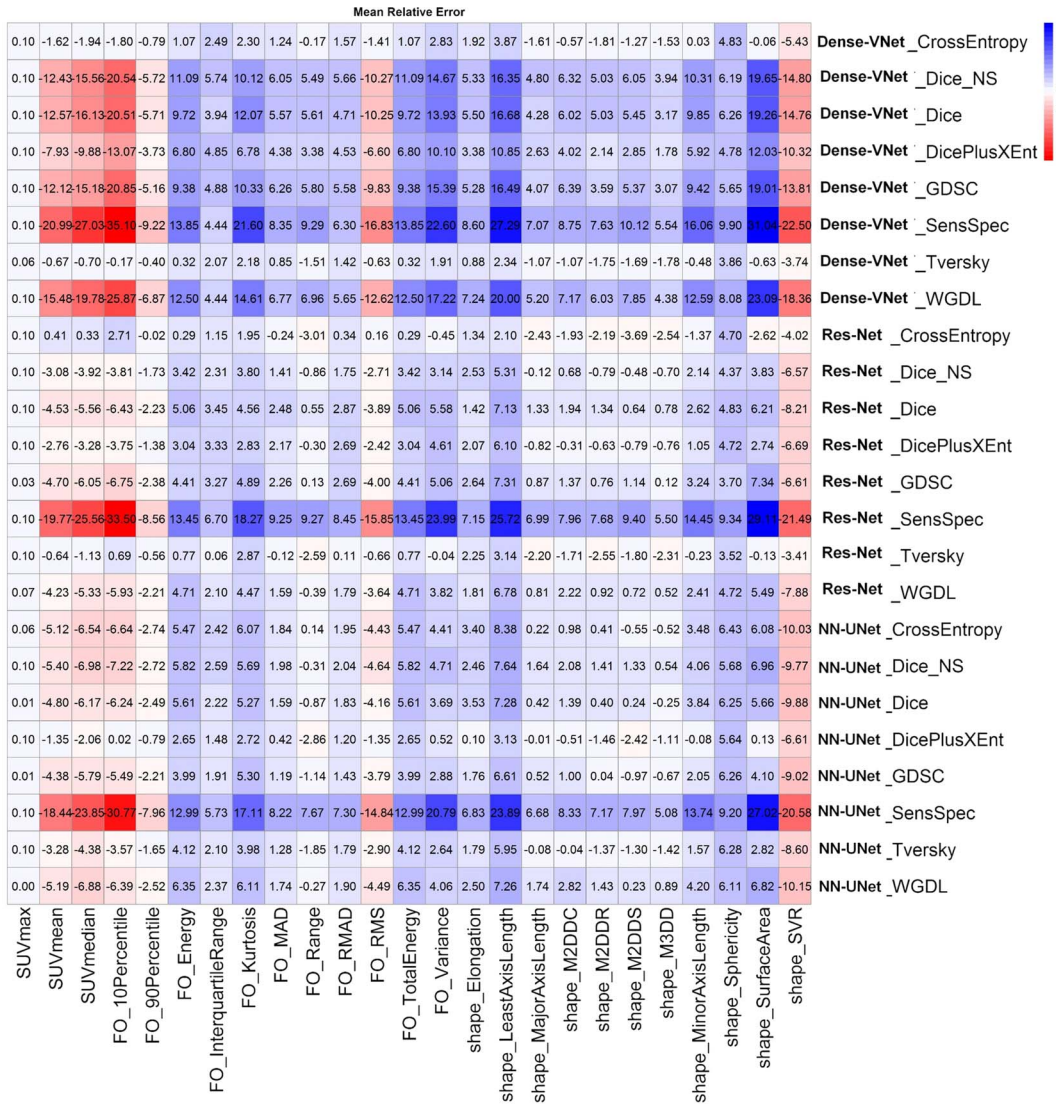

FIGURE 6. Mean relative error (MRE %) of radiomic features for the different networks and loss functions.

The different loss functions performed competitively for the different networks. The best results were achieved by cross-entropy for Res-Net and Dense-VNet and Dice plus XEnt for NN-UNet, where the differences between the 3 networks were not statistically significant.

The results of conventional image-derived PET metrics and first-order and shape features percent relative error for different networks and loss functions are depicted in Figure 6. This reveals that almost all networks correctly segmented the region contains the maximum value of GTV as RE for SUV_{max} is less than 1%. The percent relative error of SUV_{max} was less than 5% in networks with Dice coefficients more than 0.84. A low RE% (0.41%) was achieved by Res-Net with cross-entropy loss. For shape features, maximum 3D diameter and sphericity achieved REs ≤5% and ≤10%, respectively, which is considered very small with small variability in typical radiomics studies.

Figure 7 presents representative outliers where the investigated networks failed to segment properly the GTVs. As shown in this figure, the low uptake of tumors, high uptake in the background, and irregular and sparse shape of the GTVs led to outliers. The frequency of outliers' occurrence in less than 5% (4 cases) of the total number of cases. Additional examples of outliers are presented in Supplemental Figures 31–57, http://links.lww.com/CNM/A336.

DISCUSSION

This work set out to assess the potential of fully automated GTV delineation from PET images in HNC patients using deep learning algorithms. The present study was designed to assess 3 state-of-the-art image segmentation algorithms combined with various popular loss functions and evaluate their performance on external validation data sets using well-established metrics. The results of this study indicated that the different loss functions performed competitively for the different networks. The best results were also achieved by cross-entropy for Res-Net and Dense-VNet and Dice plus XEnt for NN-UNet. The differences between the 3 networks were not statistically significant.

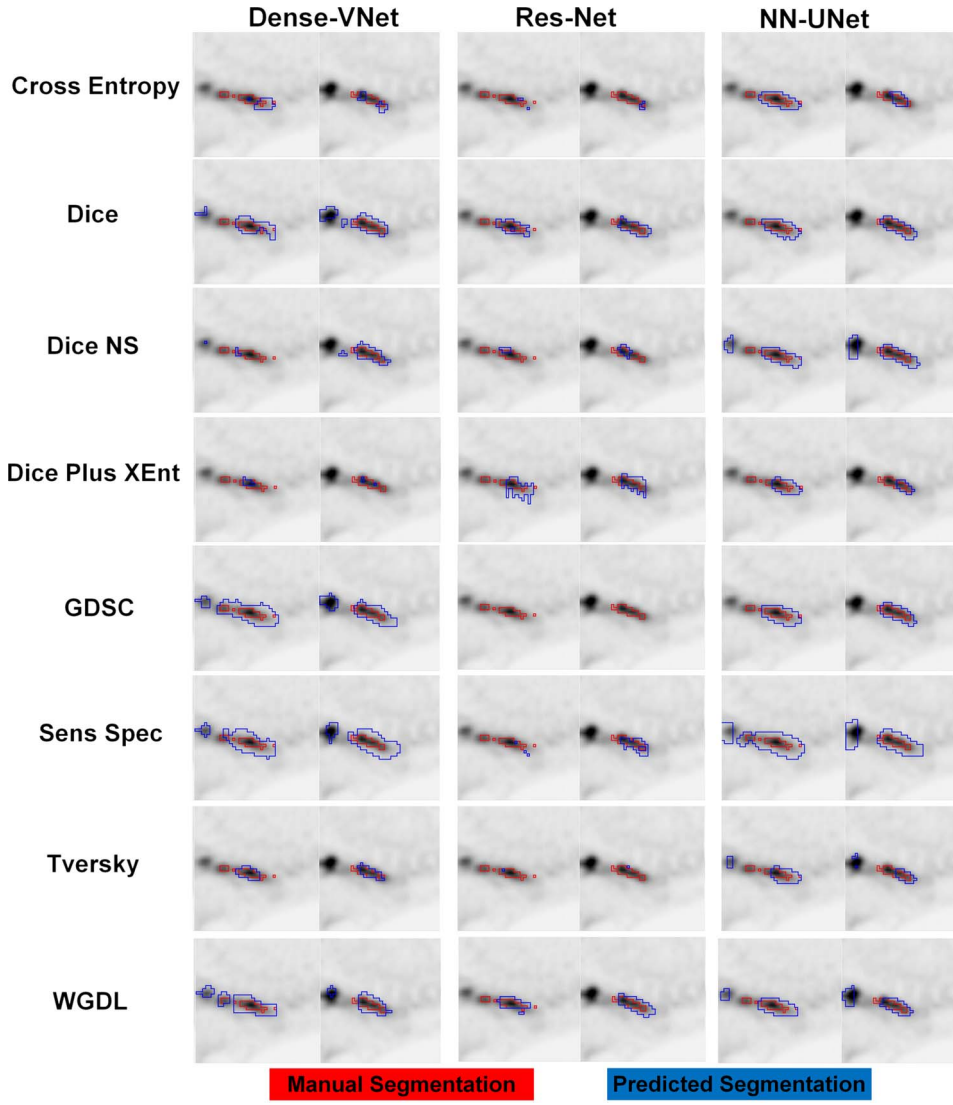

FIGURE 7. Two-dimensional views of manual and automated segmentations achieved by the different networks for different cases where failure was observed resulting in outliers.

The Medical Image Computing and Computer Assisted Intervention challenge[54] has been conducted to address the potential of automated PET image segmentation algorithms. In this challenge, the guidelines suggested by the American Association of Physicists in Medicine task group 211 were cautiously followed.[9] This study included 176 PET images of simulated, experimental phantom, and clinical studies that were divided into 17 training and 157 test data sets. Various strategies were evaluated under this framework where convolutional neural network–based algorithms yielded the highest performance (Dice = 0.80) significantly outperforming conventional techniques (9 of 12), including K-means (Dice = 0.79), Gaussian mixture model (Dice = 0.78), and fuzzy C-means (Dice = 0.73) algorithms.

Afshari et al[51] designed a new loss function that dynamically combines supervised and unsupervised components inspired from previous segmentation techniques.[68] The results achieved by this algorithm demonstrated improvement of the Dice index by 30% compared with the weakly supervised algorithm trained by only bounding boxes. Andrearczyk et al[50] evaluated 2D and 3D V-Net on PET and CT images separately and in a multichannel approach on 202 HNC patients from 4 centers and evaluated their performance using the one-leave-center-out method. They reported Dice coefficients of 0.48 and 0.58 for CT and PET-only images, respectively, which improved to 0.59 and 0.60 for late fusion approaches for 2D and 3D V-Net algorithms, respectively.

Huang et al[52] designed a dual-center study for lesion segmentation from PET/CT images of HNC patients using a U-Net architecture. The training and evaluation involved 22 patients with one-leave-out approach using 2-channel input comprising PET and CT images simultaneously. They reported a Dice coefficient of 0.73 for GTV delineation. Guo et al[48] came up with a 3D multimodality dense net for PET and CT image segmentation using 140/35 clinical studies for training/validation with further testing on an additional external set consisting of 75 HNC patients. They reported a Dice of 0.73 for the proposed architecture, thus outperforming 3D-UNet (Dice = 0.71) when using multimodality images, whereas the Dice was 0.32 and 0.67 when using only CT and PET images, respectively. Our work took advantage of the dense block implemented in Dense-VNet architectures to achieve a high Dice value compared with previous PET-only studies. Jin et al[49] developed a deep learning–based fully automated GTV segmentation of esophagus cancer using 2 streams chained fusion of PET and CT images. Their method was implemented using 110 clinical studies and evaluated using a 5-fold cross-validation scheme. They achieved a Dice of 0.76 ± 0.13, thus outperforming 3D Dense UNet (0.74 ± 0.16). Previous studies highlighted the complementary contribution of information from CT images to correctly delineate the GTVs.[48,49,52]

In this work, we evaluated different network architectures and loss functions trained on an augmented data set and evaluated on unseen external validation data set of HNC patients gathered from open-access database and from our center. We achieved the highest Dice coefficient (0.87) for PET-only images in the external validation set. Despite this high Dice score, we observed some outliers where the networks failed to properly delineate the GTVs owing to the black-box nature of deep learning algorithms. This is mainly caused by the low uptake of tumors, high uptake in the background, and irregular/sparse shape of tumors. In some outlier cases, the predicted segmentation was extended to the background, which could be handled by semiautomated approaches. We also developed different networks, which achieved various degrees of success and variable performance for different patients. We plan to use STAPLE or voting algorithms to provide more robust segmentation algorithms. Another limitation of this study was the use of PET-only images. Further work should focus on the incorporation of anatomical/structural information available from concurrent CT or MR images.

CONCLUSIONS

We assessed the performance of various deep neural networks for GTV delineation from PET images in HNC patients. Deep learning algorithms exhibited promising performance for automated GTV delineation on HNC PET images. The results demonstrated that the different loss functions performed competitively for the different networks, where cross-entropy for Res-Net and Dense-VNet and Dice plus XEnt for NN-UNet emerged as the most promising for GTV delineation. However, caution is recommended when considering deployment in the clinic of deep learning–based segmentation algorithms owing to the presence of outliers.

REFERENCES

1. Rohren EM, Turkington TG, Coleman RE. Clinical applications of PET in oncology. *Radiology*. 2004;231:305–332.
2. Lammertsma AA. Forward to the past: the case for quantitative PET imaging. *J Nucl Med*. 2017;58:1019–1024.
3. Zaidi H, Karakatsanis N. Towards enhanced PET quantification in clinical oncology. *Br J Radiol*. 2018;91:20170508.
4. Norouzi A, Rahim MSM, Altameem A, et al. Medical image segmentation methods, algorithms, and applications. *IETE Technical Review*. 2014;31:199–213.
5. Patil DD, Deore SG. Medical image segmentation: a review. *Int J Comput Sci Mob Comput*. 2013;2:22–27.
6. Xu Z, Gao M, Papadakis GZ, et al. Joint solution for PET image segmentation, denoising, and partial volume correction. *Med Image Anal*. 2018;46:229–243.
7. Foster B, Bagci U, Mansoor A, et al. A review on segmentation of positron emission tomography images. *Comput Biol Med*. 2014;50:76–96.
8. Day E, Betler J, Parda D, et al. A region growing method for tumor volume segmentation on PET images for rectal and anal cancer patients. *Med Phys*. 2009;36:4349–4358.
9. Hatt M, Lee JA, Schmidtlein CR, et al. Classification and evaluation strategies of auto-segmentation approaches for PET: report of AAPM task group no. 211. *Med Phys*. 2017;44:e1–e42.
10. Zaidi H, El Naqa I. PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques. *Eur J Nucl Med Mol Imaging*. 2010;37:2165–2187.
11. Drever L, Roa W, McEwan A, et al. Iterative threshold segmentation for PET target volume delineation. *Med Phys*. 2007;34:1253–1265.
12. Jentzen W, Freudenberg L, Eising EG, et al. Segmentation of PET volumes by iterative image thresholding. *J Nucl Med*. 2007;48:108–114.
13. Vees H, Senthamizhchelvan S, Miralbell R, et al. Assessment of various strategies for ^{18}F-FET PET-guided delineation of target volumes in high-grade glioma patients. *Eur J Nucl Med Mol Imaging*. 2009;36:182–193.
14. Brambilla M, Matheoud R, Basile C, et al. An adaptive thresholding method for BTV estimation incorporating PET reconstruction parameters: a multicenter study of the robustness and the reliability. *Comput Math Methods Med*. 2015;2015:571473.
15. Zhuang M, Dierckx RA, Zaidi H. Generic and robust method for automatic segmentation of PET images using an active contour model. *Med Phys*. 2016;43:4483–4494.
16. Abdoli M, Dierckx RA, Zaidi H. Contourlet-based active contour model for PET image segmentation. *Med Phys*. 2013;40:082507.
17. Blanc-Durand P, Van Der Gucht A, Verger A, et al. Voxel-based ^{18}F-FET PET segmentation and automatic clustering of tumor voxels: a significant association with IDH1 mutation status and survival in patients with gliomas. *PLoS One*. 2018;13:e0199379.
18. Belhassen S, Zaidi H. A novel fuzzy C-means algorithm for unsupervised heterogeneous tumor quantification in PET. *Med Phys*. 2010;37:1309–1324.
19. Hatt M, Cheze le Rest C, Turzo A, et al. A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET. *IEEE Trans Med Imaging*. 2009;28:881–893.
20. Stefano A, Vitabile S, Russo G, et al. An enhanced random walk algorithm for delineation of head and neck cancers in PET studies. *Med Biol Eng Comput*. 2017;55:897–908.

21. Geets X, Lee JA, Bol A, et al. A gradient-based method for segmenting FDG-PET images: methodology and validation. *Eur J Nucl Med Mol Imaging*. 2007;34:1427–1438.
22. Tylski P, Bonniaud G, Decenciere E, et al. ^{18}F-FDG PET images segmentation using morphological watershed: a phantom study. *IEEE Nuclear Science Symposium Conference Record*. 2006;4:2063–2067.
23. Song Q, Bai J, Han D, et al. Optimal co-segmentation of tumor in PET-CT images with context information. *IEEE Trans Med Imaging*. 2013;32: 1685–1697.
24. Beichel RR, Van Tol M, Ulrich EJ, et al. Semiautomated segmentation of head and neck cancers in ^{18}F-FDG PET scans: a just-enough-interaction approach. *Med Phys*. 2016;43:2948–2964.
25. Yang J, Beadle BM, Garden AS, et al. A multimodality segmentation framework for automatic target delineation in head and neck radiotherapy. *Med Phys*. 2015;42:5310–5320.
26. Zeng Z, Wang J, Tiddeman B, et al. Unsupervised tumour segmentation in PET using local and global intensity-fitting active surface and alpha matting. *Comput Biol Med*. 2013;43:1530–1544.
27. Montgomery D, Amira A, Zaidi H. Fully automated segmentation of oncological PET volumes using a combined multiscale and statistical model. *Med Phys*. 2007;34:722–736.
28. Andrearczyk V, Oreiller V, Depeursinge A. Oropharynx detection in PET-CT for tumor segmentation. In: *Proceedings of the 2020 Irish Machine Vision and Image Processing Conference (IMVIP 2020)*. 2020:109–112.
29. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436–444.
30. Arabi H, AkhavanAllaf A, Sanaat A, et al. The promise of artificial intelligence and deep learning in PET and SPECT imaging. *Physica Medica*. 2021;83:122–137.
31. Häggström I, Schmidtlein CR, Campanella G, et al. DeepPET: a deep encoder-decoder network for directly solving the PET image reconstruction inverse problem. *Med Image Anal*. 2019;54:253–262.
32. Shiri I, Arabi H, Geramifar P, et al. Deep-JASC: joint attenuation and scatter correction in whole-body (18)F-FDG PET using a deep residual network. *Eur J Nucl Med Mol Imaging*. 2020;47:2533–2548.
33. Cui J, Gong K, Guo N, et al. PET image denoising using unsupervised deep learning. *Eur J Nucl Med Mol Imaging*. 2019;46:2780–2789.
34. Shiri I, AmirMozafari Sabet K, Arabi H, et al. Standard SPECT myocardial perfusion estimation from half-time acquisitions using deep convolutional residual neural networks. *J Nucl Cardiol*. 2020.
35. Akhavanallaf A, Shiri I, Arabi H, et al. Whole-body voxel-based internal dosimetry using deep learning. *Eur J Nucl Med Mol Imaging*. 2021;48: 670–682.
36. Moradi S, Oghli MG, Alizadehasl A, et al. MFP-Unet: A novel deep learning based approach for left ventricle segmentation in echocardiography. *Phys Med*. 2019;67:58–69.
37. Isensee F, Kickingereder P, Wick W, et al. No new-net. In: *International MICCAI Brainlesion Workshop*. Granada, Spain: Springer; 2018:234–244.
38. Oktay O, Schlemper J, Folgoc LL, et al. Attention u-net: learning where to look for the pancreas. *arXiv*. 2018.
39. Zhou Z, Siddiquee MMR, Tajbakhsh N, et al. Unet++: A nested u-net architecture for medical image segmentation. *arXiv*. 2018;11045:3–11.
40. Yu H, Caldwell C, Mah K, et al. Automated radiation targeting in head-and-neck cancer using region-based texture analysis of PET and CT images. *Int J Radiat Oncol Biol Phys*. 2009;75:618–625.
41. Berthon B, Evans M, Marshall C, et al. Head and neck target delineation using a novel PET automatic segmentation algorithm. *Radiother Oncol*. 2017;122:242–247.
42. Kawata Y, Arimura H, Ikushima K, et al. Impact of pixel-based machine-learning techniques on automated frameworks for delineation of gross tumor volume regions for stereotactic body radiation therapy. *Phys Med*. 2017;42:141–149.
43. Grossiord E, Talbot H, Passat N, et al. Automated 3D lymphoma lesion segmentation from PET/CT characteristics. *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. 2017;174–178.
44. Leung KH, Marashdeh W, Wray R, et al. A physics-guided modular deep-learning based automated framework for tumor segmentation in PET. *Phys Med Biol*. 2020;65:245032.
45. Fu X, Bi L, Kumar A, et al. Multimodal spatial attention module for targeting multimodal PET-CT lung tumor segmentation. *IEEE J Biomed Health Inform*. 2021;PP. [Online ahead of print]. doi:10.1109/JBHI.2021.3059453.
46. Zhong Z, Kim Y, Plichta K, et al. Simultaneous cosegmentation of tumors in PET-CT images using deep fully convolutional networks. *Med Phys*. 2019; 46:619–633.
47. Chen L, Shen C, Zhou Z, et al. Automatic PET cervical tumor segmentation by combining deep learning and anatomic prior. *Phys Med Biol*. 2019; 64:085019.
48. 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;64:205015.
49. Jin D, Guo D, Ho T-Y, et al. Accurate esophageal gross tumor volume segmentation in PET/CT using two-stream chained 3D deep network fusion. *arXiv*. 2019;182–191.
50. Andrearczyk V, Oreiller V, Vallières M, et al. Automatic segmentation of head and neck tumors and nodal metastases in PET-CT scans. *Proc Mach Learning Res*. 2020;33–43.
51. Afshari S, BenTaieb A, Mirikharaji Z, et al. Weakly supervised fully convolutional network for PET lesion segmentation. *Medical Imaging 2019: Image Processing, International Society for Optics and Photonics*. 2019:109491K.
52. Huang B, Chen Z, Wu P-M, 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;2018:1–12.
53. Berthon B, Spezi E, Galavis P, et al. Towards a standard for the evaluation of PET auto-segmentation methods: requirements and implementation. *Med Phys*. 2017;44:4098–4111.
54. Hatt M, Laurent B, Ouahabi A, et al. The first MICCAI challenge on PET tumor segmentation. *Med Image Anal*. 2018;44:177–195.
55. Prior FW, Clark K, Commean P, et al. TCIA: an information resource to enable open science. *Int Conf IEEE Eng Med Biol Soc*. 2013;2013:1282–1285.
56. Gevaert O, Xu J, Hoang CD, et al. Non–small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data—methods and preliminary results. *Radiology*. 2012;264:387–396.
57. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging*. 2013;26:1045–1057.
58. Bakr S, Gevaert O, Echegaray S, et al. A radiogenomic dataset of non–small cell lung cancer. *Sci Data*. 2018;5:180202.
59. Vallières M, Kay-Rivest E, Perrin LJ, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Sci Rep*. 2017;7:10117.
60. Rosset A, Spadola L, Ratib O. OsiriX: an open-source software for navigating in multidimensional DICOM images. *J Digit Imaging*. 2004;17:205–216.
61. Li W, Wang G, Fidon L, et al. On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task. *arXiv*. 2017;348–360.
62. Gibson E, Giganti F, Hu Y, et al. Automatic multi-organ segmentation on abdominal CT with dense v-networks. *IEEE Trans Med Imaging*. 2018;37: 1822–1834.
63. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Cham, Switzerland: Springer International Publishing. 2015;234–241.
64. Jadon S. A survey of loss functions for semantic segmentation. *arXiv*. 2020.
65. Zwanenburg A, Leger S, Vallières M, et al. Image biomarker standardisation initiative. *arXiv*. 2016.
66. van Griethuysen JJ, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77: e104–e107.
67. Brosch T, Yoo Y, Tang LY, et al. Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE Trans Med Imaging*. 2015;3–11.
68. Mumford DB, Shah J. Optimal approximations by piecewise smooth functions and associated variational problems. *Commun Pure Appl Mathematics*. 1989; 42:577–685.