Neurosurgical Review (2023) 46:116
https://doi.org/10.1007/s10143-023-02014-3

RESEARCH

Fully automated imaging protocol independent system for pituitary adenoma segmentation: a convolutional neural network—based model on sparsely annotated MRI

Martin Černý[1,2] · Jan Kybic[3] · Martin Májovský[1] · Vojtěch Sedlák[4] · Karin Pirgl[1,5] · Eva Misiorzová[6] · Radim Lipina[6] · David Netuka[1]

Received: 31 January 2023 / Revised: 8 March 2023 / Accepted: 28 April 2023

Abstract

This study aims to develop a fully automated imaging protocol independent system for pituitary adenoma segmentation from magnetic resonance imaging (MRI) scans that can work without user interaction and evaluate its accuracy and utility for clinical applications. We trained two independent artificial neural networks on MRI scans of 394 patients. The scans were acquired according to various imaging protocols over the course of 11 years on 1.5T and 3T MRI systems. The segmentation model assigned a class label to each input pixel (pituitary adenoma, internal carotid artery, normal pituitary gland, background). The slice segmentation model classified slices as clinically relevant (structures of interest in slice) or irrelevant (anterior or posterior to sella turcica). We used MRI data of another 99 patients to evaluate the performance of the model during training. We validated the model on a prospective cohort of 28 patients, Dice coefficients of 0.910, 0.719, and 0.240 for tumour, internal carotid artery, and normal gland labels, respectively, were achieved. The slice selection model achieved 82.5% accuracy, 88.7% sensitivity, 76.7% specificity, and an AUC of 0.904. A human expert rated 71.4% of the segmentation results as accurate, 21.4% as slightly inaccurate, and 7.1% as coarsely inaccurate. Our model achieved good results comparable with recent works of other authors on the largest dataset to date and generalized well for various imaging protocols. We discussed future clinical applications, and their considerations. Models and frameworks for clinical use have yet to be developed and evaluated.

Keywords Pituitary adenoma · Magnetic resonance imaging · Image segmentation · Machine learning

✉ Martin Černý
dr.martin.cerny@gmail.com

Jan Kybic
jkybic@gmail.com

Martin Májovský
drvojak2@gmail.com

Vojtěch Sedlák
sedlakvojt@gmail.com

Karin Pirgl
Karin.pirgl@gmail.com

Eva Misiorzová
e.misiorzova@seznam.cz

Radim Lipina
radim.lipina@fno.cz

David Netuka
netuka.david@gmail.com

[1] Department of Neurosurgery and Neurooncology, 1st Faculty of Medicine, Charles University, Central Military Hospital Prague, U Vojenské nemocnice 1200, 169 02 Praha 6, Czech Republic

[2] 1st Faculty of Medicine, Charles University Prague, Kateřinská 1660/32, 121 08 Praha 2, Czech Republic

[3] Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Technická 2, 166 27 Praha 6, Czech Republic

[4] Department of Radiodiagnostics, Central Military Hospital Prague, U Vojenské nemocnice 1200, 169 02 Praha 6, Czech Republic

[5] 3rd Faculty of Medicine, Charles University Prague, Ruská 87, 100 00, Praha 10, Czech Republic

[6] Department of Neurosurgery, Faculty of Medicine, University of Ostrava, University Hospital Ostrava, 17. listopadu 1790/5, 708 52 Ostrava-Poruba, Czech Republic

Published online: 10 May 2023

Introduction

Pituitary adenomas (PAs) are the most common tumours of the sellar region, accounting for 10–15% of all intracranial neoplasms [1]. PAs are benign lesions; however, they can sometimes manifest clinically by compression of the optic chiasm or dysregulated hormonal production. In some patients, surgical removal is performed using a minimally invasive transsphenoidal endoscopic approach [2].

Several experimental and clinical applications of machine learning have been recently introduced into neurosurgery [3], particularly in pituitary adenomas [4]. Different models were proposed for automatic segmentation [5–8], outcome prediction [9, 10], consistency prediction [11–13], or acromegaly diagnosis from facial pictures [14, 15].

Our work presents a fully automated system for pituitary adenoma segmentation. We aim to offer a simple, ready-to-use system for clinical researchers without extensive technical knowledge. We also explore the possibility of training the model on sparsely annotated data because collecting segmentations for all slices in the training dataset would be very time-consuming and impractical. The complete source code, the trained model, and example data are provided online along with the article.

Medical professionals can benefit from automatic segmentation in multiple ways. Radiosurgical planning relies on high-precision 3D segmentations, which must be manually delineated by trained professionals [16]. Augmented reality technology can improve the precision in endoscopic skull base surgery by overlaying 3D anatomical models generated from preoperative medical images onto endoscope images [17]. Tumour volume can also be used for progression and treatment response tracking [18]. The spatial relationship of the tumour and its surrounding structures can serve as a good predictor of surgical outcome [19–23]. Automated and assisted image analysis can significantly increase efficiency and enable high throughput workflows and cost savings [24].

Multiple automatic and semi-automatic PA segmentation attempts have been attempted in recent years. First, local image pattern-based methods requiring a manual seed point initialization and parameter setting were proposed [6, 7, 25]. Although user interaction was still needed, they significantly reduced segmentation time. More recently, deep learning methods were applied to this problem [5, 8], enabling further advances in segmentation accuracy and time savings. We aimed to further develop this field by proposing an end-to-end system with no user interaction requirements and examining the utility of and multimodal imaging data and.

Methods

Patients

Preoperative contrast–enhanced T1-weighted (CE-T1) coronal magnetic resonance imaging (MRI) scans of patients who underwent primary surgery for pituitary adenoma between 2007 and 2018 were retrospectively included. T2-weighted (T2) scans were also used if available for the same patient. The scans were acquired according to various imaging protocols on multiple 1.5T and 3T devices used over 11 years. Patients were randomly assigned to training (80%) and validation (20%) datasets.

A test dataset from MRI scans of patients who underwent primary surgery for PA between January 2022 and June 2022 and consented to publish their anonymised imaging data was created prospectively. CE-T1, non-contrast T1 and T2 coronal MRI scans were acquired for each patient using a 3T MR system (GE Discovery MR750w, GE Healthcare, Chicago, ILL, USA) with a standard 32-channel head coil. The facial part of the image was manually erased to prevent reconstruction of the patient's appearance. All metadata regarding patient personal information were removed.

Data annotations

All coronal CE-T1 scans were manually segmented in ITKSnap v3.8.0 [26] by the principal author, a neurosurgery resident (MČ), and reviewed and corrected if necessary by a board-certified neurosurgeon with 10 years of experience in endoscopic pituitary surgery (MM). First, clinically relevant slices were identified. Clinical relevance was defined as the presence of four cross sections of intracavernous and supraclinoid segments of the internal carotid artery, allowing for good assessment of cavernous sinus invasion. Regions of interest (ROIs) were placed in three labels for the tumour, internal carotid artery (ICA), and normal gland on all of the clinically relevant slices and none of the clinically irrelevant slices. The tumour label was placed over all tumour pixels in clinically relevant slices including cystic parts. If a tangential cross section through a ICA segment was encountered, a circular cross section approximation ROI was placed. If the cross section through carotid syphon was encountered or one of the four cross sections could not be identified, the slice was not considered clinically relevant. Normal gland was only marked if identifiable in the slice. Figure 1 shows examples of different slice types and their annotations.

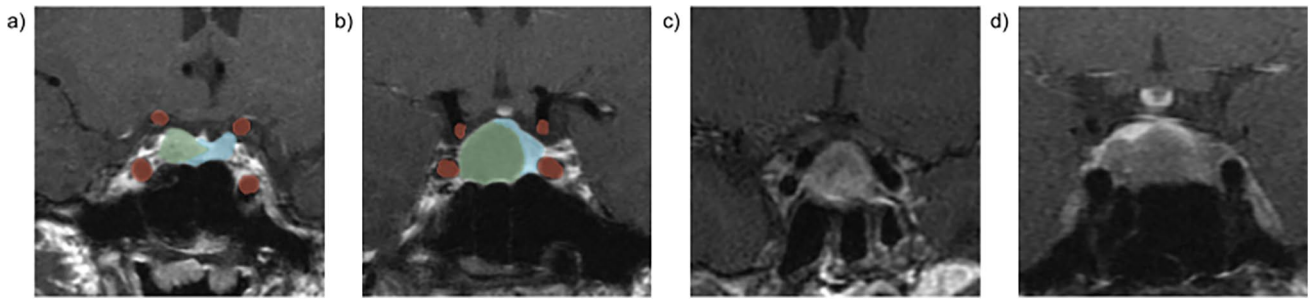

Fig. 1 Examples of annotations of different slice types in the dataset: **a** slice with four ICA cross sections; **b** a slice with tangential sections through supraclinoid ICA segments; **c** cross section through the carotid siphon; **d** a slice that does not allow for identification of the supraclinoid cross section of the left ICA. Slices **a** and **b** were annotated and considered clinically relevant, slices **c** and **d** were left unannotated as they do not allow for proper assessment of lateral invasion. Coloured areas mark the segmentations for tumour (green), ICA (red), and normal gland (blue)

Software

All data processing was performed in Python v3.8.8 [27]. Medical imaging data were processed using the SimpleITK library v2.1.0 [28–30], and the machine learning model was written in Keras v2.4.3 [31] with Tensorflow v2.3.0 [32] backend. The source code, test dataset, and trained model are available from the author's GitHub repository [33].

Data preprocessing

If available, non-contrast T1 and T2 scans were co-registered and transformed into a CE-T1 coordinate space to achieve voxelwise spatial correspondence. Scans were then cropped to 194 × 194 pixels in the coronal plane to ensure uniform input dimensions. Because MRI scans for sellar lesions are routinely centred on the sella, we could simply crop the middle part of the image. The original FOV was 14.9 × 14.9 ± 2.2 cm in the training dataset; cropped images had a size of 5.8 × 5.8 ± 1.6 cm.

Intensity normalisation was performed to ensure a per subject pixel intensity mean of 0 and a standard deviation of 1. Subsequently, all annotated slices were extracted with one preceding and one following slice with channels corresponding to each patient's three pulse sequences (channels-last). If a pulse sequence was unavailable for the patient, all pixels were replaced with 0 for the missing channel. The same number of randomly selected unannotated slices was extracted for the training of the slice relevance classifier. Some 1264 and 330 annotated and 1259 and 330 unannotated slices were extracted in the training and validation datasets, respectively. Figure 2 summarises the dataset extraction pipeline.

Segmentation model architecture

Our baseline model uses the U-Net architecture proposed by Ronnenberger [34], a convolutional neural network (CNN) variant, and accepts CE-T1 images as input. Figure 3 a presents an overview of the model architecture. Model variants accepting combinations of channels for non-contrast T1 and T2 images were also tested.

First, a 3D convolution layer with a kernel shape of 3 × 3 × 3 integrated the data from adjacent slices, creating a latent shape embedding (192, 192, 32). Then, five downsampling layers were applied, each comprising two 2D convolutional layers with leaky ReLU [35] activation followed by a max

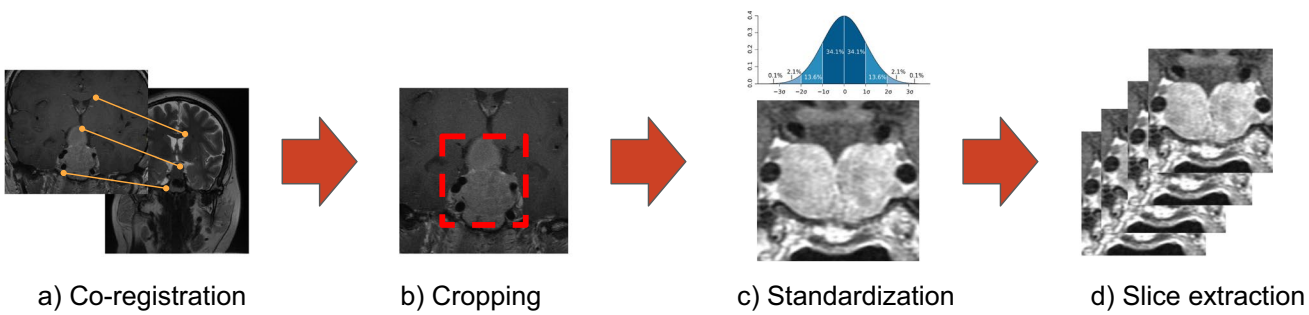

Fig. 2 Dataset extraction pipeline

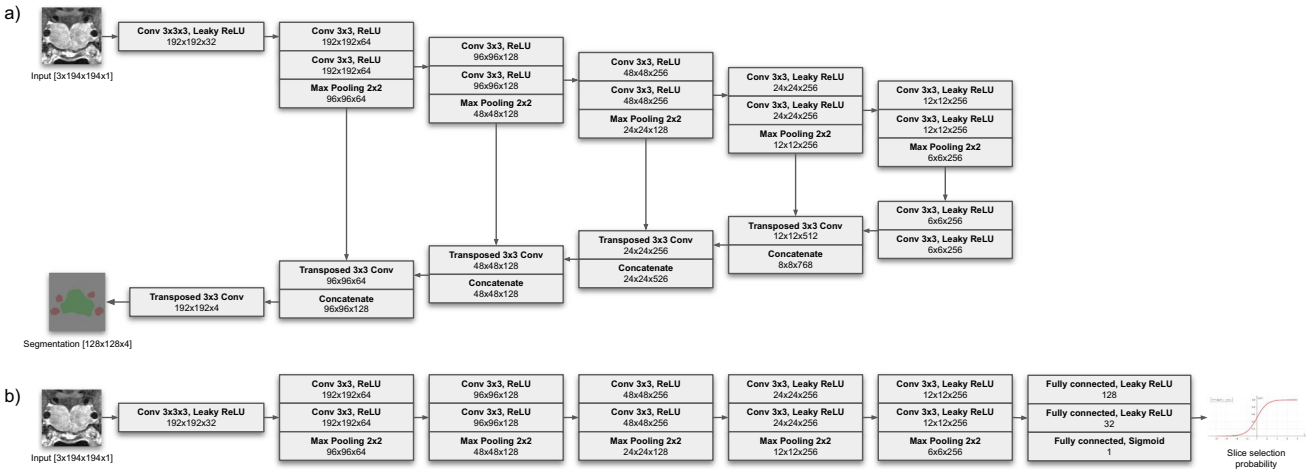

Fig. 3 **a** Schematic depiction of the segmentation model, where grey boxes signify model layers with layer type and layer output dimensions; the output of the model is a segmentation map with four labels; **b** schematic depiction of the slice selection model. The model returns a probability between 0 and 1 of the slice being a relevant slice

pooling layer achieving a final shape of (12, 12, 256). Models with different numbers of downsampling layers were also tested.

An upsampling stack with skip connections restored the resolution of (192, 192, 4) with four logits for the respective label classes. Each upsampling layer consisted of a 2D convolutional layer, batch normalisation, dropout, and a dense layer with a leaky ReLU activation. A categorical cross-entropy loss function was applied during model training.

Slice selection model architecture

An additional model was trained for relevant slice selection. The model employed a downsampling stack described in the previous section, followed by two fully connected layers. The first fully connected layer was followed by a leaky ReLU activation layer. The slice relevance is defined as the probability that the slice was annotated in the dataset and therefore deemed clinically relevant by the annotator. To estimate the slice relevance, the second fully connected layer was followed by a sigmoid activation layer returning the slice relevance probability between 0 and 1 (Fig. 3b). Binary cross-entropy loss function was applied to the model.

Architecture optimization

We performed a five-fold cross-validation testing to determine the optimal model configuration. We examined the effect of adding input channels and different numbers of downsampling layers. Models were trained for 50 epochs. The tumour Dice coefficient score (DCS) on the validation dataset was used as comparison metrics. We performed an unpaired *t*-test to compare model variants with the baseline model and to pick the best variant to be used for all further experiments in this study.

Model training

Data augmentation [36] was performed by randomly cropping the input using small random shifts. For slice relevance, the model was tasked with predicting whether the slice comes from the annotated or unannotated part of the dataset. The model parameters were randomly initialised and tuned using the Adam (adaptive moment estimation) optimizer [37]. Both models were trained independently. All training was performed on a C2 series Compute Engine instance on the Google Cloud platform with 4 vCPU and 16 GB RAM.

Model performance evaluation

For the segmentation model, a total Dice coefficient score [38] and per label Dice coefficient scores for tumour, ICA, and normal gland were recorded over training time. DSC is a quantitative assessment of overlap of two areas commonly used in segmentation studies and can be denoted as

$$\mathrm{DCS} = 2 * |\mathrm{X} \cap \mathrm{Y}||\mathrm{X}| + |\mathrm{Y}|$$

where X is the ground truth and Y is the predicted segmentation.

For the slice selection model, classification accuracy was recorded over training time while sensitivity, specificity, and AUC were calculated for the best model iteration.

To assess the applicability in clinical settings, an independent neurosurgeon from outside our institution (EM), unfamiliar with the details of this study, was asked to review

Table 1 Baseline characteristics of the patient dataset

	Train	Validation	Test
N	394	99	28
Extracted slices	1264	330	103
Age, *years* ± *SD*	53.2 ± 15.6	54.2 ± 14.8	56.8 ± 14.6
Male sex, *n* (%)	179 (45.4)	57 (57.6)	18 (64.3)
Tumour type, *n (%)*			
Non-functioning	242 (61.4)	59 (60)	25 (89.3)
GH-secreting	99 (25.1)	21 (21)	1 (3.6)
Prolactin-secreting	12 (3)	5 (5)	1 (3.6)
ACTH-secreting	39 (9.9)	14 (14)	1 (3.6)
Plurihormonal	2 (0.5)	0 (0)	0 (0)
Imaging data availability, *n (%)*			
CE-T1 only	12 (3)	1 (1)	0 (0)
CE-T1 + T1	238 (60.4)	57 (56)	0 (0)
CE-T1 + T2	10 (2.5)	5 (5)	0 (0)
CE-T1 + T1 + T2	134 (34)	36 (36)	28 (100)

all segmentations in the validation and test datasets and to rate them as accurate, slightly inaccurate and coarsely inaccurate. The rater was provided with no further instructions.

Results

Dataset

Total of 493 patients were included in the study: 13 (2.6%) only had CE-T1W scans, 295 (59.8%) had CE-T1W and non-contrast T1W scans, 15 (3%) had CE-T1W and T2W scans, and 170 (34.5%) had all three sequences available. Table 1 summarises patient baseline characteristics. For CE-T1 scans, both 2D and 3D (navigation) series were routinely available, while for the other pulse sequences, only 2D series were usually available. Table 2 presents an overview of imaging data properties. 394 (80%) patients were randomly assigned to the training dataset and 99 (20%) to the validation dataset. Total of 28 patients enrolled in the prospective test cohort.

Architecture optimization

A five-fold cross-validation ablation study was performed on different model configurations. No configuration significantly ($p < 0.05$) outperformed the baseline model in total DCS. Adding input channels for non-contrast T1 and T2 images did not result in more accurate results ($p = 0.707$ and $p = 0.96$). Limiting our

Table 2 Imaging data characteristics

		Train	Validation	Test
n	CE-T1	394	99	28
	T1	372	93	28
	T2	144	41	28
Sequence type		Spin echo	Spin echo	Spin echo
TR *ms (± SD)*	CE-T1	539.67 ± 112.2	550.15 ± 1.5	400
	T1	554.64 ± 105.66	557.27 ± 104.0	440
	T2	3819.86 ± 973.96	4218.42 ± 985.84	4000.78 ± 426.34
TE *ms (± SD)*	CE-T1	12.26 ± 1.44	12.38 ± 1.5	14
	T1	12.44 ± 1.04	12.62 ± 1.0	13
	T2	102.15 ± 11.8	100.56 ± 9.47	90.96 ± 4.03
FOV *cm (± SD)*	CE-T1	14.86 ± 2.19	14.6 ± 1.79	16
	T1	14.55 ± 1.61	14.32 ± 1.13	16
	T2	18.02 ± 3.37	18.32 ± 2.9	16
N of averages *n (± SD)*	CE-T1	1.98 ± 0.38	1.98 ± 0.36	1.5
	T1	2.01 ± 0.33	2.02 ± 0.29	1.5
	T2	3.21 ± 1.74	2.96 ± 1.07	1
Pixel bandwidth *Hz (± SD)*	CE-T1	101.2 ± 19.71	105.15 ± 41.12	97.66
	T1	100.17 ± 16.38	102.48 ± 33.94	97.66
	T2	144.87 ± 41.77	153.64 ± 50.7	122.07
Slice thickness *mm (± SD)*	CE-T1	2.26 ± 0.49	2.18 ± 0.35	2.5
	T1	2.21 ± 0.37	2.15 ± 0.31	2.5
	T2	2.95 ± 0.87	2.89 ± 0.84	2.5
Slice spacing *mm (± SD)*	CE-T1	0.73 ± 0.53	2.63 ± 0.33	2.8
	T1	2.67 ± 0.36	2.61 ± 0.29	2.8
	T2	3.33 ± 1.21	3.28 ± 1.15	2.8
Cropped area *cm (± SD)*		5.8 ± 1.6	5.69 ± 1.29	6.0

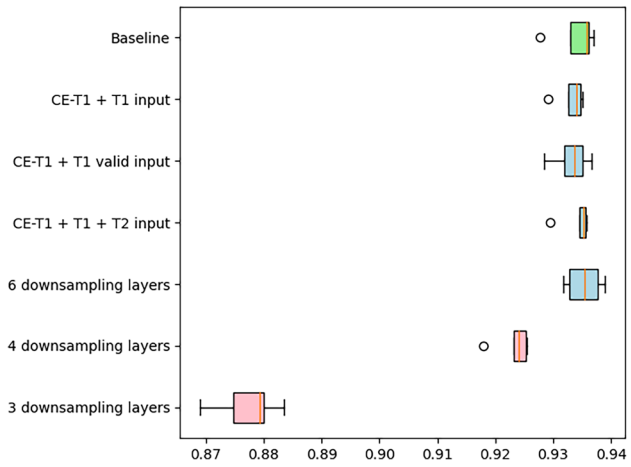

Fig. 4 Comparison of the five-fold cross-validation ablation study results of different models scored by DCS for tumour, ICA, and normal gland. The baseline model is in green, with models with insignificantly ($p > 0.05$) different results in blue and with significantly worse results in red

dataset to only the subset with both CE-T1 and non-contrast T1 images available ($n = 295$) did not improve the results ($p = 0.459$). We observed a significant drop in total DCS when decreasing the number of downsampling layers of the U-Net network from 5 to 4 and 3 ($p = 0.001$ and $p < 0.001$); however, there was no improvement when adding a 6th downsampling layer ($p = 0.68$). Similar results were observed for model loss, tumour DCS, ICA DCS, and normal gland DCS. Figure 4 shows the comparisons of five-fold cross-validation results for different models. Detailed results of the five-fold cross-validation are summarised in Table 3.

Segmentation

The segmentation model was trained for 100 epochs. The training took 8.2 h. The minimum total validation loss was reached after 18 epochs. The model achieved per class Dice coefficients of 0.889 for tumour, 0.802 for ICA, and 0.316 for normal gland labels. Dice coefficients of 0.910, 0.719, and 0.240 for the respective labels were achieved on the test dataset. Figure 5 depicts the course of the Dice coefficient during the training. Figure 6 displays ground truth and predicted segmentations in eight slices from the test dataset.

Slice selection

The slice selection model was trained for 50 epochs. The training took 5.3 h. The minimum total validation loss was reached after 10 epochs. For the validation dataset, 88.8% accuracy, 87.6% sensitivity, 90.0% specificity, and an AUC of 0.944 were achieved, and for the test dataset, 82.5% accuracy, 88.7% sensitivity, 676.9% specificity, and an AUC of 0.907 (Fig. 7).

Reviewer satisfaction

To assess the applicability in clinical settings, an independent neurosurgeon from outside of our institution (EM) and unfamiliar with the details of this study was asked to review all results in the validation and test dataset and to rate them on a three-point scale (accurate, slightly inaccurate, and coarsely inaccurate). Our model performed both slice selection and segmentation in the reviewed images. In the validation dataset, 63 (63.7%) segmentations were marked as accurate, 28 (28.3%) as slightly inaccurate, and 8 (8.1%) as coarsely inaccurate. In the test dataset, 20 (71.4%) segmentations were marked as accurate, 6 (21.4%) as slightly inaccurate, and 2 (7.1%) as coarsely inaccurate.

Discussion

MRI segmentation has multiple applications in medicine and neurosurgery in particular. However, manually creating a segmentation mask is time-consuming and tedious,

Table 3 The five-fold cross-validation of different modes, values for total DCS, tumour DCS, ICA DCS, and normal gland DCS on the validation dataset are given as mean (±SD). The *p*-value represents the significance of the difference with the baseline model. An asterisk marks significant entries

Model	Tumour DCS		ICA DCS		Normal gland DCS	
	Mean (SD)	*p*	Mean (SD)	*p*	Mean (SD)	*p*
Baseline 5 downsampling layers, CE + T1 input	0.934 (0.003)		0.843 (0.002)		0.611 (0.014)	
CE-T1 + T1 input	0.933 (0.002)	0,696	0.845 (0.003)	0,258	0.616 (0.012)	0,631
CE-T1 + T1 input (only non-zero)	0.933 (0.003)	0,74	0.846 (0.005)	0,272	0.623 (0.009)	0,208
CE-T1 + T1 + T2 input	0.934 (0.002)	0,945	0.846 (0.002)	0,071	0.62 (0.019)	0,477
6 downsampling layers	0.935 (0.003)	0,53	0.845 (0.004)	0,538	0.632 (0.014)	0,069
4 downsampling layers	0.923 (0.003)	0,001*	0.84 (0.005)	0,236	0.593 (0.013)	0,105
3 downsampling layers	0.877 (0.005)	<0.001*	0.823 (0.003)	<0.001*	0.513 (0.011)	<0.001*

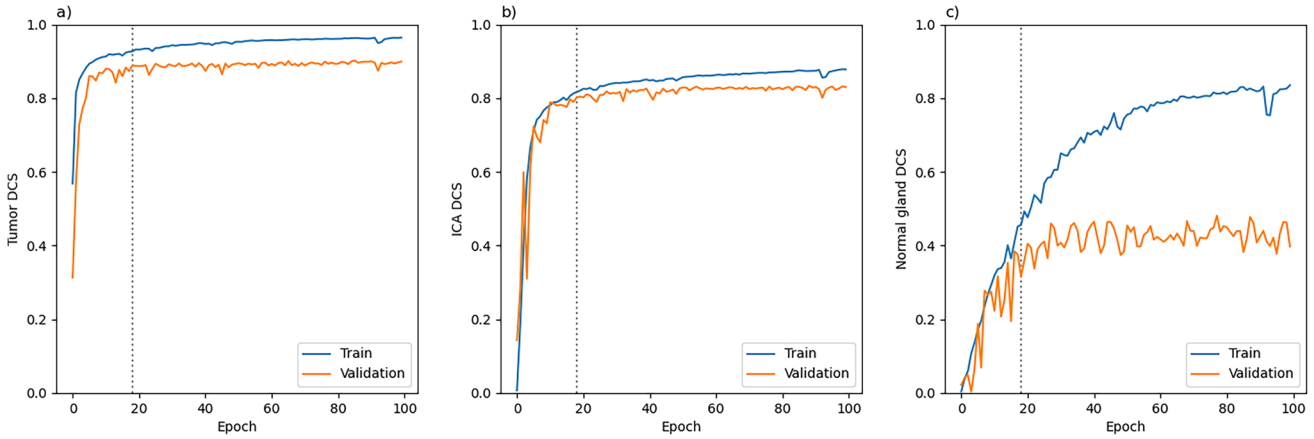

Fig. 5 Train and validation Dice coefficients over training time for the tumour, ICA, and normal gland (**a**–**c**). The grey vertical line marks the epoch of the best performing model for the validation loss

especially when working with thin-slice 3D data, which is common in many navigation and radiosurgical protocols. A reliable and accurate automatic or semi-automatic segmentation tool would benefit clinical practice.

Multiple automatic and semi-automatic PA segmentation attempts have been attempted in recent years. The first contribution by Egger [25] in 2011 used a directed 3D graph from a user-defined seed point, sending rays through the surface points of a polyhedron and sampling the graph's nodes along every ray [39]. This method achieved an average DCS of 0.775. In their follow-up work, Egger [7] examined the competitive region-growing method [40], achieving an average DSC of 0.820. The subsequent work by Egger [6] compares their previously developed graph-based method [25] to a balloon inflation method [41], yielding an average DSC of 0.775 and 0.759, respectively. All these methods required a manual seed point initialization and parameter setting; however, segmentation time was significantly reduced.

Wang [5] proposed a deep learning-based model classifying image voxels into eight classes (background, PA, normal pituitary, right ICA, right cavernous sinus, left ICA, left cavernous sinus, and optic chiasm). On clinically relevant slices,

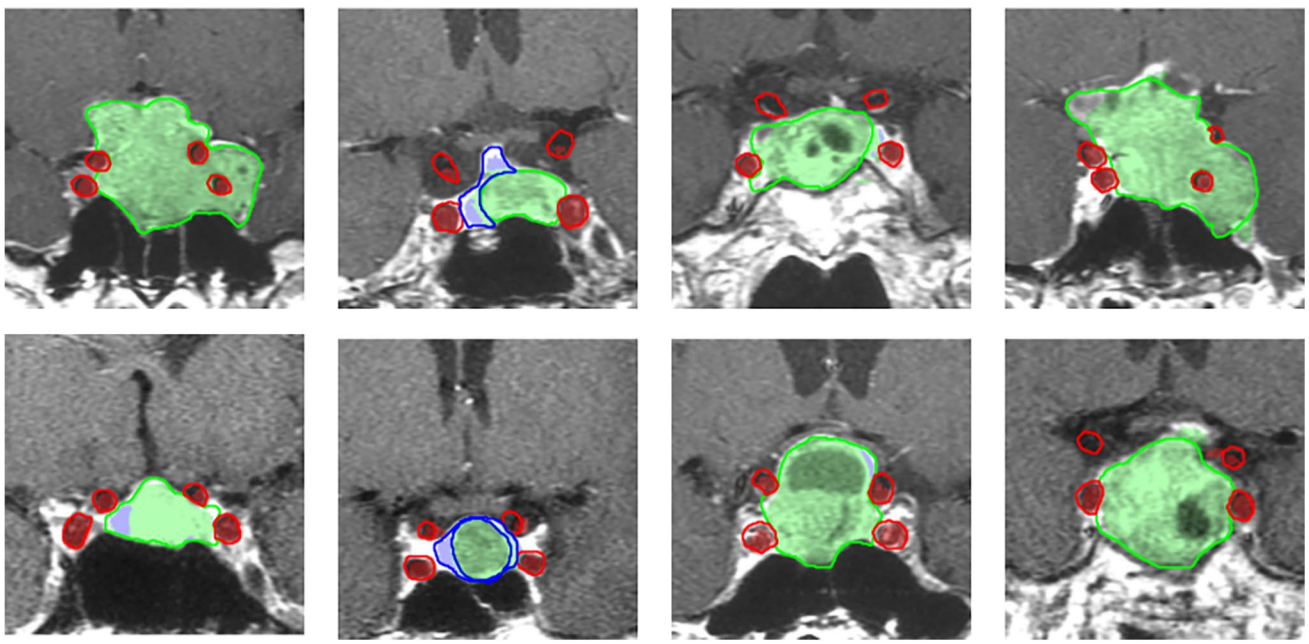

Fig. 6 Eight examples of ground truth (outline) and predicted (coloured area) segmentations for tumour (green), ICA (red), and normal gland (blue)

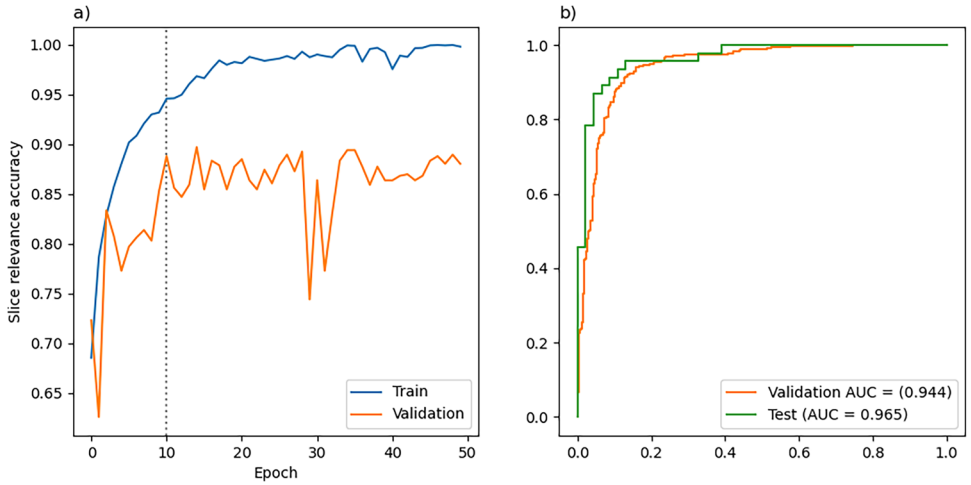

Fig. 7 **a** Train and validation slice selection accuracy over training time. The grey vertical line marks the epoch of the best performing model for total validation loss; **b** ROC for slice selection on validation and test datasets

they achieved a DCS of 0.940, 0.824, 0.847, and 0.721 for PA, right ICA, left ICA, and normal gland, respectively. Their model called GSU-net was a form of a U-net [35] with elements of a gated-shape convolutional network [42]. Their inputs consisted of CE-T1, non-contrast T1 and T2 scans concatenated with the results of an edge detector. Shu [8] used nnU-net [43], a self-configuring U-net framework, and achieved a PA DCS of 0.803 and 0.853 on their two datasets. They noticed poor performance of their model (DCS < 0.5) for tumours with a volume < 1000 mm^3.

Although all authors gave quantitative results, they cannot be directly compared, as their input data and target labels varied. We propose using our publicly available test dataset as a benchmark to compare the accuracy of pituitary adenoma segmentation models.

Machine learning models generalise poorly when processing inputs dissimilar to what they have been trained on [44]. Because it is unrealistic to provide annotations for training data for all the slices, we only focused on the clinically relevant slices, providing only sparse annotations [45, 46]. However, such a model might produce nonsense results for slices anterior or posterior to the region of clinical interest unseen during the training. Wang [5] also reported better results on relevant slices compared to all slices (PA DCS 0.940 vs. 0.898). In such settings, the user would have to select relevant slices manually. We successfully overcame this issue by training a classifier to distinguish relevant (annotated) and irrelevant (unannotated) slices with over 90% accuracy on the test dataset. It should also be noted that we trained our model on scans acquired according to various imaging protocols on 1.5T and 3T systems used over an 11-year period. This approach contributes to higher input data variance, leading to better generalisation. In contrast, previous studies used scans acquired according to a defined imaging protocol.

We were surprised that adding more pulse sequences to the input did not improve the segmentation results. Intuitively, we expected the model to benefit from complementary information. In clinical practice, surgeons often look into different pulse sequences and image planes to better understand the complex anatomical situation. For example, comparing non-contrast and CE images, primarily acquired in the venous phase, helps to identify structures with a high contrast enhancement ratio, i.e. the cavernous sinus. Similarly, examining the T2-weighted image can help identify the tumour's cystic parts.

In our work, however, adding these inputs as additional channels did not increase total DCS ($p > 0.05$). In contrast, several studies outside the domain of PA [47–49] demonstrated improved segmentation results using multimodal inputs. Generally, there are three strategies for combining multimodal inputs: fusion at the input level, intermediate layer level, and decision level [50]. There is no consensus about which of these strategies is superior. Because of its simplicity, most authors merge the inputs directly before entering the model [51–59]. Fusing the modalities halfway through the model is supposed to allow the network to process low-level features independently and high-level features combined [60–63]. Decision level fusion means training separate classifiers for individual modalities and weighting their outputs ("voting") [64]. Guo [47] performed a comparison study proving that input and layer level fusion outperformed decision layer fusion and single-modality models in soft-tissue sarcoma segmentation from PET, CT, T1, and T2 images. Le [58] examined two methods of layer level fusion which outperformed input level fusion and single modality models in the diagnosis of prostate cancer from T2-weighted and ADC images.

Our model uses input level fusion. Pixelwise correspondence of the mask with CE-T1 images is ensured, as these were used to draw the ground truth segmentation masks. The relevance of other modalities for the segmentation mask can be adversely affected when a misalignment with CE-T1 occurs. Further research is needed to clarify the effect of

misalignment on the model performance. Feature-level fusion could also compensate for the misalignment in that the images are being fused at a higher level of their receptive fields, where the images suffer less from small shifts between channels.

The proposed model achieved accurate (71.4%) or slightly inaccurate (21.4%) results in the majority of cases, as judged by an expert human rater. We noted that the model often confused tumour for normal pituitary gland and vice versa, which is sometimes difficult to differentiate even for human experts.

In further work, this model should be modified for clinical use. Radiosurgical planning is a direct application of automated segmentation [19]; however, there are more considerations than achieving anatomically precise contouring. The model has to account for a safe border zone around the lesson [65]. Also, damage to important radiosensitive structures (such as the optic chiasm) has to be avoided [66]. Cases referred for radiosurgical treatment are usually partially resected tumours or tumour recurrences, which can exhibit radiological features such as eccentric position within the sella turcica unaccounted for in our dataset. Further research directly on radiosurgical cases would be needed to assess our model's clinical utility or to propose new models directly for this clinical application. Different workflows would have to be tested including a semi-automated mode with the models suggesting a segmentation and a clinical expert correcting it if necessary. Such workflow would allow for faster case processing and higher throughput while keeping control over the quality of radiosurgical plans. Other application fields would be progression and treatment response tracking [21] and augmented reality endoscopic surgery [20].

We aimed to present a simple, ready-to-use system for clinical researchers without extensive technical knowledge and to enable further development in automatic segmentation applications not limited to the domain of pituitary adenomas. By replacing the training data, the model can be directly used on other types of medical images. Our open source code can also be modified by any researchers and serve as a starting point for their segmentation projects. We encourage further research in this field.

Conclusions

We developed and evaluated a fully automated segmentation system for pituitary adenomas. Our model achieved good results comparable to recent works of other authors, with DCS > 0.9 for tumour and DCS > 0.7 for ICA on the largest pituitary adenoma segmentation dataset to date. The model also generalised well for various imaging protocols. We found that the model often confused tumour for normal pituitary gland and vice versa, which is sometimes difficult to discriminate even for human experts. The model also achieved a relevant slice identification accuracy of 82.5%. Further development is needed for successful adaptation in clinical practice. We discussed some future applications and their considerations. Models and frameworks for these applications have yet to be developed and evaluated.

Author contribution MČ conceptualised the study, annotated the data, and wrote the source code and most parts of this article. JK advised and critically reviewed the technical implementation. MM reviewed and corrected the data annotations and contributed to the sections on clinical applications, VS contributed to the sections on magnetic resonance imaging and conducted the acquisition of prospective cohort imaging data, KP performed data analysis and literature review, EM performed an independent review of segmentation results, and DN and RL supervised and advised the work on this article as senior researchers.

Funding This work was supported by the Ministry of Defense of the Czech Republic (institutional support MO1012), the Charles University in Prague (Project Cooperation Neurosciences), and the OP VVV-funded project "CZ.02.1.01/0.0/0.0/16_019/0000765 Research Center for Informatics".

Data availability The source code, test dataset, and trained model are available online from the author's GitHub repository [33].

Declarations

Ethics approval This study was approved by the institutional ethical committee (ref. no. 108/17-9/2022 for the retrospective part, ref. no. 108/17-11/2022 for the prospective part), and data were anonymized at the time of patient inclusion and treated according to the ethical standards of the Declaration of Helsinki. Written informed consent for imaging data processing and publication was obtained from all participants in the prospective cohort. The requirement of informed consent was waived by the institutional ethical committee in the retrospective cohort because of the large-scale retrospective nature of the study and no potential harm to the study participants. A detailed methodical description of data anonymization was approved by the institutional ethical committee.

Competing interests The authors declare no competing interests.

References

1. Daly AF, Beckers A (2020) The epidemiology of pituitary adenomas. Endocrinol Metab Clin North Am 49(3):347–355. https://doi.org/10.1016/j.ecl.2020.04.002
2. Molitch ME (2017) Diagnosis and treatment of pituitary adenomas: a review. JAMA 317(5):516–524. https://doi.org/10.1001/jama.2016.19699
3. Celtikci E (2018) A systematic review on machine learning in neurosurgery: the future of decision-making in patient care. Turk Neurosurg 28(2):167–173. https://doi.org/10.5137/1019-5149.JTN.20059-17.1
4. Dai C, Sun B, Wang R, Kang J (2021) The application of artificial intelligence and machine learning in pituitary adenomas. Front Oncol 11:784819. https://doi.org/10.3389/fonc.2021.784819

5. Wang H, Zhang W, Li S, Fan Y, Feng M, Wang R (2021) Development and evaluation of deep learning-based automated segmentation of pituitary adenoma in clinical task. J Clin Endocrinol Metab 106(9):2535–2546. https://doi.org/10.1210/clinem/dgab371
6. Egger J, Zukić D, Freisleben B, Kolb A, Nimsky C (2013) Segmentation of pituitary adenoma: a graph-based method vs. a balloon inflation method. Comput Methods Programs Biomed 110(3):268–278. https://doi.org/10.1016/j.cmpb.2012.11.010
7. Egger J, Kapur T, Nimsky C, Kikinis R (2012) Pituitary adenoma volumetry with 3D Slicer. PloS One 7(12):e51788. https://doi.org/10.1371/journal.pone.0051788
8. Shu X, Zhou Y, Li F, Zhou T, Meng X, Wang F, Zhang Z, Pu J, Xu B (2021) Three-dimensional semantic segmentation of pituitary adenomas based on the deep learning framework-nnU-Net: a clinical perspective. Micromachines 12(12):1473. https://doi.org/10.3390/mi12121473
9. Voglis S, van Niftrik C, Staartjes VE, Brandi G, Tschopp O, Regli L, Serra C (2020) Feasibility of machine learning based predictive modelling of postoperative hyponatremia after pituitary surgery. Pituitary 23(5):543–551. https://doi.org/10.1007/s11102-020-01056-w
10. Laws ER, Catalino MP (2020) Editorial. Machine learning and artificial intelligence applied to the diagnosis and management of Cushing disease. Neurosurg Focus 48(6):E6. https://doi.org/10.3171/2020.3.FOCUS20213
11. Fan Y, Hua M, Mou A, Wu M, Liu X, Bao X, Wang R, Feng M (2019) Preoperative noninvasive radiomics approach predicts tumor consistency in patients with acromegaly: development and multicenter prospective validation. Front Endocrinol 10:403. https://doi.org/10.3389/fendo.2019.00403
12. Zeynalova A, Kocak B, Durmaz ES, Comunoglu N, Ozcan K, Ozcan G, Turk O, Tanriover N, Kocer N, Kizilkilic O, Islak C (2019) Preoperative evaluation of tumour consistency in pituitary macroadenomas: a machine learning-based histogram analysis on conventional T2-weighted MRI. Neuroradiology 61(7):767–774. https://doi.org/10.1007/s00234-019-02211-2
13. Zhu H, Fang Q, Huang Y, Xu K (2020) Semi-supervised method for image texture classification of pituitary tumors via CycleGAN and optimized feature extraction. BMC Med Inform Decis Mak 20(1):215. https://doi.org/10.1186/s12911-020-01230-x
14. Meng T, Guo X, Lian W, Deng K, Gao L, Wang Z, Huang J, Wang X, Long X, Xing B (2020) Identifying facial features and predicting patients of acromegaly using three-dimensional imaging techniques and machine learning. Front Endocrinol 11:492. https://doi.org/10.3389/fendo.2020.00492
15. Wei R, Jiang C, Gao J, Xu P, Zhang D, Sun Z, Liu X, Deng K, Bao X, Sun G, Yao Y, Lu L, Zhu H, Wang R, Feng M (2020) Deep-learning approach to automatic identification of facial anomalies in endocrine disorders. Neuroendocrinology 110(5):328–337. https://doi.org/10.1159/000502211
16. Jarrett D, Stride E, Vallis K, Gooding MJ (2019) Applications and limitations of machine learning in radiation oncology. Br J Radiol 92(1100):20190001. https://doi.org/10.1259/bjr.20190001
17. Bong JH, Song HJ, Oh Y, Park N, Kim H, Park S (2018) Endoscopic navigation system with extended field of view using augmented reality technology. Int J Med Robot Comput Assist Surg 14(2). https://doi.org/10.1002/rcs.1886
18. Yu YL, Yang YJ, Lin C, Hsieh CC, Li CZ, Feng SW, Tang CT, Chung TT, Ma HI, Chen YH, Ju DT, Hueng DY (2017) Analysis of volumetric response of pituitary adenomas receiving adjuvant CyberKnife stereotactic radiosurgery with the application of an exponential fitting model. Medicine 96(4):e4662. https://doi.org/10.1097/MD.0000000000004662
19. Knosp E, Steiner E, Kitz K, Matula C (1993) Pituitary adenomas with invasion of the cavernous sinus space: a magnetic resonance imaging classification compared with surgical findings. Neurosurgery 33(4):610–618. https://doi.org/10.1227/00006123-199310000-00008
20. Micko AS, Wöhrer A, Wolfsberger S, Knosp E (2015) Invasion of the cavernous sinus space in pituitary adenomas: endoscopic verification and its correlation with an MRI-based classification. J Neurosurg 122(4):803–811. https://doi.org/10.3171/2014.12.JNS141083
21. Araujo-Castro M, Pascual-Corrales E, Martínez-Vaello V, Baonza Saiz G, Quiñones de Silva J, Acitores Cancela A, García Cano AM, Rodríguez Berrocal V (2021) Predictive model of surgical remission in acromegaly: age, presurgical GH levels and Knosp grade as the best predictors of surgical remission. J Endocrinol Invest 44(1):183–193. https://doi.org/10.1007/s40618-020-01296-4
22. Hardy J, Vezina JL (1976) Transsphenoidal neurosurgery of intracranial neoplasm. Adv Neurol 15:261–273
23. Wilson G (1979) Neurosurgical management of large and invasive pituitary tumors. Clin Manag Pituit Disord:335–342
24. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D (2019) Key challenges for delivering clinical impact with artificial intelligence. BMC Med 17(1):195. https://doi.org/10.1186/s12916-019-1426-2
25. Egger J, Bauer MH, Kuhnt D, Freisleben B, Nimsky C (2011) Pituitary adenoma segmentation. arXiv preprint arXiv:1103.1778. https://doi.org/10.48550/arXiv.1103.1778
26. Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, Gerig G (2006) User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. NeuroImage 31(3):1116–1128. https://doi.org/10.1016/j.neuroimage.2006.01.015
27. van Rossum G (1995) Python reference manual. Department of Computer Science [CS] R 9525
28. Beare R, Lowekamp B, Yaniv Z (2018) Image segmentation, registration and characterization in R with SimpleITK. J Stat Softw 86:8. https://doi.org/10.18637/jss.v086.i08
29. Yaniv Z, Lowekamp BC, Johnson HJ, Beare R (2018) SimpleITK image-analysis notebooks: a collaborative environment for education and reproducible research. J Digit Imaging 31(3):290–303. https://doi.org/10.1007/s10278-017-0037-8
30. Lowekamp BC, Chen DT, Ibáñez L, Blezek D (2013) The design of SimpleITK. Front Neuroinform 7:45. https://doi.org/10.3389/fninf.2013.00045
31. Chollet F (2015) Keras. GitHub Retrieved from: https.github.com/fchollet/keras
32. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C et al (2016) Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467
33. Cerny M (2022) Fully automated imaging protocol independent system for pituitary adenoma segmentation. GitHub repository https://github.com/DrMartinCerny/pituitary_adenoma_segmentation
34. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, pp 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
35. Nair V, Hinton GE (2010) Rectified linear units improve restricted boltzmann machines. In: Icml
36. Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. J Big Data 6:60. https://doi.org/10.1186/s40537-019-0197-0
37. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980
38. Zou KH, Warfield SK, Bharatha A, Tempany CM, Kaus MR, Haker SJ et al (2004) Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports. Acad Radiol 11(2):178–189

39. Egger J, Bauer MH, Kuhnt D, Carl B, Kappus C, Freisleben B, Nimsky C (2010) Nugget-cut: a segmentation scheme for spherically-and elliptically-shaped 3D objects. In: Joint Pattern Recognition Symposium. Springer, Berlin, Heidelberg, pp 373–382
40. Ikonomakis N, Plataniotis KN, Venetsanopoulos AN (2000) Color image segmentation for multimedia applications. J Intell Robot Syst 28(1):5–20
41. Zukić D, Egger J, Bauer MH, Kuhnt D, Carl B, Freisleben B et al (2011) Glioblastoma multiforme segmentation in MRI data with a balloon inflation approach. arXiv preprint arXiv:1102.0634
42. Takikawa T, Acuna D, Jampani V, Fidler S (2019) Gated-scnn: gated shape cnns for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 5229–5238
43. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH (2021) nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 18(2):203–211
44. Zhang L, Wang X, Yang D, Sanford T, Harmon S, Turkbey B et al (2020) Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE Trans Med Imaging 39(7):2531–2540
45. Bokhorst JM, Pinckaers H, van Zwam P, Nagtegaal I, van der Laak J, Ciompi F (2018) Learning from sparsely annotated data for semantic segmentation in histopathology images. In: International Conference on Medical Imaging with Deep Learning--Full Paper Track
46. Zhang Z, Li J, Zhong Z, Jiao Z, Gao X (2019) A sparse annotation strategy based on attention-guided active learning for 3D medical image segmentation. arXiv preprint arXiv:1906.07367
47. Guo Z, Li X, Huang H, Guo N, Li Q (2019) Deep learning-based image segmentation on multimodal medical imaging. IEEE Trans Radiat Plasma Med Sci 3(2):162–169
48. Le MH, Chen J, Wang L, Wang Z, Liu W, Cheng KTT, Yang X (2017) Automated diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks. Phys Med Biol 62(16):6497
49. Wang J, Berger D, Mattie D & Levman J (2021) Multichannel input pixelwise regression 3D U-Nets for medical image estimation with 3 applications in brain MRI
50. Zhou T, Ruan S, Canu S (2019) A review: deep learning for medical image segmentation using multi-modality fusion. Array 3:100004
51. Pereira S, Pinto A, Alves V, Silva CA (2016) Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans Med Imaging 35(5):1240–1251
52. Isensee F, Kickingereder P, Wick W, Bendszus M, Maier-Hein KH (2017) Brain tumor segmentation and radiomics survival prediction: contribution to the brats 2017 challenge. In: International MICCAI Brainlesion Workshop. Springer, Cham, pp 287–297
53. Isensee F, Kickingereder P, Wick W, Bendszus M, Maier-Hein KH (2018) No new-net. In: International MICCAI Brainlesion Workshop. Springer, Cham, pp 234–244
54. Cui S, Mao L, Jiang J, Liu C, Xiong S (2018) Automatic semantic segmentation of brain gliomas from MRI images using a deep cascaded neural network. J Healthc Eng 2018:4940593. https://doi.org/10.1155/2018/4940593
55. Wang G, Li W, Ourselin S, Vercauteren T (2017) Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In: International MICCAI brainlesion workshop. Springer, Cham, pp 178–190
56. Kamnitsas K, Ledig C, Newcombe V, Simpson JP, Kane AD, Menon DK, Rueckert D, Glocker B (2017) Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med Image Anal 36:61–78. https://doi.org/10.1016/j.media.2016.10.004
57. Zhao X, Wu Y, Song G, Li Z, Zhang Y, Fan Y (2018) A deep learning model integrating FCNNs and CRFs for brain tumor segmentation. Med Image Anal 43:98–111. https://doi.org/10.1016/j.media.2017.10.002
58. Myronenko A (2018) 3D MRI brain tumor segmentation using autoencoder regularization. In: International MICCAI Brainlesion Workshop. Springer, Cham, pp 311–320
59. Clèrigues A, Valverde S, Bernal J, Freixenet J, Oliver A, Lladó X (2020) Acute and sub-acute stroke lesion segmentation from multimodal MRI. Comput Methods Programs Biomed 194:105521
60. Chen L, Wu Y, DSouza AM, Abidin AZ, Wismüller A, Xu C (2018) MRI tumor segmentation with densely connected 3D CNN. In: Medical Imaging 2018: Image Processing, vol 10574. SPIE, pp 357–364
61. Dolz J, Desrosiers C, Ben Ayed I (2018) IVD-Net: Intervertebral disc localization and segmentation in MRI with a multi-modal UNet. In: International workshop and challenge on computational methods and clinical applications for spine imaging. Springer, Cham, pp 130–143
62. Dolz J, Gopinath K, Yuan J, Lombaert H, Desrosiers C, Ben Ayed I (2019) HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation. IEEE Trans Med Imaging 38(5):1116–1126. https://doi.org/10.1109/TMI.2018.2878669
63. Rokach L (2010) Ensemble-based classifiers. Artif Intell Rev 33(1):1–39
64. Nie D, Wang L, Adeli E, Lao C, Lin W, Shen D (2019) 3-D fully convolutional networks for multimodal isointense infant brain image segmentation. IEEE Trans Cybern 49(3):1123–1136. https://doi.org/10.1109/TCYB.2018.2797905
65. Castro DG, Cecílio SA, Canteras MM (2010) Radiosurgery for pituitary adenomas: evaluation of its efficacy and safety. Radiat Oncol 5:109. https://doi.org/10.1186/1748-717X-5-109
66. Girkin CA, Comey CH, Lunsford LD, Goodman ML, Kline LB (1997) Radiation optic neuropathy after stereotactic radiosurgery. Ophthalmology 104(10):1634–1643. https://doi.org/10.1016/s0161-6420(97)30084-0