Computational and Structural Biotechnology Journal 19 (2021) 3077–3086

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Image-driven classification of functioning and nonfunctioning pituitary adenoma by deep convolutional neural networks

Hongyu Li [a,b,1], Qi Zhao [a,1], Yihua Zhang [c,1], Ke Sai [d], Lunshan Xu [c], Yonggao Mou [d], Yubin Xie [a], Jian Ren [a], Xiaobing Jiang [a,d,e,*]

[a] State Key Laboratory of Oncology in South China, Cancer Center, Collaborative Innovation Center for Cancer Medicine, School of Life Science, Sun Yat-sen University, Guangzhou, Guangdong 510060, China
[b] School of Data and Computer Science, Sun Yat-sen University, Guangzhou, Guangdong 510060, China
[c] The Department of Neurosurgery, Daping Hospital, Army Medical University, Chongqing 400042, China
[d] Department of Neurosurgery/Neuro-oncology, Sun Yat-sen University Cancer Center. State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Guangzhou, China
[e] Jiangmen Central Hospital, Affiliated Jiangmen Hospital of Sun Yat-Sen University, Jiangmen, China

ARTICLE INFO

Article history:
Received 30 October 2020
Received in revised form 5 May 2021
Accepted 13 May 2021
Available online 14 May 2021

Keywords:
Pituitary adenomas
MRI
Deep learning

ABSTRACT

The secreting function of pituitary adenomas (PAs) plays a critical role in making the treatment strategies. However, Magnetic Resonance Imaging (MRI) analysis for pituitary adenomas is labor intensive and highly variable among radiologists. In this work, by applying convolutional neural network (CNN), we built a segmentation and classification model to help distinguish functioning pituitary adenomas from non-functioning subtypes with 3D MRI images from 185 patients with PAs (two centers). Specifically, the classification model adopts the concept of transfer learning and uses the pre-trained segmentation model to extract deep features from conventional MRI images. As a result, both segmentation and classification models obtained high performance in two internal validation datasets and an external testing dataset (for segmentation model: Dice score = 0.8188, 0.8091 and 0.8093 respectively; for classification model: AUROC = 0.8063, 0.7881 and 0.8478, respectively). In addition, the classification model considers the attention mechanism for better model interpretation. Taken together, this work provides the first deep learning-based tumor region segmentation and classification models of PAs, which enables early diagnosis and subtyping PAs from MRI images.

1. Introduction

Pituitary adenomas (PAs) account for approximately 15% of all intracranial tumors [1]. Recently, the prevalence of PAs has increased to 115 cases per 100,000. The increase is probably due to the rising use of diagnostic medical imaging and enhanced awareness [2]. Clinically, PAs may cause considerable mortality due to their mass effects and the hypersecretion of one or more pituitary hormones [3]. Nonfunctioning pituitary adenomas (NFPAs) are usually associated with mass effects, including headache, visual defects, and the development of hypopituitarism. Secretory adenomas produce one or more pituitary hormones such as prolactin, growth hormone (GH), adrenocorticotropic hormone (ACTH), and thyroid-stimulating hormone (TSH), causing phenotypic clinical symptoms, including loss of libido, hyperthyroidism, acromegaly and Cushing's syndrome [2]. Treatments are determined often according to the size of the lesion and the status of the secreting function. Therefore, early detection and management of pituitary adenomas are non-negligible in improving the prognosis of patients with PAs.

Contrasted magnetic resonance imaging (MRI) is the mainstream method to evaluate the location and size of PAs. The GrowCut algorithm is freely available as a module for the medical image computing platform 3D Slicer [4] and has been used in a recent study to segment PAs based on MRI [5]. Assessment of pituitary hormones is another essential factor to determine the types and treatment modalities for adenomas. Analysis of MRI images is labor intensive and highly variable among radiologists. Besides, the clinical testing of all pituitary hormones is usually time

* Corresponding author at: Sun Yat-sen University Cancer Center, Guangzhou, China.
E-mail addresses: renjian@sysucc.org.cn (J. Ren), jiangxiaob1@sysucc.org.cn (X. Jiang).
[1] Contribute equally to this work.

https://doi.org/10.1016/j.csbj.2021.05.023

consuming, economically costly, and even remains unavailable in many local medical centers. Therefore, constructing an artificial intelligence system will assist the radiologists to obtain more reliable PAs diagnoses from the conventional MRI images, and thus led to time and cost saving.

For the past decades, deep architecture [6] has garnered a great amount of attention in various fields due to its representational power. Deep learning methods, especially convolutional neural networks, have shown great potential power in the assessment of medical problems, such as cancer classification, tumor segmentation and survival prediction [7–10]. Furthermore, reports demonstrated that a computer-aided diagnosis (CAD) system can accurately diagnose PAs through MRI images [11]. Ranging from the LeNet architecture [12] to Residual-style Networks [13–15], the network architectures have become deeper and wider for rich representations. On the large labelled datasets, CNNs have shown good performance in different computer vision tasks, such as ImageNet [16,17], Microsoft COCO [18].

However, CNNs cannot be trained efficiently from scratch for medical images due to small datasets. For the small dataset scenario, an effective method to employ CNNs to medical image classification is transfer learning [19]. Transfer learning is a deep learning approach in which a model that has been trained for one task is used as a starting point to train a model for a similar task. Usually, fine-tuning a network from pre-trained network with transfer learning is more computationally efficient than training a network from scratch. Transfer learning techniques have been shown to be successful in several medical applications, such as the diagnosis of Alzheimer's disease [20], magnetic resonance (MR) image segmentation [21] and microscopy images [22]. In the previous studies, there are two types of transfer learning approaches, such as (i) use off-the-shelf trained CNN models over a large dataset of natural images as a feature extractor and train a separate learning method for classification [23–26]. (ii) Use pre-trained CNNs and apply fine-tuning to the application of medical images database [27,28].

Although increasing the depth and width of network architecture could help improve model performance, models tend to produce many redundant features and make convergence difficult. Many researchers have investigated a different aspect of the architecture design, termed attention. The significance of attention mechanisms has been studied extensively in previous research literature [29,30]. Introducing attention mechanisms in network architecture design is more computationally efficient.

In this study, we employed a 3D deep learning algorithm to generate three fully automated segmentation models based on conventional MRI images. Then, we exploited transfer learning methodology for feature representation based on trained segmentation models to enhance classification accuracy. Our method is different from (i) discussed above. We did not agree with the idea of transfer learning from natural 2D images due to the purpose of making full use of the 3D contextual information of MRI images. We used pre-trained segmentation CNN models which are trained on relatively large 3D MRI patches to initiate the classification models and fine-tuned the classification models. Meanwhile, we adopted an attention module to automatically refine extracted features, making the classification model more concerned with features that contain significant information.

2. Materials and methods

2.1. Study cohorts

From January 2017 to March 2021, patients with PAs surgically treated in Sun Yat-sen University Cancer Center (SYSUCC) were retrospectively reviewed. The following inclusion criteria were used: (1) pathological confirmation of a pituitary adenoma, (2) a completed evaluation of pituitary hormones, and (3) availability of four contrast MRI data (T1-weighted image, contrast-enhanced T1-weighted image, T2-weighted image, and T2-weighted FLAIR image). All the procedures in the current study were approved by the ethics committee of SYSUCC. Written informed consent was obtained from all the patients. A total of 168 patients from SYSUCC were enrolled and divided into training (n = 100), internal validation 1 (n = 44) and validation 2 (n = 24) sets. For external independent testing, a total of 17 patients from Daping Hospital, Army Medical University who were treated surgically from January 2018 to December 2019 were recruited as a testing set (Fig. 1).

2.2. Clinical and laboratory evaluation

Demographic and clinical-pathological variables (see **Sup. Table 1**), including age, sex, pathological diagnoses, hormone levels and tumor size were collected using an electronic medical record system. Venous blood samples were collected in the morning after a 12-hour fast. Hormone assays including prolactin, cortisol, TSH, free T4 (fT4), GH, luteinizing hormone (LH), follicle-stimulating hormone (FSH), estradiol and total testosterone were measured with methods described previously [31]. The diagnosis of a pituitary adenoma was based on physical examination, contrast magnetic resonance imaging, hormone assay, and pathological observation. Accordingly, PAs were further classified into functioning and nonfunctioning adenomas based on their hormone-secreting status [2,32].

The diagnosis of prolactinoma was confirmed by constantly increasing prolactin levels (>100 ng/ml). The diagnosis of a GH secreting adenoma was based on an increased GH level, a lack of GH suppression under 1 ng/ml during an oral glucose load (75 g), and an IGF-1 level over reference for sex and age [33]. The invasiveness of the tumor was evaluated according to the Knosp's classification [34].

2.3. MRI data collection and process

An MRI was performed on a 1.5 or 3.0 T system. Four sequences, T1W, contrast-enhanced T1W (T1CE), T2W and T2W-FLAIR, were acquired for further analysis. The MRI images in the training, validation and testing datasets were preprocessed in the following manner:

a) DICOM files were converted to NIFTI format.

b) All MRI volumes were rigidly registered to the same T2 anatomic template and resampled to 1 mm voxel resolution through the Oxford Center for Functional MRI of the Brain's (FMRIB) Linear Image Registration Tool (FLIRT) [35,36] from the FMRIB Software Library [37–39].

c) The volumes of all modalities were skull-stripped using the Brain Extraction Tool (Bet) [40].

d) All MRI volumes used N4BiasCorrection [41] to remove random field inhomogeneity.

e) Rescaling intensity range to [0, 1].

The tumor contours for 185 subjects were manually labeled and validated by two neurosurgeons (XB Jiang and YH Zhang). All segmentation was performed using the MITK software [42], taking about 30 min per subject. To deal with ambiguities in tumor contours' definition, we had all subjects labeled by the two neurosurgeons and subsequently fused the results to obtain a single consensus tumor contour for each subject.

To take advantage of the 3D contextual information of the MRI, we randomly extracted patches from preprocessed T1W, T2W, Flair and T1CE images on the axial, sagittal and coronal views as

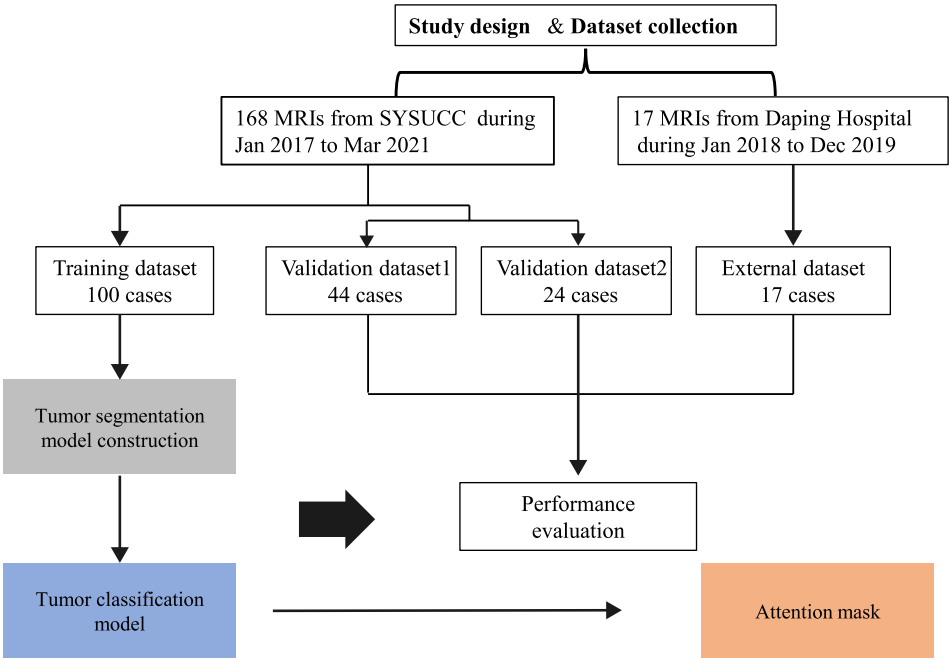


Fig. 1. Schematic overview of the study.

the input images for the segmentation task. Four cropped patches (96 × 96 × 96 voxels) were randomly extracted from preprocessed T1W, T2W, Flair and T1CE images for each subject on the fly during every training iteration. Therefore, each segmentation model totally used 4*30000 patches for training. Considering the limited contribution of normal tissues to classify secreting types of PAs, the classification model was only trained with patches (96 × 96 × 96 voxels) that were extracted based on the tumor center calculated from the tumor mask. Therefore, a total of 4*3000 patches were applied for training for each classification model.

2.4. Deep learning model design

2.4.1. CNN for PAs segmentation

Our segmentation network was based on Residual Unet architecture (Fig. 2A). Similar to a standard Unet [43], which is consisted of an analysis path (encoder part) and a synthesis path (decoder part). This network had 35 convolutional layers and was made of the following blocks: Resblock, Conv3D for down-sampling and Deconv3D for up-sampling. Each Resblock used in this study, as shown in **Sup. Fig. 1**, consisted of a shortcut and a few stacked layers: the convolutional layers and the parametric rectified linear unit (PReLU) layers. The analysis path consists of the repeated Resblock, each followed by a Conv3D block that did a $3 \times 3 \times 3$ convolution operation with stride 2 in each dimension for down-sampling. In the synthesis path, the repeated Resblock was followed by a Deconv3D block that did a $3 \times 3 \times 3$ transposed convolution operation with stride 2 in each dimension for an up-sampling of the feature map. Shortcut connections from layers of equal resolution in the analysis path provided the essential high-resolution features to the synthesis path. In the last layer, a $3 \times 3 \times 3$ convolution operation reduces the number of output channels to the number of labels, which is 2 in our case. The segmentation model would output a pixel-wise mask of the input image which 1 stands tumor tissue and 0 stands normal tissue. The segmentation model contains a total of 76,967,968 parameters for learning.

2.4.2. CNN classifier for secreting function type prediction

Transfer learning aims to transfer knowledge between related source and target domains [44]. Transfer learning methods can be divided into instance-transfer, feature-transfer, parameter-transfer and relational-knowledge-transfer approaches. These approaches [23,24] focused on feature transfer between datasets under different tasks, even from nonmedical datasets. Here, we didn't adapt the idea of transfer learning from natural 2D images due to the purpose of making full use of the third dimension of MRI images. In comparison, we assumed that the source task (segmentation task) and the target task (classification task) here shared some parameters or prior distributions with the hyper-parameters of the models. Therefore, we transferred the learned weights from the segmentation models to train the classification models.

Our proposed classification model, as shown in Fig. 2B, made use of the trained analysis path (encoder) in the segmentation model to extract the features of the MRI images. Combined with an attention module, our model learned to suppress irrelevant regions in an input image while highlighting salient features. The attention module (shown in Fig. 2C) used in our network is a Convolutional Block Attention Module (CBAM) [45]. The input feature was refined based on an attention mask generated by CBAM. The weights of the analysis path in the classification network were transferred from trained segmentation models. The decoder was modified to adapt to the classification task. The synthesis path and final layer of the segmentation network were removed. Instead, a 3D average pooling layer and a fully connected layer followed by a softmax layer, with an output size of two, were inserted. The classification would predict the probability of the patient with functioning pituitary adenomas (FPAs). A total of 38,497,810 parameters is available for learning for the classification model.

2.4.3. Multi-view model combination

To make full use of 3D contextual information, both segmentation and classification models were trained on extracted axial, sagittal, and coronal images, respectively. In the validation and testing procedure, predictions for segmentation and classification

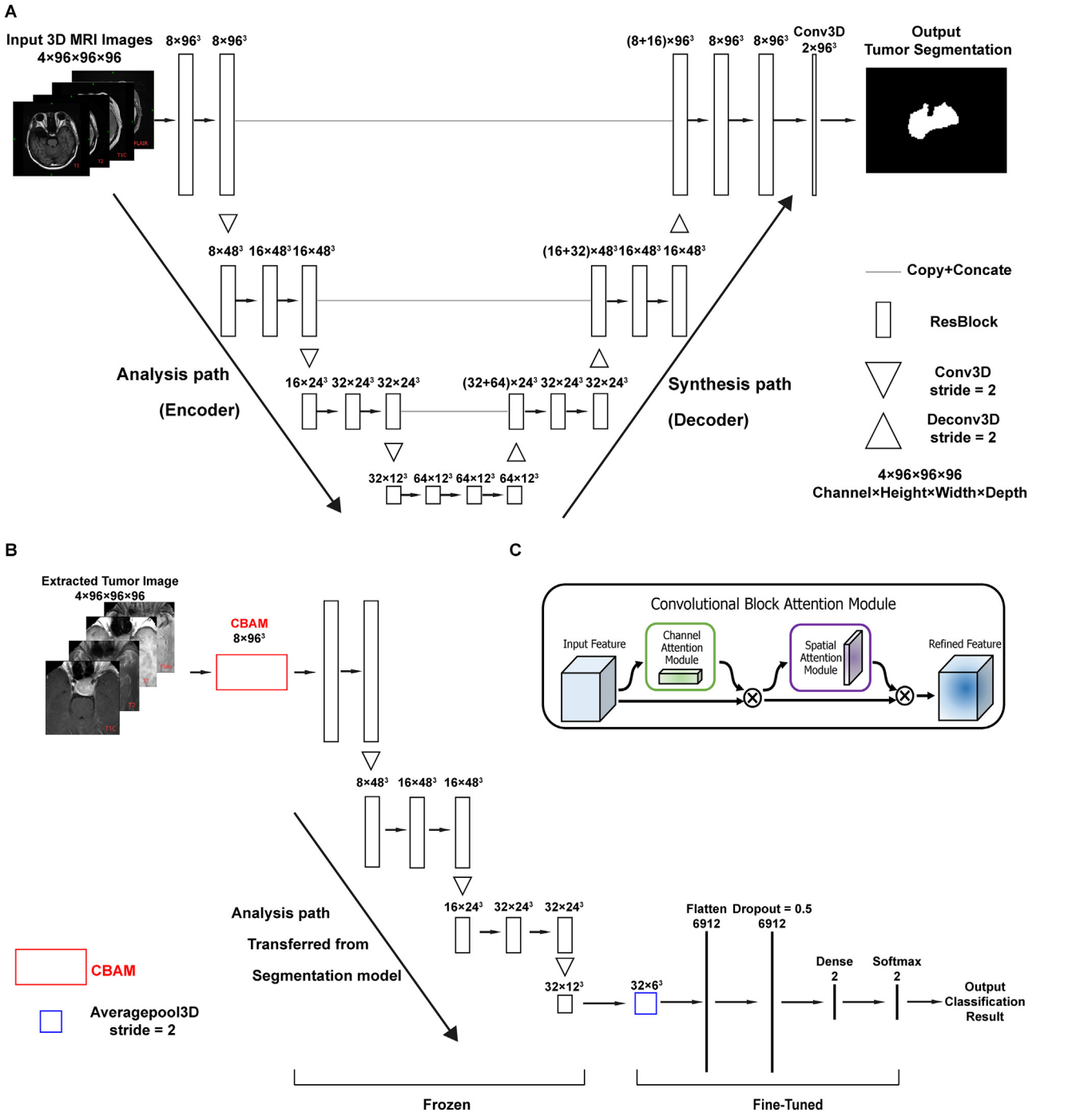


Fig. 2. Segmentation and classification network overview. (A) Segmentation network architecture. (B) Classification network architecture. (C) Convolutional block attention module used in the classification network.

on different views were combined to obtain the final predictions as combined model prediction results. At test time, for each segmentation network structure, the corresponding versions of trained models were used to obtain a segmentation result from these three views, and these softmax outputs were averaged to obtain a single fused result. The classification network structure was tested similarly.

2.5. Experimental setup

The segmentation and classification networks were implemented in the TensorFlow library and NiftyNet platform in Python [46,47] and were trained on an NVIDIA GTX2080Ti GPU. The main hyper-parameters of the two architectures are shown in **Sup. Table 2**.

For the classification task, a two-phase training was used. In the first phase, the network was trained on training patches for 2500 iterations. During the first phase, all layers except the attention module and the fully connected layers were fixed. In the second phase, the network was trained on training patches for another 500 iterations. During the second phase, all layers were trainable.

For the segmentation task, three same architecture networks (as shown in Fig. 2A) were trained based on the MRI patches extracted on the axial, sagittal and coronal views, separately. Sim-

ilarly, three same architecture classification networks (as shown in Fig. 2B) were trained on the axial, sagittal and coronal views separately based on transferring the encoder part of the three trained segmentation networks.

2.6. Statistical analysis

Statistical analysis for demographic variables was performed by using chi-square tests for categorical data and one-way ANOVA for continuous data. For the segmentation model, the experimental results were evaluated based on two main metrics, namely, the Dice similarity coefficient (DSC) and the Hausdorff distance. For tumor regions, we obtained a binary map with algorithmic predictions $P\epsilon\{0,1\}$ and the experts' consensus truth $\epsilon\{0,1\}$, and we calculated the Dice score which is defined as:

$$\mathrm{Dice}(\mathrm{P},\mathrm{T}) = \frac{2|P \cap T|}{|P| + |T|} \quad (1)$$

For surface distance evaluation, we calculated the Hausdorff distance. For two point sets X and Y, the one-sided HD from X to Y is defined as:

$$\mathrm{hd}(\mathrm{X},\mathrm{Y}) = \max_{x \in X} \min_{y \in Y} ||x - y||_2 \quad (2)$$

And similarly, for hd(Y, X):

$$\mathrm{hd}(\mathrm{Y},\mathrm{X}) = \max_{y \in Y} \min_{x \in X} ||x - y||_2 \quad (3)$$

Finally, the Hausdorff distance is defined as:

$$\mathrm{Haus}(\mathrm{P},\mathrm{T}) = \max(hd(P,T), hd(T,P)) \quad (4)$$

For the classification model, the classification performance was evaluated by generating receiver operating characteristics (ROC) and precision-recall (PR) curves. The AUROC among different models was compared by Delong's method [48].

3. Results

3.1. Patient characteristics

The flow diagram of this study is shown in Fig. 1. A total of 185 patients were included. As shown in **Sup. Table 1,** no significant differences in sex, age, Knosp's Grade, tumor type, tumor volume and diameter were observed among the training, validation 1, validation 2 and the external testing datasets.

3.2. Model construction for PAs segmentation from MRI images

We randomly selected one patient's MRI image in testing dataset and visualized the segmentation results (shown in **Sup. Fig. 2**). The MRI scan slices on the axial view, sagittal view and coronal view are visualized in **Figure S2(A)**, **Figure S2(B)** and **Figure S2 (C)**, respectively. From these example segmentations, our model had a promising performance for 3D MRI slices. Table 1 presents quantitative evaluations in the validation dataset 1. It shows that the axial, sagittal, and coronal models achieved average Dice scores of 0.7942, 0.8024 and 0.8082 for the whole tumor. Using the multi-model ensemble method, the multi-view combined model achieved the best performance in the validation dataset 1 (average Dice score of 0.8188) and was better than GrowCut algorithm (average Dice score of 0.7014). To further evaluate our proposed model, we collected another 24 samples as validation dataset 2 to validate. As shown in **Sup. Table 3**, our proposed segmentation model still achieved the better performance (average Dice score close to 0.810 and average Hausdorff distance close to 5.352 mm) than GrowCut algorithm (average Dice score close to 0.689 and average Hausdorff distance close to 33.605 mm). In the testing dataset, the axial, sagittal and coronal models achieved a similar Dice score (shown in Table 2) for the whole tumor. Similarly, the best performance was also done by the combined model with an average Dice score of 0.81. The combined model still achieved a better performance than GrowCut algorithm (average Dice score of 0.6893). These results demonstrated the potential of our segmentation models in 3D MRI segmentation tasks.

3.3. Classification model for predicting functioning and nonfunctioning PAs

Transfer learning and an attention module were applied to explore feature representations. The attention-based model (Att model) was validated and assessed by comparison with the model trained by random initialization (RI model) and the model trained by transfer learning only (TF model). The RI and TF models shared the same architecture, which only removed the attention module in comparison with our proposed attention-based classification network. Hence, the random initializing model (RI model) and the transfer-learning only model (TF model) were the baseline models.

A 4-fold cross-validation was performed in the training dataset by randomly shuffling the dataset and distributing them into 4 groups (75 samples for training and 25 samples for in-training validation). Validation and testing datasets were used to validate our proposed model with two baseline models after cross-validation.

To evaluate the prediction performance of the proposed classification model, we performed a 4-fold cross-validation in training dataset on the axial, sagittal and coronal views. Fig. 3(A-D) shows the mean values of AUC and ROC curves of the RI, TF and Att models trained on the different plane views. Fig. 3E presents AUROC comparison results for RI, TF and Att model on axial, sagittal, coronal and combined views. **Sup. Table 4** presents quantitative evaluations of an AUROC comparison for the RI, TF and Att models on the different plane views. As a result, the Att models trained on the axial, sagittal and coronal plane views showed a performance under 4-fold cross-validation with the area under the ROC curve close to 0.79. Similarly, the multi-view combined Att model achieved the best performance (AUC = 0.801; 95% CI, 0.738–0.855) and was significantly better (P < 0.0001) than the combined RI (AUC = 0.709; 95% CI, 0.639–0.772) and the combined TF (AUC = 0.713; 95% CI, 0.643–0.776) model.

To check the robustness of our proposed model, we performed the 10-fold cross-validation in training dataset on the axial, sagittal and coronal views. **Sup. Table 5** presents quantitative evaluations of an AUROC comparison for the RI, TF and Att models on different plane views. Under 10-fold cross-validation, the multi-view combined Att model achieved the best performance (AUC = 0.792; 95% CI, 0.726–0.849) and was significantly better than the combined RI (P = 0.0122, AUC = 0.724; 95% CI, 0.653–0.788) and the combined TF (P = 0.0307, AUC = 0.745; 95% CI, 0.675–0.807) model. These results suggested that our proposed classification model was reliable.

To rigorously evaluate the prediction and generalizability performance of our proposed classification model, we next compared the combined Att, RI and TF models in the validation and testing datasets. Fig. 4 presents ROC curves, PR curves for combined RI, TF, Att model and the confusion matrix, diagnostic performances for combined Att model in validation dataset 1 and testing dataset. **Supplementary Tables 6–7** present the quantitative ROC analysis and comparisons in the validation dataset 1 and testing dataset. The results show that the combined Att model achieved the best performance in the validation (AUROC = 0.8063; 95%CI, 0.708–0.883) and testing (AUROC = 0.8478; 95%CI, 0.725–0.947) datasets. The performance of the combined Att model in the validation data-

Table 1
Dice and Hausdorff measurements between the proposed method and GrowCut algorithm in validation dataset 1. Bold numbers indicate the best performance values on Dice and Hausdorff measurements.

View	Dice_mean	Dice_std	Hausdorff_mean (mm)	Hausdorff_std (mm)
Axial	0.7942	0.0895	7.9551	6.2622
Sagittal	0.8024	0.1134	7.984	8.6931
Coronal	0.8082	0.0828	7.177	3.8330
Combined	**0.8188**	**0.0763**	**6.4735**	**3.3578**
GrowCut	0.7014	0.0595	27.607	6.7506

Table 2
Dice and Hausdorff measurements between the proposed method and GrowCut algorithm in testing dataset. Bold numbers indicate the best performance values on Dice and Hausdorff measurements.

View	Dice_mean	Dice_std	Hausdorff_mean (mm)	Hausdorff_std (mm)
Axial	0.7652	0.1159	11.2054	6.5137
Sagittal	0.7792	0.0991	11.2809	8.7102
Coronal	0.7646	0.1169	12.7353	9.0150
Combined	**0.8093**	**0.0769**	**9.3599**	**5.4566**
GrowCut	0.6893	0.0653	28.2917	6.6768

set 1 and testing dataset was comparable with the performance in the training dataset (AUC = 0.801; 95% CI, 0.738–0.855). The diagnostic performance of our proposed model achieved an accuracy of 0.7083 with the Youden's Index of 0.1667) in validation dataset 2 (**Sup. Tables 8-9**).

To detect the classification performance within subgroups divided by clinical characteristics, we run the model in a combined dataset sub-grouped by gender and age. A total of 85 patients (27 FPA and 58 NFPA) were included, and 37 of them are female, with a median age of 48. As shown in **Sup. Tables 10-11**, the proposed classification model achieved similar performance (AUROC = 0.7937 in female subgroup, 0.7929 in male subgroup, 0.8108 in older subgroup, and 0.7976 in young subgroup).

3.4. Model interpretation with the attention mask

Models trained with an attention module could learn to suppress irrelevant regions in the input MRI images while highlighting the salient features. To determine how our proposed models identify the tumor region from the MRI images, attention maps were generated using contrasted T1W scan to exhibit where and what the models focus on.

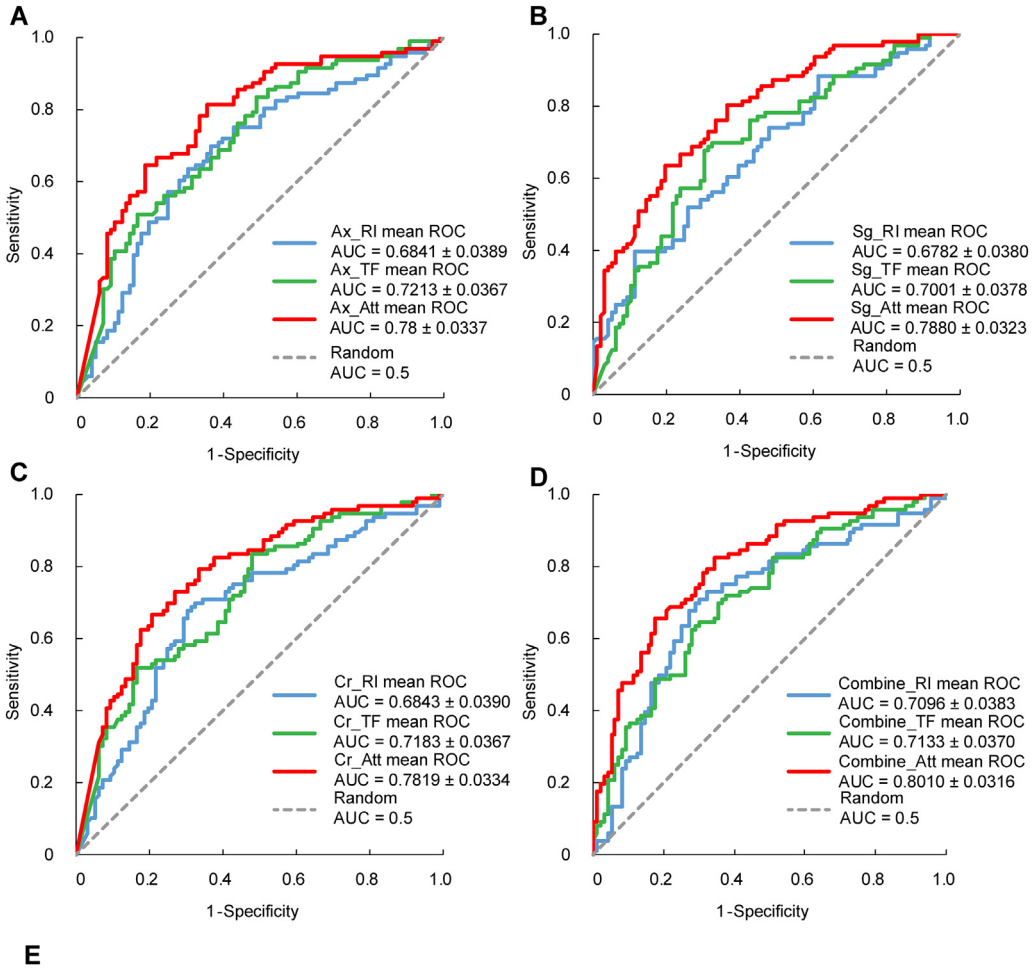
Fig. 5 showed the T1CE scan and its corresponding attention map for the patient with NFPA (Fig. 5(A-B)) and those with FPA (Fig. 5(C-D)). The degree of the attention weights is marked with different colors, where red represents the most attention paid by the model. Heterogeneous colors distributed among the contrasted T1W image indicate that the model trained with an attention mechanism pays different attention to the regions. As shown in Fig. 5, the profile of the color distribution in functioning and nonfunctioning pituitary adenomas is different, where the tumor region is marked in deep red for the nonfunctioning pituitary adenomas but in light red for functioning ones. Additionally, the tissues with the highest signal on contrasted T1W, including a cavernous and basal sinus, were marked in red, and the normal brain tissues were marked in light colors. Intriguingly, regions with the lowest signal on contrasted T1W were also in red, such as basal cisterns and the fourth ventricle.

4. Discussion

MRI is generally preferred over CT for the diagnosis of PAs because of its superior definition of small lesions in the pituitary sella and its improved anatomic definition before surgery. At present, surgery is the standard first-line therapy for the treatment of patients with non-functioning PAs. Therefore, the accurate tumor contours segmentation and precise classification of secreting function types for PAs are crucial steps in surgical and treatment planning. With the recent progress of AI algorithm on MRI images, we expected a similar application that provides a more efficient manner to diagnosis and classify PAs directly with MRI images. Eventually, we developed two CNN models for PAs segmentation and classification of functioning and nonfunctioning PAs based on conventional T1W, T2W, Flair and T1CE MRI sequences.

As a result, the combined segmentation model achieved 0.8188, 0.8091 and 0.8093 Dice scores for the whole tumor in two validation datasets and a testing dataset by using multi-model ensemble methods. For segmentation tasks, we found several factors that may explain the high performance achieved by our segmentation model. First, the use of a 3D convolutional neural network compared to a 2D convolutional neural network could make full use of the third MRI image dimension. Additionally, the Resblock used in the segmentation model allows us to build a deeper network and take advantage of the deep neural network's powerful representational ability. Finally, the 3D segmentation models were trained on image patches extracted on axial, sagittal and coronal views. The results in Tables 1-2 revealed that joint use of the three models' predictions achieved substantially improved performance after using any one model prediction in the validation and testing datasets. Moreover, the inference time was approximately 32.8 s per patient (using one RTX 2080Ti GPU). A previous study has shown that there was reduction of intra-observer variation (by 36.4%), reduction of interobserver variation by 54.5%, and time savings of 39.4% with automated segmentation model assistance for nasopharyngeal carcinoma [49]. Due to its short inference time and the accuracy of tumor segmentation, the automated segmentation model could be used as a PA computer-aided diagnosis tool for radiologists.

Additionally, the combined classification model also achieved a high predicting performance with accuracies of 72.7% (AUROC = 0.8063) in validation dataset 1, 70.8% (AUROC = 0.7881) in validation dataset 2 and 82.3% (AUROC = 0.8478) in the testing dataset. During this task, we mainly investigated whether the integration of the transfer-learning method as well as the attention mechanism could substantially improve the overall performance. Four-fold cross-validation shows that transfer-learning based models (TF and Att models) achieved higher AUC than random initial models, while the attention-based models (Att models) also



E

View	RI~TF	TF~Att	RI~Att
Axial	P = 0.2482	P = 0.0009	P = 0.0055
Sagittal	P = 0.5146	P = 0.0002	P = 0.0004
Coronal	P = 0.2941	P = 0.0100	P = 0.0028
Combined	P = 0.8527	P < 0.0001	P < 0.0001

Fig. 3. ROC analysis under 4-fold cross-validation. (A) the mean ROC curves of RI, TF and Att model trained on axial view. (B) The mean ROC curves of RI, TF and Att model trained on sagittal view. (C) The mean ROC curves of RI, TF and Att model trained on coronal view. (D) The mean ROC curves of multi-view combined of RI, TF and Att model. (E) Comparison results of averaged AUROC under 4-fold cross-validation for RI, TF and Att model on axial, sagittal, coronal and combined views.

obtained a higher AUC value than transfer-learning models themselves (TF models). Further comparisons between TF, RI and Att models in training, validation and testing datasets proved our assumption that the source task (segmentation task) and the target task (classification task) shared certain parameters and similar prior distributions with the hyper-parameters of the models. Moreover, comparing to make network architecture deeper and wider, attention mechanisms used in classification models aim to improve classification performance without increasing in models' complexity and computation. The generated attention masks allow the proposed Att model to concentrate adaptively on the abnormal regions. Experimental results in the training, validation and testing datasets (as shown in **Sup. Tables 4-6**) demonstrated that the attention module in our transfer learning-based models plays a critical role. The benefit comes from encoding a top-down attention mechanism into a bottom-up top-down feedforward convolutional structure in the classification model, so it can learn the specialized features of the input MRI images. Similarly, by using a multi-model ensemble method, the combined Att model was more robust and achieved the best performance.

Model comparisons in the validation and testing datasets demonstrated that introducing attention module enables models to perform better. By visualizing the attention masks, we found that our models pay more attention to some regions with the lowest signal. It is unclear why these areas attract attention from the machine, and much more works are warranted to investigate the underlining mechanisms and its clinical significance. In the model of attention mask, the machine may be aware of some unique features from the MRI, and thus help the radiologists to differentiate functioning and nonfunctioning preoperatively. As far as we know, there is no theoretical basis for the classification of PAs based on MRI. However, some studies have explored the correlation

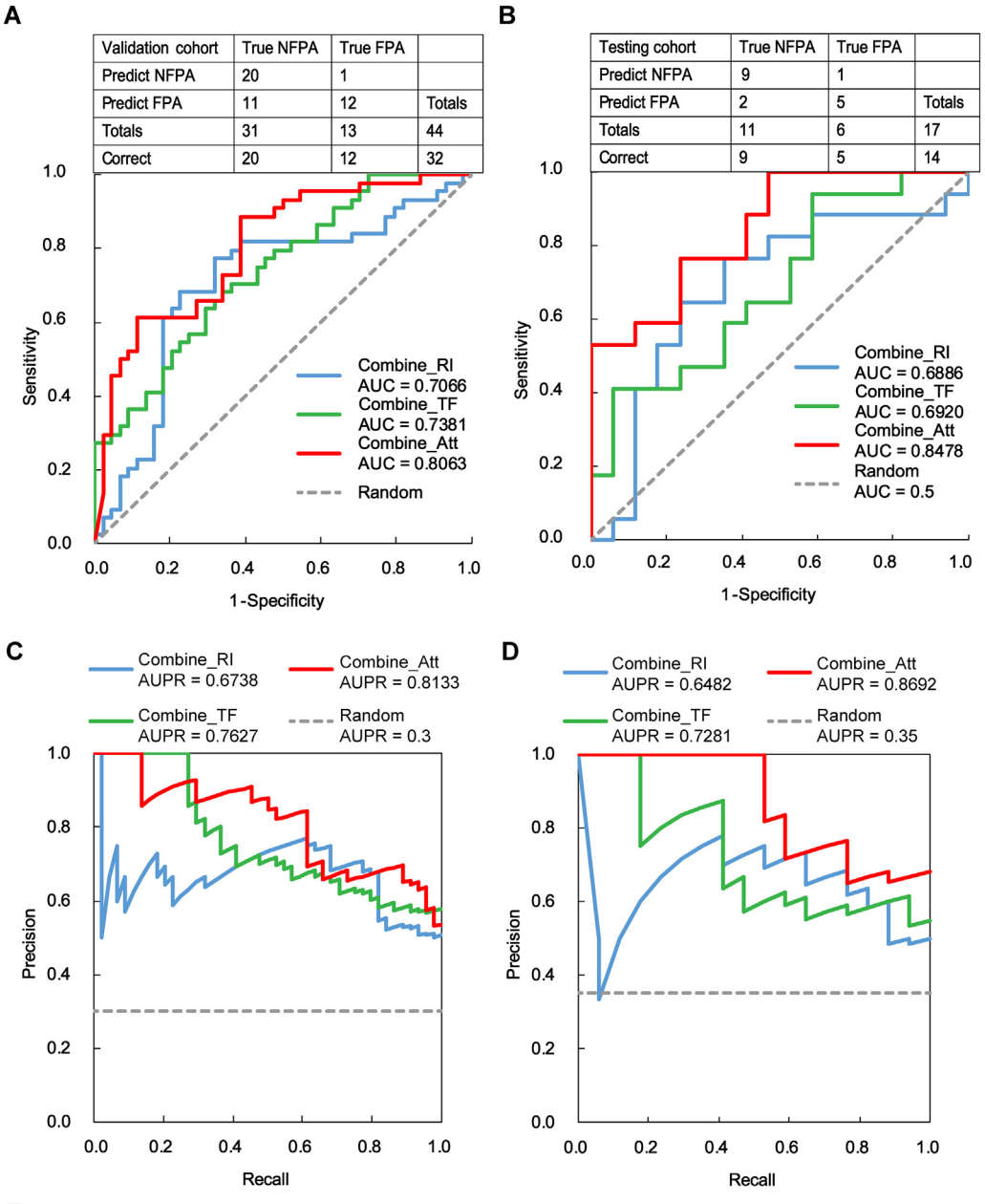


Fig. 4. Evaluation of classification model in validation and testing datasets. The ROC curves of multi-view combined RI, TF and Att model and the confusion matrix for multi-view combined Att model in the (A) validation dataset 1 and (B) testing dataset. Precision-Recall Curves of multi-view combined RI, TF and Att model in the (C) validation dataset 1 and (D) testing dataset. (E) The diagnostic performance of multi-view combined Att model in validation dataset 1 and testing dataset.

between MRI and pathological features. For example, Peng et al. suggested a machine learning model which can immunohistochemically classify PAs with an MR-based radiomic analysis [50]. Similarly, diffusion-weighted imaging (DWI) MRI was reported to differentiate functional types of pituitary macro-adenomas in a small set of patients [51]. These data indicate that an MRI-based deep convolutional neural network is potential to aid in classifying the functioning status of PAs based on preoperatively MRI.

The main limitation of this study is the sample size, resulting in fewer micro-adenomas for build segmentation models, which could cause poor generalization ability in unseen micro-adenomas data. In addition, the collected ACTH patients were relatively small comparing to GH and PRL patients, which could limit the accuracy of our classification model for ACTH patients. Our coming efforts will include more data points from those kinds of patients as well as a large sample size to further improve the accuracy of the models. Moreover, we failed to evaluate the ability segmentation model in analyzing tumor constituents. The constituent of PAs is a critical factor for surgical plans, which should be addressed in the future works. Finally, as only newly diagnosed and surgically treated PAs were recruited for the analysis, this model could only apply to patients with

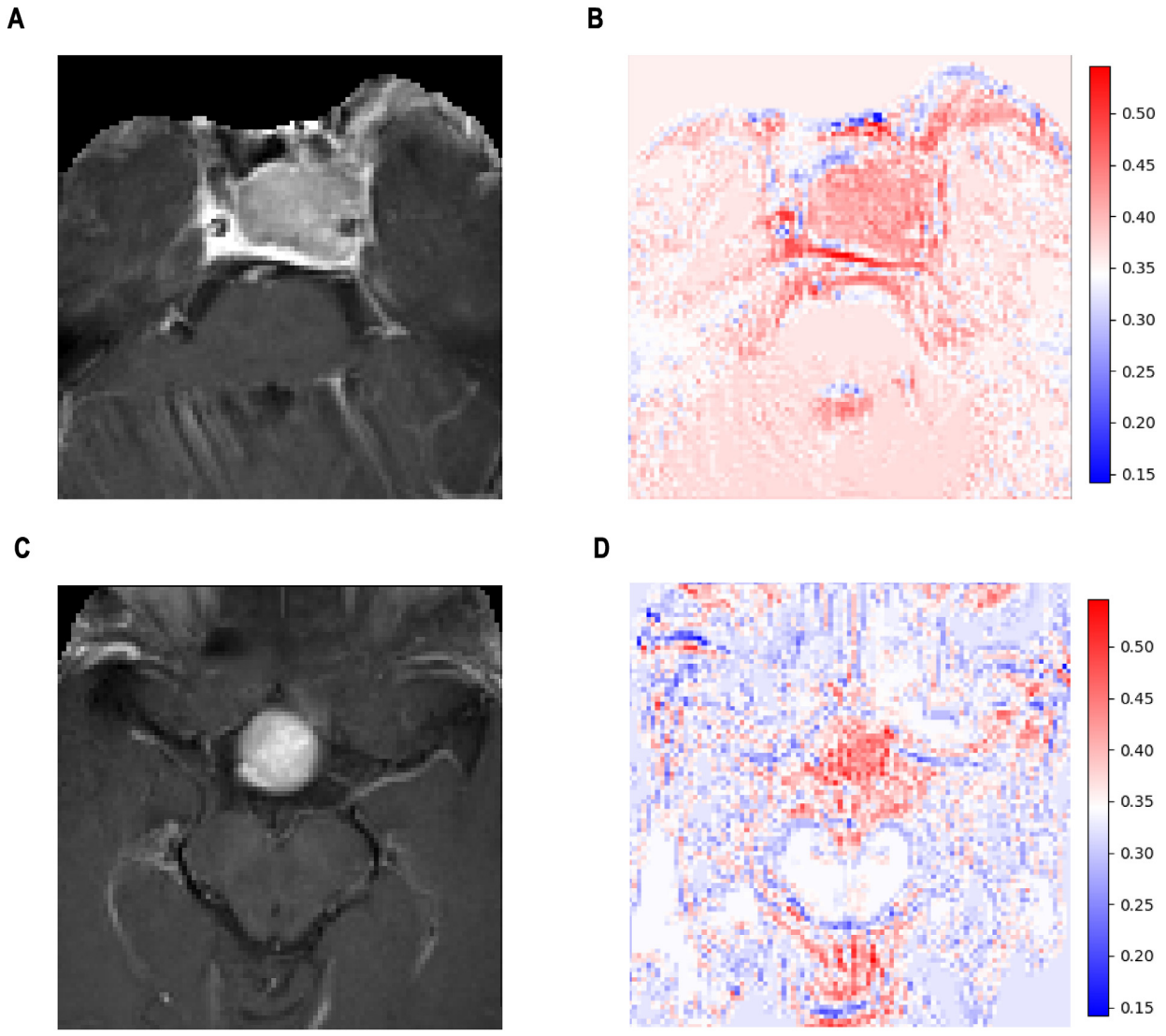


Fig. 5. (A) Original contrast enhanced T1w (T1CE) image for the patient with NFPA. (B) Attention mask of the same T1CE image for the patient with NFPA. (C) Original contrast enhanced T1w (T1CE) image for the patient with FPA. (D) Attention mask of the same T1CE image for the patient with FPA. Basal cisterns and the fourth ventricle with low signals were marked as red in attention mask. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

primary PAs which should be surgically, but not for the recurrent and/or those treated medically. Therefore, much more work is encouraged to investigate the role of deep learning in predicting the tumor constituent.

Collectively, this research was the first computer-based description to predict subtypes of PAs by conventional 3D MRIs, and the models showed preferable performance in the testing set, enabling supporting early diagnosis and treatment plan for PAs. Our models have the potential to be used more widely as a practical tool to support PA early diagnosis and treatment planning.

5. Financial disclosure

The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Funding

This work was supported by grants from the National Natural Science Foundation of China (Grant Nos. 81702479, 31471252, 31771462); National Key R&D Program of China (Grant No. 2017YFA0106700); Program for Guangdong Introducing Innovative and Entrepreneurial Teams (Grant No. 2017ZT07S096); Guangdong Basic and Applied Basic Research Foundation (Grant No. 2020A1515010280). Science and Technology Program of Jiangmen, China (Grant No. 2018630100110019805). Guangdong Natural Science Foundation (Grant No. 2018A030313323); and Fundamental Research Funds for the Central Universities (SYSU: 18ykpy34).

CRediT authorship contribution statement

Hongyu Li: Data curation, Formal analysis, Investigation, Methodology, Software, Validation. **Qi Zhao:** Conceptualization, Formal analysis, Software, Supervision, Validation, Writing - review & editing. **Yihua Zhang:** Resources, Validation, Writing - original draft, Writing - review & editing. **Ke Sai:** . **Lunshan Xu:** Data curation. **Yonggao Mou:** Resources. **Yubin Xie:** Formal analysis, Funding acquisition, Software, Supervision. **Jian Ren:** Conceptualization, Methodology, Project administration, Resources, Writing - review & editing. **Xiaobing Jiang:** Conceptualization, Data curation, Funding acquisition, Investigation, Project adminis-

tration, Resources, Supervision, Writing - original draft, Writing - review & editing.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.05.023.

References

[1] Ostrom QT, et al., CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2011-2015. Neuro Oncol, 2018;20(suppl_4):iv1-iv86.
[2] Longo DL, Melmed S. Pituitary-tumor endocrinopathies. N Engl J Med 2020;382(10):937–50.
[3] Pal A, Leaver L, Wass J. Pituitary adenomas. BMJ 2019;365:2091.
[4] Pieper S, Halle M, Kikinis R. 3D Slicer. in 2004 2nd IEEE international symposium on biomedical imaging: nano to macro (IEEE Cat No. 04EX821). 200IEEE.
[5] Egger J, et al., Pituitary adenoma volumetry with 3D Slicer. PLoS One 2012;7 (12):e51788.
[6] Bengio Y, Learning deep architectures for AI, in Foundations and trends in machine learning. 2009, Now Publishers: Hanover, Mass. p. 1 electronic text (127 p. ill. (some col.)).
[7] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542 (7639):115–8.
[8] Pereira S, Pinto A, Alves V, Silva CA. Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans Med Imaging 2016;35(5):1240–51.
[9] Nie D, et al. 3D Deep Learning for Multi-modal Imaging-Guided Survival Time Prediction of Brain Tumor Patients. in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. 2016. Cham: Springer International Publishing.
[10] Zadeh Shirazi A, Fornaciari E, Bagherian NS, Ebert LM, Koszyca B, Gomez GA. DeepSurvNet: deep survival convolutional network for brain cancer survival rate classification based on histopathological images. Med Biol Eng Comput 2020;58(5):1031–45.
[11] Qian Y et al. A novel diagnostic method for pituitary adenoma based on magnetic resonance imaging using a convolutional neural network. Pituitary 2020:1–7.
[12] Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998;86(11):2278–324.
[13] He K et al. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[14] Zagoruyko S, Komodakis N. Wide Residual Networks. 2016.
[15] Xie, S., et al., Aggregated Residual Transformations for Deep Neural Networks. 2016.
[16] Deng J, et al., ImageNet: A Large-Scale Hierarchical Image Database. Cvpr: 2009 IEEE Conference on Computer Vision and Pattern Recognition 2009;1-4:248-255.
[17] Russakovsky Olga, Deng Jia, Su Hao, Krause Jonathan, Satheesh Sanjeev, Ma Sean, et al. ImageNet large scale visual recognition challenge. Int J Comput Vision 2015;115(3):211–52.
[18] Lin TY, et al., Microsoft COCO: Common Objects in Context. Computer Vision - Eccv 2014, Pt V, 2014;8693:740-755.
[19] Shin HC, et al., Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. 2016;35(5):1285-1298.
[20] Guerrero R, Ledig C, Rueckert D, Manifold Alignment and Transfer Learning for Classification of Alzheimer's Disease. 2014;77-84.
[21] van Opbroek Annegreet, Ikram M Arfan, Vernooij Meike W, de Bruijne Marleen. Transfer learning improves supervised image segmentation across imaging protocols. IEEE Trans Med Imaging 2015;34(5):1018–30.
[22] Becker Carlos, Christoudias C Mario, Fua Pascal. Domain adaptation for microscopy imaging. IEEE Trans Med Imaging 2015;34(5):1125–39.
[23] Deepak S, Ameer PM. Brain tumor classification using deep CNN features via transfer learning. Comput Biol Med 2019;111:103345.
[24] Swati Zar Nawab Khan, Zhao Qinghua, Kabir Muhammad, Ali Farman, Ali Zakir, Ahmed Saeed, et al. Brain tumor classification for MR images using transfer learning and fine-tuning. Comput Med Imaging Graph 2019;75:34–46.
[25] Bar Y et al. Deep learning with non-medical training used for chest pathology identification. Medical Imaging 2015: Computer-Aided Diagnosis 2015;9414.
[26] van Ginneken B et al. Off-the-Shelf Convolutional Neural Network Features for Pulmonary Nodule Detection in Computed Tomography Scans. In: 2015 Ieee 12th International Symposium on Biomedical Imaging (Isbi). p. 286–9.
[27] Chen Hao, Ni Dong, Qin Jing, Li Shengli, Yang Xin, Wang Tianfu, et al. Standard plane localization in fetal ultrasound via domain transferred deep neural networks. IEEE J Biomed Health Inf 2015;19(5):1627–36.
[28] Carneiro G, Nascimento J, Bradley AP. Unregistered multiview mammogram analysis with pre-trained deep learning models. Med Image Comput Comput-Assist Intervent Pt Iii 2015;9351:652–60.
[29] Chen L et al. Attention to Scale: Scale-Aware Semantic Image Segmentation. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[30] Jaderberg M, et al., Spatial Transformer Networks. Adv Neural Inform Process Syst 28 (NIPS 2015), 2015.
[31] Jiang Xiao-Bing, Li Cui-Ling, He Dong-Sheng, Mao Zhi-Gang, Liu Dong-Hong, Fan Xiang, et al. Increased carotid intima media thickness is associated with prolactin levels in subjects with untreated prolactinoma: a pilot study. Pituitary 2014;17(3):232–9.
[32] Yavropoulou MP, et al., The natural history and treatment of non-functioning pituitary adenomas (non-functioning PitNETs). Endocr Relat Cancer 2020; 27 (10): R375-R390.
[33] Katznelson Laurence, Laws Edward R, Melmed Shlomo, Molitch Mark E, Murad Mohammad Hassan, Utz Andrea, et al. Acromegaly: an endocrine society clinical practice guideline. J Clin Endocrinol Metab 2014;99(11):3933–51.
[34] Knosp Engelbert, Steiner Erich, Kitz Klaus, Matula Christian. Pituitary adenomas with invasion of the cavernous sinus space: a magnetic resonance imaging classification compared with surgical findings. Neurosurgery 1993;33 (4):610–8.
[35] Jenkinson M, Smith S. Global optimisation method for robust affine registration of brain images. Med Image Anal 2001;5:143–56.
[36] Jenkinson Mark, Bannister Peter, Brady Michael, Smith Stephen. Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. NeuroImage 2002;17(2):825–41.
[37] Smith S, et al., Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage 2004;23 Suppl 1:S208-19.
[38] Woolrich M, et al., Bayesian analysis of neuroimaging data in FSL. NeuroImage 2008;45:S173-86.
[39] Jenkinson M, et al., FSL. NeuroImage 2011;62:782-90.
[40] Jenkinson M, BET2: MR-Based Estimation of Brain, Skull and Scalp Surfaces. Eleventh Annual Meeting of the Organization for Human Brain Mapping, 2005, 2005.
[41] Tustison Nicholas J, Avants Brian B, Cook Philip A, Yuanjie Zheng, Egan Alexander, Yushkevich Paul A, et al. N4ITK: improved N3 bias correction. Med Imaging IEEE Trans 2010;29(6):1310–20.
[42] Wolf I, et al. The medical imaging interaction toolkit (MITK): a toolkit facilitating the creation of interactive software by extending VTK and ITK. in Medical Imaging 2004: Visualization, Image-Guided Procedures, and Display. 2004. International Society for Optics and Photonics.
[43] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. Springer; 2015.
[44] Pan S, Yang Q. A survey on transfer learning. Knowl Data Eng IEEE Trans 2010;22(10):1345–59.
[45] Woo S, et al., CBAM: Convolutional Block Attention Module. 2018.
[46] Abadi M, et al., TensorFlow: A system for large-scale machine learning. 2016.
[47] Gibson Eli, Li Wenqi, Sudre Carole, Fidon Lucas, Shakir Dzhoshkun I, Wang Guotai, et al. NiftyNet: a deep-learning platform for medical imaging. Comput Methods Programs Biomed 2018;158:113–22.
[48] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837–45.
[49] Lin Li, Dou Qi, Jin Yue-Ming, Zhou Guan-Qun, Tang Yi-Qiang, Chen Wei-Lin, et al. Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma. Radiology 2019;291(3):677–86.
[50] Peng A et al. A machine learning model to precisely immunohistochemically classify pituitary adenoma subtypes with radiomics based on preoperative magnetic resonance imaging. Eur J Radiol 2020;125:108892.
[51] Sanei Taheri Morteza, Kimia Farnaz, Mehrnahad Mersad, Saligheh Rad Hamidreza, Haghighatkhah Hamidreza, Moradi Afshin, et al. Accuracy of diffusion-weighted imaging-magnetic resonance in differentiating functional from non-functional pituitary macro-adenoma and classification of tumor consistency. Neuroradiol J 2019;32(2):74–85.