Pituitary (2022) 25:842–853
https://doi.org/10.1007/s11102-022-01255-7

Generating novel pituitary datasets from open-source imaging data and deep volumetric segmentation

Rachel Gologorsky[1] · Edward Harake[2] · Grace von Oiste[3] · Mustafa Nasir-Moin[3] · William Couldwell[4] · Eric Oermann[3,5,6] · Todd Hollon[7]

Accepted: 8 July 2022 / Published online: 9 August 2022

Abstract

Purpose The estimated incidence of pituitary adenomas in the general population is 10–30%, yet radiographic diagnosis remains a challenge. Diagnosis is complicated by the heterogeneity of radiographic features in both normal (e.g. complex anatomy, pregnancy) and pathologic states (e.g. primary endocrinopathy, hypophysitis). Clinical symptoms and laboratory testing are often equivocal, which can result in misdiagnosis or unnecessary specialist referrals. Computer vision models can aid in pituitary adenoma diagnosis; however, a major challenge to model development is the lack of dedicated pituitary imaging datasets. We hypothesized that deep volumetric segmentation models trained to extract the sellar and parasellar region from existing whole-brain MRI scans could be used to generate a novel dataset of pituitary imaging.

Methods Six open-source whole-brain MRI datasets, created for research purposes, were included for model development. Deep learning-based volumetric segmentation models were trained using 318 manually annotated MRI scans from a single open-source MRI dataset. Out-of-distribution volumetric segmentation performance was then tested on 418 MRIs from five held-out research datasets.

Results On our annotated images, agreement between manual and model volumetric segmentations was high. Dice scores (a measure of overlap) ranged 0.76–0.82 for both in-distribution and out-of-distribution model testing. In total, 6,755 MRIs from six data sources were included in the final generated pituitary dataset.

Conclusions We present the first and largest dataset of pituitary imaging constructed using existing MRI data and deep volumetric segmentation models trained to identify sellar and parasellar anatomy. The model generalizes well across patient populations and MRI scanner types. We hope our pituitary dataset will be an integral part of future machine learning research on pituitary pathologies.

Keywords Pituitary gland · Magnetic resonance imaging · Volumetric segmentation · Computer vision · Dataset generation

Introduction

The pituitary gland is the central regulatory endocrine gland of the human body, controlling all the major hormonal axes. Primary tumors of the pituitary gland, or pituitary adenomas, are among the most common brain tumors, with an estimated incidence of 10–30% in the general population [1]. Some pituitary adenomas secrete abnormally high amounts of normal stimulatory hormones (i.e., functioning adenomas) and account for a variety of endocrinopathies (most commonly Cushing's disease and gigantism/acromegaly). However, the majority of pituitary adenomas do not secrete hormones (i.e., non-functioning) and are either incidentally discovered or present with symptoms of local mass effect on the optic apparatus, resulting in vision loss. Magnetic resonance imaging (MRI) is the primary imaging modality for diagnosing pituitary adenomas and is critical for guiding management decisions, evaluating treatment response, and long-term surveillance [2, 3].

Unfortunately, MRI-based diagnosis of pituitary adenomas remains a major challenge [4]. Complex normal pituitary anatomy, including different radiographic features between the anterior and posterior pituitary gland, makes distinguishing normal from pathologic features challenging. Pregnancy, puberty, medications, and aging are a few

Extended author information available on the last page of the article

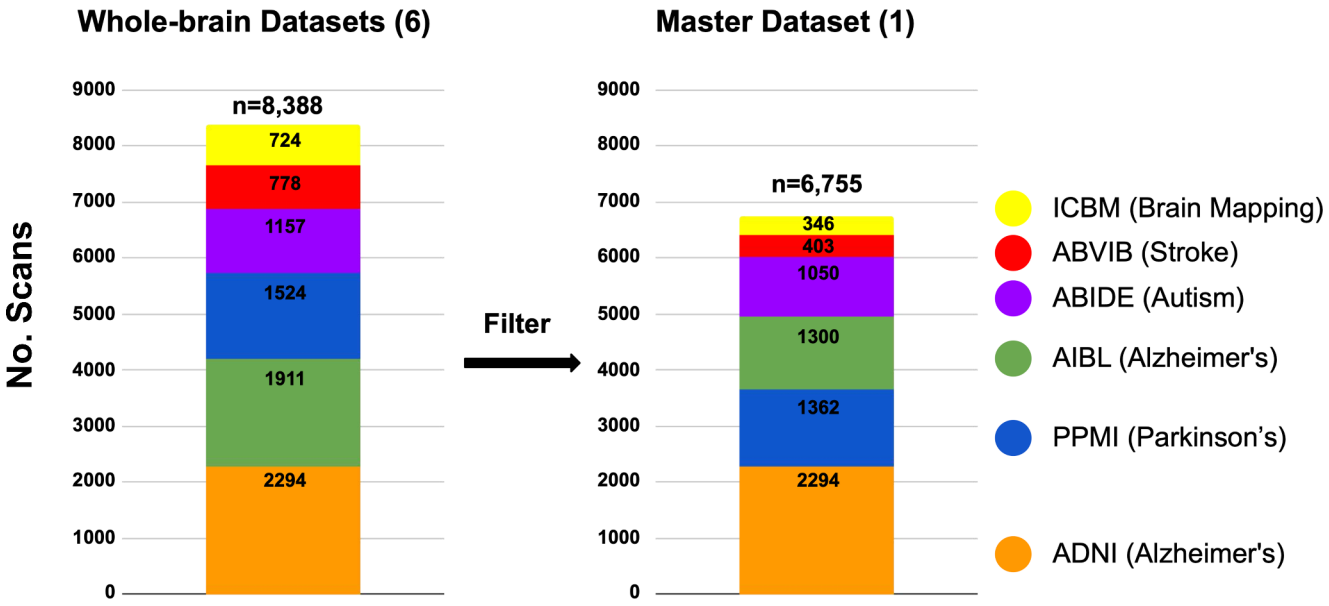


Fig. 1 Data sources. Datasets are publicly available from the Laboratory of Neuro Imaging (LONI) [10]. We included the T1-weighted MR sequences in the component datasets with slice thickness < 3 mm to capture several slices of the pituitary gland. The inclusion criteria are further described in the Methods section

examples of normal states that can result in aberrant pituitary gland anatomy and lead to false-positive pituitary adenoma diagnosis. Conversely, many functioning pituitary adenomas are small, measuring less than 5 mm, and share similar radiographic features with the normal pituitary glands, leading to false-negative diagnoses [5]. These diagnostic challenges invite an innovative computer-aided diagnostic solution for improved detection and monitoring of pituitary adenomas.

Machine learning techniques are increasingly being developed for the diagnosis and study of medical, neurological, and endocrine disorders [6–8]. As machine learning becomes democratized and computational resources are increasingly available, the barrier to adopting machine learning in medicine frequently becomes the availability of large, high-quality datasets [9]. While these datasets are beginning to emerge for some diseases and imaging modalities (e.g., pneumonia/chest radiographs, dementia/brain MRI), for many disorders the datasets do not yet exist. We hypothesized that a large dataset of pituitary glands could be constructed from a diverse group of patients by applying deep learning-based volumetric segmentation models to existing open-source MRI datasets. Here, we discuss our approach towards using computer vision to algorithmically create a novel open-source brain MRI dataset and then present and characterize the world's largest dataset of pituitary and sellar region imaging. We conclude by discussing the application of this dataset to future studies of pituitary adenomas and endocrine disorders.

Methods

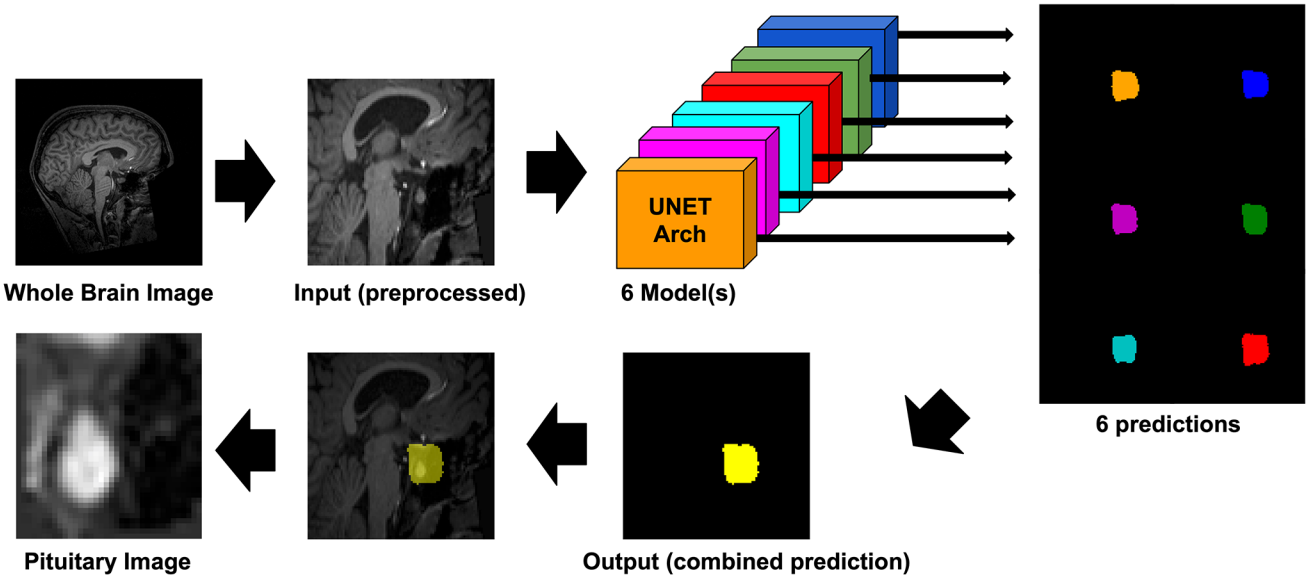
We identified multiple large open-source brain MRI datasets created for research purposes in neurological diseases. We included all studies that contained the brain and cranial base. Our whole-brain MRI data sources and collection process are described in the **Data sources** section below and illustrated in Fig. 1. Our segmentation model training and pituitary extraction process are described in the **Model architectures and training** section and illustrated in Fig. 2.

Data sources

We identified six open-source whole-brain MRI datasets for inclusion into our master pituitary dataset, representing a diverse set of patient characteristics, demographics, scanner technologies, and imaging sequences. All datasets are available from the Laboratory of NeuroImaging (LONI) Image & Data Archive [10]. These component datasets are:

ABIDE*(n = 1157) The Autism Brain Imaging Data Exchange* [11].

The ABIDE initiative has aggregated functional and structural brain imaging data from more than 24 international brain imaging laboratories around the world to accelerate the understanding of the neural bases of autism. This dataset includes data from those diagnosed with autism spectrum disorder and matched healthy controls. The median patient age is 14.7 years, with an age range of 7–64 years.



Binary Cross-Entropy Loss

- Standard loss function
- Has information-theory interpretation as the distance between two probability distributions (ground-truth vs. model predicted probability)

Dice Loss

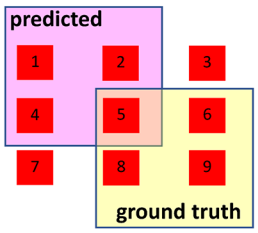
- Corresponds to metric of interest
- Has physical interpretation as the probability-weighted area of non-overlap between model prediction vs. ground truth

Conceptual Demonstration

$$\text{BCE loss} = \frac{1}{n}\sum_{i=1}^{n} -ln(\text{Prob[correct class]}) = 10.666/9 = 1.185$$

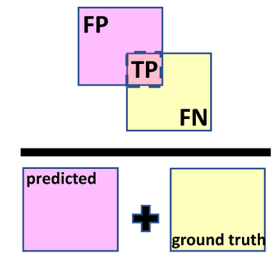
$$\text{Dice loss*} = \frac{FP+FN}{(FP+TP)+(TP+FN)} = (3+3)/(4+4) = 0.75$$

Table: Sample Model Output



Pixel No. (n = 9)	Ground Truth 0 = background 1 = pituitary	Model Prediction (Prob[class = 0], Prob[class = 1])	Model Prediction Prob[correct class]
1	0	1 (10%, 90%)	10% (FP)
2	0	1 (30%, 70%)	30% (FP)
3	0	0 (90%, 10%)	90% (TN)
4	0	1 (20%, 80%)	20% (FP)
5	1	1 (10%, 90%)	90% (TP)
6	1	0 (80%, 20%)	20% (FN)
7	0	0 (80%, 20%)	80% (TN)
8	1	0 (70%, 30%)	30% (FN)
9	1	0 (90%, 10%)	10% (FN)

FP, false positive. FN, false negative. TP, true positive. TN, false negative.



*For simplicity, the unweighted Dice loss calculation is demonstrated here.

Fig. 2 (a) Pituitary extraction. The process of generating a pituitary image involves: (i) applying a standard pre-processing pipeline to the whole-brain MRI input, (ii) applying a set of UNET-based segmentation models to output binary segmentation maps corresponding to the pituitary region of interest, (iii) combining the individual segmentation maps via a standard post-processing pipeline, and (iv) cropping the original whole-brain MRI to the region indicated by the combined segmentation map. **(b) Training loss functions.** The process of training a model requires a loss function to quantify the model's prediction error, so that this error can be minimized. The BCE loss function maximizes the probability of correctly classifying each pixel as being in the background (class = 0) or pituitary ROI (class = 1). The Dice loss function maximizes the overall overlap between the predicted and pituitary ROI. False positive, false negative, true positive, and true negative are abbreviated FP, FN, TP, and TN, respectively

ABVIB*(n = 778) The Aging Brain: Vasculature, Ischemia, and Behavior Study*.

The primary goal of ABVIB is to assess the contributions of cardiovascular risk factors (laboratory studies) and cerebrovascular disease (carotid intima-media thickness and retinal vessels) to brain structure and function. Launched in 1994, data from a second study cohort was started in 2008–2013 and is included in the current database.

Participants represent a spectrum of vascular risk and cognitive impairment.

ADNI*(n = 2294) The Alzheimer's Disease Neuroimaging Initiative* [12].

ADNI is a longitudinal multicenter study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimer's disease (AD). Specifically, we included scans from ADNI1:Complete 1Yr 1.5T collection, which included participants with mild cognitive impairment (MCI), AD, and elderly controls.

AIBL*(n = 1911) Australian Imaging Biomarkers and Lifestyle Study of Ageing* [13].

AIBL is a large-scale (1,000 + participant) cohort study of cognition with a focus on discovering the biomarkers and lifestyle factors linked to the subsequent development of Alzheimer's disease. Minimum participant age is 60 years, and the study includes data from patients with AD, MCI, and healthy volunteers.

PPMI*(n = 1524) Parkinson's Progression Markers Initiative* [14].

PPMI is a large-scale longitudinal observational multicenter study aggregating clinical data, imaging data, and biologic samples to establish markers of disease progression in Parkinson's disease (PD). Participants include 1,400 + individuals at 33 clinical sites in 11 countries, representing de novo Parkinson's, control volunteers, and at-risk populations.

ICBM*(n = 724) International Consortium for Brain Mapping* [15].

The purpose of ICBM is to create a human brain atlas based on an average space constructed from the average position, orientation, scale, and shear from all the individual subjects. The data were collected from three North American sites, representing 850 normal adult subjects ranging in age from 18 to 90 years.

Inclusion criteria

Our unified "master" dataset included the T1-weighted MR sequences in the component datasets with slice thickness < 3 mm to capture several slices of the pituitary gland. We chose the T1-weighted sequence as it was the most frequent sequence across our datasets and best captures anatomical features. Our formal inclusion criteria for studies were: (1) "T1" or "MPR" in the file name path and (2) 100–300 slices. In cases where the MRI sequence type was not available in the metadata, we used the filename as a proxy if appropriate. Likewise, we used the number of slices as a proxy for slice thickness as sequences with fewer than 100 images were likely to have poor coverage of the pituitary gland.

Table 1 Manual annotations. Annotated scans in ABIDE were split into a 60:20:20 train/valid/test set. Annotated scans from the other five datasets were used solely as a test set to evaluate model generalizability to new datasets and different patient populations.

Dataset	Scans	Labels
ABIDE	1050	333 (201/66/66)
ABVIB	403	92
ADNI	2294	90
AIBL	1300	89
PPMI	1362	95
ICBM	346	52
Total	6755	751

Annotated scans in ABIDE were split into a 60:20:20 train/valid/test set. Annotated scans from the other five datasets were used solely as a test set to evaluate model generalizability to new datasets and different patient populations

Training/Valid/Test data

The pituitary region of interest (ROI) was manually annotated in order to generate the "gold standard" ground truth labels needed for model training and evaluation. Our pituitary ROI included a margin around the pituitary gland, including the sellar and parasellar regions, to include areas of potential involvement with large pituitary pathology. Methodologically, the ROI was demarcated in 3DSlicer [16] based on anatomical boundaries. The cavernous sinuses served as the lateral boundaries. Inferior to superior, we included the region between the bottom of sella/mid-clivus to above the optic chiasm. Anterior to posterior, we included the region between the mid-planum sphenoidale to the anterior pons.

We labeled 333 scans from the ABIDE component dataset, which we split into a 60:20:20 train/valid/test set (201/66/66). We trained our pituitary ROI segmentation models only on ABIDE data. To evaluate model performance in response to distribution shift and real-world generalization across different patient populations and imaging systems, we annotated a number of ground truth labels from each of the other component datasets (Table 1). These out-of-distribution labels were used only in evaluating the model and were not part of the training process.

Model architectures and training

We trained six different model architectures available from the MONAI framework [17] and from open-source Github repositories in order to find the best-performing architecture or ensemble of architectures for our task.

Our primary performance metric was the Dice score, a measure of overlap between the algorithm-generated segmentation and the provided manual annotation. The Dice score ranges from 0 (no overlap) to 1 (100% overlap). It is equivalent to the F1 score, a measure of classification accuracy defined as the harmonic mean of precision (positive predictive value) and recall (sensitivity).

The process of training a model requires a loss function to quantify the model's prediction error, so that this error can be minimized. The choice of loss function is important, as it influences the accuracy and generalizability of the solution reached from the initial start point (initializations are randomly generated).

We trained each model architecture twice: once with a standard loss function that can be easily optimized - binary cross-entropy loss (BCE loss) - and once with Dice loss (DICE loss). The Dice loss forces the model maximize the Dice score. The BCE loss function forces the model to maximize the probability of correctly classifying each pixel as being in the background (class = 0) or part of the pituitary ROI (class = 1). BCE is the standard loss function used for binary segmentation problems.

We thus evaluated the effect of the loss function on model performance and robustness to out-of-distribution datasets. Additionally, in order to evaluate the reproducibility of model performance on the test set, we re-ran the experiment 10 times starting from ten independent random initializations.

The following model architectures were trained:

UNET3D The UNET architecture is the standard architecture for image segmentation tasks [18–22]. We implemented a 3-dimensional UNET using MONAI's UNET architecture [23] with standard parameters.

VNET The VNET architecture is a UNET variant introduced specifically for medical image segmentation tasks [24]. We leveraged MONAI's VNET with the same default parameters as for UNET3D.

UNETR This latest addition to the MONAI neural network library is a vision transformer model. This model had the highest Dice scores of any other MONAI model when tested on the image segmentation task in the Medical Segmentation Decathlon dataset [25].

OBELISKHYBRID Our OBELISKHYBRID implementation is based on the code in the Github repository [26]. OBELISKHYBRID is a UNET-based model that incorporates OBELISK layers into traditional UNET architecture [26]. Because the OBELISK kernel decouples the effective receptive field from the number of levels in the UNET encoder, models with OBELISK layers can include significantly fewer parameters, enabling us to experiment with larger dimensions as model input. We tested two OBELISKHYBRID models, OBELISK 96 and OBELISK 144.

CONDSEG: We implemented CONDSEG using MONAI's UNET architecture. The Conditional Segmentation model [27] is a UNET model modified to take in 3-channel input: the input scan, in addition to a randomly selected (atlas, atlas label) scan. In contrast to the above approaches, the model's task is now a hybrid of image registration and ROI propagation.

In addition to the above models, we also evaluated the model ensemble comprised of the three best-performing model architectures on the test set, namely, UNET3D, VNET, and CONDSEG. We evaluated three different ensemble models, corresponding to:

- **Ensemble BCE**: UNET3D + VNET + CONDSEG ensemble trained with BCE loss.
- **Ensemble DICE**: UNET3D + VNET + CONDSEG ensemble trained with DICE loss.
- **Ensemble Combined**: Ensemble BCE + Ensemble DICE (six models total, three trained with BCE loss and three with DICE loss).

Pre-processing pipeline

Our pre-processing/data augmentation pipeline included: (1) performing N4 bias correction, (2) re-orienting to a standard MRI orientation (LAS orientation), (3) standard isotropic spacing, (4) intensity normalization (scaling the voxel intensity values by the mean and standard deviation of the non-zero pixels), (5) random affine transformation with probability 0.5, and (6) center cropping to a pre-specified spatial dimension.

Inputs to all models except OBELISK 144 were resampled to standard isotropic spacing of 1.5 mm and center cropped to a standard dimension of $96 \times 96 \times 96$ voxels. We took advantage of the parsimonious OBELISK architecture to evaluate the effect of higher-resolution inputs: Inputs to OBELISK 144 were resampled to standard isotropic spacing of 1.0 mm and were center cropped to a standard dimension of $144 \times 144 \times 144$ voxels. We used the SimpleITK library [28, 29] to perform the N4 bias correction and orientation to LAS coordinates. The remaining transformations were performed using MONAI's Transform library.

Data augmentations

In addition to the above pipeline, we experimented with increased data augmentation in order to enhance our model's robustness and generalizability to out-of-distribution datasets. Specifically, we experimented with adding random intensity shifts and scaling, random Gaussian noise, random contrast adjustment, and random flips along spatial axes. However, we omit these results here as increased data augmentation did not significantly improve model performance.

Post-processing pipeline

Each machine learning model outputs a binary segmentation map, with 0 and 1 used to classify pixels in the background and in the pituitary region of interest, respectively. We post-processed the binary segmentation output to keep only the largest connected component.

In addition, to maintain consistency in evaluating across models, we resampled the OBELISK 144 model output to 1.5-mm isotropic voxel spacing and we center-cropped the output to the $96 \times 96 \times 96$ dimensions, thus matching the statistics of the other model outputs. In our ensemble model, we output the binary segmentation corresponding to the majority vote among individual binary predictions (ties are by default labeled 0, i.e., not part of the ensemble algorithm's predicted pituitary ROI).

Results

We present model results with respect to overall performance, robustness to distribution shift, and performance reproducibility. In the **Supplemental Information** section, we present results regarding the similarity between models trained with different loss functions and architectures as well as the auto-similarity resulting from re-training the same model architecture and loss function from multiple random initializations (Supplementary Fig. 1 and Supplementary Fig. 2).

Overall performance

We evaluated the overall performance of our individual and ensemble models on our collective test set of 484 annotated scans (66 test cases from ABIDE, 414 test cases from the other datasets, see Table 1).

Of the six individual models, UNET3D, VNET, and CONDSEG trained with DICE loss performed the best on the collective test set, achieving Dice scores of approximately 80% (Fig. 3). While the ensemble model performed similarly in terms of mean Dice score, ensembling resulted in fewer outliers than the individual component models. Overall, the best model is Ensemble DICE, which is the Ensemble model of UNET3D + VNET + CONDSEG trained with DICE loss. We used Ensemble DICE as our final model to generate our pituitary master dataset.

Robustness to distribution shift

We evaluated model performance on each component dataset separately (Fig. 4). With the exception of OBELISK 144, all models, including OBELISK 96, appear to be robust to out-of-distribution data, with similar performance on the ABIDE test set and on the test cases from the five out-of-distribution datasets (ABVIB, ADNI, AIBL, PPMI, ICBM). Notably, the OBELISK 144 model appears to overfit the dataset on which the model was trained (ABIDE) and suffers a performance drop in generalizing to other datasets. While we do not have an explanation for this, the result was reproducible.

Performance reproducibility

The consistency of each model's test set performance across ten random initializations is shown graphically in Fig. 5 and numerically in Table 2. The UNET3D, CONDSEG, and Ensemble models demonstrate high reproducibility (low variability) in the mean test set Dice score. Notably, with the exception of UNETR, models trained with DICE loss have higher average Dice scores and lower standard deviation than models trained with BCE loss, suggesting that they consistently converge to better optima than models trained with BCE loss. The exception is UNETR; however, UNETR performed significantly worse than other models, perhaps because vision transformer models require more training data to converge due to less inductive bias.

Table 2 is a granular version of Fig. 5, illustrating that generalization performance on out-of-distribution datasets was also reproducible, with similar model performance between component datasets that was consistent (low standard deviation across 10 runs). Notably, the Ensemble DICE model performed well on the individual datasets and was the best-performing model on the test set overall.

Error analysis

Finally, we performed an error analysis to evaluate our proposed volumetric pituitary segmentation model. We visually

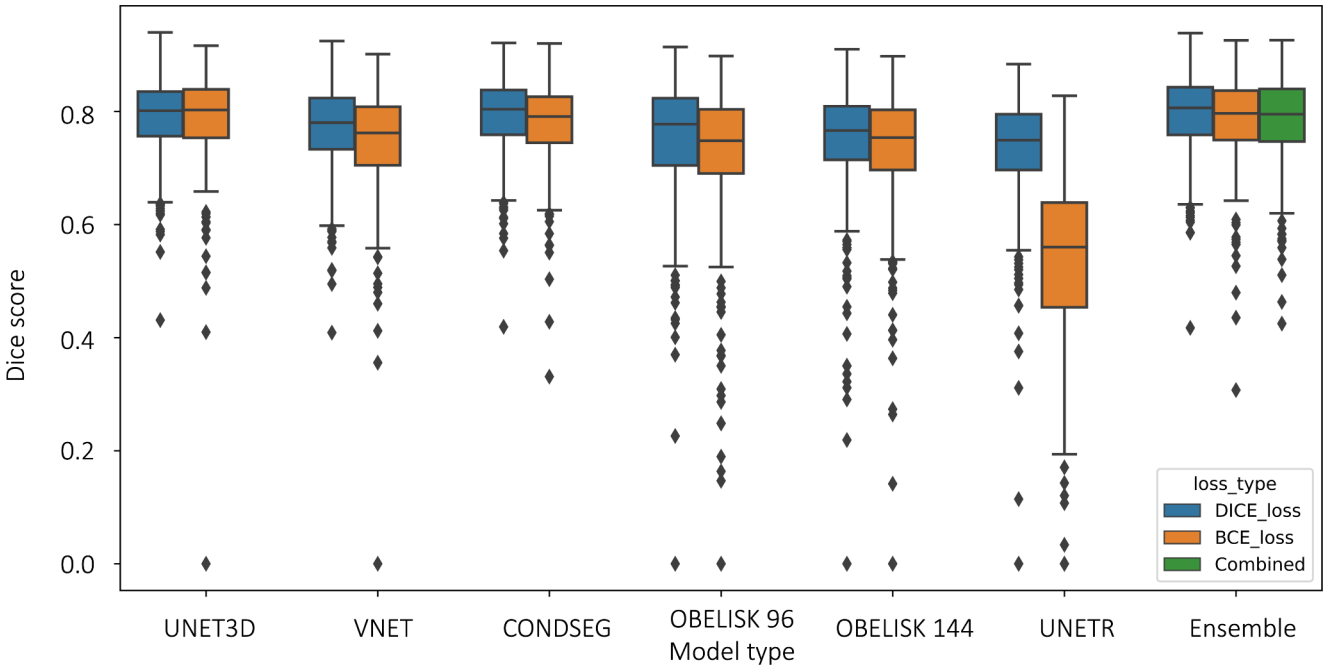


Fig. 3 Overall model performance. Each model was trained twice: once with DICE loss, once with BCE loss. We evaluated the ensemble of the three best models: UNET3D + VNET + CONDSEG, trained with the specified loss type: BCE, DICE, or both combined. The Ensemble models, particularly Ensemble DICE, had fewer outliers

inspected our Ensemble DICE model's three worst pituitary region segmentations, shown in Fig. 6. We discovered that the segmentation with the lowest Dice score, 0.41, was due to an annotation error, wherein the pituitary ROI was enlarged. In the next two poor segmentations, with Dice scores of 0.59, the error was in segmenting the periphery. However, the pituitary gland itself was always correctly segmented.

In comparison, for the best 3 segmentation instances, the Dice score was approximately 90%. As a further check on average model performance, we visually inspected our algorithm's segmentations on a randomly selected test item from each component dataset, as can be seen in Fig. 7. These visual inspections enable us to say that our model's segmentations do not appear systematically biased, an error that could have been masked by looking at Dice scores alone.

Discussion

We present the world's largest dataset of pituitary imaging constructed entirely using existing open-source datasets and deep volumetric segmentation computer vision models trained to recognize the normal sellar and parasellar anatomy. Our aim is to use this dataset of relatively normal imaging to investigate the classification of pituitary pathology, and to generate a normal control group for training machine learning models to localize ACTH-secreting microadenomas within the pituitary gland [30]. Complementary research has demonstrated the ability of machine learning to subtype pituitary adenomas as growth hormone or prolactin-secreting based on imaging alone [31]. Yet the ability of machine learning to classify and lateralize ACTH-secreting microadenomas has remained an open research question due to the lack of training data.

Additionally, for the benefit of the wider research community, we have publicly released our trained ensemble model with a detailed code walkthrough on model inference: github.com/RGologorsky/PituitaryGenerator.

Our approach towards algorithmically extracting novel datasets for medical image classification using weak annotations and volumetric semantic segmentation combined with open-source medical datasets can be applied to any number of biomedical organ systems or diseases. For example, the National Lung Screening Trial's 26,000 + patients have, in addition to their lung tumors, a vast amount of relatively normal pulmonary, cardiac, thoracic spine, and other imaging. An approach similar to ours is feasible to generate massive datasets of normal images of other intra-thoracic structures for subsequent normal samples or studies of the variation of normal structures at a population scale.

A notable challenge to the approach outlined here is the necessity of the computer vision algorithm to generalize from the dataset it was trained on to other datasets.

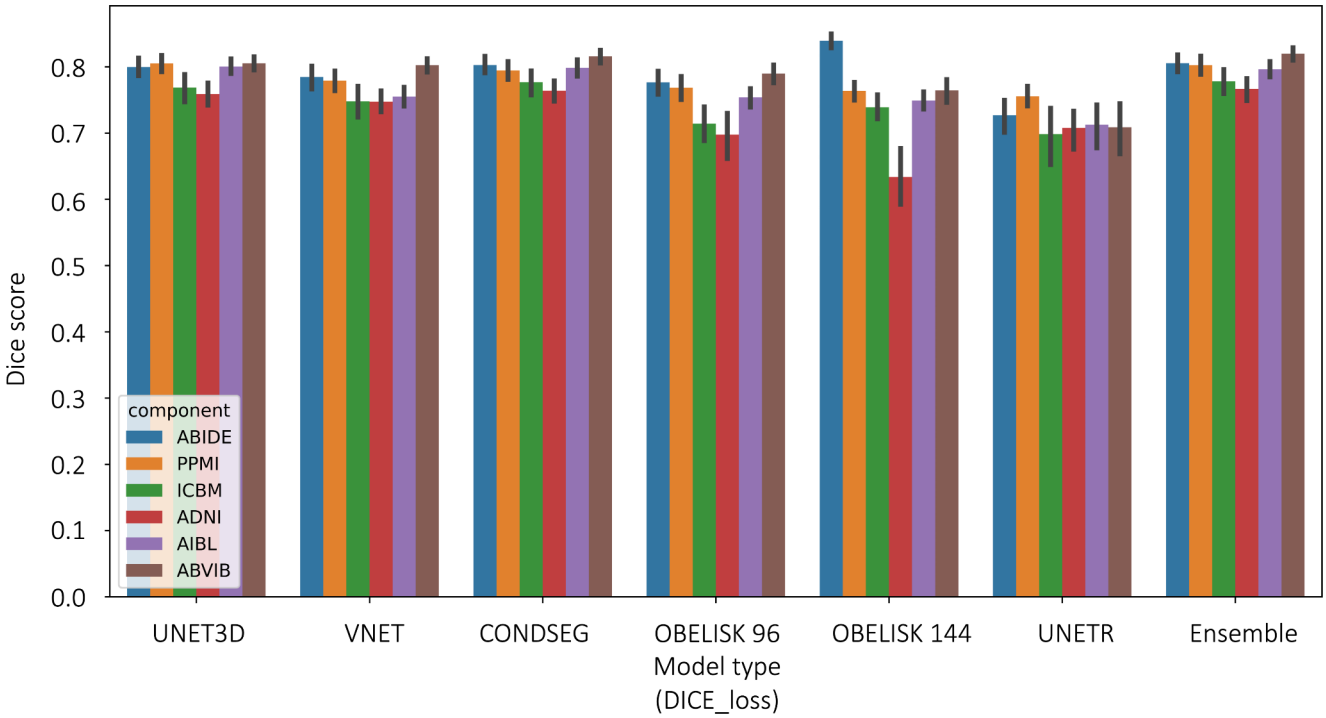


Fig. 4 Model performance by data source. Model robustness and generalization ability from ABIDE to new datasets was assessed. While OBELISK 144 performed best on ABIDE (the data source during model training), performance declined on out-of-distribution data. The UNET, VNET, CONDSEG and combined Ensemble model consistently performed well across diverse datasets

Confounding factors, such as different patient populations, imaging protocols, and systems, can impact a model's generalization performance [32]. One specific instance of this generalization challenge that we encountered is the fact that the training dataset (ABIDE) contains a much younger population (patients diagnosed with autism disorders, median age 14.7 years) with different pituitary anatomy from the out-of-distribution datasets, which included older patients diagnosed with stroke, Parkinson's, and Alzheimer's disease. Despite the different patient population, we found that the algorithm was performant across all tested datasets.

Limitations of our approach include the lack of a pathological set to test against due to the fact that we are curating the first pituitary gland dataset and no open-source pituitary adenoma datasets are available. Other limitations include that our algorithmically generated dataset could contain pathological examples due to the incidence of undiagnosed pituitary adenomas in the general population.. Lastly, although we have an excellent estimate of out-of-distribution performance, we did not directly investigate any techniques for domain adaptation, such as computed tomography (CT) segmentation, which is an interesting direction for future research.

In conclusion, we present an approach to algorithmically generate a biomedical imaging dataset from existing open-source MRI data using deep volumetric segmentation computer vision models. This is the first and largest dataset of endocrine and pituitary imaging and will provide a set of useful normal control for future pituitary and neuro-endocrine research.

Funding This work was supported by the Neurosurgery Research & Education Foundation (2021 Medical Student Summer Research Fellowship to R. Gologorsky).

Declarations

Competing interests The authors report no relevant conflicts of interest.

Compliance with ethical standards While our research was done using human MRI data, these data were anonymized and are publicly available. No informed consent was required to complete the study.

References

1. Ezzat S, Asa SL, Couldwell WT, Barr CE, Dodge WE, Vance ML, McCutcheon IE (2004) The prevalence of pituitary adenomas: a systematic review. Cancer 101(3):613–619. https://doi.org/10.1002/cncr.20412
2. Choi SH, Kwon BJ, Na DG, Kim JH, Han MH, Chang KH (2007) Pituitary adenoma, craniopharyngioma, and Rathke cleft cyst

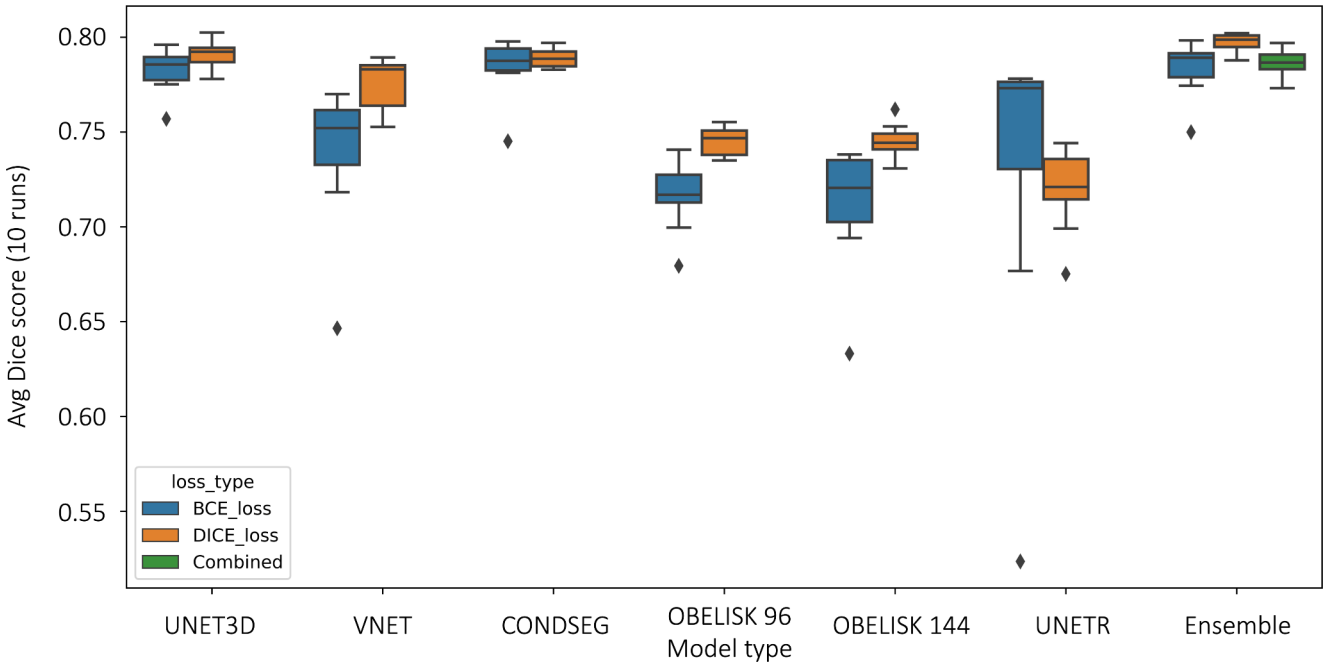


Fig. 5 Average model performance over 10 runs. Model performance was defined as the mean Dice score of the model's predictions over the test set. To assess performance reproducibility, each model was evaluated ten times (i.e. trained after ten independent initializations of the same model architecture). The Ensemble models performed best and their performance demonstrated high reproducibility

involving both intrasellar and suprasellar regions: differentiation using MRI. Clin Radiol 62(5):453–462. https://doi.org/10.1016/j.crad.2006.12.001

3. Heck A, Ringstad G, Fougner SL, Casar-Borota O, Nome T, Ramm-Pettersen J, Bollerslev J (2012) Intensity of pituitary adenoma on T2-weighted magnetic resonance imaging predicts the response to octreotide treatment in newly diagnosed acromegaly. Clin Endocrinol (Oxf) 77(1):72–78. https://doi.org/10.1111/j.1365-2265.2011.04286.x
4. Altshuler DB, Andrews CA, Parmar HA, Sullivan SE, Trobe JD (2021) Imaging errors in distinguishing pituitary adenomas from other sellar lesions. J Neuroophthalmol 41(4):512–518. https://doi.org/10.1097/WNO.0000000000001164
5. Chandler WF, Barkan AL, Hollon T, Sakharova A, Sack J, Brahma B, Schteingart DE (2016) Outcome of transsphenoidal surgery for Cushing disease: A single-center experience over 32 years. Neurosurgery 78(2):216–223. https://doi.org/10.1227/NEU.0000000000001011

Table 2 Average model performance by data source. Average model performance by data source. Model performance was defined as the mean Dice score of the model's predictions over the specified test set. To assess performance reproducibility, each model was evaluated ten times (i.e., trained after ten different initializations of the same model architecture). The Ensemble DICE model performed well on the individual datasets and was the best-performing model on the test set overall.

Model	Loss	ABIDE	PPMI	ICBM	ADNI	AIBL	ABVIB	Overall
UNET3D	BCE_loss	80.3 ± 5.3	80.2 ± 6.6	73.6 ± 13.3	76.3 ± 8.4	80.0 ± 5.1	80.6 ± 5.5	78.8 ± 7.8
	DICE_loss	80.0 ± 5.4	**80.5 ± 6.2**	76.9 ± 7.5	75.9 ± 7.9	**80.0 ± 5.1**	80.5 ± 4.9	79.1 ± 6.5
VNET	BCE_loss	76.2 ± 7.6	76.0 ± 9.3	68.7 ± 14.4	74.8 ± 8.0	73.2 ± 6.3	78.1 ± 5.6	74.9 ± 8.9
	DICE_loss	78.5 ± 7.3	77.9 ± 7.4	74.8 ± 8.3	74.7 ± 8.0	75.5 ± 6.7	80.3 ± 5.0	77.1 ± 7.4
CONDSEG	BCE_loss	80.6 ± 5.3	77.7 ± 7.2	75.9 ± 8.9	75.8 ± 7.3	78.4 ± 5.3	80.1 ± 5.4	78.1 ± 6.8
	DICE_loss	80.3 ± 5.1	79.5 ± 6.7	77.7 ± 6.8	76.4 ± 7.9	79.8 ± 5.9	81.6 ± 4.5	79.3 ± 6.5
OBELISK 96	BCE_loss	78.4 ± 7.0	74.6 ± 8.8	59.2 ± 18.7	67.4 ± 21.1	71.2 ± 13.0	75.7 ± 8.1	71.7 ± 14.7
	DICE_loss	77.7 ± 7.2	76.8 ± 8.5	71.4 ± 9.9	69.8 ± 16.7	75.4 ± 7.0	79.0 ± 6.4	75.2 ± 10.6
OBELISK 144	BCE_loss	81.7 ± 4.9	76.5 ± 7.6	69.0 ± 7.9	65.0 ± 19.4	74.1 ± 6.6	75.4 ± 7.7	73.6 ± 11.7
	DICE_loss	**83.9 ± 4.3**	76.4 ± 6.7	73.9 ± 7.0	63.4 ± 20.6	74.9 ± 6.2	76.4 ± 8.5	74.5 ± 12.4
UNETR	BCE_loss	55.7 ± 12.9	59.6 ± 10.4	46.5 ± 20.6	40.5 ± 19.7	55.3 ± 15.0	54.5 ± 15.4	52.4 ± 17.1
	DICE_loss	72.7 ± 10.4	75.5 ± 7.4	69.8 ± 16.0	70.8 ± 13.6	71.3 ± 15.2	70.9 ± 19.4	72.0 ± 14.3
Ensemble	BCE_loss	80.2 ± 5.7	79.6 ± 7.5	75.0 ± 10.1	76.6 ± 7.8	78.6 ± 5.2	80.7 ± 5.4	78.6 ± 7.2
	DICE_loss	80.5 ± 5.3	80.3 ± 6.6	**77.8 ± 6.6**	**76.7 ± 7.8**	79.6 ± 5.6	**82.0 ± 4.6**	**79.6 ± 6.4**
	Combined	80.0 ± 5.9	79.9 ± 7.5	75.2 ± 8.8	76.2 ± 8.2	78.4 ± 5.5	81.1 ± 5.3	78.7 ± 7.2

Model performance was defined as the mean Dice score of the model's predictions over the specified test set. To assess performance reproducibility, each model was evaluated ten times (i.e., trained after ten different initializations of the same model architecture). The Ensemble DICE model performed well on the individual datasets and was the best-performing model on the test set overall

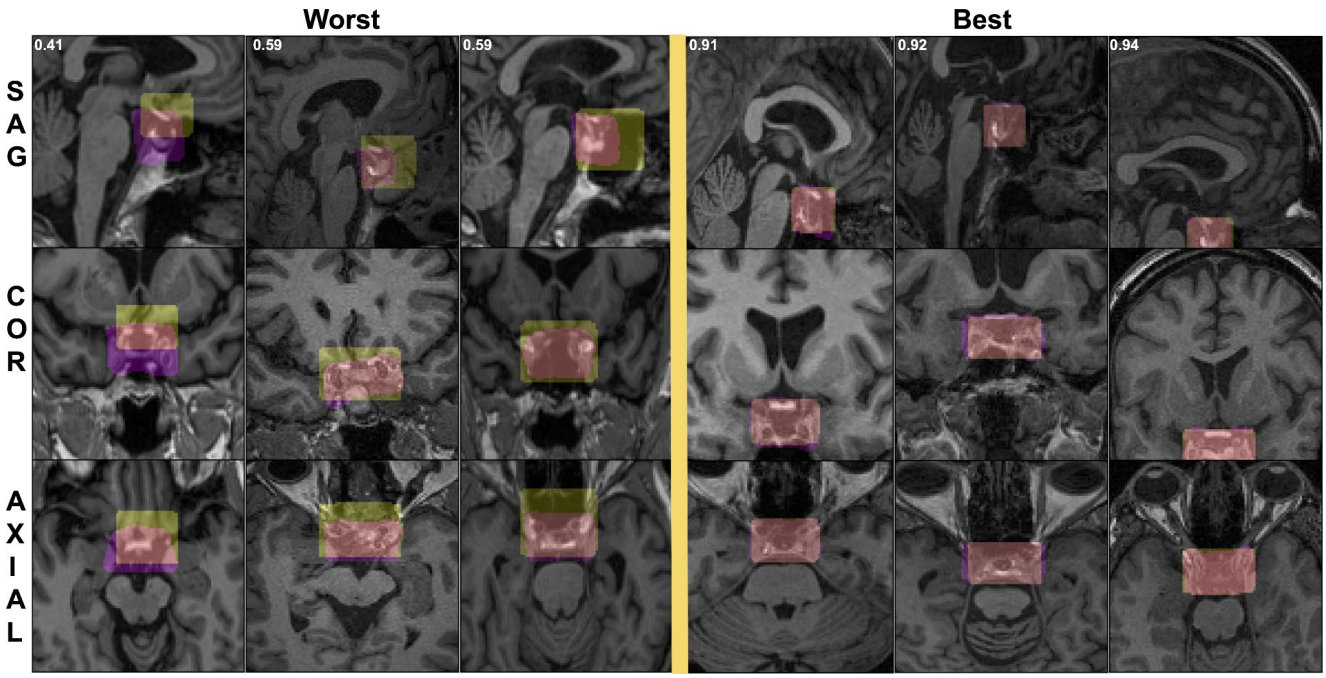


Fig. 6 Error analysis. Manual annotation (yellow), algorithm-generated segmentation (magenta), overlap (light pink). Left columns: Three worst algorithm segmentations by Dice score. Right: For comparison, the three best algorithm segmentations by Dice score. In all, the prediction error was in the periphery and the pituitary gland was captured

6. Fan Y, Jiang S, Hua M, Feng S, Feng M, Wang R (2019) Machine learning-based radiomics predicts radiotherapeutic response in patients with acromegaly. Front Endocrinol (Lausanne) 10:588. https://doi.org/10.3389/fendo.2019.00588

7. Swinburne NC, Schefflein J, Sakai Y, Oermann EK, Titano JJ, Chen I, Tadayon S, Aggarwal A, Doshi A, Nael K (2019) Machine learning for semi-automated classification of glioblastoma, brain metastasis and central nervous system lymphoma using magnetic

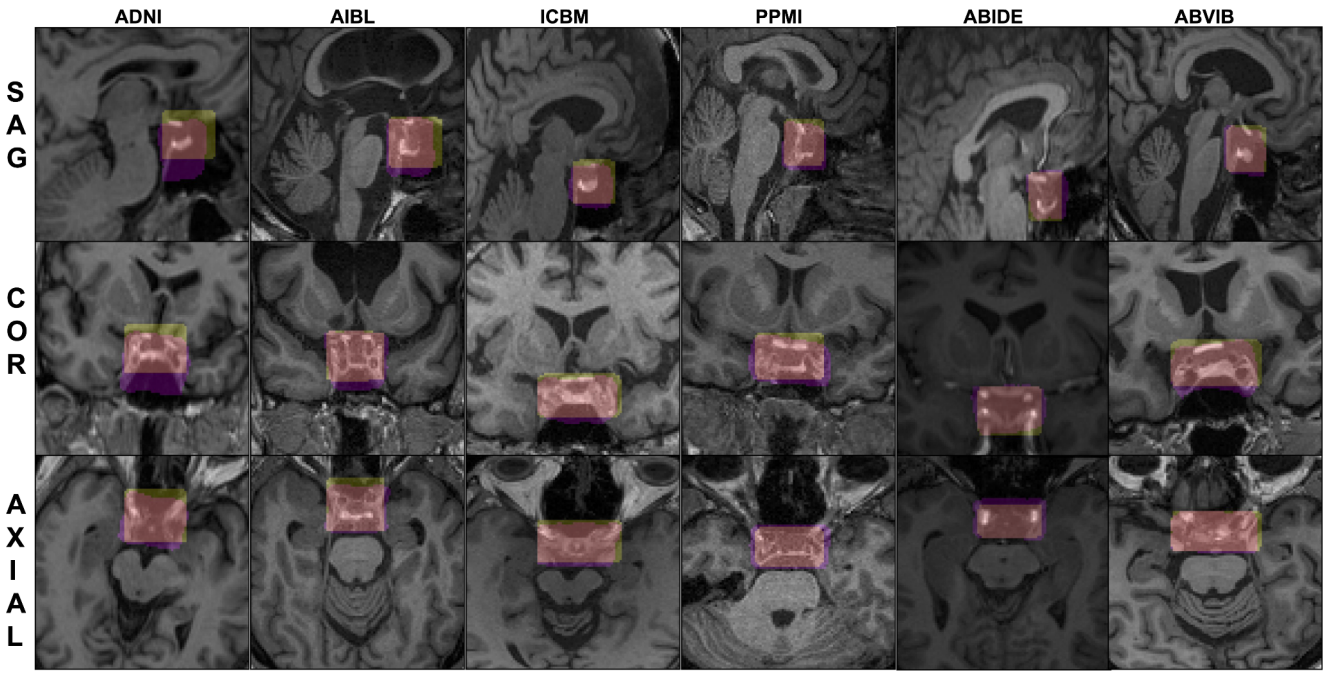


Fig. 7 Segmentation examples from component datasets. Manual annotation (yellow), algorithm-generated segmentation (magenta), overlap (light pink). Each column displays a randomly chosen example from the specified component dataset shown in the sagittal, coronal, and axial views

resonance advanced imaging. Ann Transl Med 7(11):232. https://doi.org/10.21037/atm.2018.08.05
8. Titano JJ, Badgeley M, Schefflein J, Pain M, Su A, Cai M, Swinburne N, Zech J, Kim J, Bederson J, Mocco J, Drayer B et al (2018) Automated deep-neural-network surveillance of cranial images for acute neurologic events. Nat Med 24(9):1337–1341. https://doi.org/10.1038/s41591-018-0147-y
9. Rajkomar A, Dean J, Kohane I (2019) Machine learning in medicine. N Engl J Med 380(14):1347–1358. https://doi.org/10.1056/NEJMra1814259
10. Crawford KL, Neu SC, Toga AW (2016) The Image and Data Archive at the Laboratory of Neuro Imaging. NeuroImage 124(Pt B):1080–1083. https://doi.org/10.1016/j.neuroimage.2015.04.067
11. Di Martino A, Yan CG, Li Q, Denio E, Castellanos FX, Alaerts K, Anderson JS, Assaf M, Bookheimer SY, Dapretto M, Deen B, Delmonte S et al (2014) The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol Psychiatry 19(6):659–667. https://doi.org/10.1038/mp.2013.78
12. Petersen RC, Aisen PS, Beckett LA, Donohue MC, Gamst AC, Harvey DJ, Jack CR Jr, Jagust WJ, Shaw LM, Toga AW, Trojanowski JQ, Weiner MW (2010) Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. Neurology 74(3):201–209. https://doi.org/10.1212/WNL.0b013e3181cb3e25
13. Ellis KA, Bush AI, Darby D, De Fazio D, Foster J, Hudson P, Lautenschlager NT, Lenzo N, Martins RN, Maruff P, Masters C, Milner A et al (2009) The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. Int Psychogeriatr 21(4):672–687. https://doi.org/10.1017/S1041610209009405
14. Marek K, Jennings D, Lasch S, Siderowf A, Tanner C, Simuni T, Coffey C et al (2011) The Parkinson Progression Marker Initiative (PPMI). Prog Neurobiol 95(4):629–635. https://doi.org/10.1016/j.pneurobio.2011.09.005
15. Mazziotta J, Toga A, Evans A, Fox P, Lancaster J, Zilles K, Woods R, Paus T, Simpson G, Pike B, Holmes C, Collins L et al (2001) A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). Philos Trans R Soc Lond B Biol Sci 356(1412):1293–1322. https://doi.org/10.1098/rstb.2001.0915
16. Pieper S, Halle M, Kikinis R (2004) 3D slicer. 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro, ISBI 2004, April 15–18. Vol IEEE Cat No. 04EX821. Washington, DC. IEEE; 2004:632–635
17. MONAI Consortium (2020) Project MONAI. Zenodo. https://doi.org/10.5281/zenodo.4323059
18. Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, Han X, Chen Y-W, Wu J UNet 3+: A full-scale connected UNet for medical image segmentation. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP(2020) 2020, May 4. Barcelona, Spain. IEEE; 2020:1055–1059
19. Luo Z, Zhang Y, Zhou L, Zhang B, Luo J, Wu H (2019) Micro-vessel image segmentation based on the AD-UNet model. IEEE Access 7:143402–143411. https://doi.org/10.1109/ACCESS.2019.2945556
20. Qiang Z, Tu S, Xu L(2019) A k-Dense-UNet for biomedical image segmentation. In: Cui Z, Pan J, Zhang S, Xiao L, Yang J, eds. Intelligence Science and Big Data Engineering. Visual Data Engineering. Proceedings of the 9th International Conference, IScIDE 2019, October 17–20. Nanjing, China. Springer; 2019:552–562
21. Shi T, Jiang H, Zheng B (2020) A stacked generalization U-shape network based on zoom strategy and its application in biomedical image segmentation. Comput Methods Programs Biomed 197:105678. https://doi.org/10.1016/j.cmpb.2020.105678
22. Weng Y, Zhou T, Li Y, Qiu X (2019) NAS-Unet: Neural architecture search for medical image segmentation. IEEE Access 7:44247–44257. https://doi.org/10.1109/ACCESS.2019.2908991
23. Kerfoot E, Clough J, Oksuz I, Lee J, King AP, Schnabel JA(2019) Left-ventricle quantification using residual U-Net. In: Pop M, Sermesant M, Zhao J, et al., eds. Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges. 9th International Workshop, STACOM 2018, September 16. Granada, Spain. Springer; 2019:371–380
24. Milletari F, Navab N, Ahmadi S(2016) V-Net: Fully convolutional neural networks for volumetric medical image segmentation. 2016 Fourth International Conference on 3D Vision (3DV), October 25–28. Stanford University. IEEE; 2016:565–571
25. Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, Roth H, Xu D(2021) UNETR: Transformers for 3D medical image segmentation. arXiv:2103.10504 [eess.IV]
26. Heinrich MP, Oktay O, Bouteldja N (2019) OBELISK-Net: Fewer layers to solve 3D multi-organ segmentation with sparse deformable convolutions. Med Image Anal 54:1–9. https://doi.org/10.1016/j.media.2019.02.006
27. Hu Y, Gibson E, Barratt DC, Emberton M, Alison Noble J, Vercauteren T(2019) Conditional segmentation in lieu of image registration. arXiv:1907.00438 [eess.IV].
28. Lowekamp BC, Chen DT, Ibanez L, Blezek D (2013) The design of SimpleITK. Front Neuroinform 7:45. https://doi.org/10.3389/fninf.2013.00045
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
30. Paterno V, Fahlbusch R (2014) High-field iMRI in transsphenoidal pituitary adenoma surgery with special respect to typical localization of residual tumor. Acta Neurochir (Wien) 156(3):463–474 discussion 474. https://doi.org/10.1007/s00701-013-1978-4
31. Li H, Zhao Q, Zhang Y, Sai K, Xu L, Mou Y, Xie Y, Ren J, Jiang X (2021) Image-driven classification of functioning and non-functioning pituitary adenoma by deep convolutional neural networks. Comput Struct Biotechnol J 19:3077–3086. https://doi.org/10.1016/j.csbj.2021.05.023
32. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK (2018) Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 15(11):e1002683. https://doi.org/10.1371/journal.pmed.1002683

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11102-022-01255-7.

Authors and Affiliations

Rachel Gologorsky[1] · Edward Harake[2] · Grace von Oiste[3] · Mustafa Nasir-Moin[3] · William Couldwell[4] · Eric Oermann[3,5,6] · Todd Hollon[7]

✉ Todd Hollon

tocho@med.umich.edu

Rachel Gologorsky
rachel.gologorsky@icahn.mssm.edu

Edward Harake
edsha@med.umich.edu

Grace von Oiste
gvonoiste@college.harvard.edu

Mustafa Nasir-Moin
mustafa.nasir-moin@nyulangone.org

William Couldwell
william.couldwell@hsc.utah.edu

Eric Oermann
eric.oermann@nyulangone.org

[1] Department of Medicine, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl, 10029 New York, NY, USA

[2] Department of Medicine, University of Michigan Medical School, 1500 E Medical Center Dr, 48109 Ann Arbor, MI, USA

[3] Department of Neurosurgery, NYU Langone Health System, 530 First Ave, 10016 New York, NY, USA

[4] Department of Neurosurgery, University of Utah, 201 Presidents' Cir, 84132 Salt Lake City, UT, USA

[5] Department of Radiology, NYU Langone Health System, 530 First Ave, 10016 New York, NY, USA

[6] Center for Data Science, New York University, 60 5th Ave, 10011 New York, NY, USA

[7] Machine Learning in Neurosurgery Laboratory, Department of Neurosurgery, University of Michigan, 1500 E Medical Center Dr, 48109 Ann Arbor, MI, USA