

Searching for an optimal convolutional neural network architecture to classify biological sex from structural brain images in children

Hee-Hwan Wang¹, Kakyeong Kim¹, Dong Hyeon Kim², Hyeon Jin Kim³, Woo-Young Ahn³, Joonseok Lee⁴ and Jiook Cha⁵

¹ Department of Brain and Cognitive Sciences, Seoul National University

² Department of Computer Science and Engineering, Seoul National University

³ Department of Psychology, Seoul National University

⁴ Graduate School of Data Science, Seoul National University

⁵ AI Institute, Seoul National University, Seoul, Korea, connectome@snu.ac.kr

Abstract

Sex differences in human cognition and behaviors originate from the differences in the brain structure. Conventional neuroscience studies have rarely shown clear-cut sex differences in the human brain. In stark contrast, recent advance in machine learning (e.g., convolutional neural networks (CNN)) allows an accurate classification of sex based on the brain anatomical images (e.g., T1-weighted MRI), supporting a strong case for the brain sex differences. Can neural networks classify the sex of a preadolescent brain of which the organizational effect of sex steroids has yet to be full-blown? To this end, we implement multiple CNN architectures to learn sex differences in the brain with T1-weighted brain images in prepubertal children ages 9 to 10 years old from Adolescent Brain Cognitive Development study. This study showed that brain sex could be accurately classified by several CNN architectures. In particular, DenseNet-121 was the most predictive model to classify sex (accuracy: 0.92, F1 score: 0.93). Seeking for inferences about the model predictions, we used several interpretability approaches, such as Gradient-weighted Class Activation Map (Grad-CAM) and the occlusion sensitivity. Several brain regions including the cerebellum consistently show sex differences. Taken together, this study presents optimal neural networks for classifying biological sex from the brain morphology.

Keywords— *Sex classification, Brain Imaging, Preadolescents, Convolutional Neural Networks, Explainable AI*

I. INTRODUCTION

Sex differences in the brain structure impacts those in cognition, behaviors, mentality, and psychopathology[1][2][3]. The gap in the human neuroscience literature is that, despite of the stark differences in sex in functional domains, a clear-cut difference in structure of the brain has rarely been shown[4]. Indeed, most of the human neuroimaging literature shows inability to classify the sex based on brain structure because of small sex differences[5]. Contrarily, recent seminal computational neuroimaging begins to show feasible classification of sex based on brain anatomical images (e.g., T1-weighted MRI), supporting a strong case for the brain sex differences in humans. Another important question in the field of human computational neuroscience is the possibility of a scalable increase in performance of CNN applications to brain imaging. Testing this requires rigorous experimentation. A recent study shows limited scalability of CNNs in learning representations from brain images, even lower than simple linear models[6]. However, this study lacks rigorous search for CNNs or hyperparameter optimization, presenting a major limitation of the study. Testing an optimal neural network for 3D brain imaging with a great model space will facilitate the application of big data analytics in cognitive neuroscience and medicine. In this study we search for an optimal CNN that shows a scalable increase in performance to learn the brain sex differences in prepubertal children. To this end, we conduct systematic experiments with several widely used state-of-the-art CNNs leveraging the largest brain imaging data to date in children. Furthermore, we investigate the capacity of deep neural network inferences using several interpretability approaches.

II. MATERIALS AND METHODS

A. ABCD Participants

The participants were enrolled in the Adolescent Brain Cognitive Development (ABCD) study release 2.0 (<http://abcdstudy.org>). The ABCD study is the ongoing longitudinal multi-site study of brain development and health of youth in the United States. 11,875 children with ages of 9 to 10 years were recruited from 21 research sites. After preprocessing and quality control, we used 7,088 participants (male: 3,694; female: 3,356) for analysis.

B. T1-weighted Brain Imaging Acquisition and Preprocessing

We used T1-weighted (T1w) 3D structural MRI images from the ABCD data. The images were preprocessed with fMRIprep[7] and a Nipype tool. Cortical surface reconstruction was applied using the following procedures: Structural MRI was preprocessed using fMRIprep for cortical surface reconstruction[8], which includes, white matter segmentation and initial mesh creation⁸, correction of topological defects, surface optimization [8], correction of topological defects, surface optimization[9][10] and non-linear registration to a spherical surface-based atlas[11]. Each brain anatomical image was skull-stripped and spatially normalized to the MNI pediatric asymmetrical template. The input size of images was resized into (96 x 96 x 96) (**Figure 1**). We split the data set into 80% training set, 10% validation set, and 10% test set. Test set was only used to evaluate the performance of models.

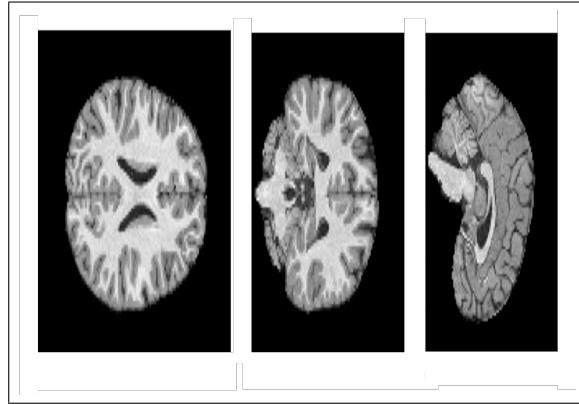


Fig. 1. Examples of 3-dimensional T1-weighted neuroimaging input data. From left to right, horizontal section, coronal section, and midsagittal section of the brain anatomical image. The input size of 3-dimensional images are (96 x 96 x 96).

C. 3D Convolutional Neural Network (3D CNN)

The 3D convolutional neural networks (3D-CNN) architectures are based on network architecture in the previous study [12]. We modified the architecture from previous study and built a simple CNN model as the baseline model

Configuration
conv3D-8
ReLU
Batch Normalization
Max pooling
conv3D-16
ReLU
Batch Normalization
Max pooling
conv3D-32
conv3D-64
ReLU
Batch Normalization
Max pooling
conv3D-128
conv3D-256
ReLU
Batch Normalization
Max pooling
FC-25
Batch Normalization
sigmoid
dropout
softmax

Table 1. The model architecture of the baseline model. The convolution layers have a 3x3x3 kernel size, a 1x1x1 stride, and a 1x1x1 zero padding. The max-pooling layers have a 2x2x2 kernel size and a 2x2x2 stride.

(**Table 1**). The baseline model consisted of six 3D convolutional layers with Rectified Linear Unit (ReLU) activation function, 3D max pooling, and 3D batch normalization . 3D convolution was applied with kernel size 3, stride 1, and zero padding 1. For a 3D max-pooling, we used kernel size 2, stride 2, and no zero padding. The dimension of the fully connected layer was set as 25, the optimal dimension for our task. The output from the fully connected layer was fed into softmax to calculate the probability of each class. We trained the baseline model with a learning rate of 1e-5, training batch size of 32, and l2 regularization of 5*1e-4. Adam optimizer was applied to training the model.

D. 3D Very Deep Convolutional Networks (3D VGGNet)

Very deep convolutional networks (VGGNet) is a CNN architecture proposed by [13] which focuses on the effect of the neural network's depth on its accuracy. VGGNet contains multiple stacking hidden layers to learn more complex hidden representations of large-scale images [13]. We compared multiple variants of VGGNet (i.e., VGGNet-11, VGGNet-13, VGGNet-16, VGGNet-19). Based on the original architecture [13], we additionally applied 3D convolution, 3D max-pooling, and 3D batch normalization for our study. Like the baseline model, 3D convolution was applied with kernel size 3, stride 1, and zero padding 1.

Also, 3D max-pooling was implemented with kernel size 2, stride 2, and no zero padding. Though the number of neurons in the first fully connected layer was the same as the original suggestion (i.e., 4,096 dimensions), the second fully connected layer had 25 neurons in our model. We trained VGGNets with Adam optimizer, a learning rate of 1e-6, training batch size of 32, the dropout rate of 0.5, and l2 regularization of 5*1e-4.

E. 3D Residual Networks (3D ResNet)

Residual networks (ResNet) is devised for solving the vanishing gradient problem of CNN during the backpropagation. ResNet has residual skip connections, which allows us to feed activation from one layer to another [14]. With the skip connections, we built deeper neural architectures (i.e., ResNet-50, ResNet-101, ResNet-152). We stacked three layers of 1x1, 3x3, 1x1 convolutions. 1x1 layers reduce and restore dimensions, and the 3x3 layer is responsible for a bottleneck with smaller input and output dimensions. We used 3D convolution with a kernel size of 7 and a feature map size of 64. We trained ResNet with SGD optimizer with training batch size of 32, momentum of 0.5 and learning rate of 5*1e-7 for ResNet-50, 0.01 for ResNet-101, and 0.001 for ResNet-152.

F. 3D Residual Networks (3D Densely Connected Convolutional Networks (DenseNet))

Densely connected convolutional networks (DenseNet)[15], another popular CNN architecture, were also tested in our study. DenseNet connects each layer to every other layer in a feed-forward way, which increases direct connections between the low and high layers [15]. With the direct connection between layers, DenseNet enhances the information flow, strengthens feature propagation, alleviates the vanishing-gradient problem, and reduces the number of parameters [15]. In particular, DenseNet-121, DenseNet-169, DenseNet-201 and DenseNet-264 were tested in our study. 3D convolution was applied with filter size 7, stride 2, and padding 3, followed by 3D batch normalization and ReLU activation function. Then, 3D max-pooling with the kernel size 3, stride 2, and padding 1 was conducted and fed into 4 dense blocks with transition blocks. In each dense block, 1x1x1 3D convolution and 3x3x3 3D convolution filter was used, and the number of filters used in dense blocks was (6, 12, 24, 16) for DenseNet-121, (6, 12, 32, 32) for DenseNet-169, and (6, 12, 64, 48) for DenseNet-264. 3D batch normalization, ReLU activation, 1x1x1 3D convolution, and 3D AvgPool were applied in the transition block. After 4 dense blocks with 3 transition blocks, a 3D adaptive AvgPool layer was added to make 1,024 1x1x1 dimensions, then flattened and fed into a softmax classification layer. We trained DenseNet with SGD optimizer: with a learning rate of 1e-3, training batch

size of 32, weight decay of 1e-4, Nesterov momentum of 0.9, and ReduceLROnPlateau learning rate scheduler (factor of 0.1, patience of 4). To reduce validation time, we only investigated validation accuracy and increment scheduler count once in two epochs.

G. Explainable AI

Explainable AI (XAI) is an approach to enable mathematical machine learning algorithms to generate explicit declarative knowledge [16]. Many complex learning algorithms such as DNN are black-box models because humans cannot understand the underlying mechanism even when they show good performances. The black-box-ness poses a challenge of trust or justification of conclusions drawn by the models. The challenge significantly limits the medical domain using brain MRI data (e.g., diagnosis) [16]. The complex DNN trained by high-dimensional brain MRI data hampers the application of the results in medical practice since medical professionals cannot understand how and why a machine decision has been made. To address the problem, we visualized Gradient-weighted Class Activation Map (Grad-CAM) [17] and the occlusion sensitivity[18] for the model predictions by using a python package, MONAI [19]. The Grad-CAM informs us to which locations on the input images the networks take attention. The occlusion map enables us to interpret which pixels affect certain decisions made by networks. The high values in the occlusion map give strong evidence to networks' decisions and vice versa.

III. EXPERIMENTS AND RESULTS

In this section, we show results from biological sex classification with 3-dimensional brain anatomical images. We conducted experiments with eleven different CNN architectures and compared the performance of tested models. Furthermore, we investigated the brain regions that most contribute to the sex classification using Gradient-weighted Class Activation Mapping (Grad-CAM) [17] and Occlusion Sensitivity.

A. Results from Sex Classification

VGGNet (VGGNet-11: 0.92; VGGNet-13: 0.92) and DenseNet (DenseNet-121: 0.92) architectures showed outperforming accuracy than the baseline model (Simple CNN: 0.87) in the test data set. . Though VGGNet-11, VGGNet-13, and DenseNet-121 showed the highest accuracy, DenseNet-121 showed the highest F1 score.

B. Grad-CAM and Occlusion Sensitivity

We used Grad-CAM to investigate specific brain regions on images that draw attention from trained models. Interestingly, trained models focused on different brain regions

	Test Accuracy	F1 Score
Simple CNN (baseline)	0.87	0.86
VGGNet-11	0.92	0.91
VGGNet-13	0.92	0.91
VGGNet-16	0.90	0.91
VGGNet-19	0.90	0.91
ResNet-50	0.91	0.90
ResNet-101	0.86	0.84
ResNet-152	0.83	0.83
DenseNet-121	0.92	0.93
DenseNet-169	0.89	0.90
DenseNet-264	0.90	0.91

Table 2. Classification performance of Neural Network Architectures. Both VGGNet and DenseNet architectures are more accurate than the baseline model in predicting biological sex from 3-dimensional brain anatomical images.

to decide whether the brain morphology was male or female (**Figure 2**). Models looked at the left hemisphere parietal lobe and temporal lobe when they correctly predicted males. When models accurately decide that brain images belong to the female, they focus more on the cerebellum, the left hemisphere, and the right hemisphere temporal lobe. In particular, models comprehensively took attention to the inside and outside the brain to include the outline of the brain into their decisions.

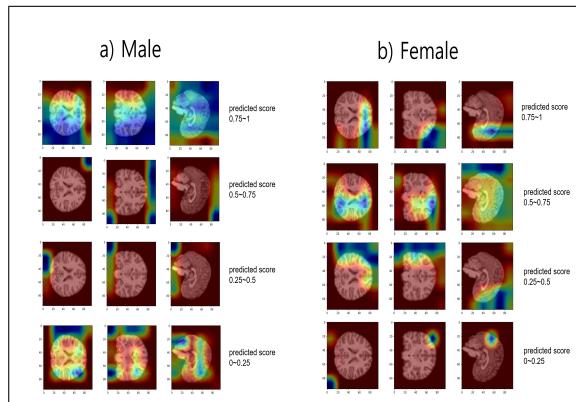


Fig. 2. Gradient-weighted class activation maps of male and female brain anatomical images drawn by VGGNet-13. Figure 2-a shows the results from male brain and Figure2-b is the results from female brain. From left to right, each panel shows mid-horizontal, mid-coronal, and midsagittal sections of the brain in order. Predicted male and female scores indicate the probability that neural networks predicted. Red areas in the two upper panel play an important role in accurate model predictions. The more contrast between the two upper panels and two lower panels show explicit sexual dimorphic locations of the brain structure. In predicting male, parietal lobe and temporal lobe draw attention. In predicting female, models take attention to the cerebellum, the left hemisphere, and the right hemisphere temporal lobe.

One step further, we used occlusion sensitivity to investigate specific brain regions critical to sexual dimorphism of the brain. As seen in (**Figure 3**), some brain regions are contributing to accurate model predictions. In the case of

the male, the right hemisphere dorsolateral prefrontal cortex, ventromedial prefrontal cortex, the left inferior temporal cortex, and posterior segments of the corpus callosum play an important role in models' right decision. Though trained models made wrong decisions, these regions were considered as contributing to prediction as male. In the prediction of the female, the dorsomedial prefrontal cortex, the anterior cingulate cortex, anterior segments of the corpus callosum, and the right hemisphere cerebellar posterior lobe play the same role as in the case of the male.

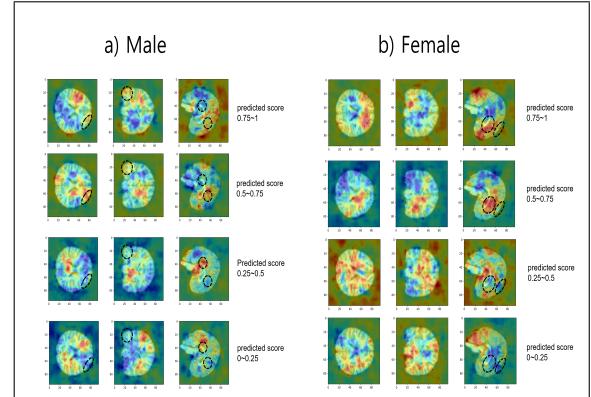


Fig. 3. Occlusion sensitivity maps of male and female brain anatomical images drawn by VGGNet-13. Figure 2-a shows the results from the male brain and Figure2-b is the results from the female brain. From left to right, each panel shows mid-horizontal, mid-coronal, and midsagittal sections of the brain in order. Predicted male and female scores indicate the probability that neural networks predicted. Red areas in the two upper panels importantly contributed to accurate predictions, while red areas in the two lower panels indicate brain regions that lead to wrong predictions. Blue areas indicate vice versa. Thus, more color contrast between the two upper panels and lower panels show explicit sexual dimorphism of the brain structure. Dashed circles point out regions where contrast between high scores and low scores are explicit.

IV. DISCUSSION

In this study, we implemented and compared various CNN architectures to search for an optimal network to classify biological sex from structural brain images in children. Several aspects of this study are worth noting. Firstly, we leveraged the largest data to date of structural brain MRI in youth. As a result, we achieved state-of-the-art performance in predicting biological sex from brain morphology in the developmental period. Secondly, we compared various CNN architectures and found for an optimal model for predicting biological sex from structural brain images. This is a meaningful contribution to the computational neuroscience literature: since the prior CNN studies used only a few hidden layers, the capacity of a deep neural network in sex classification remained unclear [20] [21]. Thirdly, our study permits neuroscientific inferences derived from the CNN using several explainable AI tools. Considering the explainability of deep neural networks is a

challenging issue and unmet needs in Computational Neuroscience, this study shows the utility of the deep neural network and interpretation tools for neuroscientific discovery [20] [21] [22]. This has implications for precision neuroscience: Explainable AI may help specify important features underlying a given outcome label from an individual structure brain image, and thus help important (e.g., clinical) decision making in human brain imaging and related fields.

Of note, our study investigates sexual dimorphism of the brain in children with largest sample sizes. Despite its potential utility as a scientific tool, the lack of sufficiently large data in human brain imaging has often hampered rigorous application of deep learning. Recently, however, several nationwide initiatives collecting unprecedentedly large-scale data (e.g., UK Biobank, Human Connectome Project, and ABCD study) have begun to permit testing scalable deep neural networks in human brain imaging. Though 7,088 images may still be a limited number of samples compared to other computer vision datasets (i.e., CIFAR, MNIST), we show the capacity of deep convolutional neural networks to apply large-scale data sets in neuroscience. This study reports greater accuracy of deeper CNN in sex classification than a recent study showing greater model predictions of shallow CNN than traditional machine learning [21]. Considering that the previous research trained a model with 1.4-fold larger samples, we expect that training deeper CNN with larger samples may increase the accuracy of the model predictions.

Comparing the results from VGGNet and DenseNet, the best performances of each type of network were the same in terms of accuracy, but DenseNet-121 shows the highest F1 score. Thus, it could be said that DenseNet-121 could classify the brain according to biological sex in the most sophisticated way reducing the effect of bias. Interestingly, shallower models show better model performance than deeper models regardless of architectures. A possible explanation is that the increasing number of parameters may result in overfitting, particularly when trainable image sets are small [20]. Considering a small size of training image set in the study, it is possible that models with relatively smaller number of parameters (i.e., VGGNet-11, VGGNet-13, and DenseNet-121) could be more generalizable and show better accuracy than deeper models (i.e., VGGNet-16, VGGNet-19, and DenseNet-169) in the model prediction with the test set. However, the relationship between the number of model parameters in various CNN architectures and the number of image sets need more to be investigated in further research.

Another noteworthy point of our study is the interpretation of deep learning models. In neuroscience, unveiling the decision processes of DNN is crucial to translating computational results to neuroscientific discovery. We used Grad-CAM and occlusion sensitivity, widely used AI interpretation methods, to interpret and verify our CNN results with a priori neuroscience literature. It should be noted

that the results from Grad-CAM support sex differences in asymmetric neurodevelopment of the temporal lobe in line with previous research [23]. The results from the occlusion sensitivity maps are also in line with previous research. Sexual dimorphism of the corpus callosum [24], dorsolateral prefrontal cortex (DLPFC) [25], anterior cingulate cortex [26] has already been reported by previous studies, and our results also support sex differences in these regions. In particular, we repeatedly observed sex differences in small segments of the corpus callosum, such as the posterior part [27] and anterior midbody [27]. Whether the cerebellum structure shows sex differences has been controversial in the human literature. A recent deep learning study reports sexual dimorphism of the cerebellum in children similarly to this study [22]. Taken together, this study shows that DNN can learn representations to classify biological sex from the brain morphology.

There is a large room for improvement in the application of DNN in neuroscience. Firstly, transfer learning could help overcome the limitation of a small sample size and achieve computational efficiency. DNN could leverage pre-trained representations and be applied to different data sets with few-shot or one-shot learning. Secondly, we need to better understand the decision-making process of DNN. Though our study shows the brain regions (i.e., voxels) important for the model to classify the sex, it is yet unclear why those regions are important. Furthermore, the relationship between morphometric features and sex (e.g., whether large total brain volume is associated with males or not) is still unclear. In further research, we aim to reveal important brain features shared by all females or males by applying machine learning interpretation to the entire samples rather than to a few typical samples.

In conclusion, the optimal DNN reported in this study could help us understand sex differences in human behaviors resulting from those in brain structure.

SUMMARY OF THIS PAPER

A. Problem Setup

Neuroscientists and clinicians presume that sexual dimorphism of the brain is one of the fundamental factor contributing to sex differences in cognition, behavior, mentality, and psychopathology. A large body of studies try to reveal sex differences in the brain, but it is still controversial. Recently, researchers begin to show feasible classification based on brain anatomical images using machine learning and deep learning framework. Yet, it is unclear that deep learning framework could show a scalable increase in sex classification with brain structure compared to conventional linear models.

B. Novelty

Firstly, we leveraged the largest data to date of structural brain MRI in youth. As a result, we could achieve state-of-the-art performance in predicting biological sex from brain morphology in the developmental period. Secondly, we compared various CNN architectures and searched for an optimal model for predicting biological sex from structural brain images. Thirdly, our study permits neuroscientific inferences derived from the black-box CNN using several explainable AI tools.

C. Algorithms

We conducted experiments with four types of most widely used CNN architectures (simple CNN, VGGNet, ResNet, DenseNet) and compared the performance in the test set. Furthermore, we investigated the brain regions that most contribute to the sex classification using Gradient-weighted Class Activation Mapping (Grad-CAM) and Occlusion Sensitivity.

D. Experiments

In terms of test accuracy, VGGNet-11, VGGNet-13, and DenseNet-121 outperformed (VGGNet-11: 0.92; VGGNet-13: 0.92; DenseNet-121: 0.92). In particular, DenseNet-121 showed the highest F1 score (DenseNet-121: 0.93). Models with shallower layers show better model performance than deeper models regardless of types of architectures. The results from Grad-CAM and Occlusion Sensitivity pointed out brain regions presumed to be associated with sex differences.

REFERENCES

- [1] Riecher-Rössler, A. Sex and gender differences in mental disorders. *The Lancet Psychiatry* vol. 4 8–9 (2017).
- [2] Ruigrok, A. N. V. et al. A meta-analysis of sex differences in human brain structure. *Neurosci. Biobehav. Rev.* **39**, 34–50 (2014).
- [3] Forde, N. J. et al. Sex Differences in Variability of Brain Structure Across the Lifespan. *Cereb. Cortex* **30**, 5420–5430 (2020).
- [4] Grabowska, A. Sex on the brain: Are gender-dependent structural and functional differences associated with behavior? *J. Neurosci. Res.* **95**, 200–212 (2017).
- [5] Joel, D. et al. Sex beyond the genitalia: The human brain mosaic. *Proceedings of the National Academy of Sciences* vol. 112 15468–15473 (2015).
- [6] Schulz, M.-A. et al. Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nat. Commun.* **11**, 4238 (2020).
- [7] Esteban, O. et al. fMRIprep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).
- [8] Dale, A. M., Fischl, B. Sereno, M. I. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage* **9**, 179–194 (1999).

- [9] Fischl, B., Liu, A. Dale, A. M. Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Trans. Med. Imaging* **20**, 70–80 (2001).
- [10] Ségonne, F., Pacheco, J. Fischl, B. Geometrically accurate topology-correction of cortical surfaces using nonseparating loops. *IEEE Trans. Med. Imaging* **26**, 518–529 (2007).
- [11] Zeineldin, R. A., Karar, M. E., Coburger, J., Wirtz, C. R. Burgert, O. DeepSeg: deep neural network framework for automatic brain tumor segmentation using magnetic resonance FLAIR images. *Int. J. Comput. Assist. Radiol. Surg.* **15**, 909–920 (2020).
- [12] Cole, J. H., Poudel, R. P., Tsagkrasoulis, D., Caan, M. W., Steves, C., Spector, T. D., Montana, G. (2017). Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, **163**, 115–124.
- [13] Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
- [14] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770–778).
- [15] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700–4708).
- [16] Holzinger, A., Biemann, C., Pattichis, C. S., Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain?. arXiv preprint arXiv:1712.09923.
- [17] Selvaraju, R. R. et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 IEEE International Conference on Computer Vision (ICCV) (2017)
- [18] Zeiler, M. D. Fergus, R. Visualizing and Understanding Convolutional Networks. *Computer Vision – ECCV 2014* 818–833 (2014)
- [19] Website. Nic Ma, Wensi Li, Richard Brown, Yiheng Wang, Behrooz, Benjamin Gorman, ... cgrain. (2021, June 1). Project-MONAI/MONAI: 0.5.3 (Version 0.5.3). Zenodo.
- [20] Lever, J., Krzywinski, M. Altman, N. Model selection and overfitting. *Nature Methods* vol. 13 703–704 (2016).
- [21] Brueggeman, L., Thomas, T. R., Koomar, T., Hoskins, B. Michaelson, J. J. Sex Differences in the Brain: Divergent Results from Traditional Machine Learning and Convolutional Networks. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI) (2020) doi:10.1109/isbi45749.2020.9098402.
- [22] Adeli, E. et al. Deep learning identifies morphological determinants of sex differences in the pre-adolescent brain. *Neuroimage* **223**, 117293 (2020).
- [23] Lenroot, R. K. et al. Sexual dimorphism of brain developmental trajectories during childhood and adolescence. *Neuroimage* **36**, 1065–1073 (2007).
- [24] Ardekani, B. A., Figarsky, K. Sidtis, J. J. Sexual dimorphism in the human corpus callosum: an MRI study using the OASIS brain database. *Cereb. Cortex* **23**, 2514–2520 (2013).
- [25] Lotze, M. et al. Novel findings from 2,838 Adult Brains on Sex Differences in Gray Matter Brain Volume. *Sci. Rep.* **9**, 1671 (2019).
- [26] Paus, T. et al. In vivo morphometry of the intrasulcal gray matter in the human cingulate, paracingulate, and superior-rostral sulci: hemispheric asymmetries, gender differences and probability maps. *J. Comp. Neurol.* **376**, 664–673 (1996).
- [27] Bermudez, P. Zatorre, R. J. Sexual dimorphism in the corpus callosum: methodological considerations in MRI morphometry. *Neuroimage* **13**, 1121–1130 (2001).