Fairness-related performance and explainability effects in deep learning models for brain image analysis

Emma A.M. Stanley, Matthias Wilms, Pauline Mouches, Nils D. Forkert

University of Calgary, AB, Canada

***Abstract*— Explainability and fairness are two key factors for the effective and ethical clinical implementation of deep learning-based machine learning models in healthcare settings. However, there has been limited work on investigating how unfair performance manifests in explainable artificial intelligence (XAI) methods, and how XAI can be used to investigate potential reasons for unfairness. Thus, the aim of this work was to analyze the effects of previously established sociodemographic-related confounders on classifier performance and explainability methods. A convolutional neural network (CNN) was trained to predict biological sex from T1-weighted brain MRI of 4,547 9–10-year-old adolescents from the Adolescent Brain Cognitive Development study. Performance disparities of the trained CNN between White and Black subjects were analyzed and saliency maps were generated for each subgroup at the intersection of sex and race. The classification model demonstrated a significant difference in the percentage of correctly classified White male (90.3±1.7%) and Black male (81.1±4.5%) children. Conversely, slightly higher performance was seen in Black female (89.3±4.8%) compared to White female (86.5±2.0) children. Saliency maps showed subgroup-specific differences, corresponding to brain regions previously associated with pubertal development. In line with this finding, average pubertal development scores of subjects used in this study were significantly different between Black and White females ($p<0.001$) and males ($p<0.001$). We demonstrate that a CNN with significantly different sex classification performance between Black and White adolescents can identify different important brain regions when comparing subgroup saliency maps. Importance scores vary substantially between subgroups within brain structures associated with pubertal development, a race-associated confounder for predicting sex. This study illustrates that unfair models can produce different XAI results between subgroups, and that these results may explain potential reasons for biased performance.**

***Index Terms*— explainable AI, fairness, bias, magnetic resonance imaging, machine learning, ABCD study**

I. Introduction

The use of artificial intelligence (AI) for analysis, screening, and diagnosis of medical image data has grown rapidly in recent years [1]. While the application of AI within radiology has led to promising results in many cases [2], several barriers to effective and ethical clinical implementation still exist. One such barrier is the so-called "black box problem", pertaining to the idea that the decision-making processes of deep learning models are typically difficult to understand and interpret for humans. To address this problem, research into explainable AI (XAI) has gained more interest in the medical imaging domain. For image classification or regression tasks, XAI techniques such as saliency maps [3], Local Interpretable Model-Agnostic Explanations (LIME) [4], and Shapley Additive Explanations [5] aim to identify regions in an image that are highly influential to the model's final prediction. The aim of implementing these techniques is to provide explanations for AI-generated decisions to patients and clinicians to improve trust in the prediction. Moreover, these techniques have also been used in research to identify regions of interest that may be associated with normal or pathological processes [6],[7].

A second major barrier to clinical implementation of deep learning models is the risk of unfair or biased performance. An unfair model can be defined as one that has significantly different error rates between, for example, subjects belonging to distinct categories of a sociodemographic group, including but not limited to race, sex, gender, and socioeconomic status. For instance, Seyyed-Kalantari et al. [8] reported that a deep learning algorithm trained on chest x-ray datasets systematically underdiagnosed traditionally underserved populations, including female patients, Black patients, and those of low socioeconomic status. Likewise, a number of other potentially biased computer-aided diagnosis (CAD) models have been reported in recent years [9][10]. However, it is difficult to assess reasons for unfairness in deep learning models due to different types of biases that could result in unfair models. For instance, medical imaging data available for training and validation of a model may suffer from a selection bias, for example, resulting from including only participants who were willing and able to contribute data for a certain study. Additionally, ground truth diagnostic labels may be affected by implicit biases from health care providers [11], which can lead to an underrepresentation of subgroups who already face disparities in accessing health care [12][13].

While XAI and subgroup fairness analysis have both gained increasing interest in the medical imaging domain, little work has been performed so far investigating how these two aspects of machine learning models interact. For example, XAI methods may be able to help identify regions linked to bias, which was investigated in [14] in the context of COVID computed tomography scan classifiers. However, this work was limited to assessing how different model architectures focus on spurious artifacts in an image, and not how these biases could relate to unfair performance. In general, research in this domain is very limited.

In this study, we use the task of sex classification using brain magnetic resonance imaging (MRI) from the Adolescent Brain Cognitive Development (ABCD) study to investigate how unfair performance in a deep learning model affects the output of XAI. We select this example problem since ground truth labels for biological sex are unaffected by diagnostic bias, total brain volume is known to vary between males and females [15], and some cortical and subcortical structural characteristics have been linked to sexual dimorphism [16]. Moreover, the stage of pubertal development also plays a role in the morphology of brain structures [17][18][19]. Adeli et al. [6] trained a deep convolutional neural network (CNN) to successfully classify sex using T1-weighted MRI datasets from 9 and 10 year-old children in the ABCD study, and used saliency maps to identify regions that were significant predictors of sex. Their study was motivated by the proposition that deep learning-identified morphological sex differences may help inform research into sex-linked neuropsychiatric diseases, especially those that commonly emerge during adolescence such as eating disorders and mood disorders [20]. Moreover, Adeli et al. [6] found that the pubertal development score was the most significant confounder for predicting sex, in addition to age and socioeconomic status, and generated a confounder-free saliency map to identify a number of brain regions informative for sex classification.

While our setup generally follows that of Adeli et al. [6] (a CNN-based sex classifier is trained to analyze sex-related morphological brain differences), instead of correcting for demographic confounders, we use them to our advantage to explicitly investigate their effects on classifier performance and how they manifest in saliency maps. Pubertal development score, the most significant confounder of sex prediction [6], is important for this task because different timings in the onset of puberty can be linked to race. For example, it has been reported that Black girls and boys on average begin puberty earlier than their White counterparts [21][22][23]. Wu et al. [23] reported that development of some early secondary sexual characteristics began at a mean age of 9.5 for Black girls, compared 10.5 for non-Hispanic White girls. Likewise in boys, Herman-Giddens et al. [22] reported a mean onset of the corresponding early secondary sexual characteristics at 10.25 years for Black boys compared to 11.47 years for non-Hispanic White boys. These findings are notable given that the age of subjects used in [6] and in our study are aged 9 and 10.

Thus, the main reason to use this data and this task is to evaluate how known effects of sex- and race-associated brain morphology manifest in the output of XAI methods that can also be applied in other medical imaging applications. Within this context, we believe that XAI can be used as a tool to identify sources of bias in subgroup level analyses. If regions linked to sociodemographic-associated morphological differences are provided as explanations in XAI, we can gain a better understanding of why a given model may exhibit performance disparities between subgroups.

Our main contributions are as follows: (1) implementation and optimization of a CNN model architecture that is highly successful at performing sex classification in adolescent subjects, (2) rigorous demographic subgroup performance analysis of the model, (3) use of saliency maps to identify important brain regions for the classification task, and (4) investigation into how those regions vary by demographic subgroup and whether they are related to established sex- and puberty-associated morphological differences.

II. Materials and Methods

2.1 ABCD Study

The data used in this work was collected from the 3.0 release of the Adolescent Brain Cognitive Development (ABCD) study (https://abcdstudy.org). The ABCD study is a large-scale longitudinal study of adolescent health and development, designed to reflect the sociodemographic variation in the United States as best as possible [24]. T1-weighted MRI, demographic information, and pubertal development measures from 4,547 subjects at the baseline timepoint in which subjects were 9-10 years old were used in this research.

The biological sex (defined as sex assigned at birth) and race information for the children were drawn from the demographics survey filled out by a parent or guardian. In this work, we restricted the racial subgroups of interest for the fairness and XAI analysis to those two with the highest amount of representation in the data, i.e., subjects who identified as only White (n=3008) or only Black (n=390). Subjects who identified as mixed race and other races were included in model training and testing but not independently analyzed for fairness.

Pubertal development was assessed according to the Pubertal Development Scale (PDS), which was completed by a parent or guardian. The PDS is a questionnaire that provides a reliable and valid assessment of pubertal development in adolescents [25]. The pubertal development score for each subject was calculated by averaging responses on a scale of 1 to 4 from five categories, with 4 representing the most advanced pubertal development [26]. Responses of “don’t know” or “refuse to answer” were not included in the average score calculation.

Detailed descriptions of study recruitment and exclusion criteria, as well as demographic and physical health assessments can be found in [24] and [27] respectively.

2.2 MRI data

T1-weighed MRI datasets were acquired across 21 sites in the United States according to protocols detailed in [28] at 1

mm isotropic resolution and processed by the ABCD minimal-processing pipeline [29]. For this work, the datasets additionally underwent skull stripping and rigid registration to the NIHPD 7.5-13.5-year-old T1-weighted asymmetric brain atlas with 1 mm isotropic voxel size [30] using Advanced Normalization Tools (ANTs) [31] version 2.3.5. The final image dimensions were 197×233×189 voxels, with voxel values linearly mapped to a range of [-0.5, 0.5].

2.3 Deep learning model

The CNN trained to predict sex from T1-weighted MRI datasets in this work was adapted from the Simple Fully Convolutional Neural Network, proposed in [32]. The architecture consisted of five convolution blocks, each containing a convolutional layer with a (3×3×3) kernel, batch normalization, (2×2×2) max pooling, and ReLU activation. A sixth convolutional block consisted of a convolutional layer with a (1×1×1) kernel, batch normalization, and ReLU activation. The convolutional filter sizes were 32, 64, 128, 256, 256, and 64 for each respective block. A seventh block consisted of an average pooling layer, a dropout layer (rate of 0.2), and a dense classification node with sigmoid activation. The model was trained with a binary cross entropy loss, the Adam optimizer (learning rate=1e-3 with decay=3e-3), and a batch size of two for 100 epochs.

2.4 Experiments

The model was trained on 4,547 MRI datasets with a 5-fold cross-validation scheme. For each fold, 80% of the data was designated for training and 20% for testing. Additionally, 10% of the training data was used as a validation set to monitor and prevent overfitting. Training and test splits in each fold were stratified by sex (male, female) and race (White, Black), to ensure equal representations. Subjects of other races were also included in each training and test set but were not independently evaluated for fairness and thus not stratified for. A detailed description of the representation from each subgroup of interest in each fold is presented in Appendix Table 4. Performance metrics are reported as the average and standard deviation over the five folds of the cross-validation, in which each subject was included in a test set exactly once.

2.5 Performance Metrics

Model performance was assessed according to the following measures: overall accuracy, subgroup true classification rate, subgroup true male classification rate, and subgroup true female classification rate (see Table 1).

Table 1 Definitions of performance metrics.

Metric	Abbreviation	Definition
Accuracy (%)	-	(correctly classified subjects) ÷ (all subjects)
True Classification Rate (%)	TCR	(correctly classified subjects in subgroup) ÷ (all subjects in subgroup)
True Male Classification Rate (%)	TMR	(correctly classified male subjects in subgroup) ÷ (all male subjects in subgroup)
True Female Classification Rate (%)	TFR	(correctly classified female subjects in subgroup) ÷ (all female subjects in subgroup)

2.6 Saliency maps

The SmoothGrad method [33] for saliency map generation using the tf-keras-vis toolkit [34] was employed in this work using the model and test subjects from a single cross-validation fold. This method adds Gaussian noise to the data before calculating the standard saliency map and averages the result over several iterations. In this work, the Gaussian noise was sampled from a distribution of N(0,0.2) and averaged over 20 iterations for each subject.

We generated average saliency maps for the four subgroups of interest (White males, Black males, White females, Black females) individually as well as for an aggregate sample. Computing saliency maps for an aggregate group of subjects without considering demographic variables such as race is typically done in other works, including the study from Adeli et al. [6], and thus provides a baseline comparison. Average SmoothGrad saliency maps were generated using 20 randomly selected subjects from each subgroup that were correctly classified by the model. In the aggregate saliency map, the 20 subjects (10 male, 10 female) used to generate the map were randomly sampled from the correctly classified test subjects without considering race. The same number of subjects were used in each average saliency map for consistency between subgroups, since White subjects significantly outnumber Black subjects in the data. Each individual saliency map was nonlinearly registered to the NIHPD atlas by applying the affine transformation matrices and deformation fields form the nonlinear transformation of the corresponding subject's T!-weighted MRI to the NIHPD atlas using ANTs. After transformation to the common atlas space, each saliency map was averaged over the 20 subjects in each subgroup. Intensity values in each subject saliency map were linearly mapped to a range of [0, 255] prior to averaging.

Extraneous noise in each resulting average saliency map was eliminated by thresholding out the lower 50% of voxel intensity values. To identify which brain regions were included in the average saliency maps, a saliency score was computed for each cortical and subcortical region in the left (LH) and right (RH) hemispheres defined in the CerebrA atlas [35]. The CerebrA atlas which is based in the MNI space was nonlinearly registered to the NIHPD atlas by applying the affine transformation matrices and deformation fields from the nonlinear transformation of the MNI ICBM152 atlas to the NIHPD atlas using ANTs. The saliency score, which represents the number of salient voxels within a given brain region, was computed by the summation of non-zero saliency map voxels in each region divided by the number of total voxels in that region and reported as a percent value. A weighted saliency score was also calculated by computing the mean saliency map intensity value in each region, min-max scaling each region's mean intensity mean to a range of [0,1], and multiplying that value by the previously mentioned saliency score as a weighting factor.

To quantify overlap in salient regions between subgroups, Dice coefficients were computed between the binarized average maps of each of the four subgroups of interest as well as the aggregate saliency map.

2.7 Statistics

Statistically significant differences in ages and PDS scores between White males, White females, Black males, and Black females were evaluated with the Mann-Whitney U-test. Differences in model performance metrics between White and Black subgroups were assessed using the two-tailed Student's t-test. A significance level of α=0.05 was set for both tests.

III. Results

The ages and PDS scores for each subgroup are reported in Table 2. The Mann-Whitney U-test revealed that pubertal development scores of Black children were significantly higher than those of White children, for both males ($p<0.001$), females ($p<0.001$), and overall ($p<0.001$). There were no significant different in ages between Black and White males ($p=0.553$), females ($p=0.381$), or overall ($p=0.297$).

Table 2 Age and pubertal development scale (PDS) scores for each subgroup.

Subgroup	Age in years (Average ± Std. Deviation)	PDS score (Average ± Std. Deviation)
White males	9.95 ± 0.64	1.37 ±0.33
Black males	9.92 ± 0.60	1.67 ± 0.49
White females	9.94 ± 0.63	1.57 ± 0.50
Black females	9.90 ± 0.62	2.10 ± 0.60

The sex classification model achieved an overall average accuracy of 87.8±0.98%, demonstrating high consistency across folds. Overall and subgroup model performance metrics for each fold are presented in Table 3. True classification rates, which includes correctly classified males and females, were 88.4±0.73% for the White subgroup and 85.1±0.73% for the Black subgroup. True male classification rates for the White subgroup were significantly higher ($p=0.03$) than the Black subgroup by an average of 9.2%. Conversely, true female classification rates for the Black subgroup were an average of 2.8% higher than the White subgroup, but not statistically significant ($p=0.260$). The model from fold 3 of the cross validation was selected for generation of saliency maps due to the overall and subgroup accuracy metrics for this fold aligning closest to the average metrics across all folds.

Table 3 Model performance (TCR=True classification rate; TMR=True male classification rate; TFR=True female classification rate).

Fold	Overall Accuracy (%)	TCR White (%)	TCR Black (%)	TMR White (%)	TMR Black (%)	TFR White (%)	TFR Black (%)
1	88.8	88.7	90.7	92.7	88.3	84.7	93.1
2	86.6	87.2	80.5	90.4	80.0	84.0	81.0
3	88.8	89.0	85.6	91.0	81.7	87.0	89.7
4	87.3	88.2	83.9	88.4	76.7	88.0	91.4
5	87.3	88.9	84.7	89.0	78.7	88.7	91.2
Average ± Std. Deviation	87.8 ± 1.0	88.4 ± 0.7	85.1 ± 3.7	90.3 ± 1.7	81.1 ± 4.5	86.5 ± 2.0	89.3 ± 4.8

Dice coefficients assessing overlap between the complete aggregate and subgroup saliency maps are visualized in Fig. 1. The saliency maps for White males, White females, and Black males all showed overlap to the aggregate saliency map in the range of 0.82-0.84, while Black females had the lowest overlap with the aggregate indicated by a Dice coefficient of 0.77. Within-sex saliency map overlap (0.84 for females and 0.83 for males) was higher than comparing overlap values between sexes. White females had similar overlap with White males and Black males, both at a Dice coefficient of 0.73, while Black females had higher overlap with Black males (0.71) than White males (0.68). Differences in average saliency maps are also seen qualitatively (Fig. 2).

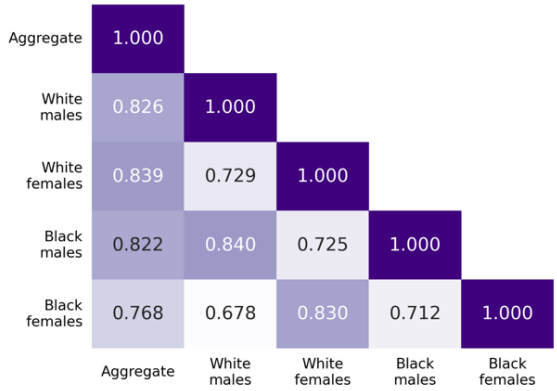

Fig. 1 Subgroup saliency map Dice coefficients.

Saliency scores, measuring the proportion of salient voxels in each brain region, and weighted saliency scores, which account for salient voxel intensity values, are reported for each subgroup in Appendix Tables 5 and 6, respectively. Top regions for the aggregate map based on weighted saliency scores included the fourth ventricle, cerebellum white matter, brainstem, amygdala, cerebellar vermal lobules VIII-X, inferior lateral ventricle, cerebellum gray matter, fusiform gyrus, and entorhinal cortex.

In several brain regions, differences are observed when comparing weighted saliency scores between sexes (see Fig. 3). For example, higher scores for females than males were found in the RH cerebellum white matter (White females=28.5%, Black females=24.5%, White males=12.3%, Black males=12.0%), LH cerebellar vermal lobules VIII-X (White females=7.3%, Black females=5.9%, White males=3.9%, Black males=3.1%), and RH cerebellum gray matter (White females=2.1%, Black females=1.5%, White males=0.6%, Black males=0.4%). Regions displaying higher scores for males than females include the LH inferior temporal lobe (White males=2.6%, Black males=2.7%, White females=1.2%, Black females=0.5%), LH entorhinal cortex (White males=4.0%, Black males=3.0%, White females=1.2%, Black females=0.4%), LH middle temporal lobe (White males=3.0%, Black males=3.0%, White females=0.3%, Black females=0.06%), LH amygdala (White males=3.0%, Black males=2.7%, White females=1.9%, Black females=0.1%), and LH superior temporal lobe (White males=2.4%, Black males=2.2%, White females=0.1%, Black females=0.03%). As seen in the numbers above and Fig. 3, some structures displaying sex differences in weighted saliency score also differ by race.

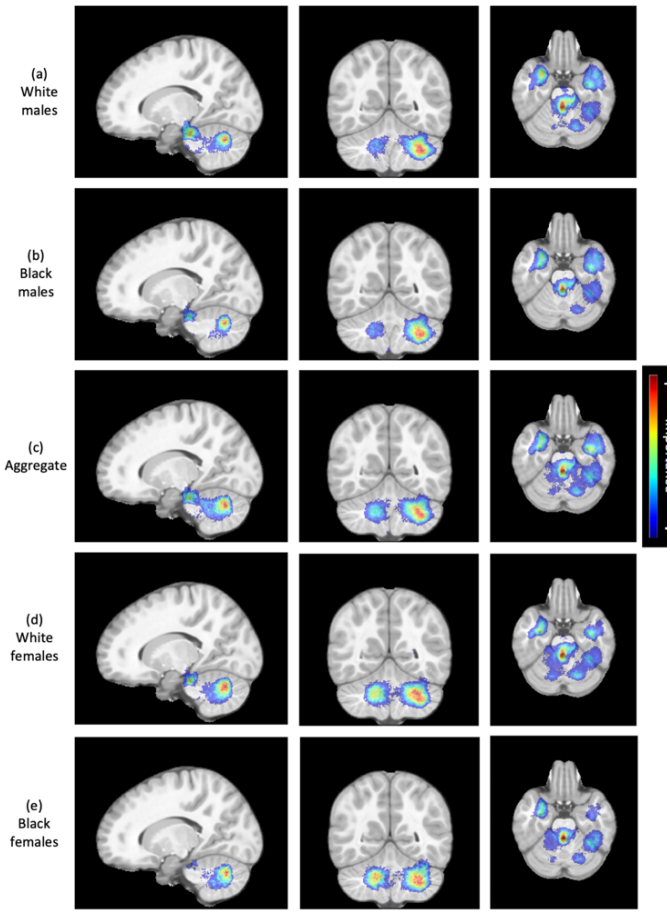

Fig. 2 Average saliency maps for (a) White males, (b) Black males, (c) aggregate sample, (d) White females, and (e) Black females overlaid on the NIHPD atlas (blue: low importance; red: high importance). The same slices are visualized for each group.

For instance, in the cerebellar vermal lobules VIII-X and LH entorhinal cortex, White children had higher weighted saliency scores than Black children of the same sex. In the cerebellum white matter, LH amygdala, and LH middle temporal lobe, White females had higher weighted saliency scores than Black females, while scores for the males appear similar in those regions. In several regions (e.g., LH inferior lateral ventricle, LH amygdala, LH entorhinal cortex), Black females had much lower scores than the remaining subgroups.

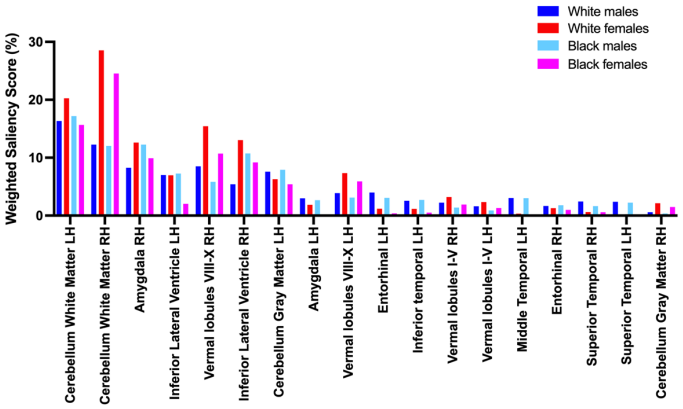

Fig. 3 Weighted saliency scores of each subgroup in selected brain regions (LH = left hemisphere; RH = right hemisphere).

IV. Discussion

The primary aim of this work was to investigate how established demographic-associated confounding effects relevant to adolescent brain sex classification manifest in model performance and saliency maps for model explainability. The results revealed that a CNN, which achieves high and consistent overall accuracy with cross validation, shows significant performance disparities between White and Black adolescents. Saliency maps generated for each subgroup demonstrated varied amounts of overlap overall, indicating that subjects of different sexes and races had different brain regions that were considered important for the model's prediction. A number of these saliency map inconsistencies were in brain regions that are known to be affected by pubertal development, the onset of which is known to vary between the racial subgroups investigated in this work.

The overall classification accuracy of 87.8% is comparable to results from Adeli et al. [6] who also used a CNN for sex classification of ABCD subjects at 9-10 years old using T1-weighted MRI datasets, achieving an accuracy of 89.6% [6]. Our CNN architecture was based on [32] who achieved 98.9% accuracy on a comparable number of subjects, although using subjects aged 44-80 years. However, it should be noted that it was not the goal of this work to develop the best performing classifier for this task but rather to investigate if XAI methods can be used to uncover reasons for biased machine learning models. When analyzing model performance by subgroup, Black and White female classification accuracy did not show significant differences, although Black females had a slightly better performance by an average of 2.8%. However, the classification accuracy of Black male children was significantly lower than that of White male children by an average of 9.2% across five folds. These differences in classification performance of Black children may be due to reasons related to both pubertal development and an imbalanced representation of Black subjects compared to White subjects. Since Black males had low representation but higher PDS scores than other males, the model may have misclassified more children in this group since male brains with higher pubertal development may appear more similar to the average brain of white females at the same age. On the other hand, Black females had much higher PDS scores than the other subgroups, so the model may have been even more confident that highly pubertally developed brains likely belonged to females.

These results highlight the problem that performance metrics on an aggregated test set can potentially mask significant disparities between sociodemographic subgroups. This implies that AI models with high overall performance that are implemented in clinical practice may systematically misdiagnose already underrepresented or at-risk subpopulations if not assessed for these disparities during development, as previously shown by [8].

In the average aggregate saliency map, regions with the highest weighted saliency scores included the fourth ventricle, cerebellum white matter, brainstem, amygdala, inferior lateral ventricle, cerebellar vermal lobules VIII-X, and cerebellum gray matter. The confounder-free saliency map proposed by Adeli et al. [6] also found the corpus medullare (cerebellum

white matter) and amygdala to be in the top regions important to sex classification. Their study additionally found the parahippocampus, hippocampus, and other cerebellar lobules (III, V, VI) to be important brain regions determined in their saliency maps. While these regions have lower weighted scores in our aggregate saliency map, unweighted saliency scores indicate that these regions did contain salient voxels (LH hippocampus=4.17%, RH hippocampus=3.78%, LH parahippocampal gyrus=9.2% RH parahippocampal gyrus=1.9%, LH vermal lobules I-V=10.9%, RH vermal lobules I-V=10.3%, LH vermal lobules VI-VII=6.8%, RH vermal lobules VI-VII=10.2%). Differences between the two studies may be partially attributed to Adeli et al. [6] assuming that sex affects the brain bilaterally, and therefore mirroring the left hemisphere for training their model and generating saliency maps. Additionally, their scans were affinely registered to an adult brain template, whereas ours were rigidly registered to an age-appropriate pediatric template.

Dice coefficients representing overlap between each of the average saliency maps indicated that the four subgroups had overlap coefficients between 0.77-0.84 with the aggregate saliency map, which was generated with a random, non-race stratified sample of correctly classified males and females. Further analysis of subgroup Dice coefficients between subgroups revealed that average maps for the same sex had higher overlap than when comparing different sexes. This underscores that different class saliency maps should be considered when proposing regions of interest, especially since standard saliency maps do not indicate whether a region contributed positively or negatively to class prediction. Moreover, we also found that within the same sex, there is only 83%-84% overlap between Black and White subgroups. This indicates that there may be factors associated to race of the subjects that lead to differences in the saliency maps, such as pubertal development stage. In the data used for training and evaluation of this model, Black subjects had significantly higher PDS scores than white subjects, with insignificant differences in age. Although the PDS score is limited to perceived rather than objective physiological measures of pubertal development, this finding is generally in agreement with the literature [22][21].

Adeli et al. [6] found that the pubertal development score was the most significant confounder for sex classification with their CNN. They reported that misclassified boys had higher PDS scores, and misclassified girls had lower PDS scores than correctly classified subjects of the same sex. Since we show that the average PDS scores of Black males and females are higher than White males and females, their finding is in line with our results showing a significantly lower true classification rate of Black males and a slightly higher true classification rate of Black females. Additionally, Adeli et al. [6] found that the cerebellum was the brain region most highly confounded by pubertal development scores. Our saliency maps also indicate the cerebellum as highly influential, with much higher weighted saliency scores for females compared to males in RH cerebellum white matter, cerebellar vermal lobules VIII-X, and RH cerebellum gray matter, all of which additionally showed higher scores for White females than Black females. In the cerebellar vermal lobules VIII-X, White males also had higher scores than Black males. These discrepancies in saliency scores within the cerebellar structures may suggest that sex and race-associated differences in pubertal stage may be contributing to the deep learning model using morphological information from this region differently for the different subgroups. Furthermore, structures included in the medial temporal lobe, such as the hippocampus and amygdala, which were identified in our saliency maps, are often reported as being influenced by pubertal development [18],[19],[36]. The Black female subgroup, who had the highest average PDS score out of the four subgroups, showed much lower weighted saliency scores in the LH amygdala and the RH hippocampus in comparison to other subgroups and the aggregate sample, potentially due to the more advanced puberty status.

Also of note is that while the saliency map for Black females has the least amount of overlap with the aggregate saliency map and shows much less salient activation in many regions that were prominent for other subgroups (e.g., LH amygdala, LH parahippocampal gyrus, LH hippocampus), Black females had the second highest subgroup classification performance with an average of 89.3%, even exceeding the model's overall accuracy of 87.8%. This suggests that poor concordance of a subgroup's saliency map to other subgroups and the aggregate does not necessarily correlate to poor classification performance.

Limitations of this study include the use of a single method for model explainability. Future work will involve the addition of other methods such as LIME, Shapley Additive Explanations, and GradCAM, which will allow for the comparison of XAI methods and how potential biases manifest differently in each. Additionally, the attributes analyzed in this work were limited to the intersection of biological sex with White and Black subjects. Future work should include the addition of more racial subgroups as well as data on the socioeconomic status into performance and explainability analysis.

V. CONCLUSION

This study used the task of sex classification based on T1-weighted MRI as a well-defined case example for investigating how potential sources of bias manifest in deep learning model performance and saliency maps. We demonstrate that presenting performance metrics and applying XAI techniques on an aggregated population may mask subgroup specific differences. We also show that a more comprehensive XAI analysis may help to identify such differences, which can help to link reasons for unfair model performance to physiological differences between sociodemographic subgroups. Further research in this area is needed to understand how biases can be revealed with other explainability methods and other medical imaging machine learning models.

ACKNOWLEDGMENTS

Data used in the preparation of this article were obtained from the Adolescent Brain Cognitive DevelopmentSM (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children age 9-10 and follow them over 10 years into early adulthood. The ABCD Study® is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041048, U01DA050989, U01DA051016, U01DA041022, U01DA051018, U01DA051037, U01DA050987, U01DA041174, U01DA041106, U01DA041117, U01DA041028, U01DA041134, U01DA050988, U01DA051039, U01DA041156, U01DA041025, U01DA041120, U01DA051038, U01DA041148, U01DA041093, U01DA041089, U24DA041123, U24DA041147. A full list of supporters is avail- able at https://abcdstudy.org/federal-partners.html. A listing of participating sites and a complete listing of the study investigators can be found at https://abcdstudy.org/consortium_members/. ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in the analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators. The ABCD data repository grows and changes over time. The ABCD data used in this report came from doi: 10.15154/1524794.

This research was enabled in part by support provided by Calcul Québec (https://www.calculquebec.ca/) and Compute Canada (www.computecanada.ca).

This work was supported by the River Fund at the Calgary Foundation, Alberta Innovates, Canada Research Chairs Program, and the Canadian Institutes of Health Research.

REFERENCES

[1] L. Lo Vercio et al., "Supervised machine learning tools: a tutorial for clinicians," J. Neural Eng., vol. 17, no. 6, p. 062001, Dec. 2020, doi: 10.1088/1741-2552/abbff2.

[2] H.-P. Chan, L. M. Hadjiiski, and R. K. Samala, "Computer-aided diagnosis in the era of deep learning," Med. Phys., vol. 47, no. 5, pp. e218–e227, 2020, doi: 10.1002/mp.13764.

[3] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," ArXiv13126034 Cs, Apr. 2014, [Online]. Available: http://arxiv.org/abs/1312.6034

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," ArXiv160204938 Cs Stat, Aug. 2016, [Online]. Available: http://arxiv.org/abs/1602.04938

[5] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems, 2017, vol. 30. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[6] E. Adeli et al., "Deep learning identifies morphological determinants of sex differences in the pre-adolescent brain," NeuroImage, vol. 223, p. 117293, Dec. 2020, doi: 10.1016/j.neuroimage.2020.117293.

[7] P. Mouches, M. Wilms, D. Rajashekar, S. Langner, and N. D. Forkert, "Multimodal biological brain age prediction using magnetic resonance imaging and angiography with the identification of predictive regions," Hum. Brain Mapp., vol. n/a, no. n/a, Epub ahead of print. doi: 10.1002/hbm.25805.

[8] L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, I. Y. Chen, and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," Nat. Med., vol. 27, no. 12, Art. no. 12, Dec. 2021, doi: 10.1038/s41591-021-01595-0.

[9] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," Proc. Natl. Acad. Sci., vol. 117, no. 23, pp. 12592–12594, Jun. 2020, doi: 10.1073/pnas.1919012117.

[10] E. Puyol-Antón et al., "Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Cham, 2021, pp. 413–423. doi: 10.1007/978-3-030-87199-4_39.

[11] W. J. Hall et al., "Implicit Racial/Ethnic Bias Among Health Care Professionals and Its Influence on Health Care Outcomes: A Systematic Review," Am. J. Public Health, vol. 105, no. 12, pp. e60–e76, Dec. 2015, doi: 10.2105/AJPH.2015.302903.

[12] R. M. Mayberry, F. Mili, and E. Ofili, "Racial and Ethnic Differences in Access to Medical Care," Med. Care Res. Rev., vol. 57, no. 1_suppl, pp. 108–145, Nov. 2000, doi: 10.1177/1077558700057001S06.

[13] A. R. Tabaac, A. L. Solazzo, A. R. Gordon, S. B. Austin, C. Guss, and B. M. Charlton, "Sexual orientation-related disparities in healthcare access in three cohorts of U.S. adults," Prev. Med., vol. 132, p. 105999, Mar. 2020, doi: 10.1016/j.ypmed.2020.105999.

[14] I. Palatnik de Sousa, M. M. B. R. Vellasco, and E. Costa da Silva, "Explainable Artificial Intelligence for Bias Detection in COVID CT-Scan Classifiers," Sensors, vol. 21, no. 16, p. 5657, Aug. 2021, doi: 10.3390/s21165657.

[15] R. K. Lenroot and J. N. Giedd, "Sex differences in the adolescent brain," Brain Cogn., vol. 72, no. 1, pp. 46–55, Feb. 2010, doi: 10.1016/j.bandc.2009.10.008.

[16] J. M. Goldstein et al., "Normal Sexual Dimorphism of the Adult Human Brain Assessed by In Vivo Magnetic Resonance Imaging," Cereb. Cortex, vol. 11, no. 6, pp. 490–497, Jun. 2001, doi: 10.1093/cercor/11.6.490.

[17] L. M. Wierenga et al., "Unraveling age, puberty and testosterone effects on subcortical brain development across adolescence," Psychoneuroendocrinology, vol. 91, pp. 105–114, May 2018, doi: 10.1016/j.psyneuen.2018.02.034.

[18] A.-L. Goddings, K. L. Mills, L. S. Clasen, J. N. Giedd, R. M. Viner, and S.-J. Blakemore, "The influence of puberty on subcortical brain development," NeuroImage, vol. 88, pp. 242–251, Mar. 2014, doi: 10.1016/j.neuroimage.2013.09.073.

[19] J. E. Bramen et al., "Puberty Influences Medial Temporal Lobe and Cortical Gray Matter Maturation Differently in Boys Than Girls Matched for Sexual Maturity," Cereb. Cortex, vol. 21, no. 3, pp. 636–646, Jan. 2011, doi: 10.1093/cercor/bhq137.

[20] S. Dalsgaard et al., "Incidence Rates and Cumulative Incidences of the Full Spectrum of Diagnosed Mental Disorders in Childhood and Adolescence," JAMA Psychiatry, vol. 77, no. 2, pp. 155–164, Feb. 2020, doi: 10.1001/jamapsychiatry.2019.3523.

[21] W. C. Chumlea et al., "Age at Menarche and Racial Comparisons in US Girls," Pediatrics, vol. 111, no. 1, pp. 110–113, Jan. 2003, doi: 10.1542/peds.111.1.110.

[22] M. E. Herman-Giddens et al., "Secondary sexual characteristics in boys: data from the Pediatric Research in Office Settings Network.," Pediatrics, vol. 130, no. 5, pp. e1058-1068, Nov. 2012, doi: 10.1542/peds.2011-3291.

[23] T. Wu, P. Mendola, and G. M. Buck, "Ethnic differences in the presence of secondary sex characteristics and menarche among US girls: the Third National Health and Nutrition Examination Survey, 1988-1994," Pediatrics, vol. 110, no. 4, pp. 752–757, Oct. 2002, doi: 10.1542/peds.110.4.752.

[24] H. Garavan et al., "Recruiting the ABCD sample: Design considerations and procedures," Dev. Cogn. Neurosci., vol. 32, pp. 16–22, Aug. 2018, doi: 10.1016/j.dcn.2018.04.004.

[25] M. E. Koopman-Verhoeff, C. Gredvig-Ardito, D. H. Barker, J. M. Saletin, and M. A. Carskadon, "Classifying Pubertal Development Using Child and Parent Report: Comparing the Pubertal Development Scales to Tanner Staging," J. Adolesc. Health, vol. 66, no. 5, pp. 597–602, May 2020, doi: 10.1016/j.jadohealth.2019.11.308.

[26] T. W. Cheng et al., "A Researcher's Guide to the Measurement and Modeling of Puberty in the ABCD Study® at Baseline," Front. Endocrinol., vol. 12, p. 608575, 2021.

[27] D. M. Barch et al., "Demographic, physical and mental health assessments in the adolescent brain and cognitive development study: Rationale and description," Dev. Cogn. Neurosci., vol. 32, pp. 55–66, Nov. 2017, doi: 10.1016/j.dcn.2017.10.010.
[28] B. J. Casey et al., "The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites," Dev. Cogn. Neurosci., vol. 32, pp. 43–54, Aug. 2018, doi: 10.1016/j.dcn.2018.03.001.
[29] D. J. Hagler et al., "Image processing and analysis methods for the Adolescent Brain Cognitive Development Study," NeuroImage, vol. 202, p. 116091, Nov. 2019, doi: 10.1016/j.neuroimage.2019.116091.
[30] V. Fonov, A. C. Evans, K. Botteron, C. R. Almli, R. C. McKinstry, and D. L. Collins, "Unbiased average age-appropriate atlases for pediatric studies," NeuroImage, vol. 54, no. 1, pp. 313–327, Jan. 2011, doi: 10.1016/j.neuroimage.2010.07.033.
[31] "Advanced Normalization Tools: V1.0," Jul. 2009, doi: 10.54294/uvnhin.
[32] H. Peng, W. Gong, C. F. Beckmann, A. Vedaldi, and S. M. Smith, "Accurate brain age prediction with lightweight deep neural networks," Med. Image Anal., vol. 68, p. 101871, Feb. 2021, doi: 10.1016/j.media.2020.101871.
[33] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: removing noise by adding noise," ArXiv170603825 Cs Stat, Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.03825
[34] keisen/tf-keras-vis: Neural network visualization toolkit for tf.keras. [Online]. Available: https://github.com/keisen/tf-keras-vis
[35] A. L. Manera, M. Dadar, V. Fonov, and D. L. Collins, "CerebrA, registration and manual label correction of Mindboggle-101 atlas for MNI-ICBM152 template," Sci. Data, vol. 7, no. 1, p. 237, Jul. 2020, doi: 10.1038/s41597-020-0557-9.
[36] N. Vijayakumar, Z. Op de Macks, E. A. Shirtcliff, and J. H. Pfeifer, "Puberty and the human brain: Insights into adolescent development," Neurosci. Biobehav. Rev., vol. 92, pp. 417–436, Sep. 2018, doi: 10.1016/j.neubiorev.2018.06.004.

Appendix: Supplementary Materials

Table 4 Demographic representation in each fold of cross validation.

Train							
Fold	**All**	**Male**	**Female**	**White**	**Black**	**White Male**	**Black Male**
1	3273	1643	1630	2165	425	1083	217
2	3273	1642	1631	2165	425	1083	217
3	3274	1643	1631	2166	425	1084	217
4	3274	1643	1631	2166	425	1084	217
5	3274	1642	1632	2165	425	1084	216

Validation							
Fold	**All**	**Male**	**Female**	**White**	**Black**	**White Male**	**Black Male**
1	364	183	181	241	47	121	24
2	364	183	181	241	47	121	24
3	364	183	181	241	47	121	24
4	364	183	181	241	47	121	24
5	364	183	181	241	47	121	24

Test							
Fold	**All**	**Male**	**Female**	**White**	**Black**	**White Male**	**Black Male**
1	910	456	454	602	118	301	60
2	910	457	453	602	118	302	60
3	909	456	453	601	118	301	60
4	909	456	453	601	118	301	60
5	909	457	452	602	118	301	61

Table 5 Unweighted saliency scores in each brain region.

Region	**Hemisphere**	**Aggregate**	**White males**	**White females**	**Black males**	**Black females**
Fourth Ventricle	Right	99.718	99.718	99.718	99.436	97.743
Fourth Ventricle	Left	99.650	99.475	99.650	98.599	99.299
Cerebellum White Matter	Right	83.921	68.639	88.444	64.534	87.027
Cerebellum White Matter	Left	83.866	67.725	78.640	56.838	67.603
Vermal lobules VIII-X	Right	51.293	42.306	63.931	21.050	44.850
Amygdala	Right	40.663	36.486	40.541	38.575	38.882
Brainstem	Right	40.125	40.450	42.176	36.278	34.297
Brainstem	Left	38.503	40.588	38.209	32.403	25.552
Fusiform	Left	33.091	22.720	29.309	27.369	25.707

Inferior Lateral Ventricle	Right	28.822	23.140	32.748	27.583	31.612
Cerebellum Gray Matter	Left	28.586	26.745	26.173	25.735	24.515
Vermal lobules VIII-X	Left	27.031	20.713	42.870	19.269	30.099
Inferior Lateral Ventricle	Left	25.750	22.168	23.621	21.781	17.522
Amygdala	Left	23.573	17.742	12.345	15.509	3.102
Inferior temporal	Left	15.995	12.676	9.053	15.336	5.740
Middle Temporal	Left	11.772	13.505	4.129	13.094	1.262
Entorhinal	Left	11.259	13.440	5.918	11.158	3.686
Vermal lobules I-V	Left	10.920	7.732	12.644	4.441	5.904
Superior Temporal	Left	10.604	14.754	2.051	13.497	0.791
Entorhinal	Right	10.424	11.289	7.858	10.424	6.705
Vermal lobules I-V	Right	10.354	7.541	16.756	4.292	7.177
Vermal lobules VI-VII	Right	10.198	9.049	18.512	3.996	2.710
Parahippocampal	Left	9.192	9.479	6.427	3.483	0.431
Cerebellum Gray Matter	Right	9.085	5.228	12.830	3.377	11.061
Ventral Diencephalon	Left	7.961	13.987	5.941	4.227	1.782
Vermal lobules VI-VII	Left	6.814	7.366	11.541	1.780	0.061
Superior Temporal	Right	5.257	10.215	3.167	5.739	2.945
Hippocampus	Left	4.175	2.835	2.747	2.966	0.659
Inferior temporal	Right	3.927	6.499	2.448	3.949	0.667
Hippocampus	Right	3.788	1.792	4.854	3.810	6.373
Rostral Middle Frontal	Left	2.852	6.110	1.341	4.392	0.073
Fusiform	Right	2.649	2.137	2.765	1.775	1.290

Lateral Occipital	Left	2.634	0.372	4.536	0.237	3.539
Lateral Orbitofrontal	Right	1.955	3.214	0.410	1.528	0.247
Parahippocampal	Right	1.872	0.183	4.073	0.110	3.156
Ventral Diencephalon	Right	1.131	1.542	0.600	0.120	0.000
Pars Triangularis	Left	1.117	4.382	0.181	3.393	0.000
Inferior Parietal	Left	0.820	0.432	1.233	0.176	0.943
Middle Temporal	Right	0.395	1.606	0.024	0.349	0.021
Insula	Left	0.241	0.000	0.050	0.000	0.000
Insula	Right	0.160	0.809	0.011	0.106	0.000
Lateral Orbitofrontal	Left	0.062	0.912	0.000	0.400	0.000
Pars Triangularis	Right	0.000	0.157	0.000	0.000	0.000
Pericalcarine	Left	0.000	0.020	0.000	0.000	0.000
Rostral Middle Frontal	Right	0.000	0.205	0.000	0.000	0.000
Superior Frontal	Left	0.000	0.003	0.000	0.003	0.002
Optic Chiasm	Right	0.000	0.118	0.000	0.000	0.000
Lateral Ventricle	Left	0.000	0.509	0.873	0.000	0.058
Thalamus	Left	0.000	0.000	0.021	0.000	0.000

Table 6 Weighted saliency scores in each brain region.

Region	Hemisphere	Aggregate	White males	White females	Black males	Black females
Fourth Ventricle	Right	99.718	99.718	99.718	99.436	97.743
Fourth Ventricle	Left	65.435	60.653	65.988	55.920	66.116
Cerebellum White Matter	Left	22.708	16.341	20.254	17.189	15.653
Cerebellum White Matter	Right	20.705	12.253	28.535	12.045	24.532
Brainstem	Right	16.234	15.665	16.584	15.829	13.889
Brainstem	Left	12.378	13.731	11.364	9.358	6.622
Amygdala	Right	12.021	8.254	12.615	12.270	9.920
Inferior Lateral Ventricle	Left	10.373	7.021	6.963	7.269	2.047
Vermal lobules VIII-X	Right	9.490	8.519	15.425	5.826	10.716
Inferior Lateral Ventricle	Right	9.427	5.412	13.033	10.742	9.182
Cerebellum Gray Matter	Left	8.099	7.585	6.281	7.920	5.421
Fusiform	Left	6.552	3.274	4.567	5.634	4.201
Amygdala	Left	5.926	2.992	1.858	2.669	0.138
Vermal lobules VIII-X	Left	4.460	3.896	7.329	3.106	5.909
Entorhinal	Left	3.623	3.981	1.183	3.057	0.383
Inferior temporal	Left	2.500	2.556	1.155	2.715	0.513
Vermal lobules I-V	Right	2.289	2.230	3.209	1.380	1.923
Vermal lobules I-V	Left	1.925	1.603	2.349	0.888	1.308
Middle Temporal	Left	1.772	3.026	0.325	3.001	0.060
Entorhinal	Right	1.478	1.657	1.289	1.786	0.998
Superior Temporal	Right	1.314	2.431	0.620	1.635	0.588

Superior Temporal	Left	1.141	2.382	0.098	2.232	0.029
Ventral Diencephalon	Left	1.140	2.092	0.864	0.760	0.165
Cerebellum Gray Matter	Right	1.067	0.578	2.142	0.370	1.497
Parahippocampal	Left	0.860	1.207	0.616	0.210	0.011
Hippocampus	Left	0.595	0.310	0.314	0.422	0.031
Vermal lobules VI-VII	Right	0.533	0.312	1.748	0.133	0.163
Inferior temporal	Right	0.407	0.928	0.366	0.368	0.035
Hippocampus	Right	0.399	0.108	0.820	0.562	0.743
Lateral Orbitofrontal	Right	0.260	0.685	0.023	0.249	0.015
Fusiform	Right	0.260	0.166	0.403	0.166	0.098
Rostral Middle Frontal	Left	0.245	1.307	0.096	0.802	0.002
Lateral Occipital	Left	0.125	0.014	0.739	0.007	0.454
Ventral Diencephalon	Right	0.083	0.140	0.031	0.000	0.000
Inferior Parietal	Left	0.046	0.028	0.213	0.004	0.150
Pars Triangularis	Left	0.022	0.813	0.007	0.552	0.000
Parahippocampal	Right	0.015	0.003	0.232	0.001	0.086
Middle Temporal	Right	0.012	0.169	0.001	0.018	0.001
Lateral Orbitofrontal	Left	0.001	0.028	0.000	0.014	0.000
Insula	Left	0.000	0.000	0.001	0.000	0.000
Insula	Right	0.000	0.070	0.000	0.006	0.000
Vermal lobules VI-VII	Left	0.000	0.183	0.569	0.023	0.003
Lateral Ventricle	Left	0.000	0.009	0.034	0.000	0.001
Optic Chiasm	Right	0.000	0.002	0.000	0.000	0.000