SEX DIFFERENCES IN THE BRAIN: DIVERGENT RESULTS FROM TRADITIONAL MACHINE LEARNING AND CONVOLUTIONAL NETWORKS

Leo Brueggeman, Taylor R. Thomas, Tanner Koomar, Brady Hoskins, Jacob J. Michaelson

Department of Psychiatry, University of Iowa, Iowa City IA

ABSTRACT

Neuroimaging research has begun adopting deep learning to model structural differences in the brain. This is a break from previous approaches that rely largely on anatomical volumetric or thickness-based features. Currently, most studies employ either convolutional deep learning based models or traditional machine learning models that use volumetric features. Because of this split, it is unclear which approach yields better predictive performance, or whether the two approaches will lead to different neuroanatomical conclusions, potentially even when applied to the same dataset. To address these questions, we present the largest single study of sex differences in the brain using 21,390 UK Biobank T1-weighted brain MRIs, which we analyzed through both traditional volumetric and 3D convolutional neural network models. Overall, we find that 3D-CNNs outperformed traditional machine learning models, with sex classification area under the ROC curve of 0.849 and 0.683, respectively. When performing sex classification using only single regions of the brain, we observed better performance from 3D-CNNs in all regions tested, indicating sex differences in the brain likely represent both structural and volumetric changes. In addition, we find little consensus in terms of brain region prioritization between the two approaches. In summary, we find that 3D-CNNs show exceptional sex classification performance, extract additional relevant structural information from brain regions beyond volume, and possibly because of this, prioritize sex differences in neuroanatomical regions differently than volume-based approaches.

Index Terms— convolutional neural networks, brain MRI, sex differences

1. INTRODUCTION

Characterizing sex differences in the human brain is relevant because it has potential to yield insight into various neuropsychiatric conditions that have strong sex biases, like eating disorders with a 8:1 female bias [1] and autism with a 4:1 male bias [2]. However, studies looking at sex differences in the human brain are usually low powered and also use derived brain measures and statistical methods that put inherent assumptions and parameters on the model. These traditional methods have had some success in detecting sex differences, especially when the study is well-powered [3]. However, the brain is complex, and these derived brain features and basic statistical tests are likely missing important aspects of sexual dimorphisms in the brain.

Deep learning has had some success in predicting sex from human brain features. A convolutional neural network (CNN) was applied to predict sex using scalp electroencephalograms and performed with an accuracy of 81% [4]. However, using deep learning to predict sex from whole brain structural MRI has yet to be done. Therefore, we developed a convolutional neural network to predict sex using T1-weighted brain MRI scans from 21,390 adults. To our knowledge, our study is the first to predict sex using a deep learning model from 3D structural brain MRIs. As a comparison, we used the same individuals' derived brain volume features as input into traditional machine learning (ML) methods and assessed accuracy across the different models (see Figure 1). Lastly, we prioritized brain regions based on detected sex differences using ML methods compared to conclusions from CNN based approaches. Ultimately, we propose a 3D-CNN architecture for accurate sex difference modeling in the brain, and highlight differences in anatomical conclusions derived from ML and 3D-CNN models.

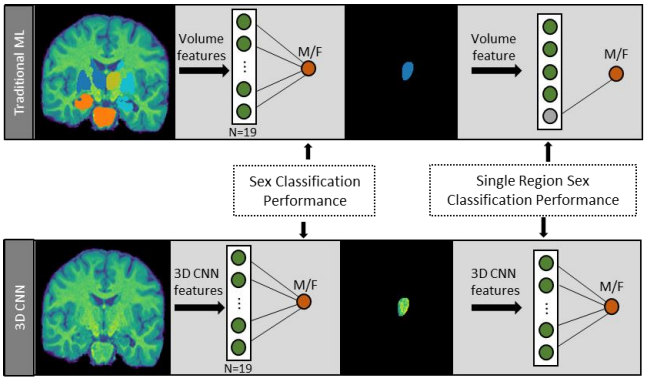

Fig. 1. Study design overview showing the performance and region prioritization comparisons made between traditional machine learning and 3D-CNN models.

2. METHODS

2.1. UK Biobank data

T1 MRI data from 21,390 participants in the UK Biobank study (release 3) was used, comprised of imaging and image-derived phenotypes. The preprocessing of UK Biobank MRI data has been previously described [5]. Briefly, the preprocessing steps performed on the images used for deep learning in this study include defacing, gradient distortion correction, and brain extraction. Brain regions used in this study were extracted as described in the UK Biobank protocol [5], and include cerebrospinal fluid (CSF), grey matter, white matter, and 15 subcortical regions including the brainstem, and the (left and right) amygdala, putamen, hippocampus, accumbens, pallidum, thalamus and caudate [6]. Total brain volume was also used as a feature in all ML models. Participant reported sex was also extracted from the UK Biobank for phenotypic modeling. To control for possible confounding, all model predictions were corrected for the date of imaging, the imaging center, subject BMI, weight, age, T1 signal to noise ratio, T1 contrast to noise ratio, scanner x,y,z position, and position on the scanning table. Specifically, model predictions were modeled using the above covariates in a linear model, and the residuals from this model were taken as the corrected predictions.

2.2. Traditional statistical models

The 19 brain region volumes (including subcortical regions, grey matter, white matter, CSF, and total brain) were obtained from the UK Biobank. Subcortical regions were processed by UK Biobank researchers using FIRST [7]. We then took each of the 19 brain region volumes and z-scaled the data by first subtracting the mean of each region and then dividing by the standard deviation. The data was then split into a train/test split of N = 19,000 training and N = 2,390 testing. Logistic regression, random forest, regularized linear discriminant analysis, and a naive bayes classifier were performed in R using the `caret` package. We used 5-fold repeated cross validation and a tune length of 10. Model accuracy was assessed in the test set by area under the ROC curve (AUROC).

2.3. 3D Convolutional neural network model

A 3D-CNN architecture was used which is similar to one previously published to detect brain age [8]. This architecture consisted of 3D convolution (Conv3D) using the Rectified Linear Unit (relu) activation function, max pooling (MaxPool3D) and batch normalization (BatchNorm) blocks. The architecture of this model is shown in Table 1 (f = filters; k = kernel size). Prior to prediction, the model passed through a 19-unit dense layer, mirroring the feature-space size of the ML models. A learning rate of 1×10^{-5} was used, which led to stable results while training. The model was fit on 15019 samples, with 4,005 validation samples, and 2366 samples held out for testing (same splits as traditional ML models). The CNN was fit for 50 epochs, with an early stopping callback with a patience of 5.

Layer type	Options	# Params
Conv3D	f=8, k=(3,3,3), relu	224
MaxPool3D	pool=(1,2,2), stride=(1,2,2)	0
BatchNorm		32
Conv3D	f=16, k=(3,3,3), relu	3472
MaxPool3D	pool=(1,2,2), stride=(1,2,2)	0
BatchNorm		64
Conv3D	f=32, k=(3,3,3), relu	13856
Conv3D	f=64, k=(3,3,3), relu	55360
MaxPool3D	pool=(1,2,2), stride=(1,2,2)	0
BatchNorm		256
Conv3D	f=128, k=(3,3,3), relu	65664
Conv3D	f=256, k=(3,3,3), relu	262400
MaxPool3D	pool=(1,2,2), stride=(1,2,2)	0
Flatten		0
Dense	nodes=19, linear	14348819
Dense	nodes=1, sigmoid	20
	Total params = 14,750,167	

Table 1: Layer representation of 3D-CNN model architecture.

2.4. Region prioritization between models

To identify brain regions of interest (ROI) from both ML and deep learning models, the parcellations based on tissue type and subcortical regions of the MNI structural atlas [6] were used. For the volume based model, volumes from these 18 different regions were corrected for the imaging covariates mentioned above, and then used as predictors for sex, separately. All AUROC values were > 0.5, indicating larger volumes of all regions were positively correlated with maleness. To individually model region-specific AUROC with the 3D-CNN, for each ROI, all voxel values outside of the ROI were set to 0. The training and testing process was repeated as described above for each region (total of 18 region-specific 3D-CNN models trained). Individual predictions were corrected for the imaging covariates and then used as predictors for sex.

3. RESULTS

3.1. Model performance comparisons

The goal of all models was to predict reported sex values of individuals using either 19 atlas-based brain region volumes or 19 3D-CNN derived brain features. A logistic regression model using only total brain volume (base model) as a predictor was included for comparison. All model predictions were corrected for imaging and subject covariates, as described in

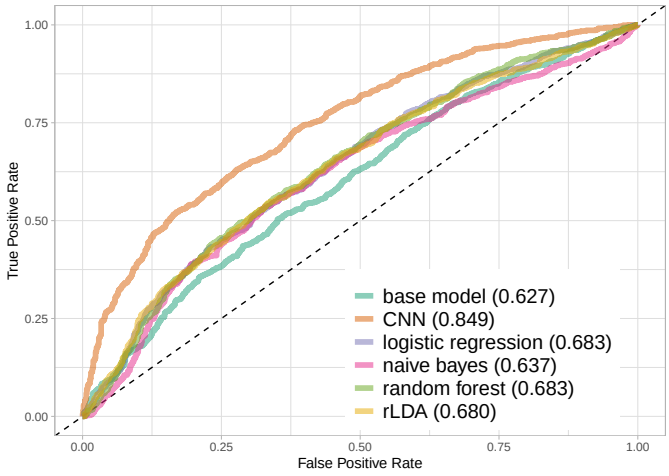

Fig. 2. Test set area under the ROC curve values of sex prediction from total brain volume (base model), traditional machine learning models, and the 3D-CNN model.

the methods. All tested ML methods showed similar performance. As shown in Figure 2, the naive bayes model had the lowest performance with a AUROC of 0.637 in the test set. In comparison, the full logistic regression and random forest models showed higher performance, both achieving an AUROC of 0.683 in the test set. Finally, the 3D-CNN model showed a significant jump in performance, with a test set AUROC of 0.849.

3.2. Sex differences in brain regions

Models were next trained on different regions of the brain to rank brain tissue types and regions by their effectiveness in predicting sex. As shown in Figure 3, the regions studied include tissue types (3.A.) and subcortical regions (3.B.).

Across all ROIs tested, the structure-based 3D-CNN predictions (3.D.) outperformed the volume-based (3.C.) predictions of sex. This difference was highly significant, even when restricted to the overall lower-performing subcortical regions (Wilcoxon rank sum test p-value = 1.2×10^{-8}). The structure-based model most predictive of sex status was the cerebrospinal fluid (CSF) tissue type (3.D.) model, with an AUROC of 0.832. The volume based model most predictive of sex status was the white matter (WM) tissue type (3.C.) model, with an AUROC of 0.623.

In addition to a large difference in magnitude between structure and volume-based sex classification AUROCs, the order of ROI informativeness was also different between these two methods. Specifically, ROI prioritization ranks by AUROC did not significantly correlate between the structure and volume-based sex classifiers (Spearman's rho = 0.1; p-value = 0.69).

4. CONCLUSION

In this work we performed the largest study of sex differences (UK Biobank, N = 21,390) in the brain using both ML and deep learning based models. After training these models to predict sex, we compared their performances and the regions of the brain which they used for their predictions.

As shown in Figure 2, the 3D-CNN method is able to derive a set of 19 brain features which are more informative for predicting sex than the atlas-based parcellations. While it is likely true that the ML methods would have higher performance with a greater number of ROIs, the difference to be made up is large, with 3D-CNNs showing a test set AUROC of 0.849, and the nearest ML model (logistic regression and random forests) showing a test set AUROC of 0.683.

When comparing the regions highlighted by both approaches, shown in Figure 3, we find that the 3D-CNN was significantly more predictive of sex than volume based models, even when restricting to individual regions of the brain. This likely indicates that sexual differences in the brain are both volumetric and structural in nature. In comparing the regions with the greatest sex differences, we find that the volume based method highly ranked the overall white matter volume (AUROC = 0.623), whereas the 3D-CNN highly ranked the CSF segmentation (AUROC = 0.832). Limiting to subcortical structures, the volume based method prioritized the right putamen (AUROC = 0.599) while the 3D-CNN highly ranked the right thalamus (AUROC = 0.660). For the 3D-CNN region rankings, the subcortical structures had a much lower AUROC than the tissue types. This could simply be due to the difference in the number of voxels available when modeling, but could also possibly indicate that sex differences in the brain are best captured through global features of the brain. Interestingly, we find that region rankings by the structure-based and volume based predictions do not correlate. This finding suggests that previously reported brain region phenotype associations may not replicate in newer studies relying on deep learning 3D-CNN based models.

In conclusion, we find that 3D-CNNs are highly capable of modeling sex differences in the brain, and that the regions prioritized by them are significantly different than regions prioritized by traditional volume-based machine learning models. All code for our analysis can be found at https://github.com/LeoBman/brain-sex-classification.

4.1. Ethics

The University of Iowa's Institutional Review Board determined this analysis not human subjects research (IRB 202001107).

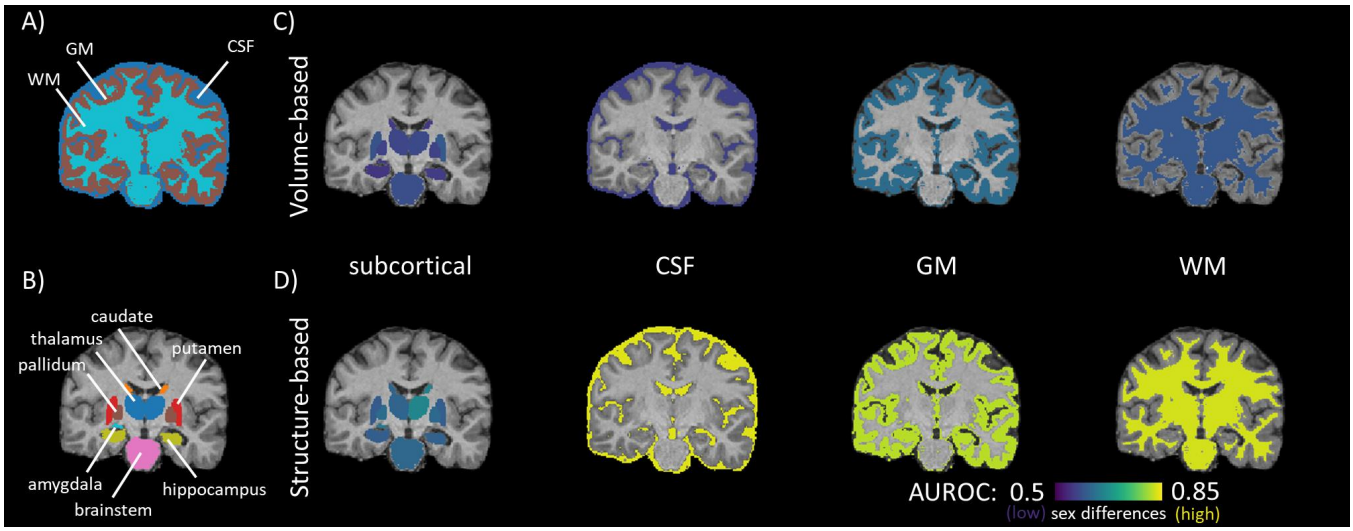

Fig. 3. Structure-based features from 3D-CNN outperform volume-only based features from the same brain regions in discriminating sex. Example tissue types (A) and subcortical regions (B). Area under the ROC curve values are shown for each regions' sex classification using either ROI volumes (C) or MRI structures (D) across subcortical, cerebrospinal fluid (CSF), grey matter (GM) and white matter (WM). structures.

5. REFERENCES

[1] Hans-Christoph Steinhausen and Christina Mohr Jensen, "Time trends in lifetime incidence rates of first-time diagnosed anorexia nervosa and bulimia nervosa across 16 years in a danish nationwide psychiatric registry study," *International Journal of Eating Disorders*, vol. 48, no. 7, pp. 845–850, Mar. 2015.

[2] Eric Fombonne, "Epidemiology of pervasive developmental disorders," *Pediatric Research*, vol. 65, no. 6, pp. 591–598, June 2009.

[3] Stuart J Ritchie, Simon R Cox, Xueyi Shen, Michael V Lombardo, Lianne M Reus, Clara Alloza, Mathew A Harris, Helen L Alderson, Stuart Hunter, Emma Neilson, David C M Liewald, Bonnie Auyeung, Heather C Whalley, Stephen M Lawrie, Catharine R Gale, Mark E Bastin, Andrew M McIntosh, and Ian J Deary, "Sex differences in the adult human brain: Evidence from 5216 UK biobank participants," *Cerebral Cortex*, vol. 28, no. 8, pp. 2959–2975, May 2018.

[4] Michel J. A. M. van Putten, Sebastian Olbrich, and Martijn Arns, "Predicting sex from brain rhythms with deep learning," *Scientific Reports*, vol. 8, no. 1, Feb. 2018.

[5] Fidel Alfaro-Almagro, Mark Jenkinson, Neal K. Bangerter, Jesper L.R. Andersson, Ludovica Griffanti, Gwenalle Douaud, Stamatios N. Sotiropoulos, Saad Jbabdi, Moises Hernandez-Fernandez, Emmanuel Vallee, Diego Vidaurre, Matthew Webster, Paul McCarthy, Christopher Rorden, Alessandro Daducci, Daniel C. Alexander, Hui Zhang, Iulius Dragonu, Paul M. Matthews, Karla L. Miller, and Stephen M. Smith, "Image processing and quality control for the first 10, 000 brain imaging datasets from UK biobank," *NeuroImage*, vol. 166, pp. 400–424, Feb. 2018.

[6] John Mazziotta, Arthur Toga, Alan Evans, Peter Fox, Jack Lancaster, Karl Zilles, Roger Woods, Tomas Paus, Gregory Simpson, Bruce Pike, Colin Holmes, Louis Collins, Paul Thompson, David MacDonald, Marco Iacoboni, Thorsten Schormann, Katrin Amunts, Nicola Palomero-Gallagher, Stefan Geyer, Larry Parsons, Katherine Narr, Noor Kabani, Georges Le Goualher, Dorret Boomsma, Tyrone Cannon, Ryuta Kawashima, and Bernard Mazoyer, "A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (ICBM)," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 356, no. 1412, pp. 1293–1322, Aug. 2001.

[7] Brian Patenaude, Stephen M. Smith, David N. Kennedy, and Mark Jenkinson, "A bayesian model of shape and appearance for subcortical brain segmentation," *NeuroImage*, vol. 56, no. 3, pp. 907–922, June 2011.

[8] James H. Cole, Rudra P. K. Poudel, Dimosthenis Tsagkrasoulis, Matthan W. A. Caan, Claire J Steves, Tim D. Spector, and Giovanni Montana, "Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker," *NeuroImage*, vol. 163, pp. 115–124, 2016.