Investigating Demographic Bias in Brain MRI Segmentation: A Comparative Study of Deep-Learning and Non-Deep-Learning Methods

Ghazal Danaee , Marc Niethammer , Jarrett Rushmore , Sylvain Bouix

Motivation

- Intrinsic biases in data when training→

  Biased models may have performance disparities based on sensitive attributes like race and sex

- Few studies done on evaluating the bias in the segmentation tasks

Introduction

Goal:

Evaluate the results of UNesT, nnU-Net, CoTr and a traditional atlas-based method (ANTs), segmenting the left and right nucleus accumbens (NAc) in MRI images

1- Segmentation performance of models

2-Volumes of the segmented structures to evaluate the effects of race, sex, and their interaction

Dataset

- T1-weighted MRIs from Human Connectome Project (HCP) Young Adult

Group	Training Images	Testing Images
Black Female	30	19
Black Male	32	20
White Female	33	19
White Male	31	20

- Groundtruth : Manually labeled gold-standard segmentations

- Why nucleus accumbens?

Biased training

- Trained each deep learning model (UNesT, etc) from scratch using only one of the 4 demographic groups (Black male, etc)
- For ANTs, 10 atlases from only one of the 4 demographic groups

Segmentation models

- We trained the default architecture of models with some modifications to hyperparameters(loss function, etc)

UNesT(Yu 2023)	Hierarchical Transformer encoder + conv decoder
nn-Unet(Isensee 2021)	3D U-Net
CoTr(Xie 2021)	CNN encoder + Deformable Transformer
ANTs(Advanced normalization tools)(Wang 2013)	Multi-atlas segmentation with joint label fusion

Evaluation Metrics

- To evaluate accuracy →
  - Dice similarity coefficient(DSC)

$$\text{DSC}(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}$$

- To evaluate Fairness and accuracy →
  - ESSP (Equity-scaled segmentation performance) (Tian et al. 2024)

$$\Delta = \sum_{a \in A} \left| \text{DSC}_{overall} - \text{DSC}_a \right|.$$

$$\text{ESSP} = \frac{\text{DSC}_{overall}}{1 + \Delta}.$$

Statistical analysis

- **Performance Bias** : Employed linear mixed models to assess bias in model performance

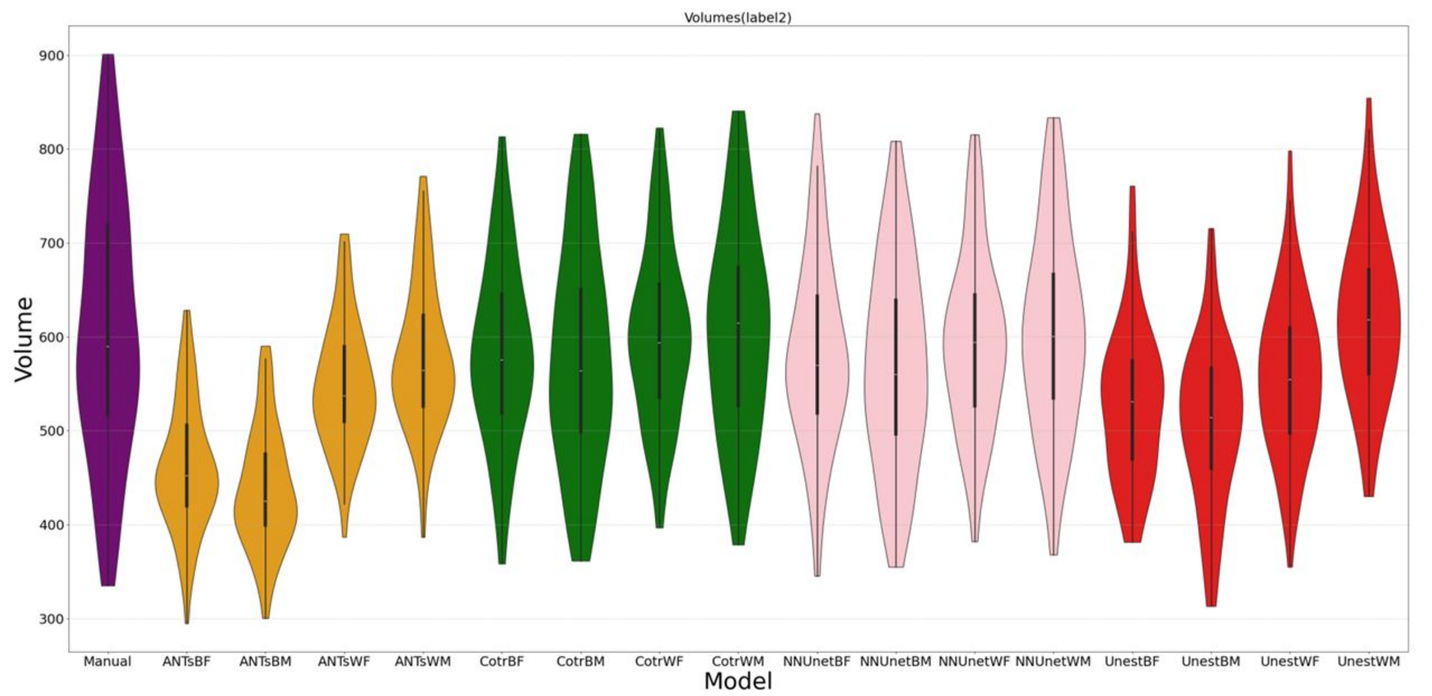
$$\text{DSC} = \beta_0 + \beta_1(\text{SameRace}) + \beta_2(\text{SameSex}) + \beta_3(\text{SameRace} \times \text{SameSex}) + \epsilon$$

- **Effect of bias on demographic analyses:**
- To investigate how race, sex, and their interaction influenced the volumes in segmentations, we used:

$$\text{Volume} = \beta_0 + \beta_1(\text{Race}) + \beta_2(\text{Sex}) + \beta_3(\text{Race} \times \text{Sex}) + \epsilon$$

Results

1. General statistics of the volumes



Results

1. **General statistics of the volumes**

-smaller standard deviation in non-manual segmentations

- right NAc volume>> left NAc volume

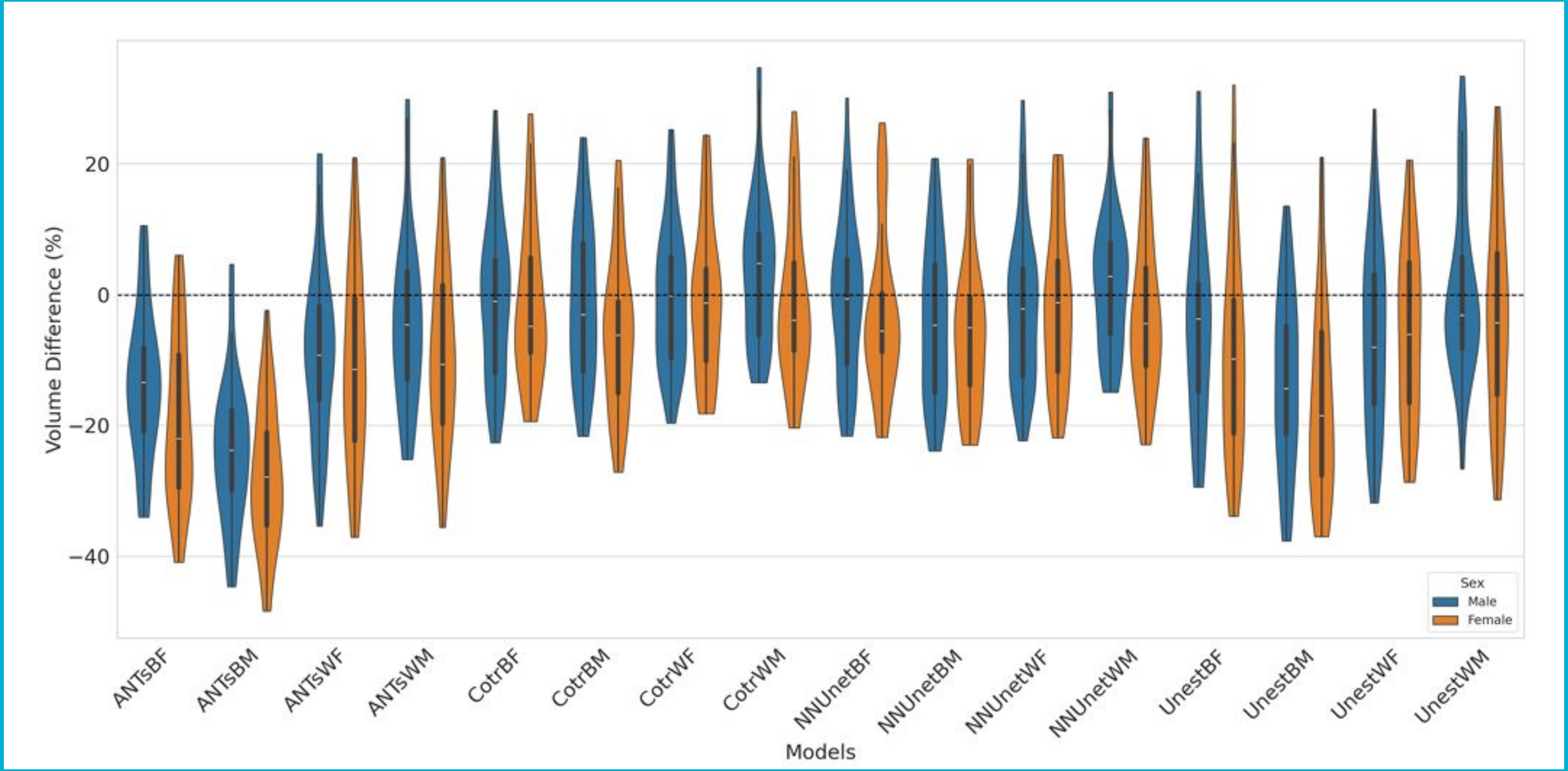
(in both manual and non manual)

-ANTsBM and UnestBM have volumes

Almost 20% smaller than the manual segmentations

Model	Right NAc		Left NAc	
	Mean	Std	Mean	Std
Manual	676.97	125.79	607.13	136.13
NNUnetBF	653.62	95.08	581.06	99.14
NNUnetBM	638.20	115.41	569.97	108.32
NNUnetWF	653.14	90.91	593.83	93.10
NNUnetWM	665.30	108.89	604.87	106.29
CotrBF	658.21	93.08	582.45	96.79
CotrBM	647.76	119.47	574.53	114.28
CotrWF	664.07	96.01	600.97	92.19
CotrWM	677.96	111.76	606.37	109.90
ANTsBF	552.27	68.63	460.58	67.22
ANTsBM	491.58	61.68	437.41	63.45
ANTsWF	595.83	70.31	548.00	66.71
ANTsWM	618.45	78.61	577.38	76.94
UnestBF	614.88	78.02	528.29	80.83
UnestBM	564.65	86.91	507.03	85.14
UnestWF	623.99	84.35	558.03	83.48
UnestWM	655.31	94.40	618.32	87.33

- Under-segmentation in ANTsBF and ANTsBM



Results

2. Bias in segmentation performance:

    nnU-net and CoTr

    ANTs and UNesT

    ANTs ESSP drop when trained on Black

Structure	TrainGp	nnU-Net			CoTr			ANTs			UNesT		
		DSC	ESSP	Δ	DSC	ESSP	Δ	DSC	ESSP	Δ	DSC	ESSP	Δ
Right NAc	WM	0.867	0.845	0.027	0.863	0.839	0.029	0.820	0.796	0.030	0.832	0.784	0.060
	WF	0.862	0.838	0.028	0.859	0.832	0.032	0.816	0.793	0.029	0.817	0.791	0.032
	BM	0.862	0.836	0.032	0.859	0.834	0.029	0.781	0.702	0.113	0.801	0.759	0.050
	BF	0.862	0.841	0.025	0.858	0.836	0.027	0.792	0.720	0.100	0.809	0.780	0.037
Left NAc	WM	0.861	0.849	0.013	0.856	0.843	0.016	0.810	0.794	0.021	0.825	0.773	0.066
	WF	0.858	0.836	0.026	0.856	0.839	0.020	0.806	0.796	0.012	0.810	0.787	0.029
	BM	0.854	0.832	0.026	0.851	0.831	0.024	0.758	0.688	0.102	0.800	0.748	0.070
	BF	0.858	0.840	0.022	0.853	0.829	0.029	0.773	0.700	0.102	0.798	0.766	0.041

Results

2. Bias in segmentation performance:

$$\text{DSC} = \beta_0 + \beta_1(\text{SameRace}) + \beta_2(\text{SameSex}) + \beta_3(\text{SameRace} \times \text{SameSex}) + \epsilon$$

Structure	Model	Same Sex			Same Race			Same Race × Same Sex		
		Coeff.	Std Err	P-value	Coeff.	Std Err	P-value	Coeff.	Std Err	P-value
Right NAc	ANTs	-0.005	0.006	0.421	0.021	0.006	**0.000**	0.008	0.008	0.451
	CoTr	0.003	0.003	0.208	0.002	0.003	0.447	0.004	0.004	0.433
	nnU-Net	-0.001	0.003	0.846	-0.000	0.003	0.979	0.006	0.004	0.117
	UNesT	0.004	0.004	0.289	0.008	0.004	**0.050**	0.012	0.006	**0.042**
Left NAc	ANTs	-0.005	0.007	0.437	0.022	0.007	**0.001**	0.011	0.010	0.269
	CoTr	-0.001	0.003	0.852	-0.000	0.003	0.986	0.009	0.004	**0.027**
	nnU-Net	0.001	0.003	0.810	0.000	0.003	0.906	0.007	0.005	0.146
	UNesT	0.002	0.005	0.682	0.011	0.005	**0.030**	0.014	0.007	**0.048**

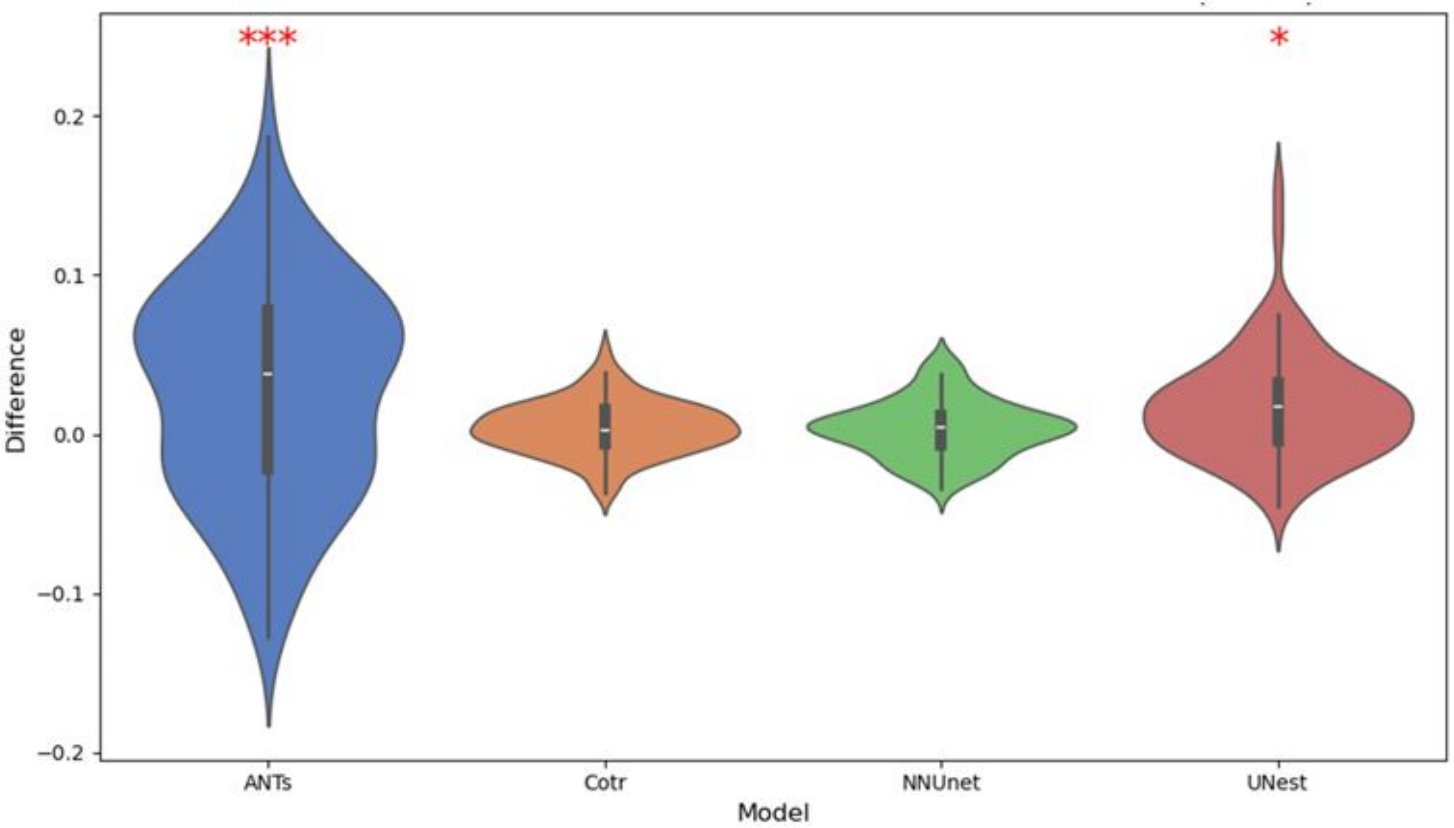


Figure 10: Difference in Dice coefficient for models of same race versus non-same race(The difference is computed as (average of same-race models) - (average of non–same-race models) for each test case)(left NAc). Significance using linear mixed effects model is denoted by *** $(1.00 \times 10^{-4} < P \leq 1.00 \times 10^{-3})$, ** $(1.00 \times 10^{-3} < P \leq 1.00 \times 10^{-2})$, and * $(1.00 \times 10^{-2} < P \leq 5.00 \times 10^{-2})$.

Results

Morphometric differences in Manual segmenations:

$$\text{Volume} = \beta_0 + \beta_1(\text{Race}) + \beta_2(\text{Sex}) + \beta_3(\text{Race} \times \text{Sex}) + \epsilon$$

Structure	Manual	Sex			Race			Race × Sex		
		Coeff.	Std Err	P-value	Coeff.	Std Err	P-value	Coeff.	Std Err	P-value
Right NAc	Manual (whole dataset)	208.63	69.06	**0.003**	225.258	69.736	**0.001**	-59.781	97.202	0.539
	Manual (Test set)	179.28	69.72	**0.010**	379.632	100.368	**0.000**	-71.332	140.284	0.611
Left NAc	Manual (whole dataset)	232.674	66.677	**0.000**	252.66	67.321	**0.000**	7.667	93.836	0.935
	Manual (Test set)	191.155	100.463	0.057	385.526	112.698	**0.001**	-53.176	155.119	0.732

Results

3. Impact of biased segmentation on morphometric analyses

$$\text{Volume} = \beta_0 + \beta_1(\text{Race}) + \beta_2(\text{Sex}) + \beta_3(\text{Race} \times \text{Sex}) + \epsilon$$

Table 5: Results for evaluating Sex effects on volumes by segmentation models for right and left NAc. Coeff. is the coefficient for a fixed factor term such as Sex that describes the effect of the factor level on the volume. Std Err is the standard error of the coefficient estimates, and P denotes P-value.

Structure	Model	Trained on BF			Trained on BM			Trained on WF			Trained on WM		
		Coeff.	Std Err	P	Coeff.	Std Err	P	Coeff.	Std Err	P	Coeff.	Std Err	P
Right NAc	ANTs	219.8	49.5	**0.000**	171	41.5	**0.000**	131	50.0	**0.009**	214	58.7	**0.000**
	CoTr	203.7	74.3	**0.006**	259	78.5	**0.001**	184	65.3	**0.005**	256	77.8	**0.001**
	nnU-Net	231.1	71.5	**0.001**	2022	74.8	**0.007**	166	74.8	**0.026**	248	78.0	**0.001**
	UNesT	246.4	59.3	**0.000**	204	65.7	**0.002**	186	65.4	**0.004**	160	71.3	**0.025**
Left NAc	ANTs	216.8	39.6	**0.000**	185	42.4	**0.000**	74.9	53.8	0.164	218	45.5	**0.000**
	CoTr	208.8	82.6	**0.012**	164	83.4	**0.049**	168	69.3	**0.015**	142	77.7	0.066
	nnU-Net	246.1	70.6	**0.000**	155	82.7	0.060	181	72.1	**0.012**	172	82.9	**0.038**
	UNesT	168.6	65.4	**0.010**	145	65.97	**0.027**	158	61.9	**0.010**	101	73.4	0.166

Results

3. Impact of biased segmentation on morphometric analyses

$$\text{Volume} = \beta_0 + \beta_1(\text{Race}) + \beta_2(\text{Sex}) + \beta_3(\text{Race} \times \text{Sex}) + \epsilon$$

Model	Trained on Black Female			Trained on Black Male			Trained on White Female			Trained on White Male		
	Coeff.	Std Err	P-val	Coeff.	Std Err	P-val	Coeff.	Std Err	P-val	Coeff.	Std Err	P-val
ANTs	29.053	52.382	0.579	34.158	52.725	0.517	-26.368	58.286	0.651	41.316	60.717	0.496
CoTr	173.632	86.213	0.044	113.263	105.153	0.281	95.737	82.276	0.245	144.842	100.016	0.148
nnU-Net	154.684	87.878	0.078	124.947	99.573	0.210	60.579	83.067	0.466	151.105	95.667	0.114
UNesT	4.000	72.013	0.956	-20.789	77.567	0.789	7.421	75.446	0.922	110.000	79.805	0.168

Discussion

- Volumes of Right NAc>>volumes of left NAc
- NAcs morphological difference in Males and females
- Race-based differences in volumes are only in manual segmentations
- The results align with previous studies (Ioannou et al. (2022)) who claimed that race bias effect was more significant than sex
- Clinical implications of biased segmentation models
- We evaluate 4 models and used gold-standard labels as ground truth

Limitations

- Small dataset size
- Biases may be different in other populations (children, elderly, etc)
- Right and left NAc are small subcortical structures
- The isolation of training set to only one demographic group may be unrealistic

Conclusion

- Results of UNesT and ANTs showed race matching improves segmentation accuracy

- nnU-net the only model that its performance is indifferent to the race-matching and sex-matching of training set and test set

- Sex differences observed with manual segmentation on the volumes can also be observed with biased models, whereas the race differences disappear in all but one model

- Most models show a lower overall Dice coefficient score and ESSP when trained on datasets from black demographic groups than those trained on white demographic groups.