Evaluation of Performance Disparities in Deep Learning Models for Neuroimaging and the Role of Entropy-Based Active Learning

by

Ghazal Danaee

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT OF A MASTER'S DEGREE
WITH THESIS IN INFORMATION TECHNOLOGY ENGINEERING
M.A.Sc.

MONTREAL, 30 APRIL 2026

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Prof. Sylvain Bouix, Thesis supervisor
Software Engineering and Information Technology department, École de technologie supérieure

Prof. Ulrich Aïvodji , Chair, Board of Examiners
Software Engineering and Information technology department, École de technologie supérieure

Prof. Christian Desrosiers, Member of the Jury
Software Engineering and Information technology department, École de technologie supérieure

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON 2026/04/20

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to everyone who supported me throughout my master's journey at ÉTS.

First, I would like to thank my supervisor, Professor Bouix, for his guidance and support. Through his supervision, I was introduced to the field of neuroimaging and to the remarkable potential of machine learning in medical imaging. His ideas, insight, and encouragement were invaluable, especially during difficult moments in my research. It was truly an honor to be his student.

I would also like to thank Dr. Jarrett Rushmore and Professor Marc Niethammer for their valuable help with my first article.

I am also sincerely thankful to Professor Desrosiers, Professor Lombaert, and Mélanie Gaillochet for their support in my second project and in the preparation of the associated article.

I am grateful to my family for their unwavering support throughout this journey.

In addition, I would like to thank my lab colleagues for their kindness and companionship. Their presence made this journey more enjoyable and helped maintain my motivation. I would also like to thank the members of my thesis defense jury for agreeing to evaluate this work.

Évaluation des disparités de performance des modèles d'apprentissage profond en neuro-imagerie et du rôle de l'apprentissage actif basé sur l'entropie

Ghazal Danaee

RÉSUMÉ

Les modèles d'apprentissage profond constituent aujourd'hui la principale approche pour la segmentation d'images médicales. Toutefois, ils peuvent produire des prédictions biaisées selon certains groupes démographiques, notamment le sexe et l'ethnoculturalité. Ces écarts sont importants, car ils peuvent influencer les analyses en aval, comme les études volumétriques, et contribuer à des inégalités dans les résultats cliniques. Cette thèse s'intéresse à ce problème en étudiant le biais démographique dans la segmentation de l'IRM cérébrale structurelle et en proposant un cadre d'apprentissage actif sensible à l'équité afin de réduire ces disparités.

Dans une première étude, publiée dans la revue Machine Learning for Biomedical Imaging, nous avons comparé plusieurs modèles d'apprentissage profond, dont nnU-Net, UNesT et CoTr, à une méthode atlasique plus classique, ANTs. À partir de segmentations de référence du noyau accumbens, annotées manuellement curées, nous avons montré que certains modèles sont beaucoup plus sensibles que d'autres au déséquilibre démographique dans les données d'entraînement. Nous avons également observé que des segmentations automatisées biaisées peuvent masquer de véritables différences volumétriques liées à la race, pourtant encore visibles dans les annotations manuelles. Ces biais peuvent ainsi modifier l'interprétation des analyses morphométriques en aval.

Dans une seconde étude, qui sera présentée à la conférence Medical Imaging with Deep Learning en 2026, nous avons proposé une stratégie d'apprentissage actif sensible à l'équité pour réduire ces disparités dès l'étape de sélection des données d'entrainement. La méthode proposée, fondée sur une entropie localisée pondérée, combine une pondération basée sur la performance spécifique à chaque groupe et une estimation localisée de l'incertitude. En concentrant l'incertitude dans une région d'intérêt, cette approche capture plus fidèlement l'incertitude épistémique pertinente et évite de confondre celle-ci avec de simples variations anatomiques plus globales. Des expériences menées sur des IRM synthétiques présentant un biais morphologique contrôlé ont montré que cette stratégie peut réduire de manière importante les écarts entre groupes, avec des améliorations allant jusqu'à 86 % par rapport à l'échantillonnage par entropie standard.
Dans l'ensemble, ces travaux montrent que l'équité en neuro-imagerie ne devrait pas être considérée comme un simple critère d'évaluation secondaire. Elle doit plutôt être pensée comme un principe central de conception. En mettant en évidence la présence de biais dans la segmentation cérébrale, à la fois dans les méthodes d'apprentissage profond et dans les approches plus traditionnelles, puis en proposant une stratégie concrète pour entraîner des modèles plus équitables, cette thèse contribue au développement de méthodes d'analyse d'images médicales plus fiables et plus justes.

Mots-clés: Apprentissage actif équitable, segmentation de l'IRM cérébrale

Evaluation of Performance Disparities in Deep Learning Models for Neuroimaging and the Role of Entropy-Based Active Learning

Ghazal Danaee

ABSTRACT

Deep learning models are now the leading approach for medical image segmentation. However, they can produce biased predictions across demographic groups such as sex and race. These disparities are important because they can affect downstream analyses, including volumetric studies, and may contribute to unequal clinical outcomes. This thesis addresses this issue by studying demographic bias in structural brain MRI segmentation and by proposing a fairness-aware active learning framework to reduce these disparities.

In the first study, published in the Machine Learning for Biomedical Imaging journal, we compared several deep-learning models, including nnU-Net, UNesT, and CoTr, as well as a traditional atlas-based method, ANTs. Using manually curated gold-standard segmentations of the nucleus accumbens, we showed that some models are much more sensitive than others to demographic imbalance in the training data. We also found that biased automated segmentations can obscure true race-related volumetric differences that remain visible in the manual annotations, thereby influencing downstream morphometric conclusions.

In the second study, which will be presented at the 2026 Medical Imaging with Deep Learning conference, we introduced a fairness-aware active learning strategy to reduce these disparities during sample selection. The proposed weighted localized entropy method combines group-specific performance weighting with localized uncertainty estimation. By focusing uncertainty within a region of interest, the method better captures meaningful epistemic uncertainty and avoids simple anatomical variation. Experiments on synthetic MRI data with controlled morphological bias showed that this strategy can substantially reduce group disparity, with improvements of up to 86% relative to standard entropy sampling.

Overall, this work shows that fairness in neuroimaging should not be treated as a secondary evaluation criterion. It should be considered a central design principle. By identifying bias in brain segmentation across both deep learning models and traditional methods, and by proposing a practical strategy for training more equitable models, this thesis advances the development of more reliable and fair medical image analysis.

Keywords: Fair Active Learning, Brain MRI Segmentation

TABLE OF CONTENTS

Page

INTRODUCTION 1

CHAPTER 1 LITERATURE REVIEW 7

CHAPTER 2 INVESTIGATING DEMOGRAPHIC BIAS IN BRAIN MRI SEGMENTATION: A COMPARATIVE STUDY OF DEEP-LEARNING AND NON-DEEP-LEARNING METHODS 13
2.1 Introduction 13
2.2 Related Works 15
2.3 Methods 17
2.3.1 Data 17
2.3.2 Biased training 17
2.3.3 Evaluation metrics 18
2.3.4 Segmentation algorithms tested 20
2.3.4.1 UNesT 20
2.3.4.2 nnU-Net 21
2.3.4.3 CoTr 22
2.3.4.4 Multi-Atlas Segmentation with Joint Label Fusion (ANTs) 22
2.3.5 Statistical analysis 24
2.3.5.1 Performance Bias 24
2.3.5.2 Effect of bias on demographic analyses 24
2.4 Results 25
2.4.1 General statistics of the volumes 25
2.4.2 Bias in volumes and segmentation performance 27
2.4.3 Impact of biased segmentation on morphometric analyses 28
2.4.4 Impact of Dataset Size, Demographics, and Atlas Selection on Bias in ANTs and UNesT 31
2.5 Discussion 34
2.6 Conclusion 39

CHAPTER 3 EXPLORING ENTROPY-BASED ACTIVE LEARNING FOR FAIR BRAIN SEGMENTATION 41
3.1 Introduction 41
3.2 Related works 43
3.3 Method 44
3.3.1 Data 44
3.3.2 Weighted localized entropy 45
3.4 Experiment and Results 46
3.4.1 Implementation details 46
3.4.2 Evaluation metrics 48

3.4.3 Baseline experiments 48
3.4.4 Active learning experiments 49
3.5 Discussion 50
3.6 Conclusion 56

CONCLUSION AND RECOMMENDATIONS 57

APPENDIX I INVESTIGATING DEMOGRAPHIC BIAS IN BRAIN MRI SEGMENTATION: A COMPARATIVE STUDY OF DEEP-LEARNING AND NON-DEEP-LEARNING METHODS 59

APPENDIX II EXPLORING ENTROPY-BASED ACTIVE LEARNING FOR FAIR BRAIN SEGMENTATION 69

BIBLIOGRAPHY 71

LIST OF TABLES

Page

Table 2.1 Training configurations for deep learning models 23

Table 2.2 Results of the linear mixed model for evaluating sex, race, age, and their interaction (sex × race) effects on volumes by **manual annotation** for right and left NAc. These results are also used to evaluate whether these relationships remain observable when tested on segmentations generated by automated methods 25

Table 2.3 Mean and standard deviation of measured volumes for the right and left NAc (mm^3). Model names indicate the demographic subgroup used for training (e.g., ANTsBF was trained on the black female group). The reported statistics are calculated from applying each trained model to all the test sets 27

Table 2.4 Segmentation performance metrics (DSC (overall Dice coefficient across all test sets), ESSP, Δ) for right and left NAc across different models and training groups. **ESSP** (Equity-Scaled Segmentation Performance) combines overall accuracy with a penalty for cross-group disparities; **higher is better** (↑). Δ quantifies differences of each demographic group from the overall mean; **lower is better** (↓) 29

Table 2.5 Segmentation performance metrics (NSD (overall NSD coefficient across all test sets), ESSP, Δ) for right and left NAc across different models and training groups. **ESSP** (Equity-Scaled Segmentation Performance) combines overall accuracy with a penalty for cross-group disparities; **higher is better** (↑). Δ quantifies differences of each demographic group from the overall mean; **lower is better** (↓) 29

Table 2.6 Effects of Same Sex, Same Race, and Interaction on Dice coefficient for right and left NAc. β_1, β_2, and β_3. are the coefficients for a fixed factor term such as sameSex that describes the effect of the factor level on the Dice coefficient. Std Err is the standard error of the coefficient estimates 30

Table 2.7 Effects of Same Sex, Same Race, and Interaction on NSD for right and left NAc. β_2, β_1, and β_3 are the coefficients for the fixed-effect terms Same Sex, Same Race, and their interaction (Same Race × Same Sex), respectively. "Std Err" denotes the standard error of the coefficient estimates 30

Table 2.8 Results for evaluating sex, race, and race × sex effects on volumes by manual annotation 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 .. 31

Table 2.9 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 31

Table 2.10 Effects of Same Sex, Same Race, and Interaction on new UNesT experiment results for right and left NAc. Left block reports linear mixed-effects coefficients for DSC $\big(\text{DSC} = \beta_0 + \beta_1(\text{SameRace}) + \beta_2(\text{SameSex}) + \beta_3(\text{SameRace} \times \text{SameSex}) + \epsilon\big)$; right block reports the corresponding NSD coefficients $\big(\text{NSD} = \gamma_0 + \gamma_1(\text{SameRace}) + \gamma_2(\text{SameSex}) + \gamma_3(\text{SameRace} \times \text{SameSex}) + \epsilon\big)$. For each factor, the table lists the coefficient, its standard error (Std Err), and the P-value 33

Table 2.11 Segmentation performance metrics and fairness metric based on Dice coefficient for biased UNesT models for right and left NAc. (1 shows best) (**ESSP** (Equity-Scaled Segmentation Performance) combines overall accuracy with a penalty for cross-group disparities; **higher is better** (↑). **Δ** quantifies differences of each demographic group from the overall mean; **lower is better** (↓)) 33

Table 2.12 Average and standard deviation of Dice coefficient (Avg±Std) for right and left NAc. Columns: UNesT trained on each subgroup (WM/WF/BM/BF – matched sample sizes) and Baselines. Superscripts rank average within each row (1 shows best) 34

Table 2.13 Average and standard deviation of NSD (Avg±Std) for right and left NAc. Columns: UNesT trained on each subgroup (WM/WF/BM/BF – matched sample sizes) and Baselines. Superscripts rank average within each row (1 shows best) 34

Table 2.14 Fairness metrics based on Dice coefficient for ANTs and baselines (1 shows the best)(**ESSP** (Equity-Scaled Segmentation Performance) combines overall accuracy with a penalty for cross-group disparities; **higher is better** (↑). **Δ** quantifies differences of each demographic group from the overall mean; **lower is better** (↓)) 35

Table 2.15 Average and standard deviation of Dice coefficient (Avg±Std) for right and left NAc. Columns: ANTs trained on each subgroup (WM/WF/BM/BF) and Baselines. Superscripts rank average within each row (1 shows best) 35

Table 2.16 Average and standard deviation of NSD (Avg±Std) for right and left NAc. Columns: ANTs trained on each subgroup (WM/WF/BM/BF) and Baselines. Superscripts rank average within each row (1 shows best) 36

Table 3.1 Summary of the network training setup, test data, and the active-learning configuration. Group 1 denotes cases with an additional localized deformation in the left caudate, while Group 2 contains only global deformation 46

Table 3.2 Training data configuration by bias strength. Group 1 denotes cases with an additional localized deformation in the left caudate, while Group 2 contains only global deformation 47

Table 3.3 Left caudate segmentation performance (DSC) stratified by training cohort and evaluated on Group 1, Group 2, and pooled test sets of the strong bias and weak bias datasets. Group 1 denotes cases with an additional localized deformation in the left caudate, while Group 2 contains only global deformation. The size of the training set is written in parentheses 49

Table 3.4 Right and left nucleaus accumbens segmentation performance (DSC) stratified by training cohort and evaluated on black and white subjects. The size of the training set is written in parentheses 49

LIST OF FIGURES

Page

Figure 0.1 Coronal, axial, and sagittal views of the brain segmentation by FreeSurfer 2

Figure 0.2 Manual Segmentation Taken from (Rushmore *et al.*, 2022a) 2

Figure 2.1 The green structure is Right NAc and the left structure is left NAc 18

Figure 2.2 Left and right NAc volumes in mm^3 across manual and automated segmentations 26

Figure 3.1 **ESSP** under different initial training set compositions. First row: balanced initialization, second row: 80/20 Group 1/Group 2 ratio, third row: 20/80 ratio. Left column: strong bias dataset, Right column: weak bias dataset 51

Figure 3.2 **Δ** and **DSC** metrics under different initial training set compositions for the strong bias experiment only. First row: balanced initialization, second row: 80/20 Group 1/Group 2 ratio, third row: 20/80 ratio. Left column: Δ, Right column: DSC 52

Figure 3.3 **Group 1 ratio** in the training set after sampling for each cycle under different initial training set. First row: balanced initialization, second row: 80/20 Group 1/Group 2 ratio, third row: 20/80 ratio. Left column: strong bias dataset, Right column: weak bias dataset 53

LIST OF ABBREVIATIONS

ETS	École de Technologie Supérieure
T1w	T1-weighted
CNN	Convolutional Neural Network
MRI	Magnetic Resonance Imaging
GAN	Generative adversarial network
NAc	Nucleus accumbens
MALF	Multi-Atlas Label Fusion
CoTr	Convolutional Transformer
ANTs	Advanced Normalization Tools
nnU-Net	no-new U-Net
UNesT	U-shape medical segmentation model with Nested Transformers (UNesT)
MELBA	Machine learning for Biomedical Imaging
MIDL	Medical Imaging with Deep learning
AL	Active learning
KL divergence	Kullback-Leibler divergence
MONAI	Medical Open Network for AI
CT	Computed Tomography
CMR	Cardiac Magnetic Resonance
MALPEM	Multi-Atlas Label Propagation with Expectation-Maximization-based refinement

HCP	Human Connectome Project
NIH	National Institutes of Health
MNI space	Montreal Neurological Institute space
DeTrans	Deformable Transformer
SGD	Stochastic Gradient Descent
RAS	Right Anterior Superior
CLIP	Contrastive Language-Image Pre-training
VLM	Vision Language Model
ROI	Region of Interest
DSC	Dice similarity coefficient
NSD	Normalized Surface Dice
ESSP	Equity Scaled Segmentation Performance
Δ	Delta measures the aggregate deviation of each demographic group's performance from the overall performance.

INTRODUCTION

Structural brain MRI (sMRI) is a core modality in modern neuroimaging, with growing impact in both clinical neurology and psychiatric research. Recent work highlights how sMRI post-processing such as volumetry, shape analysis, voxel-based morphometry, lesion mapping, is increasingly integrated into clinical workflows (Tae, Ham, Pyun & Kim, 2025). In particular, volumetric analysis is widely used to quantify differences between healthy and diseased brains and can support early diagnosis and longitudinal monitoring of neurodegenerative disorders such as hippocampal atrophy in Alzheimer's disease (Tae *et al.*, 2025). Beyond neurology, sMRI is often combined with functional MRI (fMRI) or diffusion MRI (dMRI) and has been explored for stratifying depression subtypes and predicting treatment response, with the goal of enabling more personalized interventions (Gao, Chen, Castellanos, Lu & Yan, 2025).

Segmentation of structures in medical images provides essential spatial information. This detail is crucial for tasks like anomaly segmentation (Baur, Wiestler, Albarqouni & Navab, 2020). However, supervised deep learning requires massive datasets of pixel-wise labels. Obtaining these expert-annotated masks is expensive and time-consuming. Furthermore, the heterogeneous appearance of target organs poses an additional challenge.

Manual segmentation relies on strict neuroanatomical conventions and operational definitions. These definitions serve as step-by-step instructions for delineating anatomical boundaries.

Manual annotation is highly time-consuming. For example, manual brain parcellation can take up to one week for high-resolution images (Fischl *et al.*, 2004). Consequently, both classical approaches and deep learning-based methods have been developed to mitigate the reliance on extensive manual labeling; these approaches are reviewed in the following. Classical automated methods such as FreeSurfer apply standardized procedures using probabilistic models derived from atlases (Fischl *et al.*, 2002). While effective, this pipeline has high computational costs.

An example of a full brain segmentation by FreeSurfer is shown in Figure 0.1.

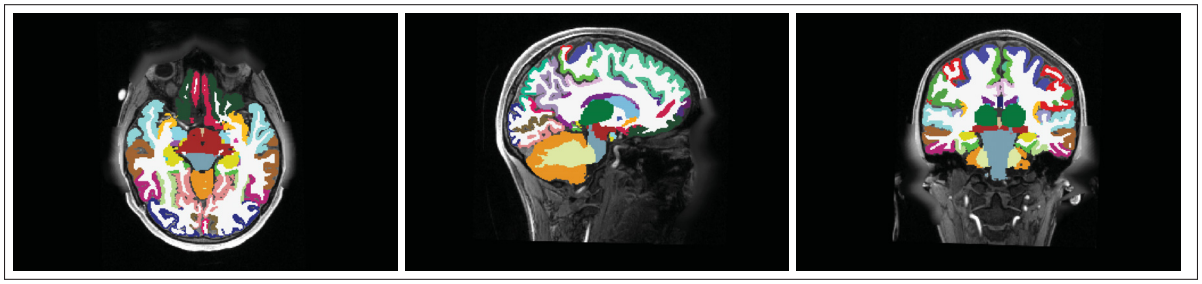

Figure 0.1 Coronal, axial, and sagittal views of the brain segmentation by FreeSurfer

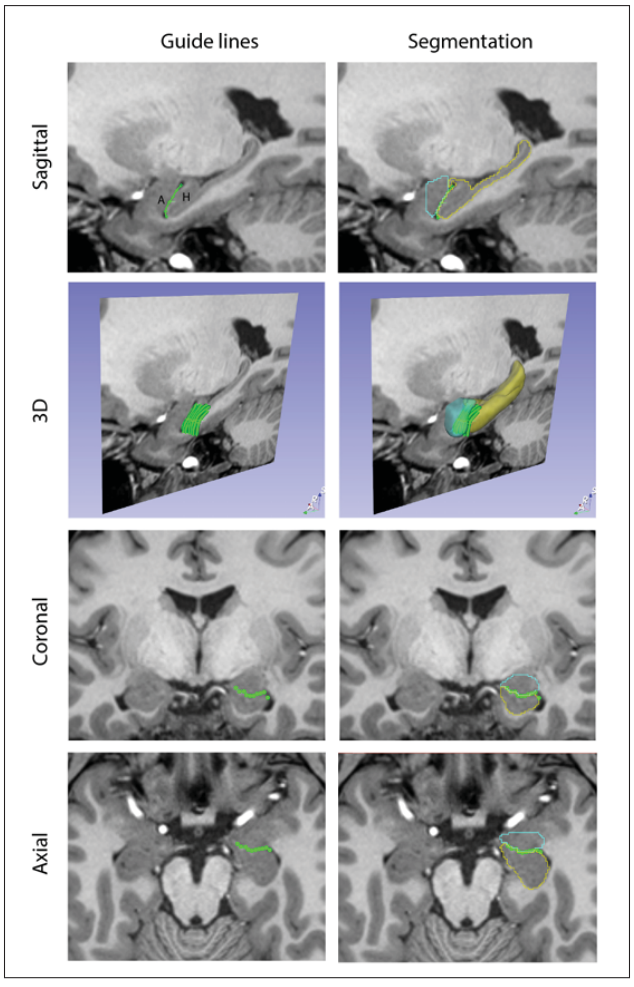

Figure 0.2 Manual Segmentation
Taken from (Rushmore *et al.*, 2022a)

Another approach is Multi-Atlas Label Fusion (MALF)
(Heckemann, Hajnal, Aljabar, Rueckert & Hammers, 2006). MALF registers multiple atlases to a target image and combines them using label fusion strategies.

Two common label fusion strategies exist. The first is majority voting, which simply counts the most frequent label. The second is weighted voting, which assigns weights based on local image similarity. Joint label fusion explicitly accounts for correlated errors between atlases to improve accuracy (Wang *et al.*, 2013a). However, these classical techniques remain highly sensitive to residual registration errors.

These limitations paved the way for deep learning architectures. Convolutional neural networks (CNNs), particularly the U-Net, are now widely used (Ronneberger, Fischer & Brox, 2015). Frameworks like nn-UNet further improved this by automatically configuring the training pipeline (Isensee *et al.*, 2021). Because CNNs are limited by localized receptive fields, transformer-based models were introduced to capture long-range dependencies. Hybrid models aim to combine the strong local feature extraction of CNNs with the global context modeling of transformers. TransUNet (Chen *et al.*, 2021) is a representative example of this idea, where CNN features are first extracted and then refined by a transformer encoder before being decoded through U-Net-style skip connections. However, applying standard transformers to 3D medical images remains computationally expensive. To address this limitation, CoTr (Xie, Zhang, Shen & Xia, 2021) efficiently integrates CNN feature extractors with deformable transformer modules, allowing the network to model long-range dependencies while reducing spatial and computational demands. Unlike vanilla transformer designs that struggle with high-resolution 3D feature maps, CoTr focuses attention on a small set of key locations, which makes transformer-based context modeling more practical for volumetric segmentation.

As segmentation models enter clinical workflows, ensuring algorithmic fairness is vital. Fairness means ensuring models perform equitably across different patient demographics (Xu *et al.*,

2024b). Key challenges include imbalanced datasets, domain differences caused by varying scanner protocols, and biological differences. Unlike simple classification, segmentation outputs are spatial maps. Therefore, bias often manifests in specific regions within an image rather than just global accuracy (Ioannou, Chockler, Hammers & King, 2022b). To combat this, current bias mitigation strategies fall into three categories (Xu *et al.*, 2024b). Pre-processing methods balance subgroup distributions before training. In-processing methods modify the model architecture during training, often using adversarial networks (Stanley, Wilms, Mouches & Forkert, 2022a; Adeli *et al.*, 2021). Finally, post-processing methods refine the trained model by removing biased components. While fairness has been extensively studied in classification settings (Mehrabi, Morstatter, Saxena, Lerman & Galstyan, 2022), fewer works have examined fairness issues in brain MRI segmentation. Recent studies have begun to highlight demographic- and acquisition-related disparities in segmentation performance (Ioannou *et al.*, 2022b; Mehta, Shui & Arbel, 2024; Svanera, Savardi, Signoroni, Benini & Muckli, 2024), but the topic remains substantially underexplored.

Our contribution

Our first article, *Investigating Demographic Bias in Brain MRI Segmentation: A Comparative Study of Deep-Learning and Non-Deep-Learning Methods* evaluates demographic bias in brain MRI segmentation by comparing three deep-learning models (UNesT, nnU-Net, CoTr) and a traditional atlas-based method (ANTs). Using manually curated, gold-standard segmentations of the nucleus accumbens, we establish a baseline for assessing bias rather than relying on automated labels. Moving beyond standard metrics, the research demonstrates how biased segmentations can distort downstream clinical volumetric analyses and alter the observed biological effects of sex and race. Ultimately, by showing that ANTs and UNesT are highly vulnerable to demographic mismatches while nnU-Net remains robust, the work highlights the critical need for diverse training datasets and systematic fairness evaluations. This suggests that practitioners should exercise caution when deploying these models in settings where demographic fairness is a

critical requirement. The article was published in the *Machine Learning for Biomedical Imaging* (MELBA) journal (Danaee, Niethammer, Rushmore & Bouix, 2025).

Author contributions

Ghazal Danaee was responsible for study design, experimental design, code implementation, statistical analysis, running experiments, and writing. Marc Niethammer contributed to study design and writing. Jarrett Rushmore contributed by providing data and helped in writing. Sylvain Bouix contributed to study design, experimental design, statistical analysis, and writing.

Additionally, in the article *Exploring Entropy-based Active Learning for Fair Brain Segmentation*, we tried to bridge the gap between active learning (AL) and algorithmic fairness in brain image segmentation by introducing a fairness-aware active learning (AL) framework. The primary contribution is a Weighted Entropy selection strategy that adaptively prioritizes under-performing patient subgroups by re-weighting uncertainty with group-specific performance scores derived from the current labeled set updated at each cycle. To ensure the model targets genuine epistemic uncertainty rather than mere anatomical volume differences, we further implemented a masked, scaled entropy restricted to the region of interest. Validated on synthetic brain MRI datasets with simulated morphological biases, the method significantly outperformed standard uncertainty and random sampling. The article was accepted at the *Medical Imaging with Deep Learning (MIDL 2026)* conference. The implementation for this work is publicly accessible at: https://github.com/Neuro-iX/Entropy-based-fair-active-learning. We also provide the code on Zenodo : https://doi.org/10.5281/zenodo.19224359. The synthetic datasets with strong and weak bias are accessible on Zenodo : https://doi.org/10.5281/zenodo.19226473.

Author contributions

Ghazal Danaee contributed to the study design and experimental design, implemented the code, conducted the experiments, and led the writing. Mélanie Gaillochet supported the study design and experimental design, assisted with code implementation, and contributed to writing.

Christian Desrosiers, Hervé Lombaert, and Sylvain Bouix contributed to the study design and experimental design and participated in writing and manuscript revisions.

CHAPTER 1

LITERATURE REVIEW

As a detailed review of the relevant literature is included in the two articles of this thesis, in this chapter, we first introduce the fundamental concepts of active learning and discuss its role in reducing annotation effort in machine learning. We then review the notion of algorithmic fairness and its growing importance in medical imaging applications. Finally, we examine the emerging topic of fairness-aware active learning and discuss the current research gap in the context of brain MRI segmentation.

Annotated medical images are often scarce because expert labeling is time-consuming and expensive (Zhou *et al.*, 2021b). To address this limitation, active learning can be used to train models with as few labeled scans as possible (Santos & Marreiros, 2025). In an active learning framework, a large pool of unlabeled data is available. At each iteration, the algorithm selects the most informative samples from this pool and requests annotations from an expert. The goal is to achieve performance comparable to training on the fully labeled dataset while minimizing the number of annotated samples (Ren *et al.*, 2021).

Several query strategies have been proposed in the active learning literature. The choice of strategy depends on the nature of the unlabeled data and the question we have from the oracle (Budd, Robinson & Kainz, 2021). The three main paradigms are stream-based selective sampling, membership query synthesis, and pool-based sampling. In stream-based selective sampling, data arrive sequentially, and the model decides whether each incoming sample should be annotated based on an informativeness criterion (Dagan & Engelson, 1995). In membership query synthesis, the model generates synthetic samples that it considers informative (Angluin, 2001). However, these synthetic queries may request annotations for data points that are not meaningful to a human oracle (Lang & Baum, 1992).

Most active learning studies rely on a pool-based strategy, where samples are selected from a large unlabeled pool (Wang *et al.*, 2024). In this setting, the model scores the unlabeled samples according to their informativeness (Lewis & Catlett, 1994). The most informative

samples are then chosen for expert annotation. This sampling method is especially effective in deep learning, as these models are inherently trained using a batch-based scheme (Budd *et al.*, 2021). In our second article, we adopt this strategy and extend it within a fairness-aware framework. Specifically, we introduce a weighting mechanism that prioritizes samples from underperforming demographic groups.

A key challenge in active learning is estimating how informative a sample is for model improvement. Several strategies have been proposed to address this problem. Common approaches include uncertainty-based sampling, representativeness-based sampling, methods based on generative adversarial networks (GANs), and learning-based active learning strategies. In uncertainty sampling, we select samples where model confidence is low which, are likely to provide the most learning benefit when annotated (Nguyen, Shaker & Hüllermeier, 2022). There are different techniques to measure uncertainty that rely on current model's predictions, including entropy (Wang & Shang, 2014), margin (Scheffer, Decomain & Wrobel, 2001), and posterior probability (Lewis & Gale, 1994). (Kirsch, van Amersfoort & Gal, 2019) proposed BatchBALD, a Bayesian active learning method that selects diverse and informative batches of samples based on their joint mutual information with model parameters. The approach estimates model uncertainty using Monte Carlo dropout as a scalable approximation. By accounting for correlations between samples, BatchBALD avoids selecting redundant data points. Some works have also considered predicting how selecting one datapoint would affect the model in the next cycle. For example, one of the methods of this kind tries to estimate expected changes in model predictions (Käding, Rodner, Freytag & Denzler, 2016).

Another class of sampling strategies focuses on diversity. These methods aim to select samples from different regions of the data distribution to increase the diversity of the labeled set. One of the early approaches in this category was proposed by (Xu, Yu, Tresp, Xu & Wang, 2003), who used k-means clustering to identify representative samples. A well-known representative-based method is Core-set (Sener & Savarese, 2018), which selects samples by minimizing the distance between the latent representations of labeled and unlabeled data using Euclidean distance. More recently, ActiveFT Xie *et al.* (2023) was proposed to select samples that best represent the

overall unlabeled distribution. This method measures representativeness in a pretrained feature space using a KL-divergence estimate tailored for transfer learning.

Hybrid-based active learning strategies are designed to overcome the limitations of using either uncertainty or diversity metrics in isolation by aiming to select data points that are simultaneously informative and representative. These approaches often achieve this balance through sophisticated joint objectives or multi-step selection processes; for instance, Nath, Yang, Landman, Xu & Roth (2021) integrates model uncertainty with image similarity or mutual information to regularize the acquisition function and ensure a diverse training pool. State-of-the-art hybrid methods like BADGE (Ash, Zhang, Krishnamurthy, Langford & Agarwal, 2020) cluster the gradient embeddings of the network's output layer to implicitly capture both sample disparity and uncertainty magnitude. Alternatively, methods such as ALFA-Mix (Parvaneh *et al.*, 2022) identify informative samples by seeking inconsistencies in predictions resulting from latent feature-space interpolations, subsequently applying clustering to these candidates to maximize batch diversity. The optimal balance between these metrics often depends on the available labeling budget. When the budget is small, different sampling strategies may be required than when more annotations are available. To address this issue, recent work such as Uncertainty Herding (UHerding) (Bae, Sutherland & Oliveira, 2025) adaptively interpolates between representation-based and uncertainty-based sampling strategies. This adaptive behavior helps maintain strong performance across both low- and high-budget scenarios.

Once new samples are annotated, the model must be updated to incorporate this information. One option is to fine-tune the current model using the newly labeled data Tajbakhsh *et al.* (2016). Another option is to retrain the model from scratch using all available labeled samples. Alternatively, the model can be retrained using all data while initializing the parameters from the previously trained model. Retraining from scratch is generally more computationally expensive than fine-tuning.

To address the prohibitive costs and time constraints of expert medical annotation, (Diaz-Pinto *et al.*, 2024) introduced MONAI Label, an open-source framework that integrates active learning

(AL) with human-in-the-loop interactive segmentation. Rather than relying solely on static automated predictions or fully manual tracing, the system employs AL strategies such as aleatoric and epistemic uncertainty estimation to query the most informative unlabeled 3D volumes for expert review. Once an uncertain volume is selected, clinicians can utilize AI-assisted refinement tools, including positive/negative click-based models (DeepGrow and DeepEdit) or free-hand scribbles, to correct the model's initial predictions. This iterative paradigm continuously updates the underlying model with the refined labels, significantly reducing the overall annotation time and demonstrating how active data selection paired with interactive refinement can scale the generation of high-quality medical datasets.

Active learning has shown strong potential in medical image segmentation, where voxel-wise annotation is particularly costly and time-consuming. This challenge is even more pronounced in 3D imaging modalities such as MRI and CT, which contain substantial redundancy across adjacent slices. To reduce annotation effort, several studies have adapted active learning to select only the most informative slices or regions within a volume. For example, Zhou, Li, Bredell, Li & Konukoglu (2021c) proposed an uncertainty-based strategy that predicts the segmentation quality of each slice and recommends those with lower predicted quality for expert annotation. Other methods explicitly account for annotation cost in medical settings. In brain tumor segmentation, Shen *et al.* (2021) estimated the cost of annotating a slice based on its distance from already labeled slices, assuming that slices similar to previously annotated ones require less effort. In addition, early frameworks such as Suggestive Annotation (SA) proposed by Yang, Zhang, Chen, Zhang & Chen (2017) showed that combining uncertainty-based selection with representative sampling can improve the segmentation of complex structures efficiently. These studies show that active learning can reduce not only the number of annotated samples, but also the overall human effort required to build accurate segmentation models.

Despite its effectiveness in reducing annotation burden, standard active learning typically focuses on improving overall accuracy. This can increase performance disparities across demographic groups inadvertently. Such a trade-off is often described as a fairness–accuracy frontier (Pang

et al., 2024). In response, fair active learning methods have been developed to reduce group-wise disparities while maintaining good predictive performance. However, many of these approaches require sensitive attribute labels for the full unlabeled pool, which may be unrealistic in practice and can raise privacy concerns. More recent methods attempt to avoid this limitation by estimating fairness effects from a small annotated validation set instead (Pang *et al.*, 2024). Overall, these works emphasize the importance of fairness-aware sampling, particularly in medical imaging, where data scarcity and demographic bias frequently overlap.

Moreover, to the best of our knowledge, fairness-aware active learning for brain segmentation has not yet been investigated in the literature. To address these gaps, this thesis presents the following contributions.

This thesis addresses important challenges at the intersection of fairness, limited annotated data, and medical image segmentation. First, we investigate how several state-of-the-art segmentation methods, including both deep learning and atlas-based approaches, are affected by demographic bias. Using a manually curated gold-standard dataset, we show that imbalanced training cohorts can lead to performance disparities across demographic groups and can influence downstream clinical conclusions related to sex and race. We then propose a fairness-aware active learning framework to reduce these disparities when annotation resources are limited. In particular, we introduce a Weighted Localized Entropy sampling strategy that gives priority to underperforming demographic subgroups based on current group-level performance. By restricting uncertainty estimation to a region-of-interest mask, the method aims to reduce the effect of anatomical volume variation and improve both accuracy and fairness.

CHAPTER 2

INVESTIGATING DEMOGRAPHIC BIAS IN BRAIN MRI SEGMENTATION: A COMPARATIVE STUDY OF DEEP-LEARNING AND NON-DEEP-LEARNING METHODS

Ghazal Danaee[1] , Marc Niethammer[2] , Jarrett Rushmore[3] , Sylvain Bouix[1]

[1] Département de génie logiciel et des technologies de l'information, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] University of California San Diego, 9500 Gilman Drive, La Jolla , San Diego, California, USA, 92093
[3] Boston University School of Medicine, 72 East Concord St., Boston, Massachusetts, USA, 02118

Article published in the MELBA journal « Machine Learning for Biomedical Imaging » in December 2025.

2.1 Introduction

Researchers have widely adopted deep-learning-based models as state-of-the-art approaches in medical image computing. However, these models may display biased predictions for individuals with different protected attributes, such as sex, age, and race (Xu *et al.*, 2024b).

When a model performs worse for specific subgroups, the downstream clinical implications can be significant, potentially leading to misdiagnosis or underdiagnosis for patients within those groups. Examining and mitigating these biases is paramount to achieving equitable healthcare outcomes. For instance, Stanley, Wilms, Mouches & Forkert (2022b) studied differences in the performances of models predicting the sex of patients using magnetic resonance imaging (MRI) and observed differences between white and black children. Frazier *et al.* (2008) studied how diagnosis and sex affect brain regions in early-onset bipolar disorder and schizophrenia. Using MRI, they found that both factors influence amygdalar and hippocampal volumes, with differences between men and women. The study underscores the critical importance of accounting for sex differences in brain studies related to mental health.

Several factors contribute to bias in medical image computing. One fact is the inherent anatomical differences between men and women. In a recent study, Dibaji, Ospel, Souza & Bento (2024) evaluated how these sex-based anatomical differences in brain MRI data influence the performance of sex classification models. They analyzed saliency maps of the models to determine the regions most influential in driving sex classification. Furthermore, Isamah *et al.* (2010) demonstrated the differences in volumes of brain structures between various racial groups.

Previous research has predominantly concentrated on fairness in classification tasks (Mehrabi *et al.*, 2022). By contrast, fairness in segmentation has received relatively little attention, despite the significant impact that segmentation bias can have on clinical decision making. The few studies conducted in the realm of segmentation have typically focused on evaluating only one type of deep learning model in their analyses. We addressed these gaps by thoroughly evaluating demographic bias in both deep-learning and non-deep-learning models for brain region segmentation. Specifically, we considered four different methods: three state-of-the-art deep-learning models with different types of architectures (UNesT (Yu *et al.*, 2023), nnU-Net (Isensee *et al.*, 2021), and CoTr (Xie *et al.*, 2021)) and a traditional atlas-based segmentation method (Multi-Atlas Segmentation with Joint Label Fusion (Wang *et al.*, 2013b)). We evaluated their bias across four demographic subgroups (black female, black male, white female, and white male). Moreover, we used manually annotated gold-standard segmentations of two subcortical structures—namely, the left and right nucleus accumbens (NAc)—as the labels for the training dataset, thus ensuring a high-quality gold standard for our evaluation. We extended our analyses beyond segmentation accuracy by investigating whether volume differences between sex, race and their interaction, observed with manual segmentation, remain consistent when using the segmentation output from biased models.

While recent studies have provided comprehensive analyses of bias across multiple attributes and mitigation strategies (Siddiqui *et al.*, 2024), our work offers a distinct contribution by comparing the performance of multiple deep-learning architectures against a traditional non-deep-learning method in the context of brain MRI segmentation.

Ultimately, our findings aim to contribute to developing more equitable and generalizable

practices in automated brain-image segmentation, thereby fostering enhanced fairness within both clinical and research environments.

2.2 Related Works

Previous studies have investigated bias in segmentation tasks in medical image computing. For example, Puyol-Antón et al. observed statistically significant differences in the performance of models in cine cardiac MR segmentation between racial groups (Puyol-Antón *et al.*, 2021, 2022). The training set was sex-balanced but not race-balanced, causing race bias in performance results. Additionally, Lee *et al.* (2025) showed racial biases in cine cardiac magnetic resonance (CMR) imaging with reduced performance on black subjects. They aimed to identify the causes of racial bias. They discovered that racial biases stemmed from non-cardiac features in MR images (areas of the image that did not include the heart), and training the model on cropped images helped narrow the performance gap between black and white patients.

In a recent study, Ioannou *et al.* (2022b) investigated the influence of sex and race on the performance of the FastSurferCNN model (Henschel *et al.*, 2020b) trained using silver standard labels derived from the Multi-Atlas Label Propagation with Expectation-Maximization-based refinement (MALPEM) algorithm
(Ledig *et al.*, 2015; Ledig, Schuh, Guerrero, Heckemann & Rueckert, 2018). The study focused on the segmentation of 78 structures in the brain and evaluated demographic biases within these regions; they found sex and race bias in some but not all structures. To assess sex bias, they trained five different models on training sets with varying ratios of white females and white males and subsequently tested these models on the test sets of white females and white males. In another experiment, they used the same trained models and tested them on black and white females to measure race biases. Their findings revealed that race bias is more significant than sex bias. In addition, they reported that specific brain regions showed a significant bias effect, indicating that the bias has a spatial component. They also observed sex biases. For example, when the model was trained on a sex-balanced dataset, its performance in segmenting three brain regions showed a statistically significant reduction in Dice coefficient for white females

compared to white males.

Alqarni *et al.* (2024) investigated racial bias in deep learning-based prostate gland segmentation from MR images by training models with varying ratios of white and black subjects. Their findings revealed that models trained on imbalanced datasets exhibited significant racial bias, while employing a race-balanced training set resulted in the best segmentation performance across both groups. In a recent study, Siddiqui *et al.* (2024) investigated biases related to age, sex, and race in the segmentation of hip and knee X-ray images. Their work demonstrates the trade-off between fairness and accuracy by comparing several bias mitigation strategies applied to U-Net models with different CNN backbones (ResNet18 and EfficientNet-B0). A study evaluating skin color bias in skin lesion segmentation algorithms (CNNs) by Benčević, Habijan, Galić, Babin & Pižurica (2024) observed lower performance on darker skin tones. Although they used mitigation methods, none of them were effective at reducing the bias. In this study, we supplement these initiatives by directly comparing multiple methods and using a manually curated gold-standard dataset, therefore providing a more comprehensive view of how different methods handle unbalanced training data.

Prior studies of bias in segmentation have often evaluated a single deep-learning architecture within a given application. Outside the brain, comparative work in cardiac MR has shown that model choice itself can affect measured sex bias and race bias (Lee *et al.*, 2023). In brain MRI, however, we are not aware of prior studies that jointly compare deep learning and atlas-based segmentation with respect to bias.

Our study offers a more comprehensive evaluation by (i) comparing two distinct types of segmentation models—deep learning–based and traditional non–deep learning approaches, a comparison not previously worked on, (ii) employing manually curated gold-standard labels for the nucleus accumbens, and (iii) examining both segmentation performance fairness and its impact on volumetric analyses through linear mixed-effects models.

2.3 Methods

2.3.1 Data

We utilized data from the Human Connectome Project (HCP) Young Adult dataset. According to the "WU–Minn HCP 1200 Subjects Data Release: Reference Manual", each HCP participant is given a participant identification number for tracking and asked a number of demographic questions, including gender, age, twin status (including self-reported zygosity), race, ethnicity, educational level, household income, and relationship status(Human Connectome Project, 2017). The T1-weighted MRIs have a resolution of $260 \times 311 \times 260$ voxels with an isotropic voxel spacing of 0.7. The subjects' ages ranged from 22 to 35. Reporting race and ethnicity in this study was mandated by the NIH, consistent with the Inclusion of Women, Minorities, and Children policy. Sex, race and ethnicity were self-reported by participants. For the training phase, we employed 30, 32, 33, and 31 images corresponding to black female, black male, white female, and white male subjects, respectively. In the testing phase, we utilized 19, 20, 19, and 20 images for black female, white male, white female, and black male subjects.

We utilized the manual segmentations of two subcortical structures, the right and left nucleus accumbens (NAc), provided by neuroanatomist Dr. Jarrett Rushmore. This structure was selected due to previously reported sex differences in microstructure (Wissman, May & Woolley, 2012). Because NAc volume is widely employed as a volumetric biomarker, demographic bias in its automated segmentation could confound clinical inference and exacerbate health disparities. Figure 2.1 shows a case of manual annotation of the right and left NAcs.

2.3.2 Biased training

For each architecture—nnU-Net, UNesT, and CoTr—four separate models were trained, with each model using data from just one of the four demographic groups: 32 black male, 30 black female, 31 white male, or 33 white female subjects. Similarly, four biased datasets were used with ANTs, leading to 4 "models" each using atlases from just one demographic subgroup:

Figure 2.1 The green structure is Right NAc and the left structure is left NAc

10 black male, 10 black female, 10 white male, or 10 white female subjects. This approach intentionally introduces bias to assess the impact of imbalanced training on segmentation fairness across demographics.

2.3.3 Evaluation metrics

We used two core metrics for evaluating raw segmentation performance. First, the Dice similarity coefficient (DSC) is an overlap-based metric that ranges from 0 with no overlap to 1 with complete overlap. Let X denote the ground truth segmentation and Y be the predicted segmentation. The Dice coefficient is computed as:

$$\mathrm{DSC}(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|} \tag{2.1}$$

Second, the normalized surface Dice (NSD) defined by Nikolov *et al.* (2021) is a boundary-based metric that measures the Dice coefficient on boundary pixels with a margin. τ is the maximum tolerated distance from the boundary that defines the border region. For two shapes X and

Y, S_X and S_Y denote their boundaries, and $\mathcal{B}_X^{(\tau)}$ and $\mathcal{B}_Y^{(\tau)}$ are the corresponding boundary regions (Maier-Hein *et al.*, 2024).

$$\mathrm{NSD}(X, Y) = \frac{|S_X \cap \mathcal{B}_Y^{(\tau)}| + |S_Y \cap \mathcal{B}_X^{(\tau)}|}{|S_X| + |S_Y|} \tag{2.2}$$

Furthermore, to evaluate fairness in the models' results, we utilized the **Equity-Scaled Segmentation Performance (ESSP) metric**, originally proposed by Tian *et al.* (2024). Computing ESSP requires two components:

1. The overall segmentation accuracy (the mean accuracy of a model on all demographic groups)
2. The deviation of a model's accuracy on each demographic group from the overall segmentation accuracy

Let A denote the set of demographic groups under consideration (in our case, white male, white female, black male, and black female), N_A denote the total size of all test sets across all subgroups in A, DSC_a represent the average Dice coefficient for subgroup $a \in A$, and n_a is the size of subgroup a. We compute the overall segmentation accuracy for each of the models by averaging the Dice coefficient over all samples in the test set:

$$\mathrm{DSC}_{overall} = \frac{1}{N_A} \sum_{a \in A} \mathrm{DSC}_a \times n_a. \tag{2.3}$$

Next, we define Δ as the sum of absolute performance discrepancies across all groups:

$$\Delta = \sum_{a \in A} \left| \mathrm{DSC}_{overall} - \mathrm{DSC}_a \right|. \tag{2.4}$$

Finally, we calculate the Equity-Scaled Segmentation Performance (ESSP) by penalizing the overall performance concerning Δ:

$$\mathrm{ESSP} = \frac{\mathrm{DSC}_{overall}}{1 + \Delta}. \tag{2.5}$$

In essence, ESSP acts as a substitute for the Dice coefficient, with a penalty for unfairness. Note that the same measure can be computed for NSD and we use the following notation $ESSP_{DSC}$ and $ESSP_{NSD}$ to respectively identify the Dice and NSD ESSP measures.

2.3.4 Segmentation algorithms tested

All deep-learning methods were used with their default configuration, as provided by their respective official code repositories. We chose the default configuration to mirror common practice, enable reproducible comparisons, and model real-world usage. There are therefore, some noticeable differences beyond network architectures, including the number of epochs and data augmentation. We summarize the training configurations of the deep learning models in Table 2.1 and review each method in details below.

2.3.4.1 UNesT

We utilized UNesT (Yu *et al.*, 2023), a segmentation model with a hierarchical transformer encoder that processes volumetric data by dividing the input into 3D patches and applying local self-attention at different scales. Unlike prior approaches relying on convolutional layers for feature extraction, UNesT leverages a transformer-based encoder to capture multi-scale features. It then uses a convolutional decoder to upsample these representations and produce the final segmentation.

Implementation details: We utilized the official implementation of the UNesT model and trained the UNesT-large version with approximately 280 million parameters from scratch on our dataset. We decided to use the default configuration of UNesT since these values worked best for our task of segmenting small subcortical structures. For example, UNesT could not segment well when trained with only 1000 epochs, which is the fixed value for the number of epochs in nnU-net. The code for UNesT is publicly accessible online[1].
Data were registered to MNI space using affine transformations for the train and test phases.

[1] Official UNesT implementation: https://github.com/MASILab/UNesT

Then, after the test set was segmented, the results were registered back to the original space. For each model trained on a specific demographic group (e.g., black females), we performed 5-fold cross-validation within that group's training data, with each fold trained for 50,000 epochs, and the ensemble of the results of models were used to produce the predictions. We used the default parameters of the UNesT-large model, as provided by the official implementation. These included a learning rate of 0.00001, the Adam optimizer, and a momentum of 0.9. We utilized Dice-cross entropy as it showed better performance in segmenting small structures.

2.3.4.2 nnU-Net

nnU-Net (Isensee *et al.*, 2021) is an adaptive model specifically designed for biomedical image segmentation. Its key advantage lies in its ability to apply to any dataset by systematizing the complex process of manual method configuration.

Implementation details: We trained the model from scratch on our training dataset, and evaluated its performance on test sets from different demographic groups. The official implementation of nnU-Net, used in our experiments, is available on their Github.[2]

We adhered to the default configuration for the nnU-Net model, as its defining characteristic is the automated optimization of the entire pipeline based on a new dataset's properties. Manually altering these systematically configured parameters would undermine the model's self-configuring design philosophy and negate its primary advantage.

nnU-Net adopts several fixed design choices, including the use of a combined cross-entropy and Dice loss function across all applications. In addition, it incorporates a set of rule-based and empirical design choices for the model's configuration. Given a new training dataset, nnU-Net automatically creates up to three pipeline configurations (2D U-Net, 3D full-resolution U-Net, and the 3D U-Net cascade) , and trains each configuration in a 5-fold cross-validation run. After training, nnU-Net empirically selects the best single configuration or an ensemble of configurations (Isensee *et al.*, 2021). In our experiments, the 3D full-resolution configuration was

[2] Official nnU-Net implementation: https://github.com/MIC-DKFZ/nnUNet/tree/master/nnunetv2

consistently selected, and its predictions were used for reporting results. nnU-Net's rule-based and empirical strategies determined the other design choices.

2.3.4.3 CoTr

The last deep-learning-based model we used is CoTr (Xie *et al.*, 2021), which leverages the strengths of both transformers and convolutional neural networks for 3D medical image segmentation. In CoTr, a CNN is designed to extract feature representations, while the authors introduced the deformable Transformer (DeTrans) to model long-range dependencies within the extracted feature maps effectively.

Implementation details: We trained CoTr from scratch on the training set and evaluated it on test sets from various demographic groups. We performed 5-fold cross-validation and used inference-time ensembling of the models. The loss function of the model is the sum of the Dice loss and cross-entropy loss. We used the official implementation of CoTr and its default configuration for training, which is available on GitHub.[3]

2.3.4.4 Multi-Atlas Segmentation with Joint Label Fusion (ANTs)

Atlas-based segmentation relies on atlases—expert-labeled sample images— for guiding segmentation. Each atlas is registered to the target image in this approach, and the warped atlases are combined using label fusion techniques, such as weighted voting. Multi-Atlas Segmentation with Joint Label Fusion (Wang *et al.*, 2013b) which incorporates dependencies between atlases, was one of the leading segmentation techniques before deep learning methods were developed. The method is quite flexible. Given a relatively small set of labeled data, one could perform segmentation with good accuracy.

Implementation details: Four variants of ANTs Joint Label Fusion were used, each using 10 atlases exclusively from training sets of each one of the demographic subgroups: 10 black male, 10 black female, 10 white male, or 10 white female subjects to segment the test set. Since

[3] Official CoTr implementation: https://github.com/YtongXie/CoTr/tree/main

Table 2.1 Training configurations for deep learning models

Configuration	nnU-Net	CoTr	UNesT
Loss function	Cross-entropy + Dice	Cross-entropy + Dice	Cross-entropy + Dice
Optimizer	SGD + Nesterov ($\mu = 0.99$)	SGD + Nesterov ($\mu = 0.99$)	Adam ($\beta_1 = 0.9$)
Learning rate (schedule)	0.01 (poly: $(1 - \frac{\text{epoch}}{\text{epoch}_{\text{max}}})^{0.9}$)	0.01 (poly: $(1 - \frac{\text{epoch}}{\text{epoch}_{\text{max}}})^{0.9}$)	1×10^{-5} (warmup–cosine)
Epochs	1000	1000	50,000
5-fold CV	Yes	Yes	Yes
Preprocessing	Crop to non-zero region	Same as nnU-Net	Convert to MNI-305 space, reorient to RAS, resample, scale intensity, spatial pad
Data augmentation	Rotations, scaling, Gaussian noise/blur, brightness, contrast, low-res simulation, gamma correction, mirroring	Same as nnU-Net	Random patch sampling, random mirror flips, random multiplicative intensity scaling
Post-processing	Remove all but largest component	Same as nnU-Net	Convert back to original space

ANTs prediction entails cost-intensive atlas registration, we decided to choose 10 for the size, as in ablation studies with varying numbers of atlas subjects $(5, 10, 15, 20)$, we found that the segmentation accuracy derived by atlases curated with 10 cases differed from those curated with larger numbers of cases by less than 0.1 in Dice coefficient. We utilized the script provided by the Advanced Normalization Tools (ANTs) ecosystem, which is available on ANTs GitHub.[4]

[4] Official ANTs implementation: https://github.com/ANTsX/ANTs/blob/master/Scripts/antsJointLabelFusion.sh

2.3.5 Statistical analysis

2.3.5.1 Performance Bias

In addition to evaluating DSC, NSD, ESSP and Δ, we employed linear mixed models to assess bias in model performance. For each subject in the test sets of different demographic subgroups, we kept the performance scores from the four models within a single architectural design (e.g., the results of four UNesT models trained on black male, black female, white male, and white female). We then used the linear mixed effects model below on these performance scores (Dice coefficient):

$$\begin{aligned}\text{DSC} = \beta_0 + \beta_1(\text{SameRace}) + \beta_2(\text{SameSex}) \\ + \beta_3(\text{SameRace} \times \text{SameSex}) + \epsilon\end{aligned} \quad (2.6)$$

where SameSex is a binary variable which defines whether the test subject has the same sex as the training dataset, and SameRace is a binary variable which defines whether the test subject has the same race as the training dataset. (e.g., coded as 1 for a match and 0 for a mismatch). The variables β_0, β_1, β_2, and β_3 are the parameters to be estimated, and ϵ is the error. This framework enabled us to quantify the contribution of each factor (as well as their interaction) to the observed Dice scores.

2.3.5.2 Effect of bias on demographic analyses

To compare the impact of a biased model on brain morphometry population analyses, we applied a linear mixed model to the volumes corresponding to the test sets of all demographic groups, produced by a single model from each architectural design and the demographic group utilized for training (e.g, UNesT model trained on black females). We then used the following linear mixed effects model on these volumes:

$$\text{Volume} = \gamma_0 + \gamma_1(\text{Race}) + \gamma_2(\text{Sex}) + \gamma_3(\text{Race} \times \text{Sex}) + \epsilon_2 \quad (2.7)$$

We can investigate how race, sex, and their interaction influenced the predicted volumes. Sex and Race are binary variables: Race indicates whether the subject is black or white, and Sex indicates whether the subject is female or male (e.g., White=0, Black=1; Male=0, Female=1). To test whether age affected the volumes of manually annotated labels, we used the linear mixed-effects model below:

$$\text{Volume} = \alpha_0 + \alpha_1(\text{Race}) + \alpha_2(\text{Sex}) + \alpha_3(\text{Sex} \times \text{Race}) + \alpha_4(\text{Age}) + \epsilon_3 \quad (2.8)$$

2.4 Results

The following sections use model-subgroup notation. For instance, 'UNesTBF' represents the UNesT model trained with the black female subset.

2.4.1 General statistics of the volumes

The results of the linear mixed model showed that after adjusting for sex and race, the estimated effect of age on the volume of the nucleus accumbens (NAc) was not statistically significant. The results are provided in Table 2.2. We therefore did not incorporate age as a variable of interest in any experiments. This is also supported by the fact that the HCP dataset has a rather narrow age range of 22 to 35 years. These results are also used in section 2.4.3 to evaluate whether these relationships remain observable when tested on segmentations generated by automated methods.

Table 2.2 Results of the linear mixed model for evaluating sex, race, age, and their interaction (sex × race) effects on volumes by **manual annotation** for right and left NAc. These results are also used to evaluate whether these relationships remain observable when tested on segmentations generated by automated methods

Structure	Method	Sex			Race			Sex × Race			Age		
		α_2	Std Err	P-value	α_1	Std Err	P-value	α_3	Std Err	P-value	α_4	Std Err	P-value
Right NAc	Manual (whole dataset)	195.38	69.39	**0.005**	231.38	69.76	**0.001**	-62.279	97.133	0.521	-11.568	6.33	0.068
	Manual (Test set)	181.554	99.73	0.069	394.84	104.80	**0.000**	-98.103	143.006	0.493	-6.422	5.960	0.281
Left NAc	Manual (whole dataset)	222.831	67.218	**0.001**	257.216	67.573	**0.000**	5.810	94.088	0.951	-8.598	6.141	0.161
	Manual (Test set)	194.65	104.15	0.062	409.01	115.12	**0.000**	-94.511	162.562	0.561	-9.916	6.733	0.141

Table 2.3 displays volume statistics for all models. The ANTsBF and ANTsBM methods demonstrate greater under-segmentation than others. Notably, these are the only models with median volume differences below 20% for the left NAc. This is also observable in Figure 2.2 visualizing the volumes of the structures by manual annotations and models.

Most segmentation models exhibit smaller standard deviations compared to the manual approach. It could suggest that they are under-representing outliers, and perhaps also that the manual annotations are noisier than automated ones. For example, the ANTsBM model shows the lowest variability, with standard deviations of 61.68 mm^3 (right NAc) and 63.45 mm^3 (left NAc), in contrast to the values from manual annotation, which are 125.79 mm^3 and 136.13 mm^3. Comparing the volumes of the left and right NAc, we observe that in the results of all models, the volume corresponding to the right NAc is greater than that of the left NAc, reflecting an anatomical trend preserved in both manual and automated segmentations. One can observe a general underestimation of the NAc size, but also some patterns of bias with ANTsBM and UNesTBM displaying volumes almost 20% smaller than the manual segmentations (Fig. 2.2).

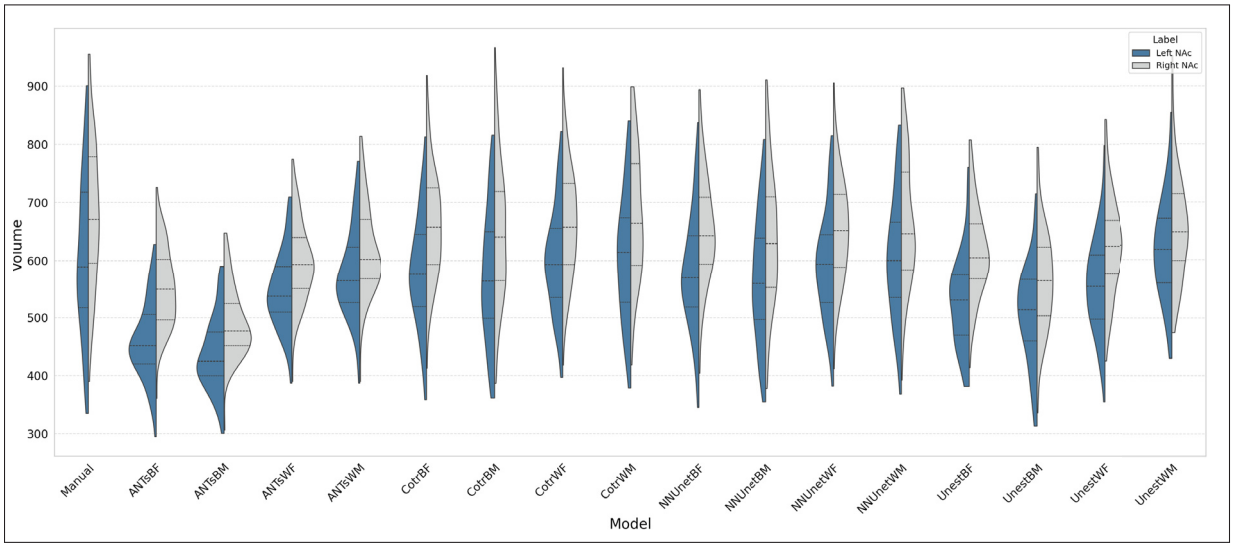

Figure 2.2 Left and right NAc volumes in mm^3 across manual and automated segmentations

Table 2.3 Mean and standard deviation of measured volumes for the right and left NAc (mm^3). Model names indicate the demographic subgroup used for training (e.g., ANTsBF was trained on the black female group). The reported statistics are calculated from applying each trained model to all the test sets

Model	**Right NAc**		**Left NAc**	
	Mean	**Std**	**Mean**	**Std**
Manual	676.97	125.79	607.13	136.13
nnU-NetBF	653.62	95.08	581.06	99.14
nnU-NetBM	638.20	115.41	569.97	108.32
nnU-NetWF	653.14	90.91	593.83	93.10
nnU-NetWM	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

2.4.2 Bias in volumes and segmentation performance

In Tables 2.4 and 2.5, we observe that nnU-Net and CoTr consistently yield the highest ESSP values in Dice coefficient and NSD, in addition to relatively balanced results across demographic groups in both metrics. However, ANTs and UNesT generally perform worse than nnU-Net and CoTr, often by a noticeable margin. ANTs shows huge drops in ESSP for both Dice coefficient and NSD when trained on the black male or black female groups. For example, ANTs exhibits a 13% reduction in ESSP measured by the Dice coefficient when trained on white females compared to black males. ANTs reaches its highest Δ values with 0.11 in Dice coefficient and 0.2 in NSD, respectively, when trained on black male cases in the right NAc.

We evaluated the influence of the model being same-sex, same-race, and both same-sex and race on the Dice and NSD metrics using linear mixed models. The results can be found in Tables 2.6 and 2.7. The difference between same-sex and non–same-sex performance results are not statistically significant for any of the models. However, when matching race, ANTs and UNesT demonstrate significantly higher Dice coefficient than when the test and training data are not matched. When evaluated with NSD, the race-matching effect persists for ANTs, but did not reach significance for UNesT. While sex is a robust factor in volumetry (Table 2.2, sex matching does not significantly affect segmentation accuracy. In contrast, race matching substantially impacts segmentation. Evaluating results between same-sex and race models (trained and tested on the same subgroup) and non-same models (race or sex mismatches), UNesT and CoTr show significantly better Dice coefficients when trained on identical race and sex sub-groups ($p = 0.027$, 0.048 respectively). This effect is also observable for UNesT when evaluated with NSD, but not for CoTr. nnU-Net was the only model that did not exhibit any changes in segmentation accuracy, considering both the Dice coefficient and NSD, across any of the three comparisons, including same race versus non-same race, same sex versus non-same sex, and the same sex and race versus non-same sex and race.

Although some models perform best on the subgroup they were trained on, there are several instances where they perform better on a different subgroup. For example, when segmenting the right Nucleus Accumbens (NAc), the UNesT model trained on the white female (WF) dataset achieves its highest average Dice score (0.83) on the black female (BF) test set, which is higher than its performance on the WF test set (0.81). The same case can be found in NSD results; For instance, nnU-Net trained on white male yields an NSD of 0.56 on black female subjects versus 0.54 on white male subjects. Detailed results for all models can be found in the supplementary material.

2.4.3 Impact of biased segmentation on morphometric analyses

We evaluated the influence of sex, race, and their interaction on the right NAc and left NAc volumes with a linear mixed model using manual segmentation and the different models. As

Table 2.4 Segmentation performance metrics (DSC (overall Dice coefficient across all test sets), ESSP, Δ) for right and left NAc across different models and training groups. **ESSP** (Equity-Scaled Segmentation Performance) combines overall accuracy with a penalty for cross-group disparities; **higher is better** (↑). Δ quantifies differences of each demographic group from the overall mean; **lower is better** (↓)

Structure	Train	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

Table 2.5 Segmentation performance metrics (NSD (overall NSD coefficient across all test sets), ESSP, Δ) for right and left NAc across different models and training groups. **ESSP** (Equity-Scaled Segmentation Performance) combines overall accuracy with a penalty for cross-group disparities; **higher is better** (↑). Δ quantifies differences of each demographic group from the overall mean; **lower is better** (↓)

Structure	Train	nnU-Net			CoTr			ANTs			UNesT		
		NSD	ESSP (↑)	Δ (↓)	**NSD**	ESSP (↑)	Δ (↓)	**NSD**	ESSP (↑)	Δ (↓)	**NSD**	ESSP (↑)	Δ (↓)
Right NAc	WM	0.527	0.483	0.090	0.512	0.469	0.090	0.430	0.405	0.060	0.428	0.387	0.105
	WF	0.527	0.492	0.070	0.525	0.468	0.120	0.432	0.412	0.050	0.407	0.382	0.064
	BM	0.517	0.457	0.070	0.510	0.455	0.120	0.380	0.316	0.200	0.392	0.341	0.1500
	BF	0.529	0.500	0.060	0.525	0.486	0.080	0.422	0.364	0.1600	0.387	0.357	0.084
Left NAc	WM	0.538	0.511	0.052	0.515	0.500	0.031	0.419	0.411	0.020	0.428	0.387	0.106
	WF	0.522	0.495	0.055	0.517	0.495	0.044	0.424	0.416	0.020	0.404	0.371	0.089
	BM	0.517	0.474	0.090	0.505	0.459	0.100	0.395	0.338	0.170	0.392	0.341	0.1500
	BF	0.539	0.509	0.060	0.509	0.472	0.079	0.387	0.337	0.150	0.392	0.358	0.094

shown in Table 2.8, sex and race effects can be observed in the full manual dataset (includes training and test data) on both sides, whereas in the smaller test datasets the sex effect on the right NAc volume loses significance (p=0.057). No sex-by-race interaction was observed. When turning to the automated biased models, one can observe a similar sex effect for all models (Table 2.9). The race effect, however, disappears for all models except for CoTrBF in the Left NAc (P-value=0.04). No automated method identified a sex-by-race interaction, which is in line with the manual segmentation results. Detailed tables with the race (γ_1) and sex-by-race

Table 2.6 Effects of Same Sex, Same Race, and Interaction on Dice coefficient for right and left NAc. β_1, β_2, and β_3. are the coefficients for a fixed factor term such as sameSex that describes the effect of the factor level on the Dice coefficient. Std Err is the standard error of the coefficient estimates

Structure	Model	Same Sex			Same Race			Same Race × Same Sex		
		β_2	Std Err	P-value	β_1	Std Err	P-value	β_3	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**

Table 2.7 Effects of Same Sex, Same Race, and Interaction on NSD for right and left NAc. β_2, β_1, and β_3 are the coefficients for the fixed-effect terms Same Sex, Same Race, and their interaction (Same Race × Same Sex), respectively. "Std Err" denotes the standard error of the coefficient estimates

Structure	Model	Same Sex			Same Race			Same Race × Same Sex		
		β_2	Std Err	P-value	β_1	Std Err	P-value	β_3	Std Err	P-value
Right NAc	ANTs	-0.001	0.010	0.889	0.040	0.010	**0.000**	0.003	0.015	0.818
	CoTr	0.007	0.007	0.353	0.009	0.007	0.248	0.008	0.010	0.467
	nnU-Net	-0.005	0.008	0.533	0.005	0.008	0.498	0.017	0.011	0.129
	UNesT	0.014	0.009	0.105	0.012	0.009	0.174	0.022	0.013	0.077
Left NAc	ANTs	-0.003	0.009	0.717	0.039	0.009	**0.000**	-0.001	0.010	0.907
	CoTr	0.002	0.007	0.776	0.004	0.007	0.522	0.018	0.010	0.061
	nnU-Net	0.006	0.008	0.436	0.009	0.008	0.259	0.010	0.011	0.392
	UNesT	0.010	0.010	0.298	0.017	0.010	0.070	0.031	0.013	**0.021**

interaction (γ_3) factors can be found in the supplementary material. In summary, the sex effect observed in the manual segmentation remains even in the most biased models, whereas the race effect observed in the manual segmentation generally cannot be observed when segmentation is performed by highly biased models.

Table 2.8 Results for evaluating sex, race, and race × sex effects on volumes by manual annotation 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

Structure	Manual	Sex			Race			Race × Sex		
		γ_2	Std Err	P-value	γ_1	Std Err	P-value	γ_3	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

Table 2.9 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		
		γ_2	Std Err	P	γ_2	Std Err	P	γ_2	Std Err	P	γ_2	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**	259.2	78.5	**0.001**	184	65.3	**0.005**	256	77.8	**0.001**
	nnU-Net	231.1	71.5	**0.001**	202.4	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

2.4.4 Impact of Dataset Size, Demographics, and Atlas Selection on Bias in ANTs and UNesT

In order to better understand potential sources of bias in ANTs and UNesT, we performed two additional sets of experiments. First, we exactly matched training dataset sizes to n=30 to rule out dataset size as a source of bias. Second, we established baseline settings where training data had a balanced representation of each subgroup.

For the equal sample size experiment, we mimicked the same design as above but only included 30 subjects per biased training set for training UNesT. The race bias effect in the left NAc was statistically significant in both NSD and Dice coefficient results. The results can be found in

Table 2.10. Notably, the β factors are very similar to those observed in the original experiments with unequal sample sizes (Tables 2.6 and 2.7).

For the UNesT baseline datasets, the first training set comprised 30 subjects with balanced demographics. Five-fold cross-validation was conducted using five folds of size six including all subgroups and two extra subjects of both of the white and black races so that the races are split evenly. In the second baseline experiment, UNesT was trained on 120 training subjects comprising 30 subjects from each subgroup. We evaluated two ANTs baselines. The first one was a balanced baseline using 10 atlases: eight atlases (two per subgroup: black female, black male, white female, white male) plus two additional atlases (one black, one white) to preserve race balance. The second baseline with 40 atlases was composed of the exact 10 atlases from each subgroup that were used in the original biased ANTs variants. We compared the performance of the baseline models with the biased models.

Table 2.11 shows that the UNesT Baseline 120 model, trained on the largest and demographically diverse dataset, is the top performer, achieving the best accuracy and ESSP in the Dice coefficient and NSD. Increasing the size of a balanced training set effectively reduces bias, as evidenced by Δ for the right NAc dropping from 0.02 (30 subjects) to 0.01 (120 subjects). The same trend can also be observed in Δ by NSD dropping from 0.07 (30 subjects) to 0.03 (120 subjects). In contrast, models trained on single demographic subgroup data from black subjects consistently ranked last on all metrics. Furthermore, a clear pattern emerges where models trained on data from white subjects outperform those trained on data from black subjects, and Δ is consistently and dramatically higher for models trained on black subjects. Consequently, the ESSP is significantly lower for models trained on black subgroups in both Dice coefficient and NSD. Notably, the tables 2.12 and 2.13 show that the assumption that a model performs best on its matched demographic was not always true; for instance, the UNesTWF model achieved its top Dice coefficient and NSD for the right NAc when tested on the black female (BF) subgroup.

The results in table 2.14, show that ANTs variants using atlases from white subjects achieve higher accuracy and drastically lower Δ than those using atlases from black subjects in both

Dice coefficient and NSD. Surprisingly, simply increasing the size of a diverse atlas set does not guarantee a fairer outcome for this traditional method. While accuracy may improve as we observe with NSD values of Baseline 40 reaching 0.45, performance disparities can worsen, leading to a lower ESSP. This is in contrast to our findings with UNesT baselines, where larger, more diverse datasets typically mitigate bias. The segmentation performance for all the methods in Dice coefficient and NSD are shown in Tables 2.12, 2.13, 2.15 and 2.16.

Table 2.10 Effects of Same Sex, Same Race, and Interaction on new UNesT experiment results for right and left NAc. Left block reports linear mixed-effects coefficients for DSC $\big(\text{DSC} = \beta_0 + \beta_1(\text{SameRace}) + \beta_2(\text{SameSex}) + \beta_3(\text{SameRace} \times \text{SameSex}) + \epsilon\big)$; right block reports the corresponding NSD coefficients $\big(\text{NSD} = \gamma_0 + \gamma_1(\text{SameRace}) + \gamma_2(\text{SameSex}) + \gamma_3(\text{SameRace} \times \text{SameSex}) + \epsilon\big)$. For each factor, the table lists the coefficient, its standard error (Std Err), and the P-value

Structure	DSC (Dice) coefficients									NSD coefficients								
	Same Sex			Same Race			Same Race × Same Sex			Same Sex			Same Race			Same Race × Same Sex		
	(β_2)	Std Err	P	(β_1)	Std Err	P	(β_3)	Std Err	P	(γ_2)	Std Err	P	(γ_1)	Std Err	P	(γ_3)	Std Err	P
Right NAc	0.005	0.004	0.185	0.007	0.004	0.083	0.003	0.006	0.584	0.014	0.009	0.103	0.008	0.009	0.375	0.008	0.012	0.526
Left NAc	0.002	0.004	0.587	0.014	0.004	**0.001**	0.004	0.006	0.546	0.009	0.008	0.231	0.022	0.008	**0.005**	0.007	0.011	0.542

Table 2.11 Segmentation performance metrics and fairness metric based on Dice coefficient for biased UNesT models for right and left NAc. (1 shows best) (**ESSP** (Equity-Scaled Segmentation Performance) combines overall accuracy with a penalty for cross-group disparities; **higher is better** (↑). **Δ** quantifies differences of each demographic group from the overall mean; **lower is better** (↓))

Structure	Train	DSC	ESSP(↑)	Δ(↓)	NSD	ESSP(↑)	Δ(↓)
Right NAc	WM	0.81^2	0.79^2	0.02^2	0.41^2	0.39^2	0.04^2
	WF	0.81^2	0.78^3	0.03^3	0.41^2	0.37^4	0.08^4
	BM	0.80^3	0.76^4	0.04^4	0.37^5	0.34^6	0.09^5
	BF	0.80^3	0.78^3	0.03^3	0.39^4	0.36^5	0.07^3
	Baseline30	0.81^2	0.79^2	0.02^2	0.40^3	0.38^3	0.07^3
	Baseline120	$\mathbf{0.82}^1$	$\mathbf{0.81}^1$	$\mathbf{0.01}^1$	$\mathbf{0.42}^1$	$\mathbf{0.40}^1$	$\mathbf{0.03}^1$
Left NAc	WM	$\mathbf{0.81}^1$	$\mathbf{0.80}^1$	$\mathbf{0.00}^1$	0.40^2	0.39^2	$\mathbf{0.02}^1$
	WF	0.80^2	0.79^2	0.02^3	0.40^2	0.38^3	0.06^3
	BM	0.79^3	0.74^4	0.06^5	0.39^3	0.34^5	0.15^5
	BF	0.79^3	0.75^3	0.05^4	0.38^4	0.34^5	0.11^4
	Baseline30	0.80^2	0.79^2	0.01^2	0.40^2	0.38^4	0.06^3
	Baseline120	$\mathbf{0.81}^1$	$\mathbf{0.80}^1$	0.01^2	$\mathbf{0.42}^1$	$\mathbf{0.40}^1$	0.04^2

Table 2.12 Average and standard deviation of Dice coefficient (Avg±Std) for right and left NAc. Columns: UNesT trained on each subgroup (WM/WF/BM/BF – matched sample sizes) and Baselines. Superscripts rank average within each row (1 shows best)

Structure	Test	UNesTWM	UNesTWF	UNesTBM	UNesTBF	Baseline (30)	Baseline (120)
Right NAc	WM	**0.83**^{1} ± 0.02	0.81^{3} ± 0.02	0.80^{4} ± 0.03	0.80^{4} ± 0.04	0.81^{3} ± 0.04	0.82^{2} ± 0.03
	WF	0.81^{2} ± 0.04	0.81^{2} ± 0.03	0.78^{4} ± 0.04	0.79^{3} ± 0.04	0.81^{2} ± 0.04	**0.82**^{1} ± 0.03
	BM	0.80^{3} ± 0.05	0.81^{2} ± 0.04	0.81^{2} ± 0.03	0.80^{3} ± 0.04	**0.82**^{1} ± 0.03	0.81^{2} ± 0.04
	BF	0.81^{3} ± 0.04	**0.83**^{1} ± 0.03	0.80^{4} ± 0.03	0.82^{2} ± 0.03	0.82^{2} ± 0.03	**0.83**^{1} ± 0.03
Left NAc	WM	**0.82**^{1} ± 0.02	0.80^{3} ± 0.04	0.79^{4} ± 0.04	0.79^{4} ± 0.03	0.80^{3} ± 0.04	0.81^{2} ± 0.03
	WF	0.81^{1} ± 0.03	0.80^{2} ± 0.02	0.77^{5} ± 0.03	0.78^{4} ± 0.03	0.79^{3} ± 0.03	**0.81**^{1} ± 0.02
	BM	0.80^{2} ± 0.05	0.80^{2} ± 0.05	**0.81**^{1} ± 0.03	0.80^{2} ± 0.04	0.80^{2} ± 0.05	**0.81**^{1} ± 0.05
	BF	0.81^{2} ± 0.03	0.81^{2} ± 0.03	0.81^{2} ± 0.03	**0.82**^{1} ± 0.03	0.80^{3} ± 0.05	**0.82**^{1} ± 0.03

Table 2.13 Average and standard deviation of NSD (Avg±Std) for right and left NAc. Columns: UNesT trained on each subgroup (WM/WF/BM/BF – matched sample sizes) and Baselines. Superscripts rank average within each row (1 shows best)

Structure	Test	UNesTWM	UNesTWF	UNesTBM	UNesTBF	Baseline (30)	Baseline (120)
Right NAc	WM	**0.42**^{1} ± 0.05	0.39^{2} ± 0.05	0.37^{4} ± 0.06	0.38^{3} ± 0.08	0.38^{3} ± 0.07	**0.42**^{1} ± 0.05
	WF	0.40^{2} ± 0.08	0.38^{4} ± 0.05	0.33^{6} ± 0.05	0.36^{5} ± 0.07	0.39^{3} ± 0.07	**0.41**^{1} ± 0.07
	BM	0.39^{4} ± 0.08	0.40^{3} ± 0.08	0.41^{2} ± 0.06	0.38^{5} ± 0.08	**0.42**^{1} ± 0.06	0.41^{2} ± 0.08
	BF	0.42^{4} ± 0.07	**0.45**^{1} ± 0.07	0.38^{5} ± 0.08	0.43^{3} ± 0.07	0.42^{4} ± 0.08	0.44^{2} ± 0.07
Left NAc	WM	0.39^{2} ± 0.06	0.39^{2} ± 0.08	0.35^{3} ± 0.08	0.35^{3} ± 0.07	0.39^{2} ± 0.06	**0.40**^{1} ± 0.07
	WF	0.40^{2} ± 0.05	0.40^{2} ± 0.06	0.35^{4} ± 0.04	0.35^{4} ± 0.05	0.38^{3} ± 0.06	**0.42**^{1} ± 0.04
	BM	0.40^{4} ± 0.09	0.39^{5} ± 0.09	**0.43**^{1} ± 0.08	0.38^{6} ± 0.07	0.41^{3} ± 0.08	0.42^{2} ± 0.08
	BF	0.42^{4} ± 0.06	0.44^{2} ± 0.06	0.43^{3} ± 0.07	0.44^{2} ± 0.08	0.41^{5} ± 0.09	**0.45**^{1} ± 0.07

2.5 Discussion

Our investigation highlights how an imbalance in demographic factors such as race and sex influences the segmentation quality and the volumetric measurements of the NAcs. While all models preserved the anatomical trend of larger right NAc volumes compared to the left NAc, aligning with manual segmentations, most models exhibited narrower volume standard deviations than manual annotations. An important finding in this study is that while most models faithfully preserve the sex-based volume differences seen in the manually labeled ground-truth

Table 2.14 Fairness metrics based on Dice coefficient for ANTs and baselines (1 shows the best)(**ESSP** (Equity-Scaled Segmentation Performance) combines overall accuracy with a penalty for cross-group disparities; **higher is better** (↑). **Δ** quantifies differences of each demographic group from the overall mean; **lower is better** (↓))

Structure	Train	DSC	ESSP(↑)	Δ(↓)	NSD	ESSP(↑)	Δ(↓)
Right NAc	WM	**0.82**[1]	**0.79**[1]	0.03[2]	0.43[2]	0.40[2]	0.06[2]
	WF	0.81[2]	**0.79**[1]	**0.02**[1]	0.43[2]	**0.41**[1]	**0.05**[1]
	BM	0.78[4]	0.70[3]	0.11[5]	0.38[6]	0.31[6]	0.20[6]
	BF	0.79[3]	0.72[2]	0.10[4]	0.42[4]	0.36[4]	0.16[5]
	Baseline10	0.78[4]	0.72[2]	0.08[3]	0.40[5]	0.35[5]	0.12[3]
	Baseline40	0.81[2]	0.72[2]	0.12[6]	**0.45**[1]	0.40[2]	0.13[4]
Left NAc	WM	**0.81**[1]	**0.79**[1]	0.02[2]	0.41[3]	**0.41**[1]	**0.02**[1]
	WF	0.80[2]	**0.79**[1]	**0.01**[1]	0.42[2]	0.41[2]	**0.02**[1]
	BM	0.75[5]	0.68[4]	0.10[4]	0.39[4]	0.33[5]	0.17[6]
	BF	0.77[4]	0.70[3]	0.10[4]	0.38[6]	0.33[5]	0.15[5]
	Baseline 10	0.78[3]	0.72[2]	0.07[3]	0.39[4]	0.35[4]	0.14[4]
	Baseline 40	0.80[2]	0.72[2]	0.11[6]	**0.45**[1]	0.39[3]	0.13[3]

Table 2.15 Average and standard deviation of Dice coefficient (Avg±Std) for right and left NAc. Columns: ANTs trained on each subgroup (WM/WF/BM/BF) and Baselines. Superscripts rank average within each row (1 shows best)

Structure	Test	ANTsWM	ANTsWF	ANTsBM	ANTsBF	Baseline (10)	Baseline (40)
Right NAc	WM	**0.82**[1]±0.02	0.81[3]±0.03	0.76[6]±0.05	0.77[5]±0.06	0.78[4]±0.04	0.81[2]±0.04
	WF	**0.81**[1]±0.04	0.80[2]±0.04	0.74[6]±0.04	0.75[3]±0.05	0.74[5]±0.05	0.74[4]±0.18
	BM	0.80[5]±0.04	0.81[2]±0.04	0.81[2]±0.03	0.81[2]±0.05	0.79[6]±0.03	**0.83**[1]±0.03
	BF	0.82[2]±0.04	0.82[2]±0.04	0.80[6]±0.03	0.82[2]±0.04	0.81[5]±0.04	**0.84**[1]±0.03
Left NAc	WM	**0.82**[1]±0.03	0.81[2]±0.03	0.73[6]±0.06	0.76[5]±0.06	0.77[4]±0.05	0.80[3]±0.05
	WF	**0.80**[1]±0.04	0.80[1]±0.03	0.72[5]±0.05	0.72[5]±0.07	0.74[3]±0.05	0.74[4]±0.18
	BM	0.80[3]±0.06	0.80[3]±0.06	0.78[6]±0.06	0.79[5]±0.06	0.80[2]±0.06	**0.82**[1]±0.05
	BF	0.80[3]±0.06	0.80[3]±0.06	0.77[6]±0.03	0.80[3]±0.03	0.80[2]±0.03	**0.83**[1]±0.03

data, race-based differences present in the manually annotated data vanish in all automated biased models.

From a fairness perspective, ESSPs measured by both Dice coefficient and NSD indicate that nnU-Net and CoTr often achieve the highest accuracy, combined with smaller inter-group disparities. In contrast, ANTs is highly sensitive to the race of the training set, with significantly

Table 2.16 Average and standard deviation of NSD (Avg±Std) for right and left NAc. Columns: ANTs trained on each subgroup (WM/WF/BM/BF) and Baselines. Superscripts rank average within each row (1 shows best)

Structure	Test	ANTsWM	ANTsWF	ANTsBM	ANTsBF	Baseline (10)	Baseline (40)
Right NAc	WM	0.44^{2}±0.05	0.43^{3}±0.08	0.35^{6}±0.07	0.38^{5}±0.12	0.38^{4}±0.08	$\mathbf{0.44}^{1}$±0.08
	WF	$\mathbf{0.43}^{1}$±0.07	0.41^{2}±0.07	0.31^{6}±0.06	0.36^{4}±0.06	0.35^{5}±0.06	0.40^{3}±0.06
	BM	0.40^{6}±0.05	0.44^{2}±0.07	0.43^{4}±0.07	0.44^{2}±0.10	0.41^{5}±0.07	$\mathbf{0.47}^{1}$±0.08
	BF	0.45^{3}±0.08	0.45^{3}±0.08	0.43^{6}±0.08	0.46^{2}±0.09	0.44^{5}±0.08	$\mathbf{0.50}^{1}$±0.09
Left NAc	WM	$\mathbf{0.42}^{1}$±0.07	$\mathbf{0.42}^{1}$±0.07	0.35^{6}±0.07	0.36^{5}±0.10	0.37^{4}±0.09	0.41^{3}±0.10
	WF	0.42^{2}±0.07	$\mathbf{0.43}^{1}$±0.06	0.33^{6}±0.06	0.34^{5}±0.08	0.35^{4}±0.06	0.41^{3}±0.07
	BM	0.41^{6}±0.09	0.42^{5}±0.08	0.43^{3}±0.07	0.43^{3}±0.10	0.43^{2}±0.10	$\mathbf{0.48}^{1}$±0.10
	BF	0.43^{3}±0.09	0.43^{3}±0.09	0.42^{5}±0.06	0.42^{5}±0.07	0.43^{2}±0.07	$\mathbf{0.48}^{1}$±0.07

lower Dice coefficient and NSD and larger Δ values measured by both Dice coefficient and NSD for models trained on black subgroups compared to white subgroups. While UNesT outperforms ANTs in terms of delivering higher overall Dice coefficient and NSD and more consistent performance across demographic groups, it has inferior performance compared to nnU-Net and CoTr in achieving consistent accuracy across all demographic groups. The linear mixed model results further show that among the factors influencing segmentation accuracy race-matching between training and test datasets provides a substantial performance benefit for the ANTs and UNesT models. Perhaps surprisingly, sex-matching had far less effect on performance. nnU-Net is the only model whose performance was not influenced by race-matching or sex and race-matching. Our results of a strong race bias effect align with the insights in the study by Ioannou *et al.* (2022b), which reported that the race bias effect was more significant than the sex bias effect. We also compared the performance of methods using NSD and found that the overall ranking of the models remained similar to the Dice results in terms of fairness, with nnU-Net as the top performer and ANTs and UNesT showing the most vulnerability to bias. The magnitude of bias was amplified with NSD. For example, Δ values for ANTs soared to 0.20. For the CoTr model, while it appeared both accurate and fair, its performance disparities were much more pronounced when evaluated with NSD. NSD amplifies boundary-level inequities that the Dice coefficient can conceal. Our analysis also revealed some counterintuitive patterns where some models trained on one demographic subgroup occasionally perform better on another.

For instance, the UNesT model trained on white females (WF) achieved a higher average Dice coefficient on the black female (BF) test set than on its own WF test set.

For the UNesT model, we conducted additional experiments with balanced-size training sets, and the race bias effect was observed in both NSD and Dice coefficient results. Further baseline experiments with demographically balanced datasets suggest that race-balanced, larger datasets significantly mitigate unfairness. The results of similar ANTs baseline experiments with balanced atlases highlight the complex nature of bias and suggest that simply balancing demographic attributes in the atlases does not equate to improving fairness. This is in contrast to our findings with UNesT baselines, where larger, more diverse datasets typically mitigate bias.

These findings can have important implications. Biased segmentation models can misrepresent brain structures. For example, ANTs trained on black males shows substantial under-segmentation of the left NAc, with volumes nearly 28% smaller than manual annotations. This difference can influence clinical applications as right and left NAc can serve as a biomarker. For example, Major depressive disorder has been associated with persistent reductions in NAc volume (Ancelin *et al.*, 2019).

It is worth noting that while the existence of bias in deep learning models is well-established, its manifestation within brain segmentation can be subtle and highly variable across different methods. This variability necessitates quantitative assessment of fairness in neuroimaging studies. Our findings corroborate this, demonstrating that bias can be pronounced in certain architectures like UNesT, yet not observable in frameworks such as nnU-Net. In brain MRI segmentation, prior bias studies typically assess a single deep-learning model, using low-quality labels as ground truth for training datasets (Ioannou *et al.*, 2022b). Outside the brain, multi-model comparisons have been reported (Lee *et al.*, 2023). Our study is, to our knowledge, the first in brain MRI to compare deep learning and a classical atlas-based method with respect to bias.

One of the challenges of our approach is that it is difficult to pinpoint the source of the observed biases. Indeed, we chose to evaluate how models perform, as recommended “off-the-shelf" by their creators, as our goal was to investigate bias in methods that are not just standalone

architectures, but complete pipelines, each with hundreds of parameters and author-recommended configurations that are integral to their performance. This is in contrast to using a very restricted framework in which we changed only a few parameters and made an inference about the source of these biases. While pinpointing the source of bias in segmentation models is complex, we can compare the methods' inherent characteristics. The ANTs method is vulnerable to bias because its weighted-voting strategy produces systematic errors when the atlas set can be dominated by a single demographic's anatomy. Conversely, nnU-Net's superior fairness likely stems from its adaptive data augmentation, which forces the model to learn generalizable anatomical patterns instead of demographic-related features. We hypothesize that UNesT, despite its powerful transformer architecture, was more susceptible to overfitting on the demographic traits of its small training set due to a lack of such rigorous data augmentation.

A limitation of this work is its relatively small dataset size within each demographic subgroup, which may restrict generalizability and more nuanced biases. Moreover, while our study focused primarily on adult populations, biases may appear differently in children or older people. Additionally, our study focused solely on the right and left nucleus accumbens which are small structures. Future research should examine a broader range of structures and also other datasets to confirm whether this trend persists. We also acknowledge that our study is restricted to a single structure. Further study will need to extend bias evaluation beyond the NAc. The main challenge here is the lack of benchmark datasets with ground truth segmentations across the brain. Although our study focused solely on the effects of demographic attributes such as sex and race, we recognize that these are not the only potential sources of bias within a dataset. Other factors can also have a considerable impact on brain volumes and model segmentation performance. For instance, the HCP Young Adult dataset includes participants aged 22 to 35 and therefore excludes other age groups, such as children and older adults. As noted in our limitations, biases may manifest differently in children or older adults. In addition, the HCP dataset comprises only healthy subjects, which means that the findings may not generalize to individuals with psychiatric disorders, congenital abnormalities in brain anatomy, or other clinical conditions. Furthermore, social status and education level can also be confounding

factors that are not considered in our study. Technical biases stemming from variations in scanner hardware, software, or acquisition parameters also introduce systematic variations. Additionally, although we intentionally designed our study to isolate demographic effects, this approach does not reflect real-world scenarios where training sets are more heterogeneous, potentially amplifying the observed biases.

Finally, our study is diagnostic in nature and does not test potential bias mitigation strategies suitable for unbalanced data regimes. For instance, sensitive class-aware data augmentation (Xu *et al.*, 2024b) could be employed. This technique involves applying more aggressive data augmentation to underrepresented demographic groups within the training set, thereby encouraging the model to learn more robust and generalizable features that are not dependent on sensitive attributes. To address subgroup imbalances, data synthesis can be utilized. This approach leverages generative models to augment the training dataset with synthetic data for the minority class, ensuring a more balanced distribution for model training (Pombo *et al.*, 2023). Incorporating such prescriptive strategies in future work is a critical next step toward developing segmentation tools that are not only accurate but also fair across diverse populations. Finally, evaluating changes in bias under different network architectures or training procedures could inform best practices for equitable brain MRI segmentation.

2.6 Conclusion

This study provides insights into demographic biases in brain MRI segmentation. Our results show a nuanced picture with different methods displaying different levels of sensitivity to demographic biases in their training data. ANTs and UNesT were most affected while nnU-Net seemed to be the most robust to biases. In terms of the relative importance of demographic variables, race seemed to impact segmentation performance more than sex, and most models show a lower overall segmentation accuracy and ESSP in both Dice coefficient and NSD when trained on datasets from black demographic groups than those trained on white demographic groups. Additionally, we found that these performance biases impact morphometric studies. Notably, a race effect on NAc volumes was observed with manual segmentations, but was not

observed with automated methods trained with biased models. These findings underscore the need for diverse training sets and rigorous model assessments across multiple demographic strata to achieve truly equitable and clinically reliable brain MRI segmentation. Finally, our study remains limited in scope as our results are based on studying one anatomical structure and a single dataset. Further research is required to conduct a more comprehensive investigation to determine whether these results are generalizable across diverse structures and datasets.

CHAPTER 3

EXPLORING ENTROPY-BASED ACTIVE LEARNING FOR FAIR BRAIN SEGMENTATION

Ghazal Danaee[1] , Mélanie Gaillochet[1,2,3] , Christian Desrosiers[1] , Hervé Lombaert[1, 2, 3] , Sylvain Bouix[1]

[1] Département de génie logiciel et des technologies de l'information, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Polytechnique Montréal, 2500, chemin de Polytechnique, Montréal, Québec, Canada H3T 0A3

[3] Mila - Quebec AI Institute, 6666 Rue Saint-Urbain, Montréal, Québec, Canada H2S 3H1

The article was accepted to the conference « Medical Imaging with Deep Learning (MIDL) » in February 2025.

3.1 Introduction

Active learning (AL) has become a key strategy for addressing the problem of annotation in medical image segmentation. The success of deep learning models for segmentation relies heavily on high-quality data labeled voxel by voxel by experts, which is time-consuming and labor-intensive to obtain. Active learning targets this bottleneck by treating annotation as a limited resource. The process begins with a small labeled set and a large unlabeled pool. At each iteration, a model is trained on the current labeled data, an informativeness score is computed for each unlabeled sample, and a small batch of high-scoring samples is selected for expert annotation and added to the training set (Budd *et al.*, 2021). This loop repeats until the labeling budget is exhausted. AL can substantially reduce annotation effort while maintaining segmentation accuracy, especially in the low-label regime (Camilleri, Wagenmaker, Morgenstern, Jain & Jamieson, 2024b).

However, medical image segmentation poses specific challenges for active learning. Unlike natural image datasets, medical image annotation cannot be easily crowdsourced; annotators must have substantial expertise, and privacy concerns further constrain data sharing (Wang *et al.*, 2024). The task is intrinsically high-dimensional, and naive uncertainty-based sampling tends to

select many highly similar, redundant images or outliers (Munjal, Hayat, Hayat, Sourati & Khan, 2022). Representative-based and hybrid AL strategies mitigate this by encouraging diversity, but since they require computing distances or distributions in a learned feature space, they can be computationally expensive (Gaillochet, Desrosiers & Lombaert, 2023). Theoretically, an advantage of active learning could be the mitigation of bias. In a simulated fraud-detection task, Weerts, van der Waa, van den Bosch & van der Aalst (2023) showed that standard uncertainty-based active learning can mitigate selection bias and improve fairness even without a fairness-specific design. By querying uncertain samples, the model explored underrepresented groups, reducing false positive disparities and yielding fairer predictions as a side effect.

Beyond pure performance, a critical emerging concern is fairness. Fairness in segmentation entails ensuring that the quality of the segmentation is comparable across groups defined by sensitive attributes, including race and sex. Although most of the fairness evaluation work in medical imaging has focused on classification (Mehrabi *et al.*, 2022), recent segmentation studies show clear demographic disparities. In cardiac MR and orthopedic imaging, models trained on racially imbalanced data exhibit significantly different performance across racial groups (Puyol-Antón *et al.*, 2021, 2022; Lee *et al.*, 2025; Siddiqui *et al.*, 2024). Similar effects have been reported for prostate and skin-lesion segmentation, where race and skin-tone imbalance in training data leads to reduced performance on black patients and darker skin types (Alqarni *et al.*, 2024; Benčević *et al.*, 2024). In brain MRI, Ioannou, Chockler, Hammers & King (2022a) demonstrated that FastSurferCNN exhibits region-specific sex and race biases, noting that race-related disparities can exceed sex-related ones. Furthermore, our prior work showed that race matching between training set and test sets can substantially improve performance for some architectures but not others when segmenting the nucleus accumbens (Danaee *et al.*, 2025). These findings build on evidence that anatomical differences across sex and racial groups shape brain volumes and model behavior (Frazier *et al.*, 2008; Dibaji *et al.*, 2024; Isamah *et al.*, 2010). Fair segmentation, therefore, requires demographically balanced training cohorts, and equity-aware metrics such as Equity-Scaled Segmentation Performance (ESSP) (Tian *et al.*, 2024).

To our knowledge, fair active learning for segmentation in a single domain remains unexplored. Wang *et al.* (2025) addressed a related but distinct problem: fairness in cross-domain medical image segmentation. Their method leveraged CLIP Radford *et al.* (2021) to encode target-domain images and sensitive attributes. It also introduced an attribute-aware sampling strategy, coined FairAP, that enforces balanced annotation quotas across subgroups and selects representative samples in the VLM latent space.

Our contribution: In this work, we address group-wise fairness within the AL acquisition process for brain MRI segmentation. We introduce a novel weighted entropy strategy which modulates voxel-wise uncertainty with group-specific performance weights. These weights are derived from the current Dice score on the labeled set to prioritize samples from under-performing groups. To ensure the acquisition score reflects true epistemic uncertainty and not anatomical variance, we compute a scaled entropy within a dilated region of interest (ROI) mask. The scaled entropy focuses on boundary uncertainty and prevents larger structures from systematically biasing the selection process across demographic groups.

3.2 Related works

Active Learning in Segmentation. We next summarize recent AL approaches for medical image segmentation. Atzeni *et al.* (2022) targeted expected dice gain per unit of manual contour length (tracing effort) for histology, and suggested selecting a specific region of interest (ROI) in one of the images for manual delineation. Kim *et al.* (2024) used image-level uncertainty with redundancy control for brain tumors, while Boehringer, Sanaat, Arabi & Zaidi (2023) prioritized the most difficult BraTS cases while pseudo-labeling the easier ones. Additionally, Qu, Jin, Fu, Wang & Song (2024), with their DifABAL, selected a compact, representative labeled core in a diffusion-learned latent space. To address the issue of redundancy in uncertainty sampling, Gaillochet *et al.* (2023) proposed active learning with stochastic batches. This simple but powerful add-on leverages randomness by generating batches of samples randomly and choosing the batch with the highest mean uncertainty, effectively improving diversity without complex computations. **Fair Active Learning.** While AL for segmentation is well-studied,

previous studies regarding fair active learning focus mainly on classification tasks. Anahideh, Asudeh & Thirumuruganathan (2022) introduced an expected fairness metric to estimate the impact of each unlabeled sample on group-wise disparity. Their acquisition strategy prioritizes high-entropy instances while favoring those expected to reduce unfairness. Similarly, Yang *et al.* (2023) attempted to balance model utility and fairness by querying informative instances for both class and group labels and utilizing a sensitive learner to infer missing attributes. Wang, Du, Liu, Zou & Hu (2022) sought to improve classifier fairness by penalizing differences in true positive/false positive rates between groups and requesting more annotated samples from the worst-performing group. More recently, Fajri, Saxena, Pei & Pechenizkiy (2024) applied fair k-means to the most uncertain points to obtain clusters that reflect the overall group distribution. They then selected candidates across clusters using a composite score that combines uncertainty and representativeness. Pang *et al.* (2024) aimed to mitigate unfairness among groups without compromising accuracy, using group annotations only on a validation set to preserve privacy. They achieved this by evaluating the impact of each new case on validation accuracy and fairness through an expected risk analysis. Their overall goal was to construct a fairness-aware dataset after active sampling.

3.3 Method

3.3.1 Data

To have control over the level of unfairness in the data, we generated synthetic T1-weighted brain MR images using the SimBA framework (Stanley, Wilms & Forkert, 2023). In this framework, images are derived by applying non-linear diffeomorphic transformations sampled from a learned space of deformations to a template image. The global deformations can be used to mimic "regular" anatomical variation and localized deformations can be utilized to show localized "bias or disease" effects. The deformation(s) applied to each case is unique and controlled by sampling from a principal component (PC) representation of deformation fields. For our experiment, we combined the global transformation with an additional localized deformation in

the left caudate. We denote as Group 1 the cases generated with both the localized deformation and the global deformation, and as Group 2 the cases generated with the global deformation only. The amount of localized deformation used to have a bias effect is varied by scaling the first component of the PC representation by a scalar sampled from $\mathcal{N}(\mu, \sigma)$. This procedure enabled us to construct two bias-strength conditions across Groups 1 and 2: the "strong bias" dataset with $\mu = 4$ and $\sigma = 2$ and the "weak bias" dataset with $\mu = 2$ and $\sigma = 2$. Ultimately, the weak bias dataset comprised 312 T1-weighted MRIs, including 156 cases exhibiting the bias effect and 156 cases without it. We generated a strong bias dataset of the same size (312 images), likewise balanced between biased and non-biased cases (156 each). All images had the resolution of $170 \times 170 \times 76$ voxels with an isotropic voxel spacing of 1mm.

For our experiment, we combined the global transformation with an additional localized deformation in the left caudate. We label the cases with the localized deformation Group 1 and the cases with only global deformation as Group 2.

3.3.2 Weighted localized entropy

We introduce two main modifications to the naive use of entropy. First, we limit the computation of entropy within a mask around the ROI. For each unlabeled candidate volume, let $\mathcal{R}$ denote the ROI and $H(v)$ the voxel-wise predictive entropy at voxel v. We define a masked, scaled entropy:

$$\hat{H} = \frac{1}{|\mathcal{R}|} \sum_{v \in \mathcal{R}} H(v), \tag{3.1}$$

which averages uncertainty over the ROI while normalizing by the region size. This reduces the influence of trivial anatomical volume differences on the overall uncertainty score. In our case, $\mathcal{R}$ is a dilated mask around the predicted segmentation of the left caudate.

Second, a group-aware weighting scheme that re-weights the scaled entropy with group-specific weights, thereby prioritizing samples from groups on which the model underperforms. For each group g, we first compute a standardized performance score based on the Dice coefficient (DSC) on the labeled set: $z_g = \frac{\overline{\text{DSC}}_{\text{all}} - \overline{\text{DSC}}_g}{\sigma_{\text{all}}}$, where $\overline{\text{DSC}}_g$ denotes the mean dice for group g, $\overline{\text{DSC}}_{\text{all}}$ is

the mean dice computed over all labeled set, and σ_{all} is the corresponding standard deviation across the labeled set. Thus, groups with worse segmentation performance (lower $\overline{\text{DSC}}_g$) yield larger z_g. We then transform these standardized scores into normalized group weights via a softmax: $w_g = \frac{\exp(z_g)}{\sum_j \exp(z_j)}$. The final acquisition score used for selection is then given by

$$\text{score}(x_g) = w_g \cdot \hat{H}, \tag{3.2}$$

so that uncertainty is explicitly re-weighted toward groups with relatively poorer DSC, ensuring that active learning focuses on groups where the model is currently less reliable.

3.4 Experiment and Results

3.4.1 Implementation details

3.4.1.0.1 Model architecture and training setup.

We have summarized detailed information about the network, active learning setup, and test data in Table 3.1.

Table 3.1 Summary of the network training setup, test data, and the active-learning configuration. Group 1 denotes cases with an additional localized deformation in the left caudate, while Group 2 contains only global deformation

Network	**Test data**	**Active learning**
3D U-Net (GroupNorm, 3D convolution, and ReLU blocks, with a sigmoid activation, 2-level configuration with feature maps of size 8, 16) **Optimizer:** Adam **Epochs:** 200 **Learning rate:** 10^{-4} **Loss:** binary cross-entropy	**Test sets:** **Group 1:** $N = 30$ **Group 2:** $N = 30$ **Combined (Group 1 ∪ Group 2):** $N = 30$ (15 from Group 1, 15 from Group 2)	**Strategies:** • Random sampling • Mean entropy sampling • Localized entropy sampling • Weighted localized entropy sampling **AL schedule:** 5 full AL cycles (10–86 labeled) **Batch size:** $b = 4$

3.4.1.0.2 Baseline Experiments.

We first establish baseline fairness by training our model under three training set compositions. The first one is a balanced cohort comprising 63 cases from Group 1 and 63 from Group 2, the second one is exclusively made of 126 Group 1 cases, and the last one is a Group 2-only cohort of 126 cases. For both the strong and weak bias datasets, the test set comprised 30 cases, including 15 cases from Group 1 and 15 cases from Group 2.

3.4.1.0.3 Active Learning.

In all AL experiments, we start with 10 labeled data, select 4 new samples to label at each AL iteration, and perform five complete AL cycles (from 10 to 86 labeled images). We investigate three initial scenarios: (*i*) a balanced initial dataset with 5 Group 1 and 5 Group 2 labeled images, (*ii*) an unbalanced initial dataset with 8 Group 1 and 2 Group 2 labeled images, and (*iii*) an unbalanced initial dataset with 2 Group 1 and 8 Group 2 labeled images.

Table 3.2 Training data configuration by bias strength. Group 1 denotes cases with an additional localized deformation in the left caudate, while Group 2 contains only global deformation

Bias strength	**Training data**
Strong bias	**Initial labeled set:** $N_0 = 10$ **Initial group proportions (Group 1/Group 2):** • 50/50 • 80/20 • 20/80
Weak bias	**Initial labeled set:** $N_0 = 10$ **Initial group proportions (Group 1/Group 2):** • 50/50 • 80/20 • 20/80

We implemented and tested four different AL strategies on the same test set used in the baseline experiments. We compared our fairness-weighted localized entropy sampling with random

sampling (RS), mean entropy sampling, and localized entropy sampling (eq.(3.1)). We report DSC, ESSP, and Δ for each experiment.

3.4.2 Evaluation metrics

We used DSC to evaluate raw performance. Furthermore, to evaluate fairness in the model's results, we utilized the Equity-Scaled Segmentation Performance (ESSP) metric, originally proposed by Tian *et al.* (2024). Given $\mathrm{DSC}_{overall}$, the average DSC over all cases, and DSC_g, the average DSC of group g, we first define Δ as the sum of absolute performance discrepancies across all groups:

$$\Delta = \sum_{g \in G} \left| \mathrm{DSC}_{overall} - \mathrm{DSC}_g \right|. \tag{3.3}$$

ESSP is then computed by penalizing the overall performance with Δ:

$$\mathrm{ESSP} = \frac{\mathrm{DSC}_{overall}}{1 + \Delta}. \tag{3.4}$$

In essence, ESSP acts as a substitute for DSC, with a penalty for unfairness.

3.4.3 Baseline experiments

In the baseline experiments, we use all the training data available to get baseline Dice score (DSC), ESSP and Δ. Results are shown in Table 3.4. Surprisingly, the balanced dataset does not lead to the best ESSP. Instead, a model trained exclusively on Group 1 (the deformed dataset) leads to better ESSP than the other scenarios. The baseline experiments characterize how different training cohort compositions (balanced, Group 1-only, Group 2-only) affect overall accuracy. They also quantify equity using ESSP and Δ in the absence of active selection. This provides a reference point to assess whether subsequent AL strategies offer fairness gains beyond what simple cohort design can achieve.

Table 3.3 Left caudate segmentation performance (DSC) stratified by training cohort and evaluated on Group 1, Group 2, and pooled test sets of the strong bias and weak bias datasets. Group 1 denotes cases with an additional localized deformation in the left caudate, while Group 2 contains only global deformation. The size of the training set is written in parentheses

Training	**Test**				
	DSC(G1)	**DSC(G2)**	**DSC(G1∪G2)**	**ESSP**	Δ
Strong Bias					
Pooled (63 G1 + 63 G2)	0.88	0.93	0.91	0.91	0.04
Group 1 (126)	0.89	0.88	0.89	0.89	0.01
Group 2 (126)	0.75	0.93	0.84	0.71	0.18
Weak Bias					
Pooled (63 G1 + 63 G2)	0.90	0.91	0.90	0.89	0.01
Group 1(126)	0.90	0.90	0.90	0.90	0.002
Group 2 (126)	0.86	0.91	0.88	0.84	0.04

Table 3.4 Right and left nucleaus accumbens segmentation performance (DSC) stratified by training cohort and evaluated on black and white subjects. The size of the training set is written in parentheses

Training	**Test**				
	DSC(W)	**DSC(B)**	**DSC(W∪B)**	**ESSP**	Δ
Right accumben					
Pooled (38 W + 38 B)	0.770	0.726	0.748	0.716	0.04
Black (76)	0.758	0.729	0.744	0.723	0.02
White (76)	0.766	0.733	0.749	0.727	0.03
Left accumben					
Pooled (38 W + 38 B)	0.790	0.724	0.757	0.714	0.06
Black(76)	0.749	0.725	0.737	0.719	0.02
White (76)	0.799	0.728	0.764	0.714	0.07

3.4.4 Active learning experiments

Results for the experiments starting from a balanced dataset (5 samples from each group) are shown in the first row of Fig. 3.1 (ESSP). The 80/20 and 20/80 Group 1/Group 2 initializations

are shown in rows 2 and 3, respectively. All methods start from the same training set and thus exhibit identical performance at the first cycle (10 labeled). We also show Δ and DSC curves for the strong bias experiment (Fig. 3.2). Results are very similar for the weak bias experiment and can be found in the appendix. Overall, weighted localized entropy outperforms all methods in terms of ESSP in all scenarios, followed closely by localized entropy, then random sampling. Global entropy performs worse than all methods by a relatively large margin. In terms of raw performance as measured by DSC (Fig. 3.2), random sampling is often the best strategy, especially at the early stages of AL. However, this naive strategy fails to perform as well in reducing Δ effectively, compared to localized entropy and the proposed weighted localized entropy (Fig. 3.2). The significant lowering of Δ by weighted localized entropy, while still remaining highly competitive in terms of DSC, allows it to achieve the top ESSP scores across most experiments.

3.4.4.0.1 Selection dynamics and group composition.

As demonstrated in the baseline experiments (Table 3.4), the best ESSP is likely achieved by over-representing Group 1 in the training dataset. This is illustrated in Fig. 3.3, where, under all scenarios, the weighted localized entropy consistently favors adding Group 1 samples to the training dataset. One can also observe a link between localized entropy and fairness as this strategy also tends to select samples from Group 1, even though it does not explicitly account for fairness. Random sampling behaves as expected, balancing data 50/50 over time, while global entropy behaves counterintuitively by adding more samples from Group 2.

3.5 Discussion

In this work, we investigated the intersection of active learning and fairness in medical image segmentation. We designed an acquisition algorithm to improve group-wise fairness rather than solely optimizing accuracy. The proposed Weighted Localized Entropy consistently achieved the strongest equity-aware performance across initialization regimes and bias strengths. Unsurprisingly, localized entropy also reduced bias, although applying the group-fairness

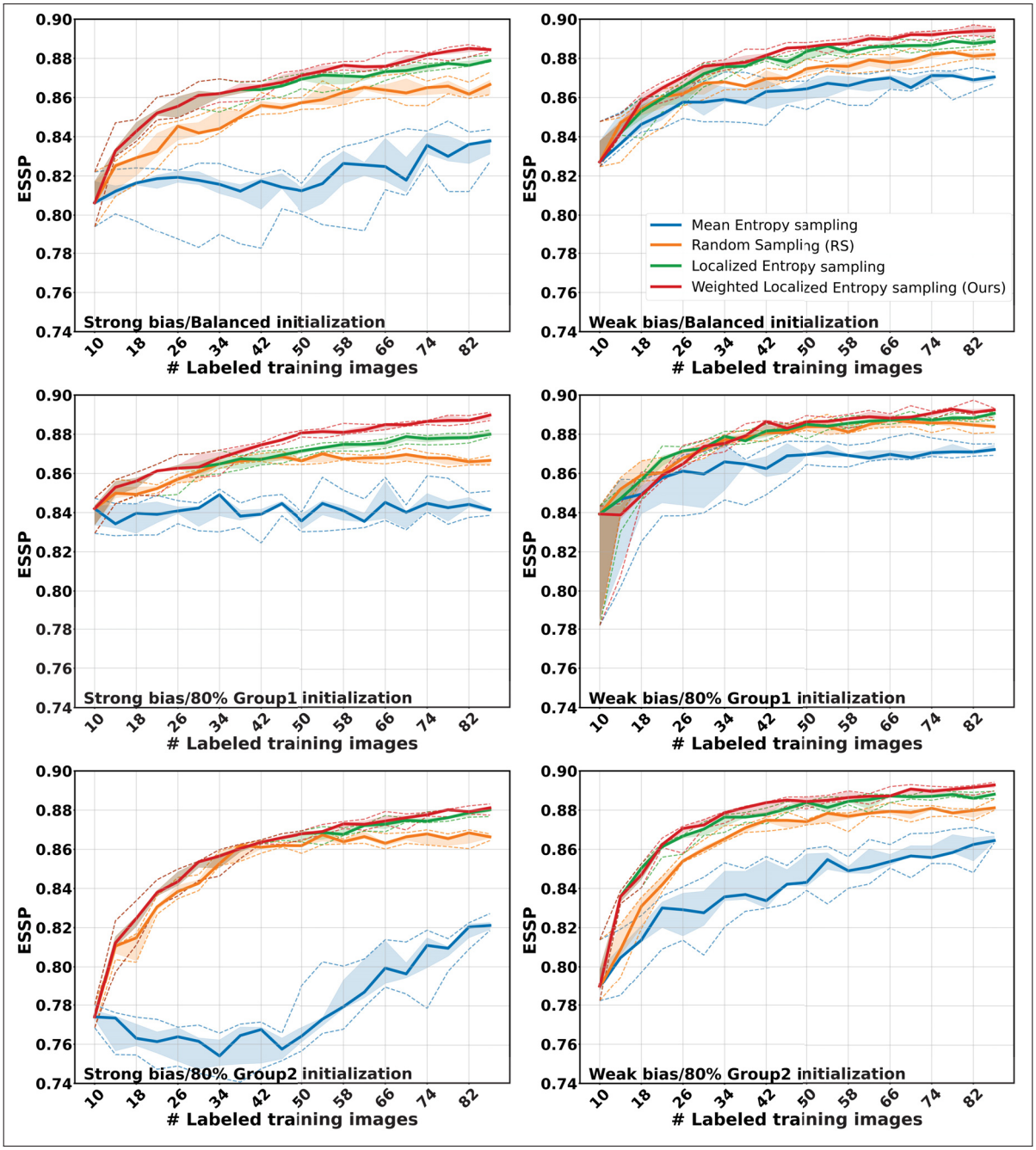

Figure 3.1 **ESSP** under different initial training set compositions. First row: balanced initialization, second row: 80/20 Group 1/Group 2 ratio, third row: 20/80 ratio. Left column: strong bias dataset, Right column: weak bias dataset

weights further yielded improvement in both fairness and accuracy. We note that random sampling usually acquired the best accuracy results, but the equity-scaled performance was

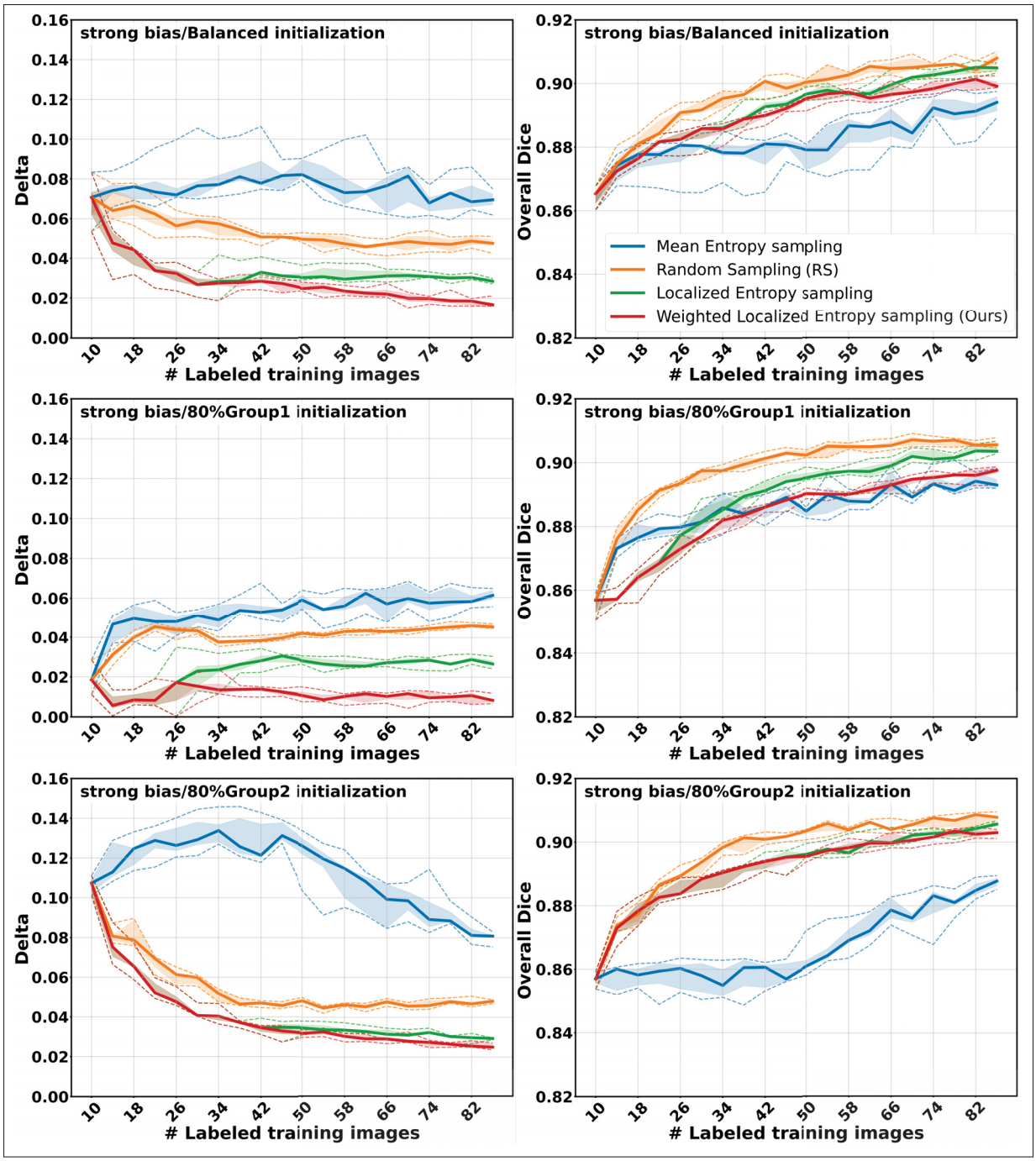

Figure 3.2 **Δ** and **DSC** metrics under different initial training set compositions for the strong bias experiment only. First row: balanced initialization, second row: 80/20 Group 1/Group 2 ratio, third row: 20/80 ratio. Left column: Δ, Right column: DSC

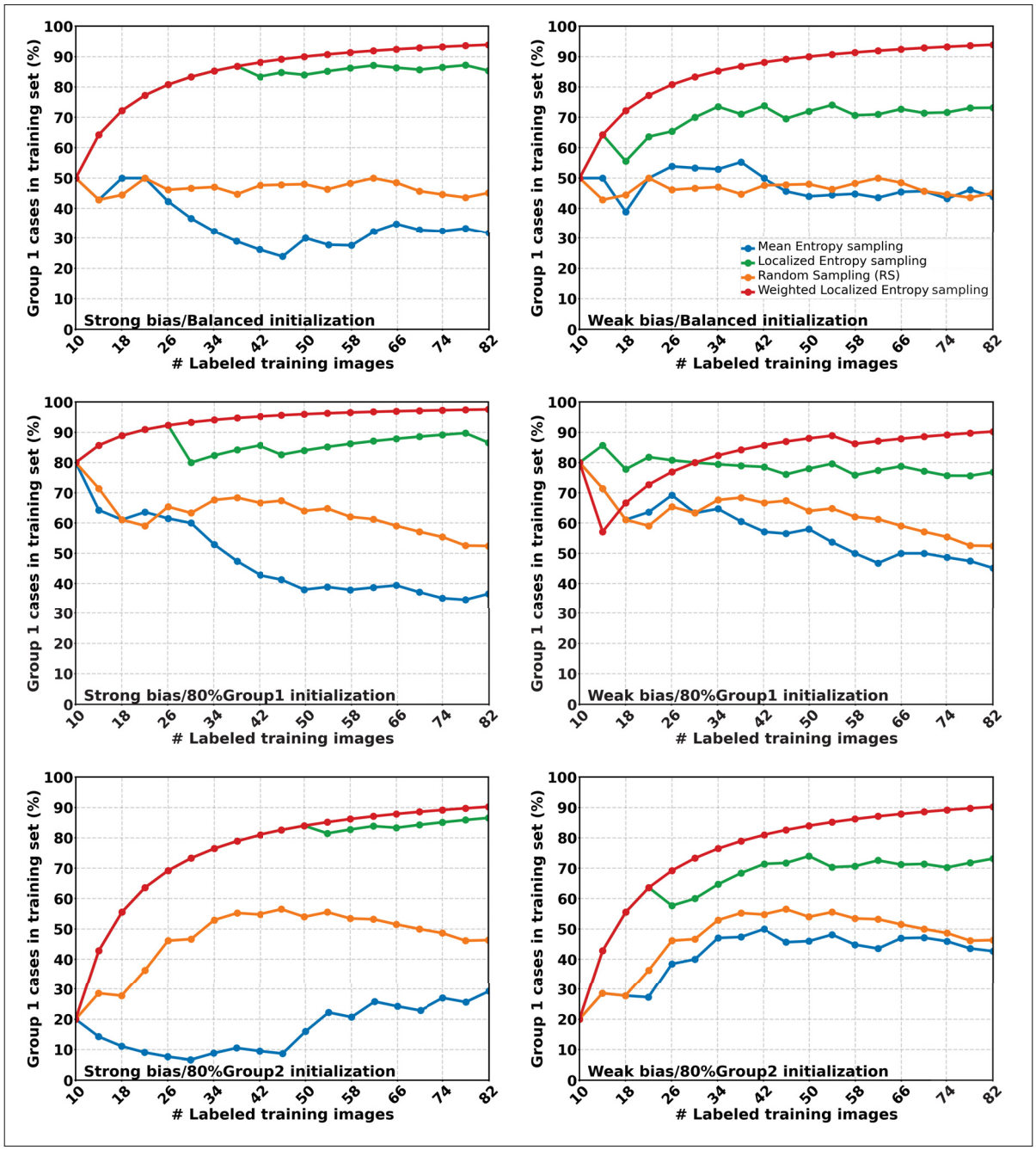

Figure 3.3 **Group 1 ratio** in the training set after sampling for each cycle under different initial training set. First row: balanced initialization, second row: 80/20 Group 1/Group 2 ratio, third row: 20/80 ratio. Left column: strong bias dataset, Right column: weak bias dataset

harmed by the substantial group-wise performance disparity. Global entropy frequently became the worst-performing method in terms of accuracy and fairness as it tended to add more Group 2 cases over time, despite Group 1 being the morphology-challenged subgroup. This behavior can be because whole-volume uncertainty may be dominated by structural extent and incidental variability rather than meaningful model confusion at the target boundary. By computing entropy inside an ROI mask and normalizing by the ROI size (Eq. 3.1), the localized entropy better isolates the uncertainty.

The proposed method demonstrated robustness in both strong and weak bias scenarios. In the strong bias setting, we observed a reduction in Δ of approximately 75% (0.0176 vs. 0.0692) relative to standard entropy at the final cycle. Notably, in the weak bias setting in which morphological differences are harder to detect, our method was even more effective relative to baselines, reducing disparity by approximately 86%. This indicates that the weighted entropy signal is sensitive enough to detect and correct minor performance drifts.

As it was observed in Table 3.4, the relationship between training composition and equitable performance is non-trivial. In the weak bias dataset, the Group 1-only configuration became the best fairness baseline (DSC=0.90, ESSP=0.90, $\Delta = 0.002$), outperforming training on all groups (DSC=0.90, ESSP=0.89, Δ=0.01). Overall, these outcomes suggest that the subgroup associated with more challenging morphology (Group 1) acts as a fairness anchor and overrepresenting it can reduce group disparity without severely compromising overall utility. This observation motivates the core design of our AL strategy, which adaptively increases the selection pressure toward the currently under-performing group.

We tackle the AL for segmentation with a lightweight mechanism that uses only labeled-set performance to estimate group weights and requires no additional fairness classifier or expensive representativeness modeling. Our approach is close to performance-driven group reweighting strategies proposed in classification (Wang *et al.*, 2022), but adapted to voxel-wise uncertainty.

Although one might suspect overfitting to Group 1, this is not supported in the classical sense: when trained only on Group 1, the model still generalizes well to Group 2, with Group 2

performance remaining close to Group 1. Additionally, the performance differences when predicting Group 2 across all baseline experiments are only minor.

We argue that the higher ESSP achieved by training on Group 1 only is not driven by overfitting, but by robust generalization from training with the more morphologically challenging Group 1 dataset. Group 1 includes both global inter-subject variability and additional localized deformations. A model trained on Group 1 learns features that remain valid when evaluated on Group 2, where the task is effectively easier because the localized deformations are absent. In contrast, training on pooled Group 1 and Group 2 data can encourage shortcut learning, where the model preferentially fits the easiest, most frequent patterns (Group 2) and underfits the complex Group 1 cases, consistent with the gaps observed in Table 3.4.

Importantly, Group 1-only training is not presented as a universally optimal deployment strategy. ESSP is not a pure accuracy metric: it explicitly penalizes between-group disparity through (Δ), defined as the sum of absolute deviations from the overall dice. Since Group 1 is the morphology-challenged subgroup by construction, the fairness gain under Group 1-only training increases. Group 2 performance remains high while the inter-group gap shrinks, and this is exactly what ESSP rewards. Moreover, the Table 1 results are computed on a held-out, balanced test set (15 subjects per group), so the effect is not memorization-based overfitting.

We agree that one limitation of this study is the use of a synthetic dataset, which we selected to explicitly control the presence and magnitude of morphological bias. This controlled setting enables a clear link between sampling strategy and performance bias. Real-world medical data contains complex biases (e.g., scanner artifacts correlated with hospital demographics) that are harder to disentangle than anatomical deformations. *Evaluating* our framework in real-world data scenarios requires (i) identification of a dataset with known biases to confirm that a measurable disparity exists, and (ii) reliable reference segmentations for training, ideally generated without performance biases. In practice, finding such datasets is extremely challenging. Automatic segmentations may carry performance biases making fairness evaluation challenging, and manual annotations by experts are scarce. Moreover, real cohorts may exhibit subtle or

region-specific morphometric differences (or none at all), and the existence of such differences cannot be assumed a priori. These challenges led us to perform these experiments solely with synthetic data. Our results support the use of a weighted sampling strategy to avoid performance bias associated with group level attributes. If one were to *translate* or apply our framework in a real-world scenario, one would need to identify one or more sensitive attributes based on apriori hypotheses (structure X has been reported to be larger in sub-population Y) and guide the sampling strategy using the weighted localized entropy (eq. 2 3.2).

Another limitation of our work is that our weighting strategy relies on the availability of group labels for the labeled set. While it is a reasonable assumption in a controlled AL setup, extending this to scenarios where sensitive attributes are missing is a necessary future step.

3.6 Conclusion

We presented a fairness-aware active learning framework for brain MRI segmentation. Through using a performance-based weighting scheme and localized entropy, the proposed algorithm actively constructs a training set that prioritizes equity. This can be especially practical for deploying segmentation models in settings where both labeling budgets and fairness requirements are critical. Our study provides a robust foundation for advancing fair active-learning approaches in medical image segmentation.

CONCLUSION AND RECOMMENDATIONS

The first paper shows that demographic bias is method-dependent. Some models were relatively robust, while others were much more sensitive to demographic mismatch. Although we did not observe extreme bias, we found subtler effects, including the loss of race-related volumetric patterns that were still visible in the manual ground-truth annotations. At the same time, these findings must be interpreted with caution, since our analysis focused only on the right and left nucleus accumbens. Broader studies across more brain structures will be needed to understand the full extent of demographic bias in brain segmentation.

The second paper shows that fairness can be improved during the active learning process itself. By combining group-aware weighting with localized uncertainty, the proposed method reduced group disparities in the strong-bias and weak-bias settings. Together, the two studies support a broader shift in medical image analysis: fairness should not be considered a secondary concern, but a central design principle.

A natural next step would be to translate the Weighted Localized Entropy framework from synthetic experiments to real clinical datasets with measurable bias. Another important direction is to design an attribute-agnostic, fair active learning framework that does not require explicit group labels during training. This would make the approach more practical in real-world settings where sensitive attributes are missing, incomplete, or noisy. Future work should also extend the evaluation beyond small subcortical structures to larger-scale and whole-brain segmentation tasks.

It is also important to acknowledge a key nuance. In medical diagnosis, strict mathematical fairness is not always beneficial. Recent perspectives (Sabuncu, Wang & Nguyen, 2025) caution against enforcing performance parity across protected groups as an absolute objective. Such constraints may unintentionally reduce diagnostic accuracy in certain subpopulations. A more appropriate goal is to achieve the highest possible performance within each demographic group

while reporting residual disparities transparently. Thus, although this work highlights equity as an important principle, future clinical deployment must carefully balance fairness with diagnostic precision.

APPENDIX I

INVESTIGATING DEMOGRAPHIC BIAS IN BRAIN MRI SEGMENTATION: A COMPARATIVE STUDY OF DEEP-LEARNING AND NON-DEEP-LEARNING METHODS

Table-A I-1 Dice coefficients of the UNesT model for different training and testing datasets. The columns represent the subgroup on which the model was trained, while the rows indicate the subgroup on which the model was evaluated. (right NAc)

Train / Test	UNesTWM		UNesTWF		UNesTBM		UNesTBF	
DSC	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**
WM	0.86	0.03	0.80	0.03	0.79	0.04	0.80	0.03
WF	0.82	0.04	0.81	0.03	0.78	0.03	0.79	0.03
BM	0.81	0.04	0.81	0.04	0.81	0.03	0.80	0.04
BF	0.82	0.03	0.83	0.04	0.81	0.03	0.82	0.03

Table-A I-2 Dice coefficients of the UNesT model for different training and testing datasets. The columns represent the subgroup on which the model was trained, while the rows indicate the subgroup on which the model was evaluated (left NAc)

Train / Test	UNesTWM		UNesTWF		UNesTBM		UNesTBF	
DSC	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**
WM	0.85	0.03	0.81	0.04	0.78	0.058	0.79	0.04
WF	0.81	0.03	0.80	0.02	0.78	0.03	0.78	0.03
BM	0.80	0.05	0.80	0.05	0.81	0.04	0.79	0.05
BF	0.81	0.03	0.82	0.03	0.81	0.03	0.81	0.03

Table-A I-3 Dice coefficients of the nnU-Net model for different training and testing datasets. The columns represent the subgroup on which the model was trained, while the rows indicate the subgroup on which the model was evaluated (right NAc)

Train / Test	nnUNetWM		nnUNetWF		nnUNetBM		nnUNetBF	
DSC	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**
WM	0.87	0.02	0.85	0.03	0.85	0.02	0.85	0.03
WF	0.86	0.03	0.85	0.02	0.85	0.02	0.85	0.02
BM	0.85	0.02	0.85	0.03	0.86	0.02	0.85	0.04
BF	0.87	0.01	0.87	0.02	0.87	0.02	0.87	0.02

Table-A I-4 Dice coefficients of the nnU-Net model for different training and testing datasets. The columns represent the subgroup on which the model was trained, while the rows indicate the subgroup on which the model was evaluated (left NAc)

Train / Test	nnUNetWM		nnUNetWF		nnUNetBM		nnUNetBF	
DSC	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**
WM	0.86	0.03	0.85	0.03	0.84	0.03	0.85	0.03
WF	0.86	0.02	0.85	0.03	0.85	0.02	0.85	0.02
BM	0.85	0.04	0.85	0.04	0.86	0.02	0.85	0.04
BF	0.86	0.02	0.87	0.03	0.85	0.04	0.86	0.02

Table-A I-5 Dice coefficients of the CoTr model for different training and testing datasets. The columns represent the subgroup on which the model was trained, while the rows indicate the subgroup on which the model was evaluated (right NAc)

Train / Test	CoTrWM		CoTrWF		CoTrBM		CoTrBF	
DSC	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**
WM	0.86	0.02	0.85	0.03	0.85	0.02	0.85	0.03
WF	0.85	0.02	0.85	0.02	0.85	0.02	0.85	0.02
BM	0.85	0.03	0.85	0.03	0.86	0.02	0.85	0.04
BF	0.87	0.02	0.87	0.02	0.86	0.02	0.87	0.02

Table-A I-6 Dice coefficients of the CoTr model for different training and testing datasets. The columns represent the subgroup on which the model was trained, while the rows indicate the subgroup on which the model was evaluated (left NAc)

Train / Test	CoTrWM		CoTrWF		CoTrBM		CoTrBF	
DSC	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**
WM	0.86	0.03	0.85	0.03	0.84	0.04	0.84	0.03
WF	0.85	0.02	0.85	0.03	0.84	0.02	0.84	0.02
BM	0.85	0.04	0.85	0.03	0.86	0.02	0.84	0.04
BF	0.85	0.03	0.86	0.03	0.85	0.04	0.86	0.02

Table-A I-7 Dice coefficients of the ANTs model for different training and testing datasets. The columns represent the subgroup on which the model was trained, while the rows indicate the subgroup on which the model was evaluated (right NAc)

Train / Test	ANTsWM		ANTsWF		ANTsBM		ANTsBF	
Dice	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**
WM	0.82	0.02	0.81	0.03	0.76	0.05	0.77	0.06
WF	0.81	0.04	0.80	0.04	0.74	0.04	0.75	0.05
BM	0.80	0.04	0.81	0.04	0.81	0.03	0.81	0.05
BF	0.82	0.04	0.82	0.04	0.80	0.03	0.82	0.04

Table-A I-8 Dice coefficients of the ANTs model for different training and testing datasets. The columns represent the subgroup on which the model was trained, while the rows indicate the subgroup on which the model was evaluated (left NAc)

Train / Test	ANTsWM		ANTsWF		ANTsBM		ANTsBF	
Dice	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**
WM	0.82	0.03	0.81	0.03	0.73	0.06	0.76	0.06
WF	0.80	0.04	0.80	0.03	0.72	0.05	0.72	0.07
BM	0.80	0.06	0.80	0.06	0.78	0.06	0.79	0.06
BF	0.80	0.06	0.80	0.06	0.77	0.03	0.80	0.03

Table-A I-9 NSD of the nnU-Net model for different training and testing datasets. The columns represent the subgroup on which the model was trained, while the rows indicate the subgroup on which the model was evaluated (right NAc)

Train / Test	nnUNetWM		nnUNetWF		nnUNetBM		nnUNetBF	
NSD	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**
WM	0.54	0.05	0.53	0.08	0.49	0.05	0.53	0.09
WF	0.51	0.07	0.51	0.06	0.48	0.05	0.52	0.05
BM	0.50	0.06	0.51	0.08	0.54	0.06	0.51	0.10
BF	0.56	0.05	0.56	0.04	0.56	0.05	0.56	0.05

Table-A I-10 NSD of the nnU-Net model for different training and testing datasets. The columns represent the subgroup on which the model was trained, while the rows indicate the subgroup on which the model was evaluated (left NAc)

Train / Test	nnUNetWM		nnUNetWF		nnUNetBM		nnUNetBF	
NSD	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**
WM	0.52	0.07	0.52	0.08	0.48	0.07	0.51	0.08
WF	0.53	0.05	0.52	0.07	0.51	0.05	0.53	0.06
BM	0.52	0.09	0.50	0.08	0.55	0.05	0.52	0.09
BF	0.53	0.06	0.55	0.07	0.53	0.08	0.56	0.06

Table-A I-11 NSD scores of the CoTr model for different training and testing datasets. The columns represent the subgroup on which the model was trained, while the rows indicate the subgroup on which the model was evaluated (right NAc)

Train / Test	CoTrWM		CoTrWF		CoTrBM		CoTrBF	
NSD	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**
WM	0.53	0.05	0.51	0.08	0.49	0.04	0.51	0.09
WF	0.50	0.06	0.49	0.06	0.47	0.05	0.51	0.05
BM	0.48	0.08	0.49	0.08	0.53	0.07	0.49	0.09
BF	0.54	0.05	0.56	0.04	0.55	0.05	0.54	0.04

Table-A I-12 NSD scores of CoTr model for different training and testing datasets. The columns represent the subgroup on which the model was trained, while the rows indicate the subgroup on which the model was evaluated (left NAc)

Train / Test	CoTrWM		CoTrWF		CoTrBM		CoTrBF	
NSD	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**
WM	0.51	0.07	0.51	0.08	0.47	0.07	0.49	0.08
WF	0.51	0.05	0.51	0.06	0.49	0.05	0.50	0.06
BM	0.51	0.09	0.51	0.06	0.54	0.06	0.50	0.09
BF	0.53	0.07	0.54	0.07	0.52	0.08	0.55	0.07

Table-A I-13 NSD scores of ANTs model for different training and testing datasets. The columns represent the subgroup on which the model was trained, while the rows indicate the subgroup on which the model was evaluated (right NAc)

Train / Test	ANTsWM		ANTsWF		ANTsBM		ANTsBF	
NSD	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**
WM	0.44	0.05	0.43	0.08	0.35	0.07	0.38	0.12
WF	0.43	0.07	0.41	0.07	0.31	0.06	0.36	0.06
BM	0.4	0.05	0.44	0.07	0.43	0.07	0.44	0.10
BF	0.45	0.08	0.45	0.08	0.43	0.08	0.46	0.09

Table-A I-14 NSD scores of ANTs model for different training and testing datasets. The columns represent the subgroup on which the model was trained, while the rows indicate the subgroup on which the model was evaluated (left NAc)

Train / Test	ANTsWM		ANTsWF		ANTsBM		ANTsBF	
NSD	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**
WM	0.42	0.07	0.42	0.07	0.35	0.07	0.36	0.10
WF	0.42	0.07	0.43	0.06	0.33	0.06	0.34	0.08
BM	0.41	0.09	0.42	0.08	0.43	0.07	0.43	0.10
BF	0.43	0.09	0.43	0.09	0.42	0.06	0.42	0.07

Table-A I-15 Results for evaluating race effects on volumes by segmentation models for right and left NAc (Coeff. is coefficient and its value is rounded to upper value, Std Err is standard deviation, P is p-value)

Structure	Model	Trained on BF			Trained on BM			Trained on WF			Trained on WM		
		γ_1	Std Err	P	γ_1	Std Err	P	γ_1	Std Err	P	γ_1	Std Err	P
Right NAc	ANTs	31	52.1	0.552	10	50.5	0.835	-32	59.3	0.583	40	63.8	0.530
	CoTr	136	84.1	0.104	120	105.4	0.253	153	85.3	0.073	124	96.2	0.197
	nnU-Net	137	84.4	0.103	113	104.5	0.277	125	82.0	0.127	118	94.3	0.209
	UNesT	9	64.8	0.881	-65	74.5	0.381	70	77.3	0.365	13	80.7	0.867
Left NAc	ANTs	29	52.3	0.579	34	52.7	0.517	-26	58.2	0.651	41	60.7	0.496
	CoTr	173	86.2	**0.044**	113	105.1	0.281	95	82.2	0.245	144	100.0	0.148
	nnU-Net	154	87.8	0.078	124	99.5	0.210	60	83.0	0.466	151	95.6	0.114
	UNesT	4	72.0	0.956	-20	77.5	0.789	7	75.4	0.922	110	79.8	0.168

Table-A I-16 Results for evaluating race × sex effects on volumes by segmentation models for right and left NAc (Coeff. is coefficient and its value is rounded to the upper value, Std Err is standard deviation, P is p-value)

Structure	Model	Trained on BF			Trained on BM			Trained on WF			Trained on WM		
		γ_3	Std Err	P	γ_3	Std Err	P	γ_3	Std Err	P	γ_3	Std Err	P
Right NAc	ANTs	38	72.6	0.596	23	71.1	0.739	90	84.2	0.283	40	88.5	0.644
	CoTr	-94	112.4	0.399	-42	149.3	0.775	-53	120.3	0.657	-19	136.5	0.889
	nnU-Net	-103	116.1	0.374	-44	147.2	0.766	-46	114.1	0.681	-7	131.5	0.955
	UNesT	-39	90.7	0.661	-13	106.0	0.896	-97	108.3	0.367	131	111.9	0.240
Left NAc	ANTs	8	74.0	0.911	-4	73.6	0.956	127	81.7	0.118	42	83.8	0.614
	CoTr	-94	120.9	0.432	-2	146.3	0.984	3	112.8	0.972	7	139.8	0.958
	nnU-Net	-107	122.4	0.378	-38	138.7	0.781	21	115.3	0.855	-22	134.2	0.869
	UNesT	20	99.4	0.836	-38	109.3	0.722	18	105.3	0.862	17	111.1	0.877

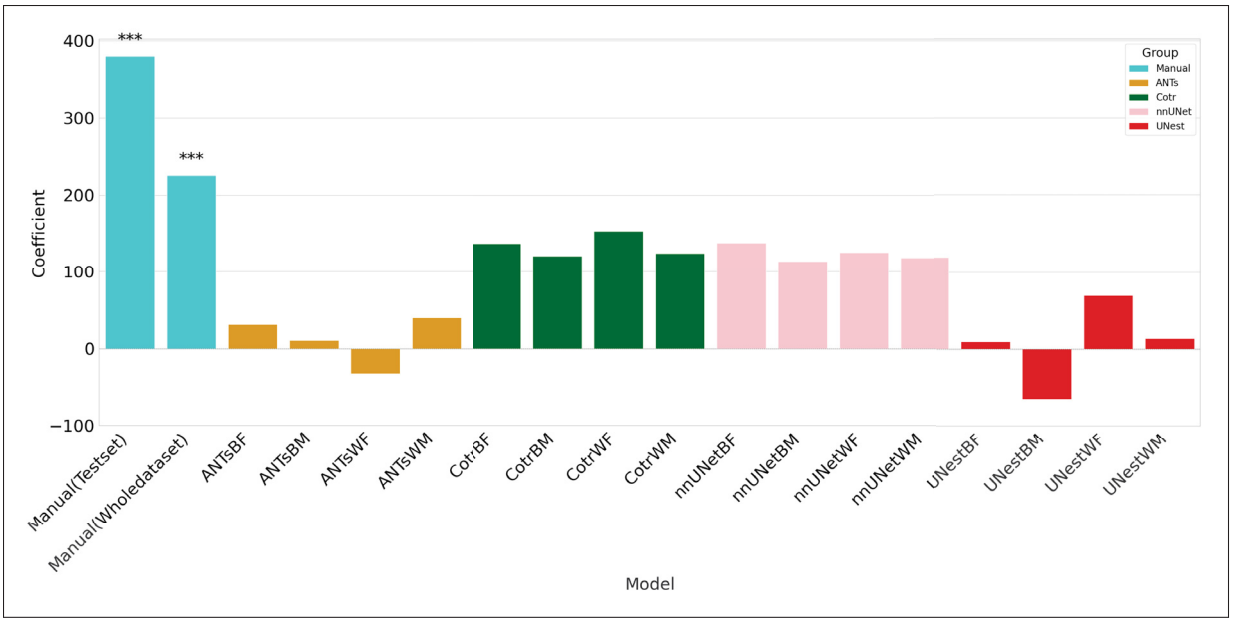

Figure-A I-1 Race coefficients in mixed linear models' results for measuring its influence on volume(right NAc). Significance using linear mixed effects model is denoted by *** ($1.00 \times 10^{-4} < P \leq 1.00 \times 10^{-3}$), ** ($1.00 \times 10^{-3} < P \leq 1.00 \times 10^{-2}$), and * ($1.00 \times 10^{-2} < P \leq 5.00 \times 10^{-2}$)

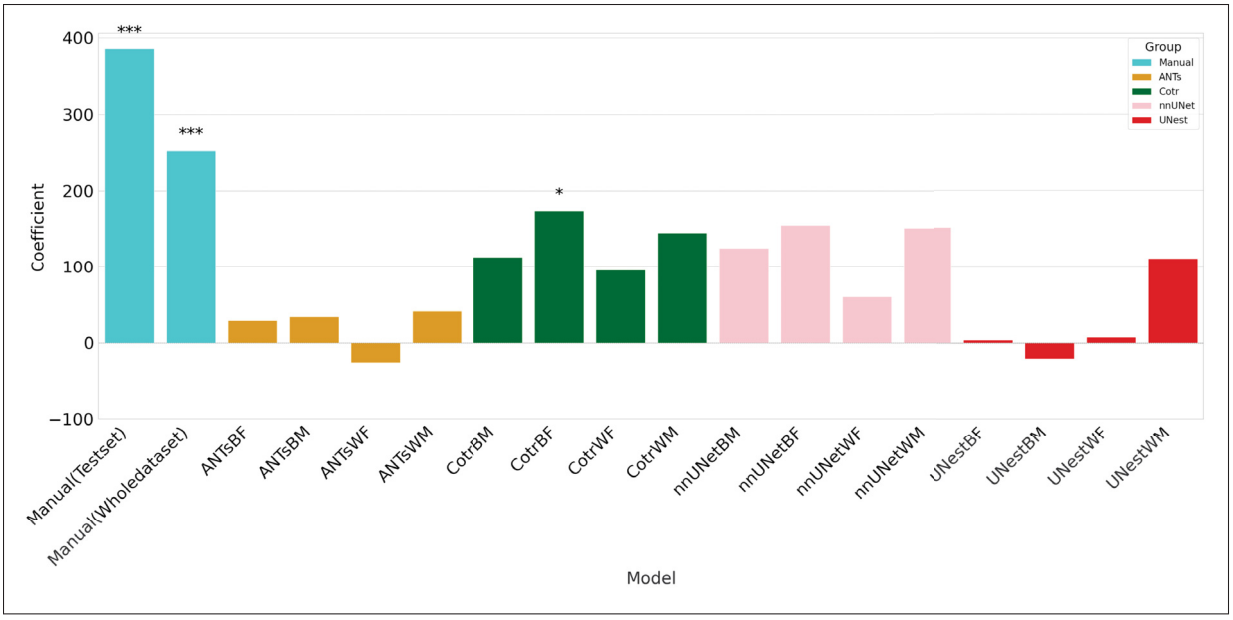

Figure-A I-2 Race coefficients in mixed linear models' results for measuring its influence on volume(left NAc). Significance using linear mixed effects model is denoted by *** ($1.00 \times 10^{-4} < P \leq 1.00 \times 10^{-3}$), ** ($1.00 \times 10^{-3} < P \leq 1.00 \times 10^{-2}$), and * ($1.00 \times 10^{-2} < P \leq 5.00 \times 10^{-2}$)

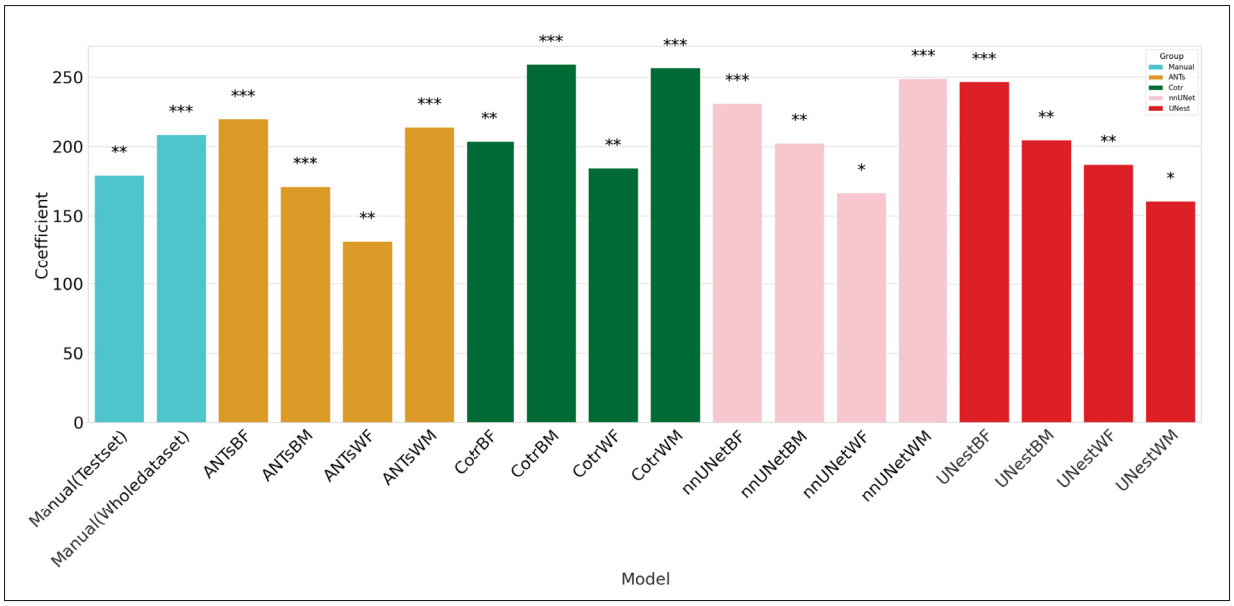

Figure-A I-3 Sex coefficients in mixed linear models' results for measuring its influence on volume(right NAc). Significance using linear mixed effects model is denoted by *** ($1.00 \times 10^{-4} < P \leq 1.00 \times 10^{-3}$), ** ($1.00 \times 10^{-3} < P \leq 1.00 \times 10^{-2}$), and * ($1.00 \times 10^{-2} < P \leq 5.00 \times 10^{-2}$)

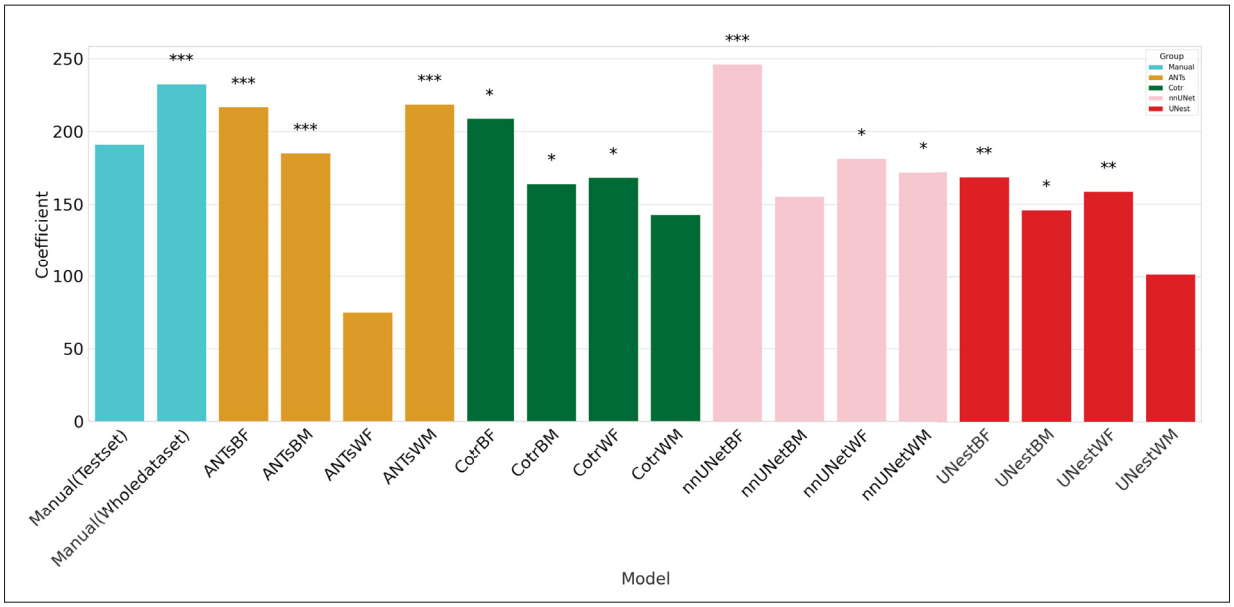

Figure-A I-4 Sex coefficients in mixed linear models' results for measuring its influence on volume(left NAc). Significance using linear mixed effects model is denoted by *** ($1.00 \times 10^{-4} < P \leq 1.00 \times 10^{-3}$), ** ($1.00 \times 10^{-3} < P \leq 1.00 \times 10^{-2}$), and * ($1.00 \times 10^{-2} < P \leq 5.00 \times 10^{-2}$)

APPENDIX II

EXPLORING ENTROPY-BASED ACTIVE LEARNING FOR FAIR BRAIN SEGMENTATION

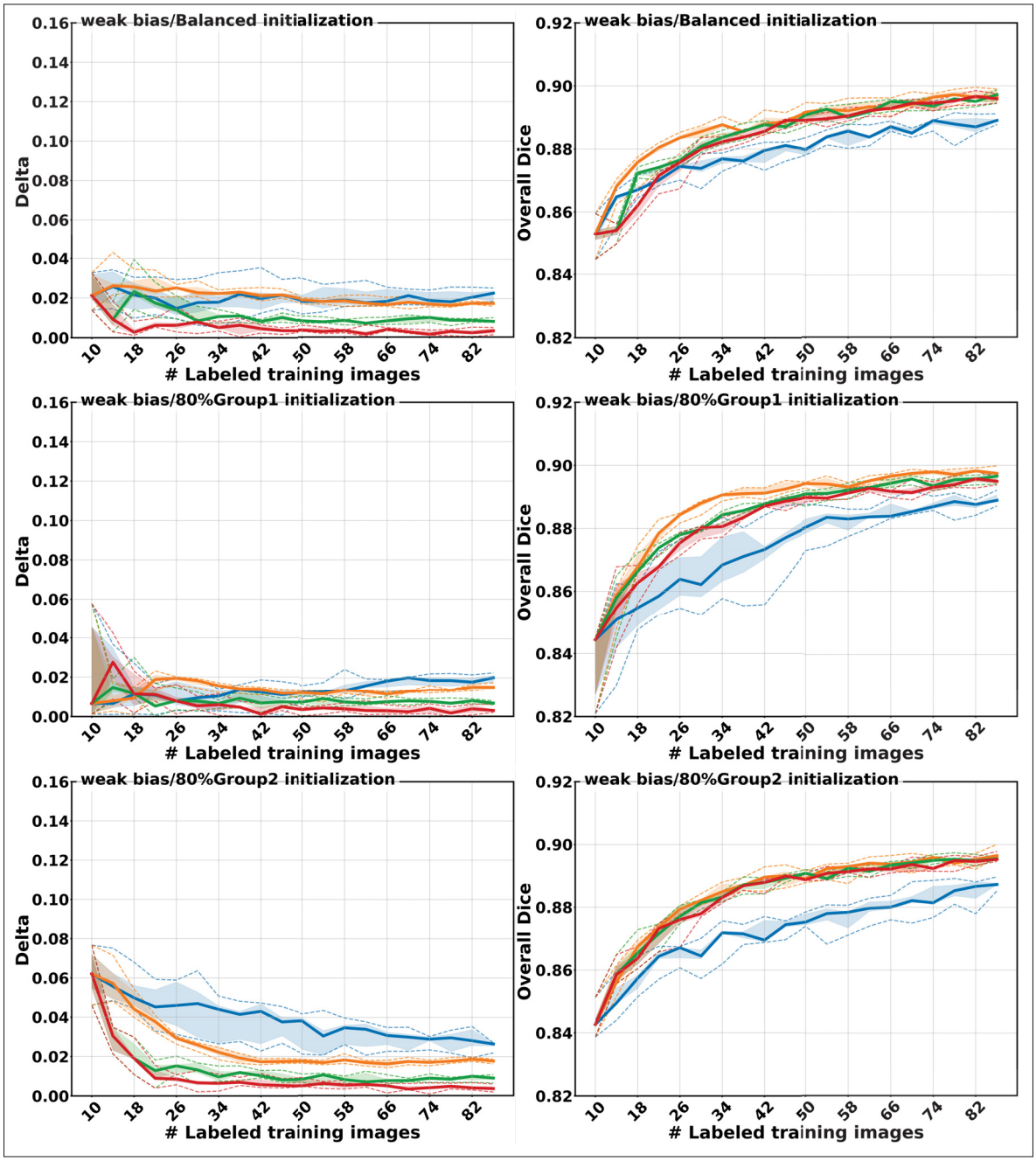

Figure-A II-1 Δ and dice for the weak bias experiments. First row: balanced initialization, second row: 80/20 Group 1/Group 2 ratio, third row: 20/80 ratio

BIBLIOGRAPHY

Adeli, E., Zhao, Q., Pfefferbaum, A., Sullivan, E. V., Fei-Fei, L., Niebles, J. C. & Pohl, K. M. (2021). Representation Learning with Statistical Independence to Mitigate Bias. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Retrieved from: https://openaccess.thecvf.com/content/WACV2021/papers/Adeli_Representation_Learning_With_Statistical_Independence_to_Mitigate_Bias_WACV_2021_paper.pdf.

Alqarni, M., Jones, E., Ribeiro, L., Hema, V., Cooper, S., Mullassery, V., Morris, S., Guerrero-Urbano, T. & King, A. (2024). *An Investigation of Race Bias in Deep Learning-Based Segmentation of Prostate MRI Images*. SSRN Preprint.

Anahideh, H., Asudeh, A. & Thirumuruganathan, S. (2022). Fair active learning. *Expert Systems with Applications*, 199, 116981.

Ancelin, M.-L., Carrière, I., Artero, S., Maller, J., Meslin, C., Ritchie, K., Ryan, J. & Chaudieu, I. (2019). Lifetime major depression and grey-matter volume. *Journal of Psychiatry and Neuroscience*, 44(1), 45–53. doi: 10.1503/jpn.180026.

Angluin, D. (2001). Queries Revisited. In *Algorithmic Learning Theory* (vol. 2225, pp. 12–31). Springer. doi: 10.1007/3-540-45583-3_3.

Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J. & Agarwal, A. (2020). Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. *International Conference on Learning Representations (ICLR)*. Retrieved from: https://openreview.net/forum?id=ryghZJBKPS.

Atzeni, A., Peter, L., Robinson, E., Blackburn, E., Althonayan, J., Alexander, D. C. & Iglesias, J. E. (2022). Deep active learning for suggestive segmentation of biomedical image stacks via optimisation of Dice scores and traced boundary length. *Medical Image Analysis*, 81, 102549.

Bae, W., Sutherland, D. J. & Oliveira, G. L. (2025). Uncertainty Herding: One Active Learning Method for All Label Budgets. *The Thirteenth International Conference on Learning Representations (ICLR)*. Retrieved from: https://openreview.net/forum?id=UgPoHhYQ2U.

Baur, C., Wiestler, B., Albarqouni, S. & Navab, N. (2020). Scale-Space Autoencoders for Unsupervised Anomaly Segmentation in Brain MRI. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, 12264(Lecture Notes in Computer Science), 552–561. doi: 10.1007/978-3-030-59719-1_54.

Benčević, M., Habijan, M., Galić, I., Babin, D. & Pižurica, A. (2024). Understanding skin color bias in deep learning–based skin lesion segmentation. *Computer Methods and Programs in Biomedicine*, 245, 108044.

Boehringer, A. S., Sanaat, A., Arabi, H. & Zaidi, H. (2023). An active learning approach to train a deep learning algorithm for tumor segmentation from brain MR images. *Insights into Imaging*, 14(1), 141.

Budd, S., Robinson, E. C. & Kainz, B. (2021). A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71, 102062.

Camilleri, R., Wagenmaker, A., Morgenstern, J., Jain, L. & Jamieson, K. (2024a). Fair Active Learning in Low-Data Regimes. *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, 244(Proceedings of Machine Learning Research), 517–531.

Camilleri, R., Wagenmaker, A., Morgenstern, J., Jain, L. & Jamieson, K. (2024b). Fair Active Learning in Low-Data Regimes. *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, 244(Proceedings of Machine Learning Research), 517–531.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L. & Zhou, Y. (2021). TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv preprint arXiv:2102.04306*. doi: 10.48550/arXiv.2102.04306.

Dagan, I. & Engelson, S. P. (1995). Committee-Based Sampling For Training Probabilistic Classifiers. *Machine Learning Proceedings 1995*, pp. 150–157. doi: 10.1016/B978-1-55860-377-6.50027-X.

Danaee, G., Niethammer, M., Rushmore, J. & Bouix, S. (2025). Investigating Demographic Bias in Brain MRI Segmentation: A Comparative Study of Deep-Learning and Non-Deep-Learning Methods. *arXiv preprint arXiv:2510.17999*.

Diaz-Pinto, A., Alle, S., Nath, V., Tang, Y., Ihsani, A., Asad, M., Pérez-García, F., Mehta, P., Li, W., Flores, M., Roth, H. R., Vercauteren, T., Xu, D., Dogra, P., Ourselin, S., Feng, A. & Cardoso, M. J. (2024). MONAI Label: A framework for AI-assisted interactive labeling of 3D medical images. *Medical Image Analysis*, 95, 103207. doi: https://doi.org/10.1016/j.media.2024.103207.

Dibaji, M., Ospel, J., Souza, R. & Bento, M. (2024). Sex differences in brain MRI using deep learning toward fairer healthcare outcomes. *Frontiers in Computational Neuroscience*, 18(1452457).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*. Retrieved from: https://arxiv.org/abs/2010.11929.

Fajri, R. M., Saxena, A., Pei, Y. & Pechenizkiy, M. (2024). FAL-CUR: Fair Active Learning using Uncertainty and Representativeness on Fair Clustering. *Expert Systems with Applications*, 242, 122842.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B. & Dale, A. M. (2002). Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain. *Neuron*, 33(3), 341–355. doi: 10.1016/S0896-6273(02)00569-X. Neurotechnique.

Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D. H., Busa, E., Seidman, L. J., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B. & Dale, A. M. (2004). Automatically Parcellating the Human Cerebral Cortex. *Cerebral Cortex*, 14(1), 11–22. doi: 10.1093/cercor/bhg087.

Frazier, J. A., Hodge, S. M., Breeze, J. L., Giuliano, A. J., Terry, J. E., Moore, C. M., Kennedy, D. N., Lopez-Larson, M. P., Caviness, V. S., Seidman, L. J., Zablotsky, B. & Makris, N. (2008). Diagnostic and sex effects on limbic volumes in early-onset bipolar disorder and schizophrenia. *Schizophrenia Bulletin*, 34(1), 37–46.

Gaillochet, M., Desrosiers, C. & Lombaert, H. (2023). Active learning for medical image segmentation with stochastic batches. *Medical Image Analysis*, 90, 102958.

Gao, Q.-L., Chen, X., Castellanos, F. X., Lu, B. & Yan, C.-G. (2025). Towards closed-loop precision psychiatry: Integrating MRI biomarkers for individualized care of major depressive disorder. *Psychoradiology*, 5, kkaf024. doi: 10.1093/psyrad/kkaf024.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C. & Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, 80, 105–124.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R. & Xu, D. (2022). UNETR: Transformers for 3D Medical Image Segmentation. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1748–1758. doi: 10.1109/WACV51458.2022.00181.

Heckemann, R. A., Hajnal, J. V., Aljabar, P., Rueckert, D. & Hammers, A. (2006). Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1), 115–126. doi: 10.1016/j.neuroimage.2006.05.061.

Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B. & Reuter, M. (2020a). FastSurfer – A fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*, 219, 117012. doi: 10.1016/j.neuroimage.2020.117012.

Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B. & Reuter, M. (2020b). FastSurfer – A fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*, 219, 117012. Retrieved from: https://doi.org/10.1016/j.neuroimage.2020.117012. doi:10.1016/j.neuroimage.2020.117012.

Hu, H., Zhang, Z., Xie, Z. & Lin, S. (2019). Local Relation Networks for Image Recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3464–3473.

Human Connectome Project. [Accessed: 2025-09-09]. (2017). WU-Minn HCP 1200 Subjects Data Release: Reference Manual. Retrieved from: Onlinemanual.

Ioannou, S., Chockler, H., Hammers, A. & King, A. P. (2022a). A Study of Demographic Bias in CNN-Based Brain MR Segmentation. *Machine Learning in Clinical Neuroimaging (MLCN 2022)*, 13596(Lecture Notes in Computer Science), 13–22.

Ioannou, S., Chockler, H., Hammers, A. & King, A. P. [doi:10.48550/arXiv.2208.06613]. (2022b). A Study of Demographic Bias in CNN-based Brain MR Segmentation. Retrieved from: https://arxiv.org/abs/2208.06613.

Isamah, N., Faison, W., Payne, M. E., MacFall, J., Steffens, D. C., Beyer, J. L., Krishnan, K. R. & Taylor, W. D. (2010). Variability in frontotemporal brain structure: the importance of recruitment of African Americans in neuroscience research. *PLoS One*, 5(10), e13642.

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P. F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S. & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2), 203–211. Retrieved from: https://www.nature.com/articles/s41592-020-01008-z. doi:10.1038/s41592-020-01008-z.

Joshi, A. J., Porikli, F. & Papanikolopoulos, N. (2009). Multi-Class Active Learning for Image Classification. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2372–2379. doi: 10.1109/CVPR.2009.5206627.

Käding, C., Rodner, E., Freytag, A. & Denzler, J. (2016). Active and Continuous Exploration with Deep Neural Networks and Expected Model Output Changes. *NIPS Workshop on Continual Learning and Deep Networks*.

Kim, D. D., Chandra, R. S., Yang, L., Wu, J., Feng, X., Atalay, M., Bettegowda, C., Jones, C., Sair, H., Liao, W. H., Zhu, C., Zou, B., Kazerooni, A. F., Nabavizadeh, A., Jiao, Z., Peng, J. & Bai, H. X. (2024). Active Learning in Brain Tumor Segmentation with Uncertainty Sampling and Annotation Redundancy Restriction. *Journal of Imaging Informatics in Medicine*, 37(5), 2099–2107.

Kirsch, A., van Amersfoort, J. & Gal, Y. (2019). BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*. Retrieved from: https://papers.nips.cc/paper/2019/hash/95323660ed2124450caaac2c46b5ed90-Abstract.html.

Lang, K. & Baum, E. (1992). Query Learning Can Work Poorly When a Human Oracle Is Used. *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pp. 335–340.

Ledig, C., Heckemann, R. A., Hammers, A., Lopez, J. C., Newcombe, V. F., Makropoulos, A., Lötjönen, J., Menon, D. K. & Rueckert, D. (2015). Robust whole-brain segmentation: Application to traumatic brain injury. *Medical Image Analysis*, 21(1), 40–58. Retrieved from: https://doi.org/10.1016/j.media.2014.12.003. doi:10.1016/j.media.2014.12.003.

Ledig, C., Schuh, A., Guerrero, R., Heckemann, R. A. & Rueckert, D. (2018). Structural brain imaging in Alzheimer's disease and mild cognitive impairment: biomarker analysis and shared morphometry database. *Scientific Reports*, 8(1), 11258. Retrieved from: https://doi.org/10.1038/s41598-018-29295-9. doi:10.1038/s41598-018-29295-9.

Lee, T., Puyol-Antón, E., Ruijsink, B., Aitcheson, K., Shi, M. & King, A. P. (2023). An Investigation into the Impact of Deep Learning Model Choice on Sex and Race Bias in Cardiac MR Segmentation. *Clinical Image-Based Procedures, Fairness of AI in Medical Imaging, and Ethical and Philosophical Issues in Medical Imaging*, pp. 215–224.

Lee, T., Puyol-Antón, E., Ruijsink, B., Roujol, S., Barfoot, T., Ogbomo-Harmitt, S., Shi, M. & King, A. (2025). An investigation into the causes of race bias in artificial intelligence–based cine cardiac magnetic resonance segmentation. *European Heart Journal - Digital Health*, ztaf008.

Lewis, D. D. & Catlett, J. (1994). Heterogeneous Uncertainty Sampling for Supervised Learning. *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 148–156.

Lewis, D. D. & Gale, W. A. (1994). A Sequential Algorithm for Training Text Classifiers. *SIGIR '94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 3–12.

Liu, Z., Zhang, X. & Jiang, B. (2023). Active learning with fairness-aware clustering for fair classification considering multiple sensitive attributes. *Information Sciences*, 647, 119521. doi: 10.1016/j.ins.2023.119521.

Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M. D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M. A., Wiesenfarth, M., Kavur, A. E., Sudre, C. H., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Rädsch, T., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Benis, A., Blaschko, M., Cardoso, M. J., Cheplygina, V., Cimini, B. A., Collins, G. S., Farahani, K., Ferrer, L., Galdran, A., van Ginneken, B., Haase, R., Hashimoto, D. A., Hoffman, M. M., Huisman, M., Jannin, P., Kahn, C. E., Kainmueller, D., Kainz, B., Karargyris, A., Karthikesalingam, A., Kenngott, H., Kofler, F., Kopp-Schneider, A., Kreshuk, A., Kurc, T., Landman, B. A., Litjens, G., Madani, A., Maier-Hein, K., Martel, A. L., Mattson, P., Meijering, E., Menze, B., Moons, K. G., Müller, H., Nichyporuk, B., Nickel, F., Petersen, J., Rajpoot, N., Rieke, N., Saez-Rodriguez, J., Sánchez, C. I., Shetty, S., van Smeden, M., Summers, R. M., Taha, A. A., Tiulpin, A., Tsaftaris, S. A., Calster, B. V., Varoquaux, G. & Jäger, P. F. (2024). Metrics reloaded: recommendations for image analysis validation. *Nature Methods*, 21(2), 195–212. doi: 10.1038/s41592-023-02151-z.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. (2022). A Survey on Bias and Fairness in Machine Learning.

Mehta, R., Shui, C. & Arbel, T. (2024). Evaluating the Fairness of Deep Learning Uncertainty Estimates in Medical Image Analysis. *Medical Imaging with Deep Learning*, 227(Proceedings of Machine Learning Research), 1453–1492. Retrieved from: https://proceedings.mlr.press/v227/mehta24a.html.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R. et al. (2015). The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10), 1993–2024. doi: 10.1109/TMI.2014.2377694.

Munjal, P., Hayat, N., Hayat, M., Sourati, J. & Khan, S. (2022). Towards Robust and Reproducible Active Learning Using Neural Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 223–232.

Nath, V., Yang, D., Landman, B. A., Xu, D. & Roth, H. R. (2021). Diminishing Uncertainty Within the Training Pool: Active Learning for Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 40(10), 2534–2547. doi: 10.1109/TMI.2020.3048055.

Nguyen, V.-L., Shaker, M. H. & Hüllermeier, E. (2022). How to Measure Uncertainty in Uncertainty Sampling for Active Learning. *Machine Learning*, 111, 89–122. doi: 10.1007/s10994-021-06003-9.

Nikolov, S., Blackwell, S., Zverovitch, A., Mendes, R., Livne, M., Fauw, J. D., Patel, Y., Meyer, C., Askham, H., Romera-Paredes, B., Kelly, C., Karthikesalingam, A., Chu, C., Carnell, D., Boon, C., D'Souza, D., Moinuddin, S. A., Garie, B., McQuinlan, Y., Ireland, S., Hampton, K., Fuller, K., Montgomery, H., Rees, G., Suleyman, M., Back, T., Hughes, C. O., Ledsam, J. R. & Ronneberger, O. (2021). Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study. *Journal of Medical Internet Research*, 23(7), e26151. doi: 10.2196/26151.

Oguguo, T., Zamzmi, G., Rajaraman, S., Yang, F., Xue, Z. & Antani, S. (2023). A Comparative Study of Fairness in Medical Machine Learning. *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. doi: 10.1109/ISBI53787.2023.10230368.

Pang, J., Wang, J., Zhu, Z., Yao, Y., Qian, C. & Liu, Y. (2024). Fairness Without Harm: An Influence-Guided Active Sampling Approach. *Advances in Neural Information Processing Systems (NeurIPS)*.

Parvaneh, A., Abbasnejad, E., Teney, D., Haffari, G. R., van den Hengel, A. & Shi, J. Q. (2022). Active Learning by Feature Mixing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12237–12246.

Pombo, G., Gray, R., Cardoso, M. J., Ourselin, S., Rees, G., Ashburner, J. & Nachev, P. (2023). Equitable modelling of brain imaging by counterfactual augmentation with morphologically constrained 3D deep generative models. *Medical Image Analysis*, 84, 102723. doi: https://doi.org/10.1016/j.media.2022.102723.

Puyol-Antón, E., Ruijsink, B., Piechnik, S. K., Neubauer, S., Petersen, S. E., Razavi, R. & King, A. P. (2021). Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation. In de Bruijne, M., Cattin, P. C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y. & Essert, C. (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* (vol. 12903, pp. 413–423). Springer, Cham.

Puyol-Antón, E., Ruijsink, B., Piechnik, S. K., Neubauer, S., Petersen, S. E., Razavi, R. & King, A. P. (2022). Fairness in Cardiac Magnetic Resonance Imaging: Assessing Sex and Racial Bias in Deep Learning-Based Segmentation. *Frontiers in Cardiovascular Medicine*, 9.

Qu, L., Jin, Q., Fu, K., Wang, M. & Song, Z. (2024). Rethinking deep active learning for medical image segmentation: A diffusion and angle-based framework. *Biomedical Signal Processing and Control*, 96, 106493.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. arXiv. Retrieved from: arXivpreprintarXiv: 2103.00020.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X. & Wang, X. (2021). A Survey of Deep Active Learning. *ACM Computing Surveys*, 54(9), 180:1–180:40.

Ronneberger, O., Fischer, P. & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, (Lecture Notes in Computer Science), 234–241.

Rushmore, R. J., Sunderland, K., Carrington, H., Chen, J., Halle, M., Lasso, A., Papadimitriou, G., Prunier, N., Rizzoni, E., Vessey, B., Wilson-Braun, P., Rathi, Y., Kubicki, M., Bouix, S., Yeterian, E. & Makris, N. (2022a). Anatomically Curated Segmentation of Human Subcortical Structures in High Resolution Magnetic Resonance Imaging: An Open Science Approach. *Frontiers in Neuroanatomy*, 16, 894606. doi: 10.3389/fnana.2022.894606.

Rushmore, R., Sunderland, K., Carrington, H., Chen, J., Halle, M., Lasso, A., Papadimitriou, G., Prunier, N., Rizzoni, E., Vessey, B., Wilson-Braun, P., Rathi, Y., Kubicki, M., Bouix, S., Yeterian, E. & Makris, N. (2022b). Anatomically curated segmentation of human subcortical structures in high resolution magnetic resonance imaging: An open science approach. *Frontiers in Neuroanatomy*, 16, 894606. doi: 10.3389/fnana.2022.894606.

Sabuncu, M. R., Wang, A. Q. & Nguyen, M. (2025). Ethical Use of Artificial Intelligence in Medical Diagnostics Demands a Focus on Accuracy, Not Fairness. *NEJM AI*, 2(1), AIp2400672. doi: 10.1056/AIp2400672.

Santos, M. & Marreiros, G. (2025). A systematic review of active learning approaches in the selection of medical images. *Procedia Computer Science*, 256, 843–851. doi: 10.1016/j.procs.2025.02.186.

Scheffer, T., Decomain, C. & Wrobel, S. (2001). Active Hidden Markov Models for Information Extraction. *Advances in Intelligent Data Analysis*, 2189(Lecture Notes in Computer Science), 309–318.

Sener, O. & Savarese, S. (2018). Active Learning for Convolutional Neural Networks: A Core-Set Approach. *International Conference on Learning Representations (ICLR)*. Retrieved from: https://openreview.net/forum?id=H1aIuk-RW.

Shen, M., Zhang, J. Y., Chen, L., Yan, W., Jani, N., Sutton, B. & Koyejo, O. (2021). Labeling Cost Sensitive Batch Active Learning for Brain Tumor Segmentation. *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1269–1273.

Siddiqui, I. A., Littlefield, N., Carlson, L. A., Chhabra, A., Menezes, Z., Thakar, S. M., Abedian, M., Mastorakos, G. M., Moradi, H., Amirian, S., Gong, M., Lohse, I., Weiss, K. R., Plate, J. F. & Tafti, A. P. (2024). Fair AI-powered orthopedic image segmentation: addressing bias and promoting equitable healthcare. *Scientific Reports*.

Sizonenko, S. V., Babiloni, C., de Bruin, E. A., Isaacs, E. B., Jönsson, L. S., Kennedy, D. O., Latulippe, M. E., Mohajeri, M. H., Moreines, J., Pietrini, P., Walhovd, K. B., Winwood, R. J. & Sijben, J. W. (2013). Brain imaging and human nutrition: which measures to use in intervention studies? *British Journal of Nutrition*, 110(S1), S1–S30. doi: 10.1017/S0007114513001384.

Smailagic, A., Costa, P., Noh, H. Y., Walawalkar, D., Khandelwal, K., Galdran, A., Mirshekari, M., Fagert, J., Xu, S., Zhang, P. & Campilho, A. J. C. (2018). MedAL: Accurate and Robust Deep Active Learning for Medical Image Analysis. *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 481–488. doi: 10.1109/ICMLA.2018.00078.

Stanley, E. A., Wilms, M., Mouches, P. & Forkert, N. D. (2022a). Fairness-related performance and explainability effects in deep learning models for brain image analysis. *Journal of Medical Imaging*, 9(6), 061102.

Stanley, E. A. M., Wilms, M., Mouches, P. & Forkert, N. D. (2022b). Fairness-related performance and explainability effects in deep learning models for brain image analysis. *J. Med. Imaging (Bellingham)*, 9(6), 061102.

Stanley, E. A. M., Wilms, M. & Forkert, N. D. (2023). A Flexible Framework for Simulating and Evaluating Biases in Deep Learning-Based Medical Image Analysis. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2023*, 14221(Lecture Notes in Computer Science), 488–498.

Svanera, M., Savardi, M., Signoroni, A., Benini, S. & Muckli, L. (2024). Fighting the scanner effect in brain MRI segmentation with a progressive level-of-detail network trained on multi-site data. *Medical Image Analysis*, 93, 103090. doi: 10.1016/j.media.2024.103090.

Tae, W.-S., Ham, B.-J., Pyun, S.-B. & Kim, B.-J. (2025). Current Clinical Applications of Structural MRI in Neurological Disorders. *Journal of Clinical Neurology*, 21(4), 277–293. doi: 10.3988/jcn.2025.0185. PMID: 40635533; PMCID: PMC12303675.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B. & Liang, J. (2016). Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Transactions on Medical Imaging*, 35(5), 1299–1312. doi: 10.1109/TMI.2016.2535302.

Tian, Y., Shi, M., Luo, Y., Kouhana, A., Elze, T. & Wang, M. (2024). FairSeg: A Large-Scale Medical Image Segmentation Dataset for Fairness Learning Using Segment Anything Model with Fair Error-Bound Scaling. *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S. W. et al. (2013). The WU-Minn Human Connectome Project: An overview. *NeuroImage*, 80, 62–79.

Wang, D. & Shang, Y. (2014). A New Active Labeling Method for Deep Learning. *2014 International Joint Conference on Neural Networks (IJCNN)*, pp. 112–119. doi: 10.1109/IJCNN.2014.6889457.

Wang, G., Du, M., Liu, N., Zou, N. & Hu, X. (2022). Mitigating Algorithmic Bias with Limited Annotations. *arXiv preprint arXiv:2207.10018*.

Wang, H., Jin, Q., Li, S., Liu, S., Wang, M. & Song, Z. (2024). A comprehensive survey on deep active learning in medical image analysis. *Medical Image Analysis*, 95, 103201.

Wang, H., Chen, W., Luo, X., Xing, Z., Liu, L., Qin, J., Wu, S. & Zhu, L. (2025). Toward Fair and Accurate Cross-Domain Medical Image Segmentation: A VLM-Driven Active Domain Adaptation Paradigm. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24102–24112.

Wang, H., Suh, J. W., Das, S. R., Pluta, J. B., Craige, C. & Yushkevich, P. A. (2013a). Multi-Atlas Segmentation with Joint Label Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3), 611–623. doi: 10.1109/TPAMI.2012.143.

Wang, H., Suh, J. W., Das, S. R., Pluta, J. B., Craige, C. & Yushkevich, P. A. (2013b). Multi-Atlas Segmentation with Joint Label Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3), 611-623. Retrieved from: https://ieeexplore.ieee.org/document/6226425. doi:10.1109/TPAMI.2012.143.

Warthen, K. G., Boyse-Peacor, A., Jones, K. G., Sanford, B., Love, T. M. & Mickey, B. J. (2020). Sex differences in the human reward system: convergent behavioral, autonomic and neural evidence. *Social Cognitive and Affective Neuroscience*, 15(7), 789-801. doi: 10.1093/scan/nsaa104.

Weerts, H., van der Waa, J., van den Bosch, A. & van der Aalst, W. M. P. (2023). Active Learning Mitigates Selection Bias and Group Disparities. *Proceedings of the AAAI Workshop on Algorithmic Fairness through the Lens of Causality and Interpretability (AFUCI)*, 3470(CEUR Workshop Proceedings), 74–88.

Wissman, A. M., May, R. M. & Woolley, C. S. (2012). Ultrastructural analysis of sex differences in nucleus accumbens synaptic connectivity. *Brain Struct Funct*, 217(2), 181-90. doi: 10.1007/s00429-011-0353-6.

Xie, Y., Lu, H., Yan, J., Yang, X., Tomizuka, M. & Zhan, W. (2023). Active Finetuning: Exploiting Annotation Budget in the Pretraining-Finetuning Paradigm. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xie, Y., Zhang, J., Shen, C. & Xia, Y. [doi:10.48550/arXiv.2103.03024]. (2021). CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation. Retrieved from: https://arxiv.org/abs/2103.03024.

Xu, Z., Yu, K., Tresp, V., Xu, X. & Wang, J. (2003). Representative Sampling for Text Classification Using Support Vector Machines. *Advances in Information Retrieval: 25th European Conference on IR Research (ECIR 2003)*, 2633(Lecture Notes in Computer Science), 393–407.

Xu, Z., Li, J., Yao, Q., Li, H., Zhao, M. & Zhou, S. K. (2024a). Addressing fairness issues in deep learning-based medical image analysis: a systematic review. *npj Digital Medicine*, 7(1), 286. doi: 10.1038/s41746-024-01276-5.

Xu, Z., Li, J., Yao, Q., Li, H., Zhao, M. & Zhou, S. (2024b). Addressing fairness issues in deep learning-based medical image analysis: a systematic review. *npj Digital Medicine*, 7(1), 276. Retrieved from: https://www.nature.com/articles/s41746-024-01276-5. doi:10.1038/s41746-024-01276-5.

Yang, L., Zhang, Y., Chen, J., Zhang, S. & Chen, D. Z. (2017). Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, pp. 399–407. doi: 10.1007/978-3-319-66179-7_46.

Yang, Z., Zhang, J., Feng, F., Gao, C., Wang, Q. & He, X. (2023). Interactive active learning for fairness with partial group label. *AI Open*, 4, 175-182.

Yao, W., Bai, J., Liao, W., Chen, Y., Liu, M. & Xie, Y. (2024). From CNN to Transformer: A Review of Medical Image Segmentation Models. *Journal of Imaging Informatics in Medicine*, 37(4), 1529–1547. doi: 10.1007/s10278-024-00981-7. Epub 2024-03-04.

Yu, X., Yang, Q., Zhou, Y., Cai, L. Y., Gao, R., Lee, H. H., Li, T., Bao, S., Xu, Z., Lasko, T. A. et al. (2023). UNesT: local spatial representation learning with hierarchical transformer for efficient medical segmentation. *Medical Image Analysis*, 102939. Retrieved from: https://arxiv.org/abs/2209.14378. doi:10.48550/arXiv.2209.14378.

Zhou, H.-Y., Guo, J., Zhang, Y., Yu, L., Wang, L. & Yu, Y. (2021a). nnFormer: Interleaved Transformer for Volumetric Segmentation. *arXiv preprint arXiv:2109.03201*. doi: 10.48550/arXiv.2109.03201.

Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., van Ginneken, B., Madabhushi, A., Prince, J. L., Rueckert, D. & Summers, R. M. (2021b). A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises. *Proceedings of the IEEE*, 109(5), 820–838. doi: 10.1109/JPROC.2021.3054390.

Zhou, T., Li, L., Bredell, G., Li, J. & Konukoglu, E. (2021c). Quality-Aware Memory Network for Interactive Volumetric Image Segmentation. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 560–570. doi: 10.1007/978-3-030-87196-3_52.