A Journey Through Model Debiasing: from methods to applications

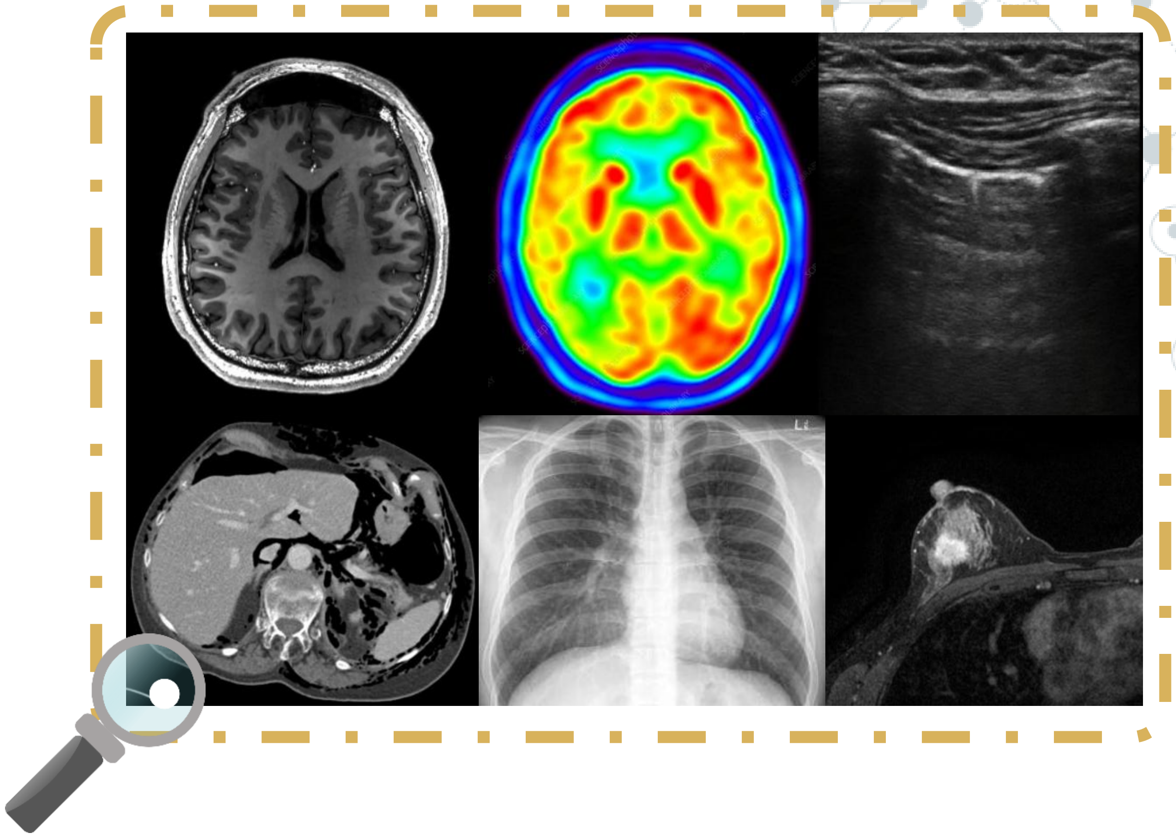
Bias Analysis in Artificial Intelligence for Medical Imaging

23rd International Conference on Image Analysis and Processing (ICIAP 2025)
September 15, 2025 – Rome, Italy

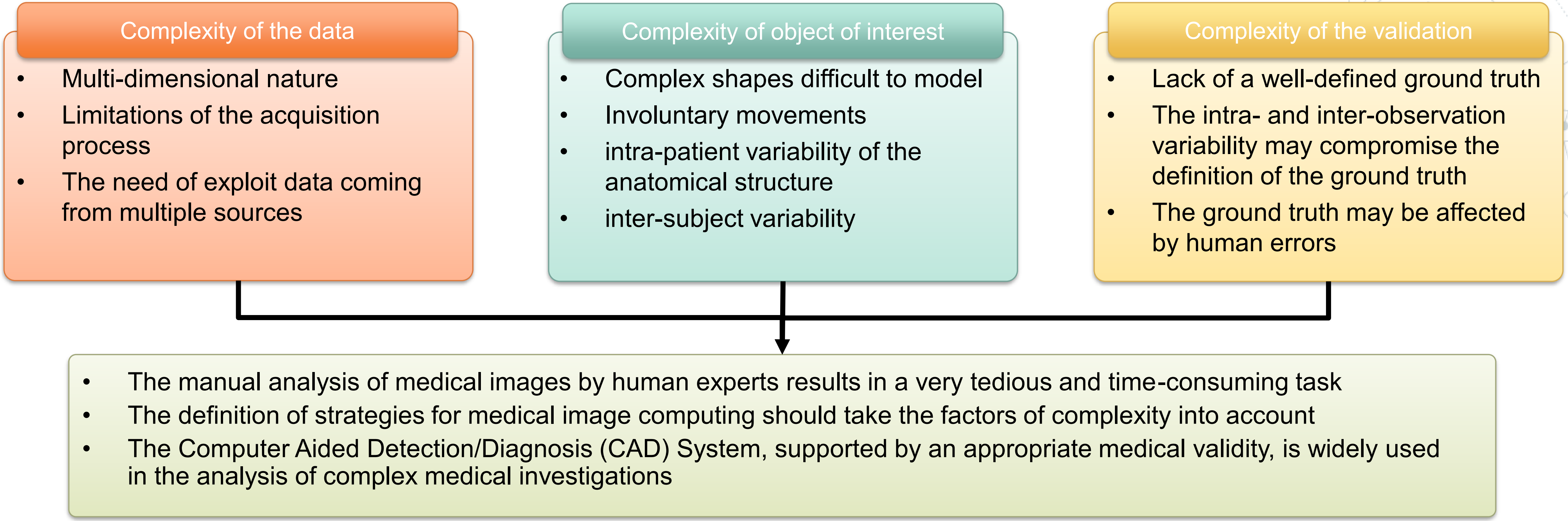
Medical Image Computing

- Medical Imaging refers to a field of medicine that deals with the visualization areas of a body and structures normally concealed by the sight
  - Novel imaging modalities are being introduced in medical practices as a result of ongoing technological advancements in image acquisition

- *Medical image analysis* or *medical image computing* refers to the process of extracting relevant information or knowledge from medical images with the aim of developing potential non-invasive biomarkers for the detection and characterization of the disease



Challenges in Medical Image Computing



Complexity of the data

- Multi-dimensional nature
- Limitations of the acquisition process
- The need of exploit data coming from multiple sources

Complexity of object of interest

- Complex shapes difficult to model
- Involuntary movements
- intra-patient variability of the anatomical structure
- inter-subject variability

Complexity of the validation

- Lack of a well-defined ground truth
- The intra- and inter-observation variability may compromise the definition of the ground truth
- The ground truth may be affected by human errors

- The manual analysis of medical images by human experts results in a very tedious and time-consuming task
- The definition of strategies for medical image computing should take the factors of complexity into account
- The Computer Aided Detection/Diagnosis (CAD) System, supported by an appropriate medical validity, is widely used in the analysis of complex medical investigations

The need for Artificial Intelligence

- The large amount of information to consider, and the high variability and complexity of medical images have prompted research into proposing solutions to automate the analysis of radiological acquisitions

- Artificial Intelligence (AI) refers to the simulation of human intelligence in machines and includes a set of strategies and algorithms that are able to discover hidden patterns in data while learning how to perform a specific task

- Many AI applications in medical field:
  - ✓ show very promising performance and cover all the steps implemented in a CAD system (pre-processing, segmentation, classification)
  - ✓ provide a way of finding non-invasive and quantitative assessments of diseases
  - ✓ might highlight pattern changes or intrinsic characteristics that are hidden from the human eye

- Radiomics is one of the most advanced applications for AI

Bias in AI for Healthcare

- Bias can be defined as the distance (or error) between the prediction and the actual target variable, whereas variance signifies the dependence of predictions on the randomness in the training data sampling
  - Bias refers to consistent deviations in AI predictions that result in unfair or unequal outcomes across different patient populations

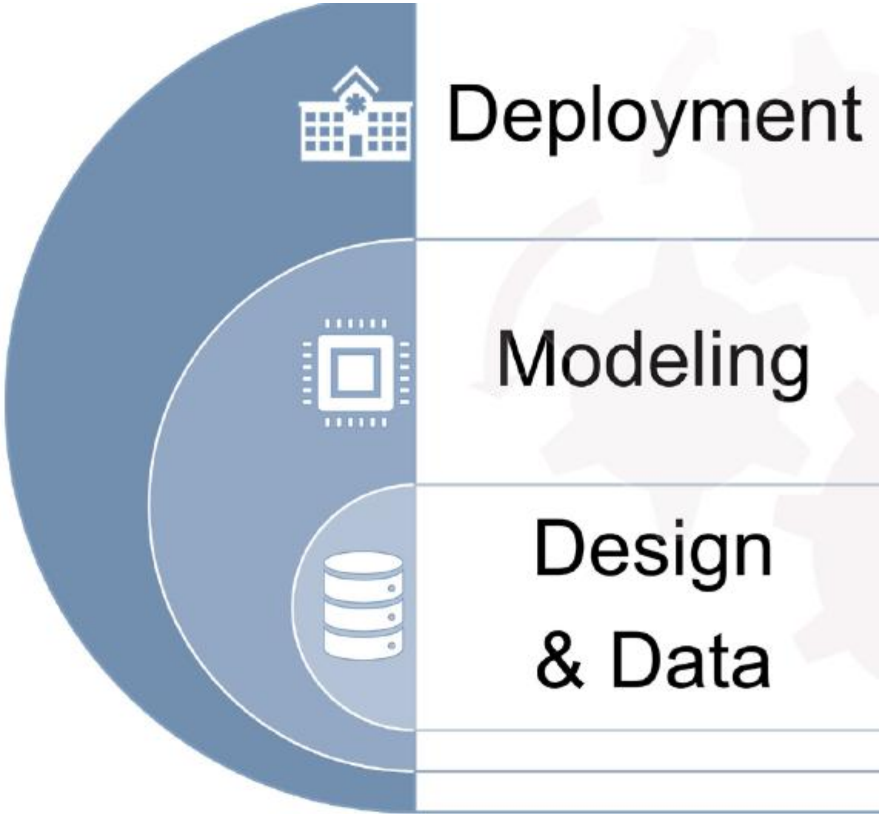
- Bias in healthcare AI is critical, as it can directly impact patient safety and equity.

- Common manifestations:
  - Worse performance on underrepresented populations
  - Amplification of existing health disparities
  - Reduced trust in AI systems

- Tackling bias is essential for building fair, reliable, and ethical healthcare AI.

Types of Bias

- Overview of potential biases and where they are most likely to occur along the medical imaging AI/ML pipeline.
- The dark shading with white dot indicates the most likely occurrence and lighter shading indicates additional potential occurrences.



Bias	Data collection	Data preparation	Model development	Model evaluation	Model deployment
Data acquisition and aggregation bias	●				
Biased synthetic data	●				
Exclusion bias	●	▨		▨	
Institutional/systemic bias	●	▨	▨		
Popularity/patient-based bias	●				▨
Population bias	●		▨	▨	
Temporal bias	●		▨	▨	▨
Sampling/representation/selection bias	●		▨	▨	▨
Activity bias	●		▨	▨	
Annotator bias		●		▨	▨
Content production bias		●			
Presentation bias		●		▨	
Inherited/error propagation bias			●		
Reference standard bias		▨	●	▨	▨
Membership bias	▨	▨	●	▨	▨
Historical bias	▨	▨	●	▨	▨
Training data bias	▨		●		
Cognitive bias		▨	●		▨
Evaluation bias			▨	●	▨
Detection bias	▨		▨	●	▨
Amplification bias	▨		▨	●	▨
Statistical bias				●	
Deployment bias					●
Concept drift/emergent bias					●
Behavioral bias	▨	▨	▨	▨	●
Uncertainty bias/epistemic uncertainty				▨	●
Funding/publication bias			▨	▨	●
Automation complacency/loss of situational awareness bias		▨			●
User interaction bias					●

[1] Koçak, Burak, et al. "Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects." Diagnostic and interventional radiology 31.2 (2025): 75.

[2] Drukker, Karen, et al. "Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment." Journal of Medical Imaging 10.6 (2023): 061104-061104.

Bias in Data Collection

Bias source	Definition
Data acquisition and generation bias	Is introduced when data (i) comes from limited acquisition sources, (ii) was collected under different standard processes, or (iii) was duplicated due to repeat collection or acquisition.
Biased synthetic data	Arises from the addition of biased synthetic data to a dataset.
Exclusion bias	Is introduced (i) when specific population groups are excluded from data collection, training, testing or subsequent analyses, or (ii) when some features from the dataset are excluded in AI/ML model training.
Institutional/systemic bias	Institutional/systemic bias occurs when the procedures and practices of institutions result in certain social groups being advantaged or favored and others being disadvantaged, devalued, or treated differently.
Popularity/patient-based bias	Occurs when current trends influence patients’ decision-making whether to undergo a specific test, which subsequently affects data collection.
Population bias	Arises when statistics, demographics, and characteristics differ between the original target population and the population represented in the actual dataset or platform.
Temporal bias	Arises from (i) differences in populations and behaviors over time, (ii) the use of data that is not representative of diagnostic clinical data, or (iii) the correlation of clinician/reader performance and state of knowledge of the disease.
Sampling/representation/ selection bias	Occurs when patient data used for training/tuning/testing an AI/ML model is not representative of the patient population to which the algorithm is intended to be applied.
Activity bias	Occurs when models are trained with data from regions or clinical sites that are active in using certain modalities (e.g., imaging specialties), archiving data, and developing models.

Bias in Data Preparation

Bias source	Definition
Annotator bias	Occurs when human annotators, or human–computer assisted systems, apply subjective, selective, and/or biased labels in the annotation process.
Content production bias	A form of behavioral bias that is expressed as lexical, syntactic, semantic, and structural differences in the content generated by users. These differences may impact the generalizability of research that utilizes user-generated content like annotations or patient-reported information.
Presentation bias	Results from the way in which images, AI/ML output, or other data are presented to the user or the annotator.

Bias in Model Development

Bias source	Definition
Inherited/error propagation bias	Occurs when machine learning models are used to generate inputs for other machine learning algorithms or trained incrementally.
Reference standard bias	Occurs when there are inconsistent reference test methods, inconsistent procedures in which a given test is performed, inconsistent ways in which results are interpreted, or ignoring indeterminate findings.
Membership bias	Occurs when membership in particular groups present systemic differences that do not necessarily correspond with to the outcome of prediction being pursued in the target population.
Historical bias	Arises from systemic societal, institutional, and individual, engrained biases and impacts prioritization of problems to pursue.
Training data bias	Occurs when there is a mismatch between the training set and the intended use.
Cognitive bias	Arises when a system of belief, typically built upon data of limited validity and sets of heuristic, subjective assessments of physical quantities or outcomes, used to reduce the complexity of tasks produces systematic bias/errors in judgement of the underlying reality.

Bias in Model Evaluation

Bias source	Definition
Evaluation bias	Arises through improper benchmark datasets, improper use of data or performance metrics.
Detection bias	Refers to systematic differences between different groups in the detection rate or severity evaluation for a disease or condition.
Amplification bias	Occurs when an AI/ML algorithm learns to predict output/classes with a greater disparity than what is in the underlying ground truth.
Statistical bias	Is the average difference between a quantity we estimate from data and the actual value of the quantity.

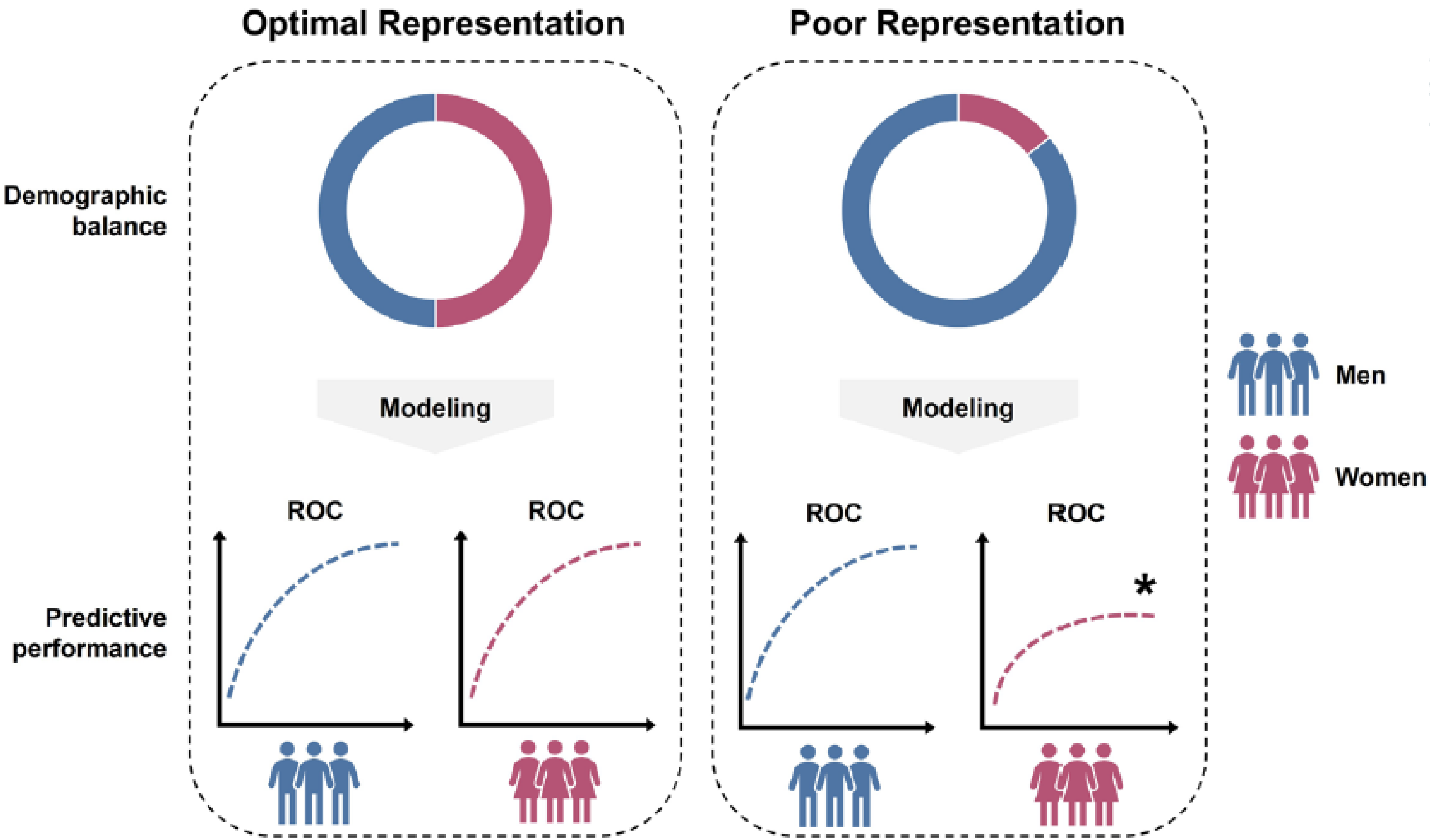
Bias in Model Deployment

Bias source	Definition
Deployment bias	Arises when there is a mismatch between the intended use of a system or algorithm and the way it is used in practice. This misuse may cause harmful decisions or consequences.
Concept drift/emergent bias	Occurs when the performance of machine learning models estimated in the laboratory setting degrades over time in the real world when the image acquisition equipment, clinical conditions, and patient population characteristics change.
Behavioral bias	Arises through systematic distortions in user behavior across platforms or contexts, or across users represented in different datasets.
Uncertainty bias/epistemic uncertainty	Is the influence of both reducible (epistemic) and irreducible (aleatoric) uncertainty on decision making drawn from AI/ML models.
Funding/publication bias	Arises through selective reporting of results.
Automation complacency/loss of situational awareness bias	Caused by over-reliance on automation.
User interaction bias	Can occur when users interact with data and algorithm outputs based on their inherent biases or a biased user-interface, impacting end user choices and decisions.

Impacts of Bias

- **Patient Harm.** Biased models can misdiagnose diseases or delay treatment, leading to avoidable health risks.
- **Health Inequities.** Bias disproportionately affects already vulnerable groups (by race, gender, age, or socioeconomic status), widening existing gaps in healthcare access and outcomes.
- **Lack of Generalization.** Models trained on narrow or homogeneous datasets fail when applied to different populations, hospitals, or devices.
- **Loss of Trust.** If AI systems are perceived as unfair or unsafe, clinicians and patients may resist adoption, slowing down innovation in healthcare.
- **Regulatory and Legal Risks**. Failure to address bias can lead to non-compliance with frameworks like the EU AI Act, with consequences for certification, liability, and deployment

Example of Bias



Koçak, Burak, et al. "Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects." *Diagnostic and interventional radiology* 31.2 (2025): 75.

Challenges in Handling Bias

- Ambiguities in interpreting results can pose significant challenges in the development and clinical use of AI software. These refer to situations where the interpretation of the results is not unique or is open to multiple meanings by the users

- Limited diversity in benchmark datasets can represent a significant challenge in AI development and generalizability. This can occur when some diseases or events are collected with underrepresentation or overrepresentation compared with their prevalence in the general population or clinical practice due to the limited patient diversity included in the training data

- Publicly accessible benchmarks are essential for comparison for AI models and represent a crucial element of open science. Multicentric databases can potentially overcome this challenge by collecting a large number of diverse and representative data in rarer conditions

- Subjectivity in the detection of bias can be related to personal interpretation and individual perspectives related to the identification of the bias itself.

Bias Evaluation: a case of study

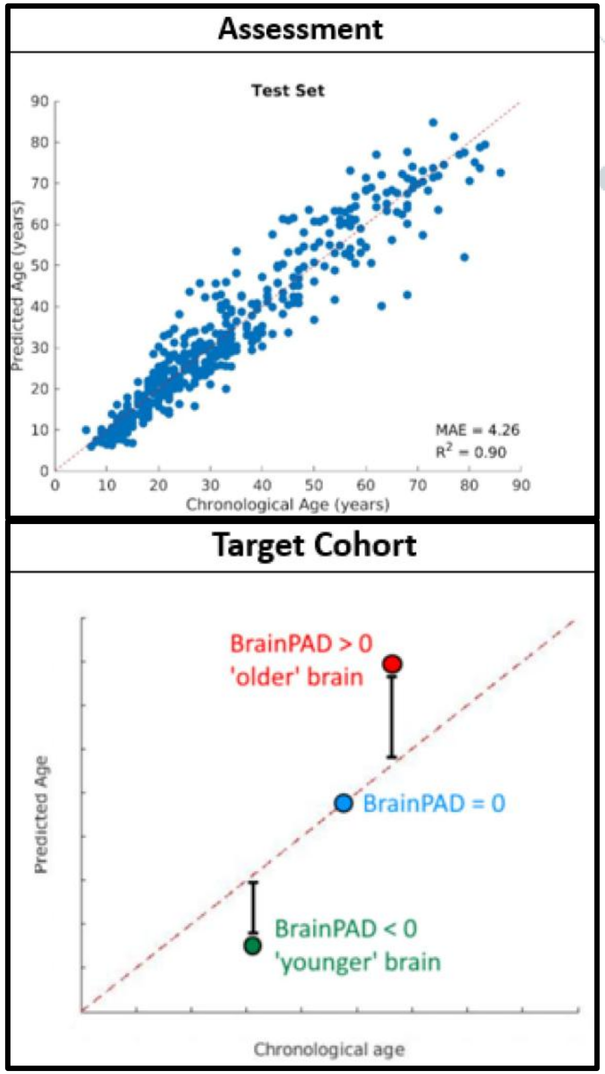
Brain Aging and the Brain-Age Paradigm

- Brain aging involves structural changes associated with functional deterioration and neurodegenerative diseases
- Biological "brain-age" may diverge from chronological age, potentially indicating age-related health risks more accurately

- **Brain-Age Paradigm for Brain Health**
  - Developing imaging-derived markers for brain health and pathology using Machine Learning (ML)
  - Modelling chronological age based on brain MRI scans in healthy people to create a baseline for "healthy" brain aging

- **Brain-Predicted Age Difference (Brain-PAD)**
  - The difference between predicted and chronological age severs as an index of structural brain health, detecting pathology across neurological and psychiatric disorders

- **Deep Learning in Brain Age Prediction**
  - It leverages neural networks to learn high-level representations of brain images, achieving high performance



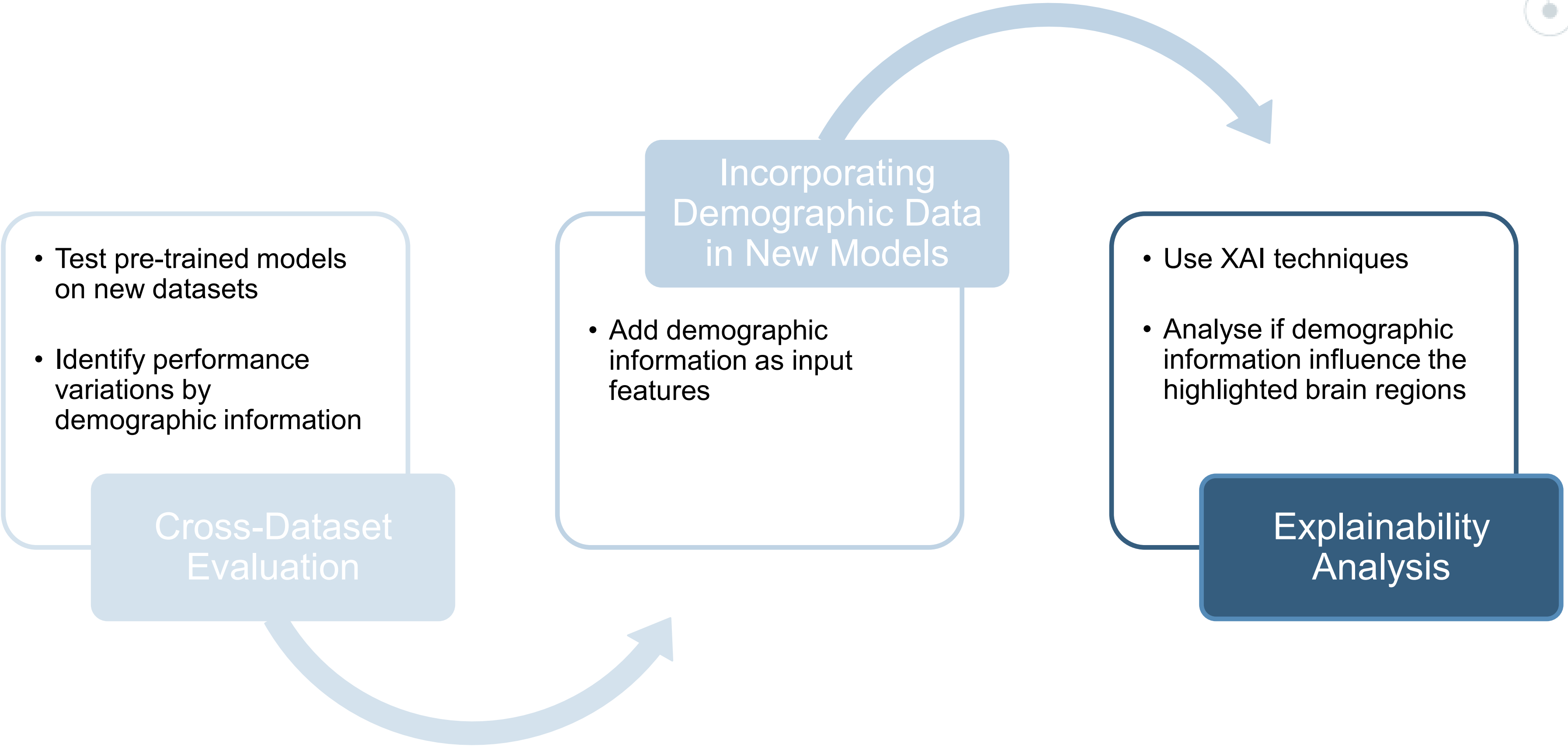
Importance of Fairness in Brain Age Prediction

- ML algorithms may underperform or behave unfairly in populations with demographic differences from training data

- **Ethical and Practical Implications**:
  - Biased predictions can perpetuate societal biases, leading to ethical issues in diagnosis
  - Inaccurate results may result in misdiagnoses or incorrect medical interventions

- **Promoting Fairness in DL Models**
  - Developing strategies to ensure consistent model performance across diverse populations
  - Ensuring fairness can enhance real-world applicability and foster equitable healthcare

Proposed Evaluation Schema



Cross-Dataset Evaluation

- Test pre-trained models on new datasets
- Identify performance variations by demographic information

Incorporating Demographic Data in New Models

- Add demographic information as input features

Explainability Analysis

- Use XAI techniques
- Analyse if demographic information influence the highlighted brain regions

Data Collection

- 10 publicly available dataset have been collected:
  - Cognitive Normal (CN) subjects
  - Patients with different neurodegenerative disease:
    - Alzheimer’s Disease (AD)
    - Parkinson’s Disease (PD)
    - Mild Cognitive Impairment (MCI)

  - Three different ethnic groups (White, Black, Asian)

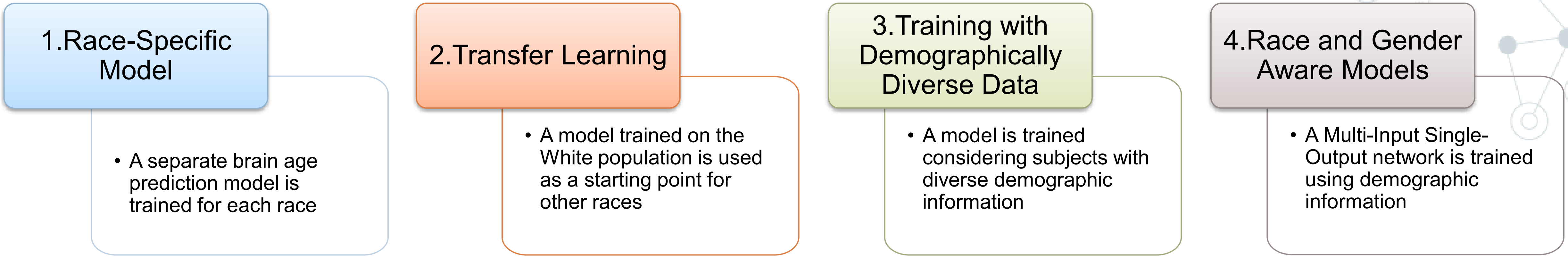
- 2 datasets are used for external evaluation

	NAME	GROUP	ETHNICITY	Images	Subjects	GENDER	Images	Subjects
External	CORR	CN	White			Male	425	213
			Black			Female	441	230
			Asian	866	443			
	HABS	CN	White	3435	2162	Male	2631	1052
			Black	708	694	Female	1512	1804
			Asian					

	NAME	GROUP	ETHNICITY	Images	Subjects	GENDER	Images	Subjects
Internal	ADNI	CN	White	970	201	Male	530	112
			Black	69	16	Female	521	108
			Asian	12	3			
		PATIENT	White	2433	534	Male	1557	340
			Black	92	23	Female	1026	229
			Asian	58	12			
	CamCAN	CN	White	625	625	Male	325	325
			Black	2	2	Female	312	312
			Asian	10	10			
	ICBM	CN	White	614	134	Male	356	78
			Black	101	21	Female	428	92
			Asian	69	15			
	IXI	CN	White	450	450	Male	238	238
			Black	15	15	Female	291	291
			Asian	64	64			
	MCSA	CN	White	1466	1466	Male	763	763
			Black			Female	703	703
			Asian					
		PATIENT	White	210	210	Male	125	125
			Black			Female	85	85
			Asian					
	NKI	CN	White	131	129	Male	120	117
			Black	61	58	Female	88	84
			Asian	16	14			
	OASIS	CN	White	2557	805	Male	1156	397
			Black	362	150	Female	1781	565
			Asian	18	7			
		PATIENT	White	384	224	Male	239	139
			Black	57	45	Female	203	131
			Asian	1	1			
	PPMI	CN	White	152	133	Male	107	91
			Black	11	10	Female	58	54
			Asian	2	2			
		PATIENT	White	1433	1045	Male	873	605
			Black	27	20	Female	610	475
			Asian	23	15			

Methodology

- Different Training Paradigms (TPs) are proposed:



1.Race-Specific Model

- A separate brain age prediction model is trained for each race

2.Transfer Learning

- A model trained on the White population is used as a starting point for other races

3.Training with Demographically Diverse Data

- A model is trained considering subjects with diverse demographic information

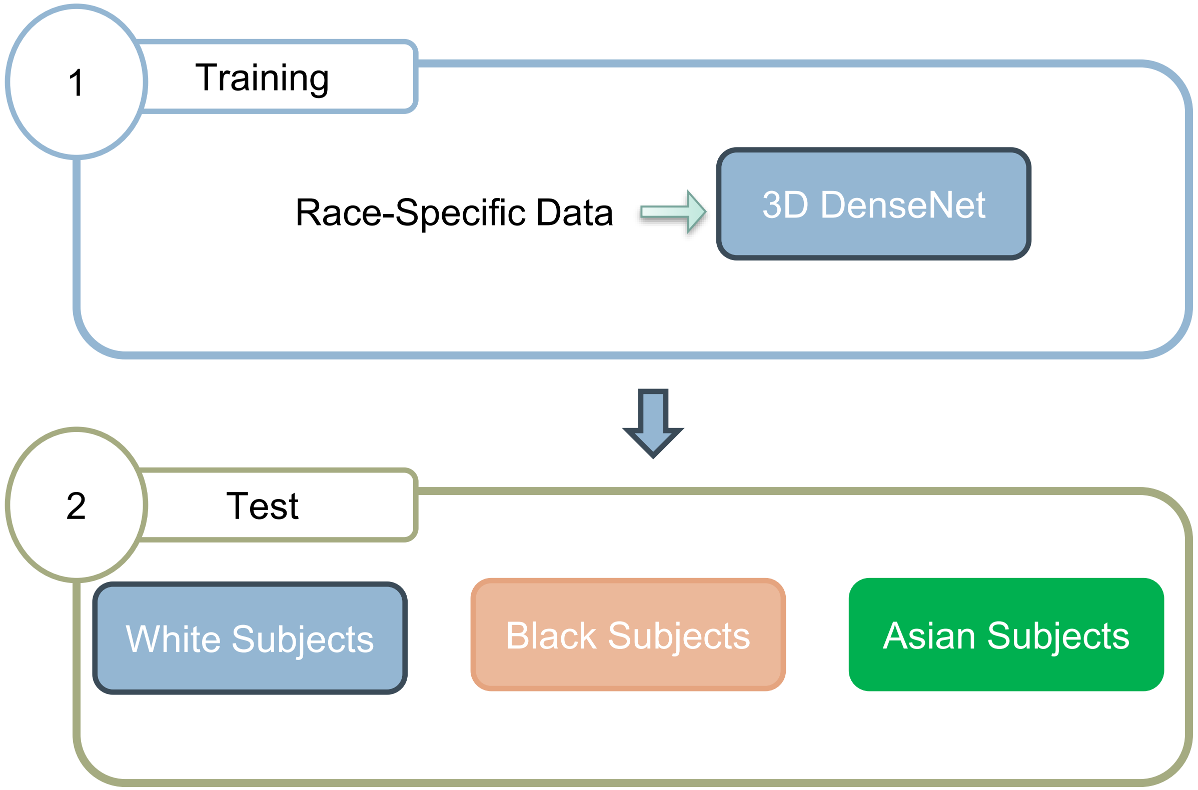
4.Race and Gender Aware Models

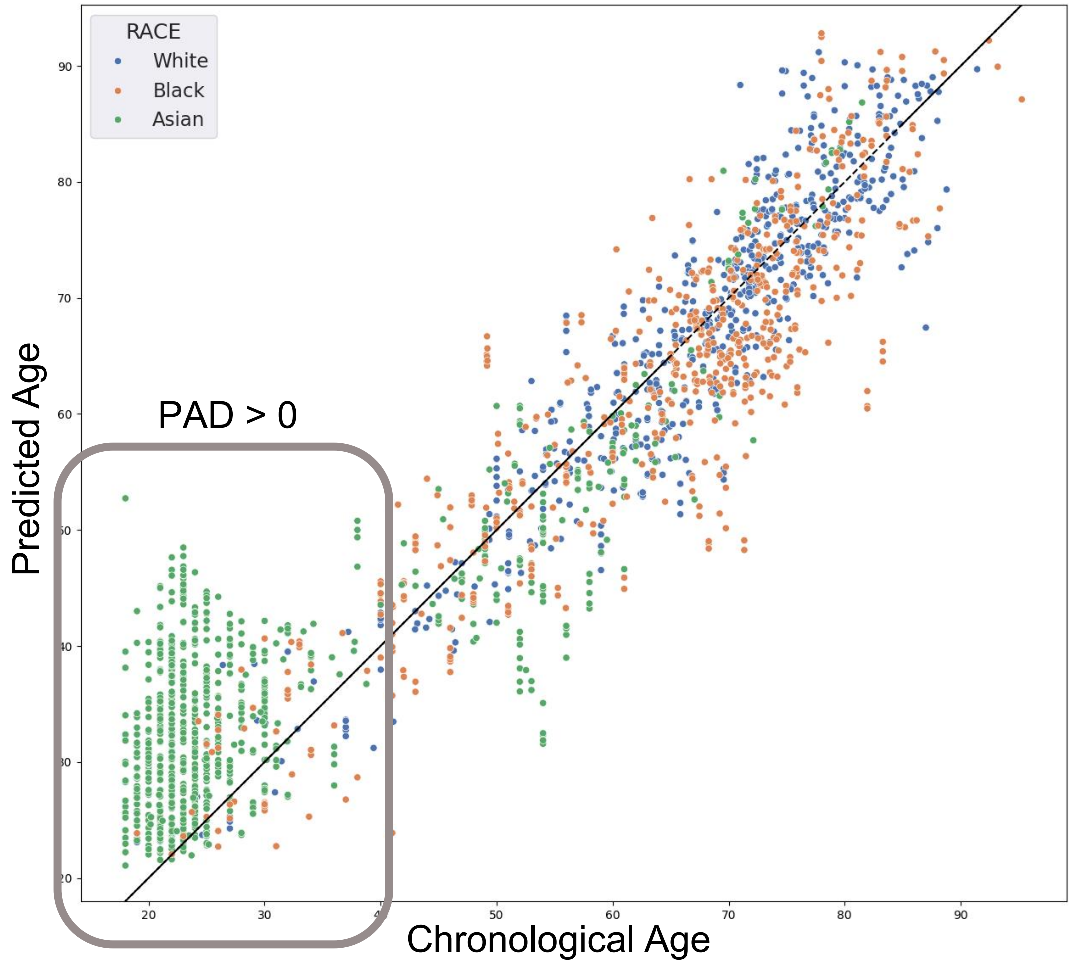
- A Multi-Input Single-Output network is trained using demographic information

- 3D DenseNet architecture is trained for the task of brain age prediction using a carefully stratified dataset

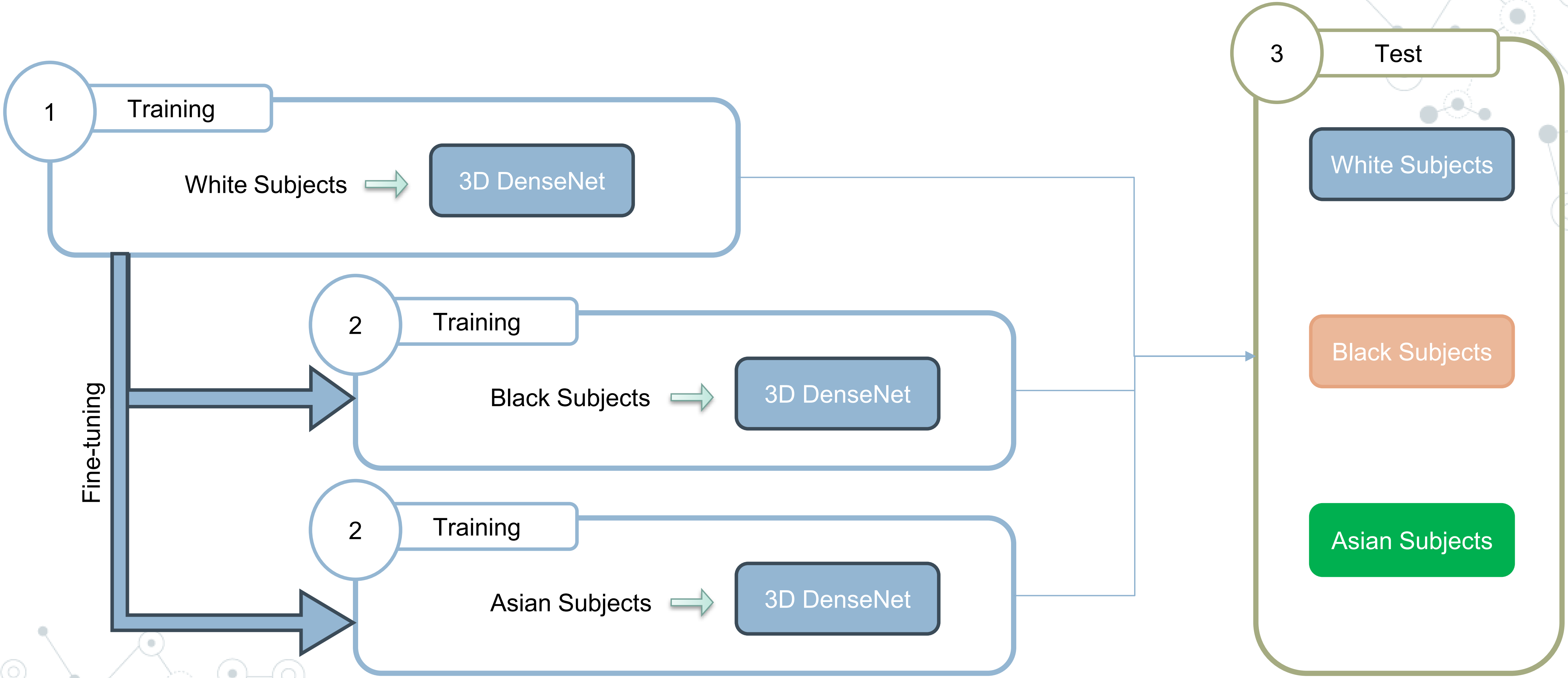
- The performance of the models was evaluated using several key metrics
  - ➢ **Brain-PAD**: the difference between the predicted brain age and the actual chronological age
  - ➢ **Mean Absolute Error (MAE)**: the average magnitude of errors
  - ➢ **R^2 coefficient**: the proportion of variance in the target variable explained by the model
  - ➢ **Correlation coefficient (C)**: the strength and direction of the linear relationship between predicted and actual brain ages

Race-Specific Model (TP1)

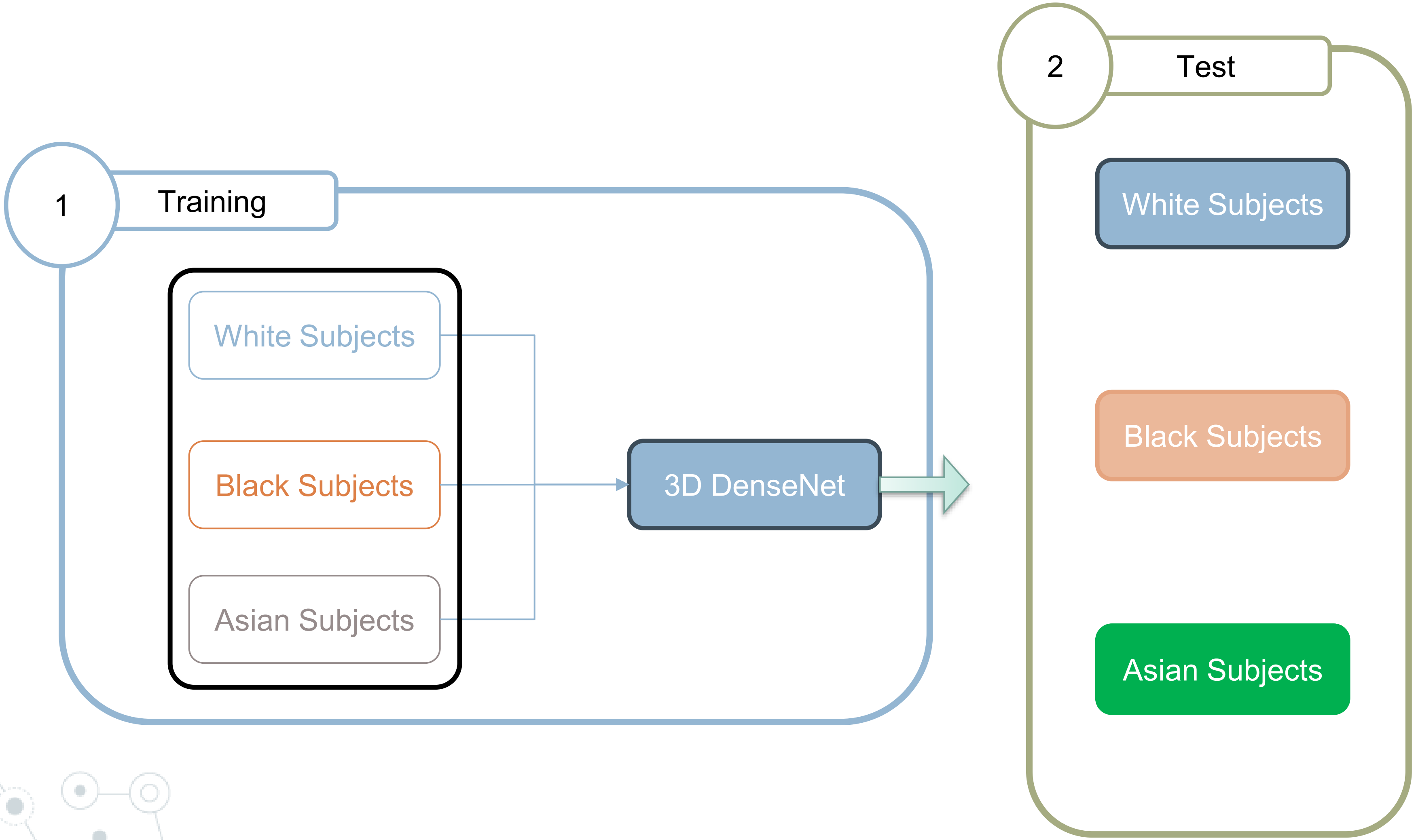




Transfer Learning (TP2)



Training with Demographically Diverse Data (TP3)



1 Training

White Subjects

Black Subjects

Asian Subjects

3D DenseNet

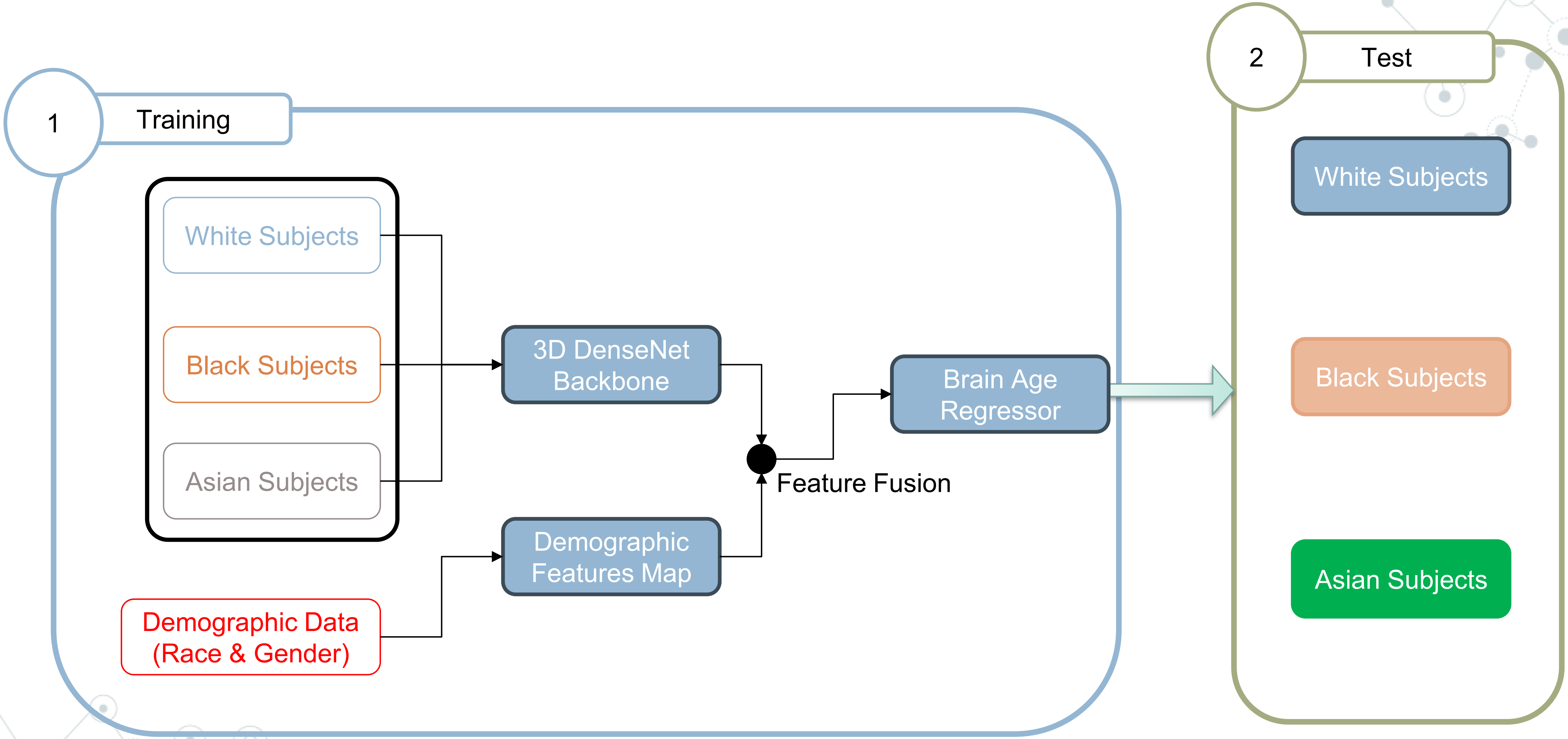
2 Test

White Subjects

Black Subjects

Asian Subjects

Race and Gender Aware Models (TP4)



Results on External Test-set

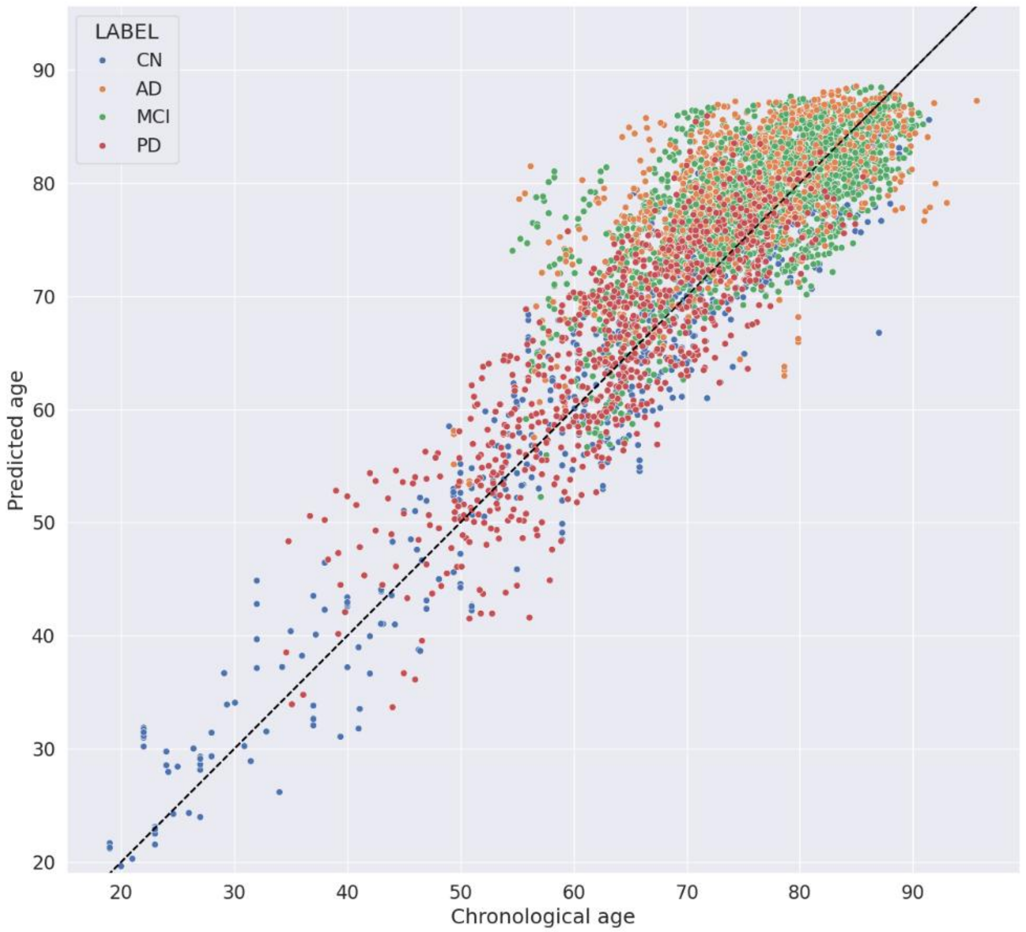
Train	Method	Test Race	#Sam	MAE	PAD	R^2	C
White	1.Race-Specific Model	White	3435	2.96	**-0.02**	0.83	0.91
Asian	1.Race-Specific Model	White	3435	6.79	-0.89	0.02	0.51
Black	1.Race-Specific Model	White	3435	5.14	-0.52	0.43	0.75
Asian	2.Transfer Learning	White	3435	2.99	0.25	0.83	**0.92**
Black	2.Transfer Learning	White	3435	3.01	0.96	0.83	**0.92**
Balanced	3. Demographically Diverse Data	White	3435	2.96	-0.14	0.83	0.91
Balanced	4.Race and Gender Aware Models	White	3435	**2.91**	0.66	**0.84**	**0.92**
White	1.Race-Specific Model	Black	708	3.85	0.93	0.59	0.81
Asian	1.Race-Specific Model	Black	708	7.92	-2.40	-0.64	0.47
Black	1.Race-Specific Model	Black	708	3.20	**-0.01**	0.76	0.89
Asian	2.Transfer Learning	Black	708	4.01	1.26	0.58	0.80
Black	2.Transfer Learning	Black	708	**2.82**	0.77	**0.81**	**0.92**
Balanced	3. Demographically Diverse Data	Black	708	3.30	1.17	0.76	0.89
Balanced	4.Race and Gender Aware Models	Black	708	2.89	1.20	0.79	0.91
White	1.Race-Specific Model	Asian	727	11.87	10.36	-0.30	0.69
Asian	1.Race-Specific Model	Asian	727	19.14	13.55	-2.43	-0.17
Black	1.Race-Specific Model	Asian	727	13.62	11.86	-0.80	0.51
Asian	2.Transfer Learning	Asian	727	**2.64**	1.40	**0.93**	**0.97**
Black	2.Transfer Learning	Asian	727	6.63	4.35	0.45	0.77
Balanced	3. Demographically Diverse Data	Asian	727	3.59	2.20	0.87	0.96
Balanced	4.Race and Gender Aware Models	Asian	727	3.03	**1.22**	0.90	0.95

Results on External Test-set

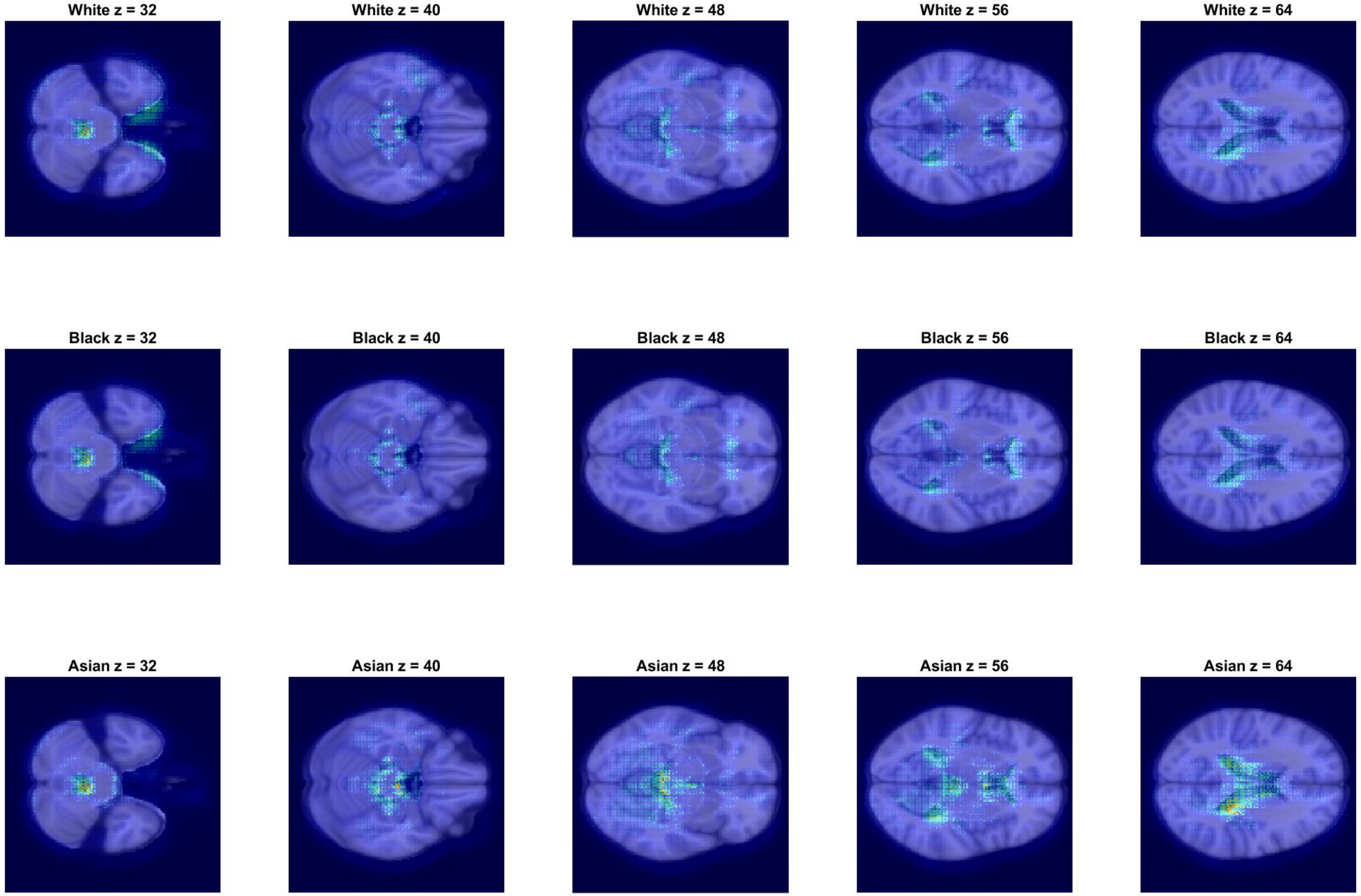
Train	Method	Test Race	#Sam	MAE	PAD	R^2	C
White	1.Race-Specific Model	White	3435	2.96	**-0.02**	0.83	0.91
Asian	1.Race-Specific Model	White	3435	6.79	-0.89	0.02	0.51
Black	1.Race-Specific Model	White	3435	5.14	-0.52	0.43	0.75
Asian	2.Transfer Learning	White	3435	2.99	0.25	0.83	**0.92**
Black	2.Transfer Learning	White	3435	3.01	0.96	0.83	**0.92**
Balanced	3. Demographically Diverse Data	White	3435	2.96	-0.14	0.83	0.91
Balanced	4.Race and Gender Aware Models	White	3435	**2.91**	0.66	**0.84**	**0.92**
White	1.Race-Specific Model	Black	708	3.85	0.93	0.59	0.81
Asian	1.Race-Specific Model	Black	708	7.92	-2.40	-0.64	0.47
Black	1.Race-Specific Model	Black	708	3.20	**-0.01**	0.76	0.89
Asian	2.Transfer Learning	Black	708	4.01	1.26	0.58	0.80
Black	2.Transfer Learning	Black	708	**2.82**	0.77	**0.81**	**0.92**
Balanced	3. Demographically Diverse Data	Black	708	3.30	1.17	0.76	0.89
Balanced	4.Race and Gender Aware Models	Black	708	2.89	1.20	0.79	0.91
White	1.Race-Specific Model	Asian	727	11.87	10.36	-0.30	0.69
Asian	1.Race-Specific Model	Asian	727	19.14	13.55	-2.43	-0.17
Black	1.Race-Specific Model	Asian	727	13.62	11.86	-0.80	0.51
Asian	2.Transfer Learning	Asian	727	**2.64**	1.40	**0.93**	**0.97**
Black	2.Transfer Learning	Asian	727	6.63	4.35	0.45	0.77
Balanced	3. Demographically Diverse Data	Asian	727	3.59	2.20	0.87	0.96
Balanced	4.Race and Gender Aware Models	Asian	727	3.03	**1.22**	0.90	0.95

Analysis of patients with neurodegenerative diseases (Internal Test-set)

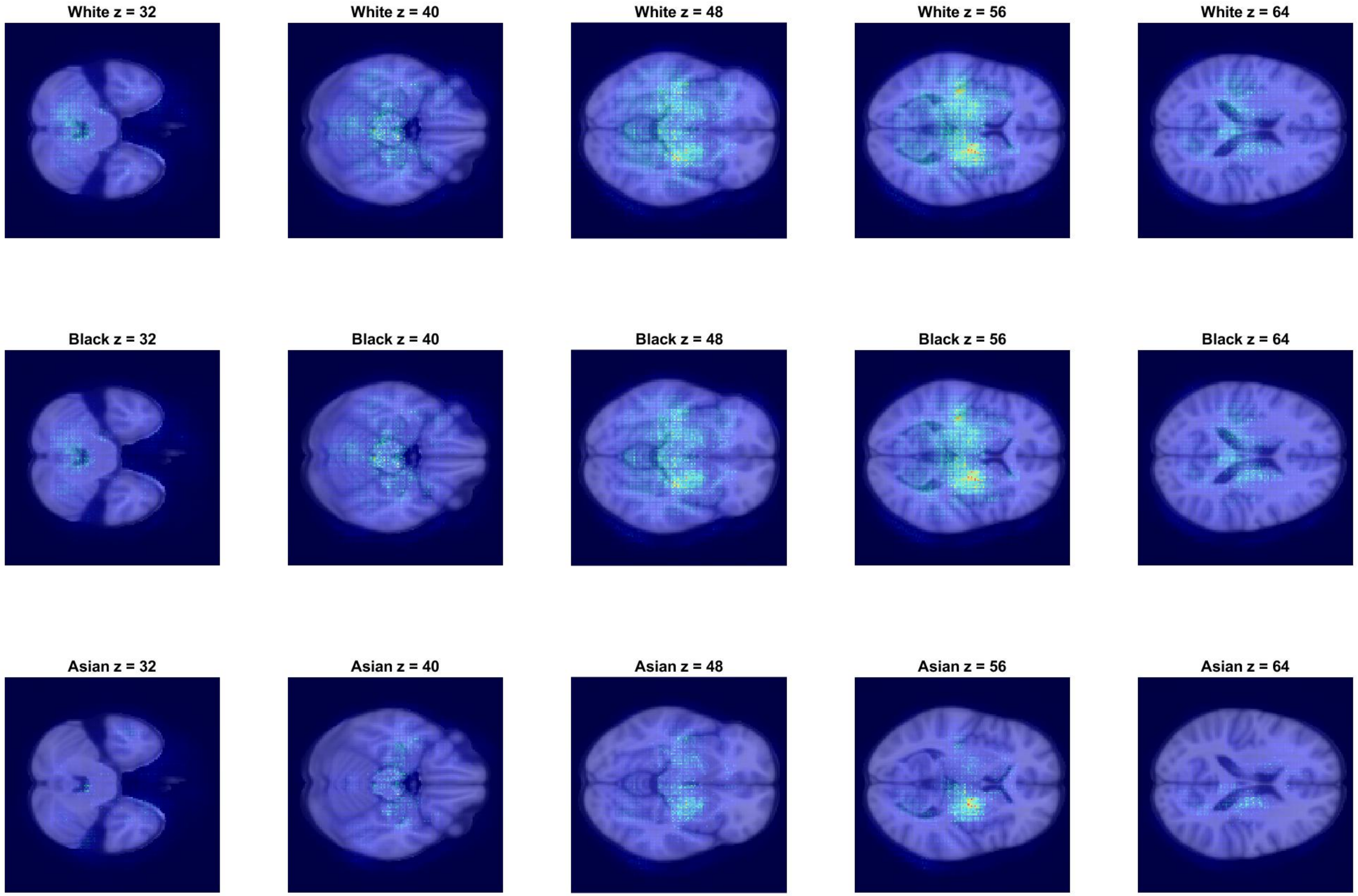
Train	Method	Race	Label	#Samples	MAE	PAD	R^2	C
White	1.Race-Specific Model	White	CN	689	3.69	-0.33	0.91	0.96
			AD	945	5.55	3.75	0.18	0.65
			MCI	1872	4.56	2.29	0.36	0.69
			PD	775	4.36	1.29	0.68	0.85
		Black	CN	611	4.70	-1.61	0.84	0.92
			AD	82	5.15	1.69	0.48	0.74
			MCI	67	3.77	1.17	0.39	0.68
			PD	23	8.08	6.28	0.14	0.67
		Asian	CN	185	4.69	2.15	0.91	0.96
			AD	8	3.64	3.04	-41.01	0.51
			MCI	51	3.47	1.05	0.72	0.88
			PD	17	5.09	2.26	0.40	0.76
Balanced	4.Race and Gender Aware Models	White	CN	689	3.62	-0.27	0.91	0.96
			AD	945	5.61	3.77	0.17	0.66
			MCI	1872	4.51	2.59	0.36	0.70
			PD	775	4.43	1.45	0.65	0.84
		Black	CN	46	4.06	0.09	0.79	0.90
			AD	82	5.43	2.96	0.42	0.75
			MCI	67	3.98	2.09	0.38	0.72
			PD	23	6.72	4.86	0.49	0.82
		Asian	CN	20	3.24	2.11	0.96	0.99
			AD	8	9.66	9.66	-167.30	0.58
			MCI	51	8.82	8.62	-1.18	0.42
			PD	17	4.92	3.31	0.35	0.64



Explainability Analysis – Guided Backpropagation

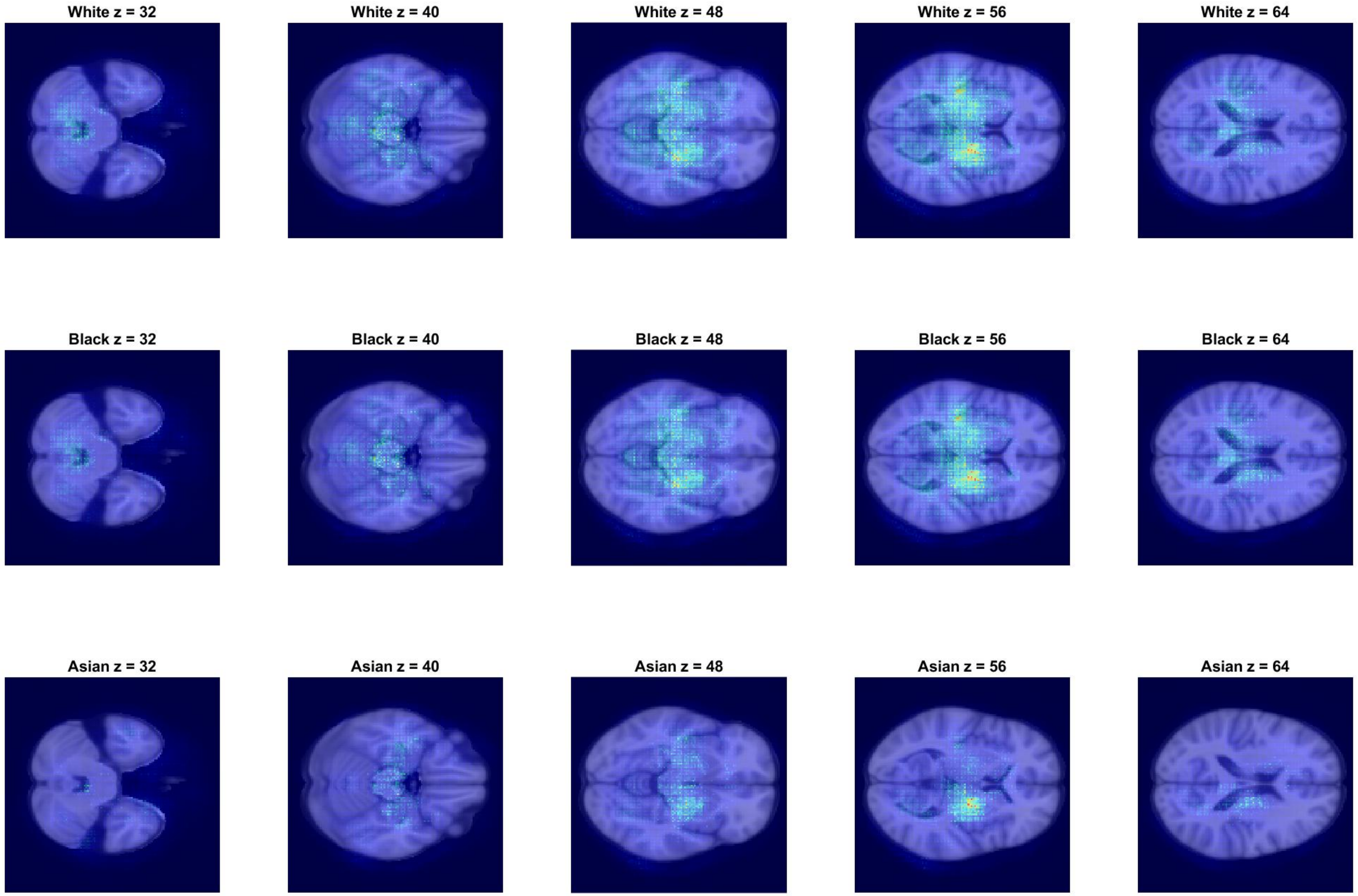


Explainability Analysis – Saliency Maps

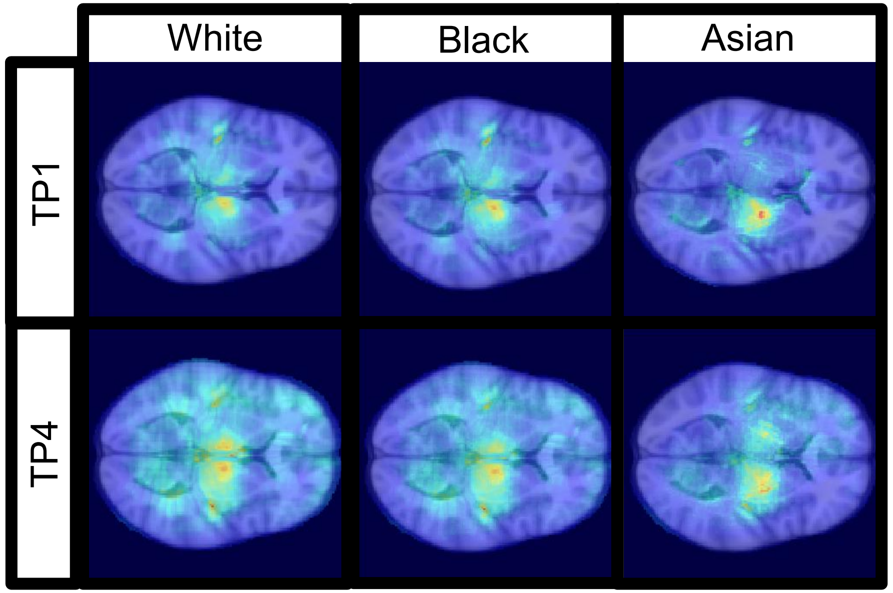


White z = 32
White z = 40
White z = 48
White z = 56
White z = 64
Black z = 32
Black z = 40
Black z = 48
Black z = 56
Black z = 64
Asian z = 32
Asian z = 40
Asian z = 48
Asian z = 56
Asian z = 64

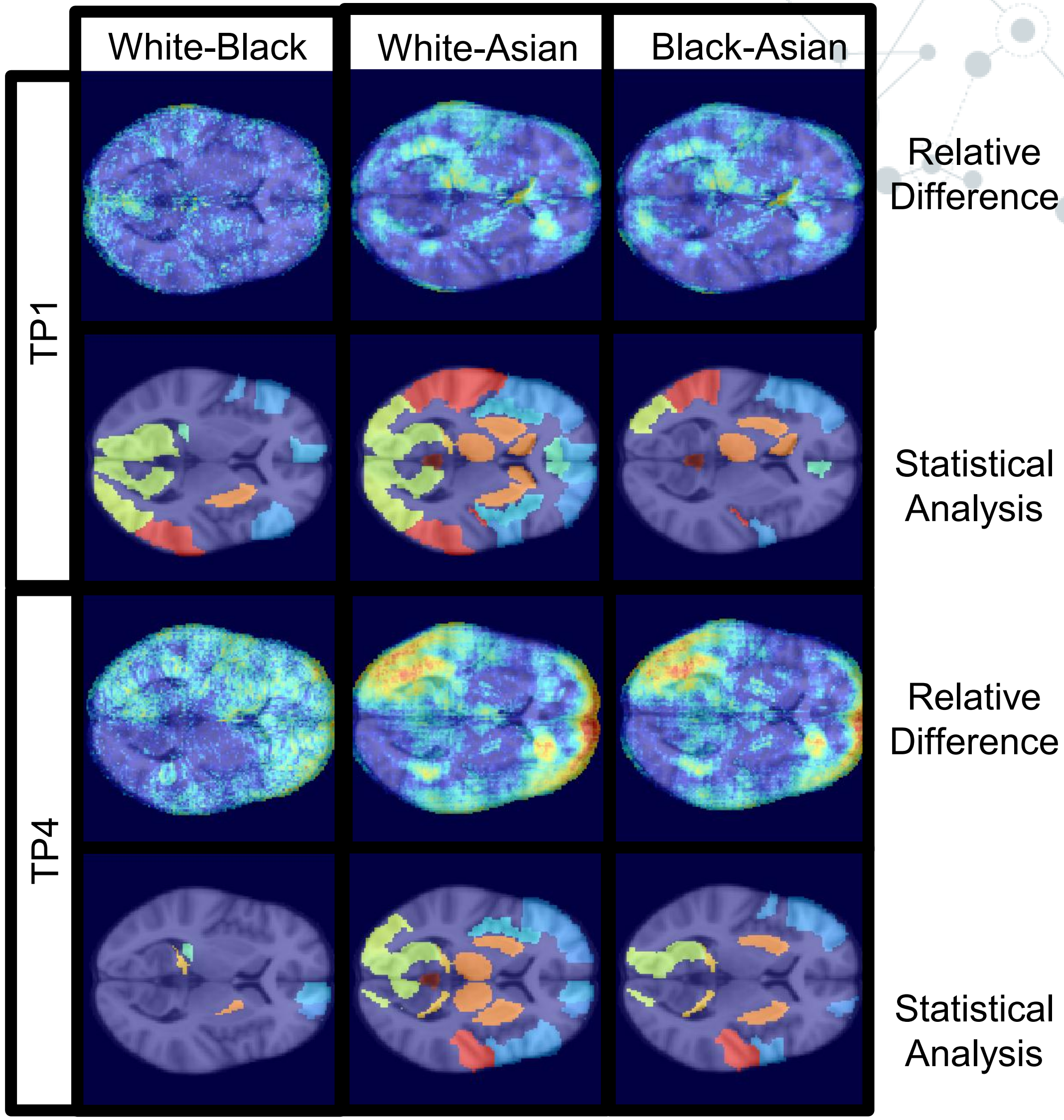
Explainability Analysis – Saliency Maps



Explainability Analysis – Statistical Analysis



Population Specific Maps



Conclusions

- **Insight into Model Bias**
  - Comprehensive evaluation of ethnicity influence on brain prediction
  - Quantitative assessment of model generalizability and fairness

- **Focus on Bias Mitigation**
  - Developing strategies to reduce bias in brain-age prediction
  - Promoting fair and ethical outcomes in clinical applications

michela.gravina@unina.it