The Impact of MRI Data Harmonization on Brain Age Prediction

Junhyeok Lee
Department of Software Convergence
Kyung Hee University
Republic of Korea
bluehyena123@khu.ac.kr

Juhyuk Han
Department of Software Convergence
Kyung Hee University
Republic of Korea
hhannn@khu.ac.kr

Minjae Kim
Department of Software Convergence
Kyung Hee University
Republic of Korea
kmj5596@khu.ac.kr

Yeonwoo Kim
Department of Software Convergence
Kyung Hee University
Republic of Korea
dusdn@khu.ac.kr

Tae-Seong Kim
Department of Biomedical Engineering
Kyung Hee University
Republic of Korea
tskim@khu.ac.kr

Won Hee Lee*
Department of Software Convergence
Kyung Hee University
Republic of Korea
whlee@khu.ac.kr

***Abstract*—Machine learning (ML) techniques are increasingly used in brain age prediction to assess brain health and detect neurological and psychiatric disorders. The availability of large, publicly accessible imaging datasets has accelerated the adoption of ML-based methods. However, multi-site MRI datasets present challenges due to site effects, which can introduce biases and affect the accuracy of brain age prediction. In this study, we examined the influence of MRI data harmonization on brain age prediction by comparing models trained with and without harmonization across a large-scale, multi-site dataset of 10,938 healthy individuals aged 5 to 95 years. Using automated ML approaches, we trained various models and computed SHapley Additive exPlanations (SHAP) values to identify the key features driving brain age predictions. Our results showed that while a weighted ensemble method achieved high prediction accuracy (MAE = 7.013; R = 0.860), data harmonization reduced prediction performance, indicating that site-related variability contains valuable information influencing model predictions. SHAP analysis also revealed substantial site-specific biases impacting the predictions. These findings suggest the need to account for site-specific factors in multi-site MRI studies. Understanding the impact of site harmonization is crucial for developing robust and generalizable brain age prediction models that can be applied across diverse populations and imaging settings.**

***Index Terms*—brain age, machine learning, magnetic resonance imaging, site effects, feature importance**

I. Introduction

Brain age prediction, a rapidly evolving field in neuroscience, aims to estimate an individual's chronological age based on neuroimaging data [1]. This technique holds significant promise for early detection of brain disorders and understanding the underlying mechanisms of brain aging [2]–[4]. However, the accuracy and generalizability of brain age prediction models can be hindered by the challenges posed by multi-site neuroimaging data [5].

Multi-site studies often involve data collected from different institutions using varying imaging protocols, scanner types, and acquisition parameters. These differences can introduce biases and artifacts, known as site effects or batch effects, that can confound the analysis and reduce the model's ability to accurately predict brain age across diverse populations [5].

To address these challenges, MRI data harmonization techniques have been developed. These methods aim to reduce the impact of site-specific variations, enabling more consistent and comparable analyses across different datasets. By harmonizing data, researchers can improve the generalizability and reliability of brain age prediction models, leading to more accurate and meaningful results [5].

This study leverages automated machine learning (AutoML) to optimize brain age prediction models and investigates the impact of MRI data harmonization on their performance. AutoML streamlines the model development process by automating tasks such as model selection and hyperparameter tuning, enabling efficient and effective model building. We aim to examine the impact of MRI data harmonization on brain age prediction by comparing the performance of brain age prediction models trained with and without harmonization. We also compute SHapley Additive exPlanations (SHAP) values to identify influential features driving brain age prediction. This study contributes to the advancement of brain age prediction by providing valuable insights into the role of site harmonization and the potential benefits of using AutoML in this field.

II. Methods

A. Datasets

We compiled a large-scale T1-weighted MRI dataset by integrating data from 20 independent, publicly available sources. The final dataset includes 10,938 healthy individuals (5,692 female) with an age range of 5 to 95 years [6]–[25]. Detailed demographic and dataset-specific information are provided in Table I.

B. Data Processing and Feature Extraction

We used FreeSurfer (version 7.2.0) [26] to extract structural features from the MRI scans. A total of 215 features were

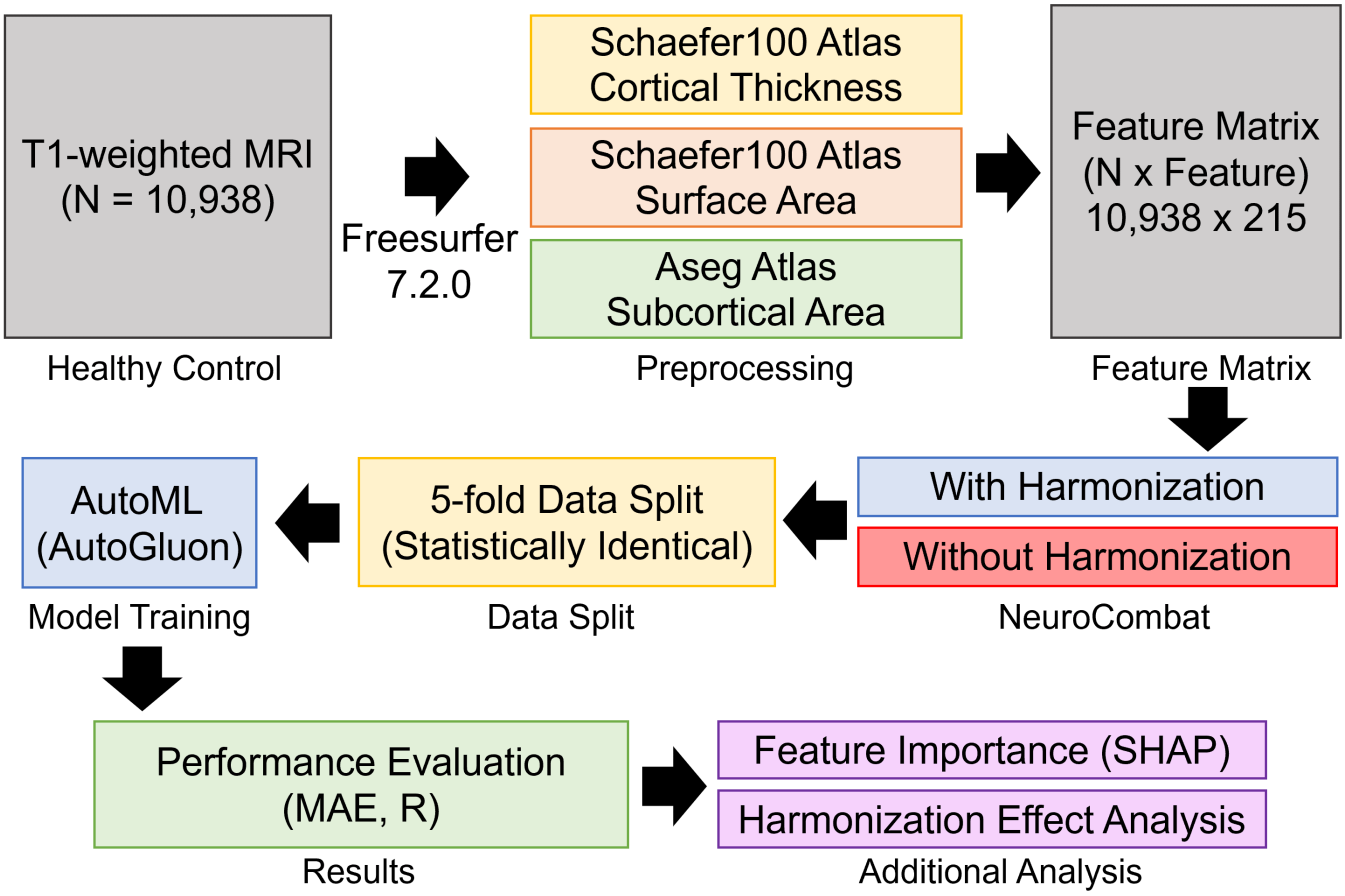

Fig. 1. Overview of the workflow for brain age prediction using AutoML.

TABLE I
SUMMARY OF DATASETS WITH PARTICIPANT DETAILS

Dataset	N	M/F	Mean Age±S.D.	Age Range
DLBS	314	117/197	54.00 ± 20.04	20–89
fcon1000	987	554/491	30.10 ± 14.40	7–85
HBN	971	628/343	10.33 ± 3.37	5.05–21.22
HCP	1,061	486/575	28.75 ± 3.67	22–37
IXI	562	250/312	48.65 ± 16.47	19.98–86.32
MCIC	94	64/30	32.63 ± 11.97	18–60
AOMIC	209	89/120	22.18 ± 1.79	18.25–26.25
BGSP	1,493	632/861	21.53 ± 2.89	19–35
BNU	180	73/107	21.22 ± 1.93	17–28
Cam-CAN	631	312/318	54.93 ± 18.38	18–88
COBRE	93	67/26	37.63 ± 11.66	18–65
CoRR	1,373	692/681	24.60 ± 13.66	6–84
DecNef	949	538/411	36.59 ± 15.52	18–80
NARPS	108	48/60	25.54 ± 3.59	18–37
NPC	65	29/36	26.55 ± 4.30	20–35
NUSDAST	98	53/45	31.96 ± 13.76	14–67
OASIS-3	714	306/408	68.68 ± 8.91	42–95
SALD	494	187/307	45.18 ± 17.44	19–80
SLIM	387	244/305	20.07 ± 1.26	17–27
UCLACNP	125	66/59	31.52 ± 8.79	21–50
Total	**10,938**	**5,246/5,692**	**32.61 ± 19.14**	**5.05–95**

selected for analysis, including 100 cortical thickness (CT) and 100 surface area (SA) features derived from the Schaefer100 atlas [27], along with 14 subcortical volume features from the Aseg atlas [28] and intracranial volume as an additional feature. To mitigate site-specific effects from the 20 independent datasets, we applied the neuroCombat harmonization technique [29], which effectively removed confounding influences related to different scanning sites. The dataset was then split into five folds for cross-validation, and we used the Kolmogorov-Smirnov test to ensure that each fold was stratified by age, with an even age distribution across the folds.

C. Machine Learning Models

We employed a variety of machine learning models for brain age prediction, including KNeighborsUnif [30], KNeighborsDist [31], ExtraTrees [32], RandomForest [33], XGBoost [34], LightGBM [35], LightGBMXT [35], LightGBMLarge [35], CatBoost [36], NeuralNetFastAI [37], NeuralNetTorch [38], and WeightedEnsemble [38]. Each of these models was selected for its distinct capabilities in handling complex, high-dimensional data, and the ensemble method was designed to further improve accuracy by leveraging the strengths of multiple models.

TABLE II
MODEL PERFORMANCE COMPARISON BETWEEN HARMONIZED AND NON-HARMONIZED SAMPLES

Model	Non-harmonized		Harmonized	
	MAE	R	MAE	R
KNeighborsUnif	10.573	0.643	12.821	0.482
KNeighborsDist	10.511	0.646	12.795	0.485
ExtraTrees	7.630	0.848	9.277	0.776
RandomForest	7.408	0.846	9.035	0.779
XGBoost	6.580	0.874	8.050	0.823
LightGBMLarge	6.543	0.875	8.000	0.825
LightGBMXT	6.230	0.888	7.643	0.843
LightGBM	6.362	0.882	7.795	0.835
CatBoost	6.302	0.887	7.673	0.841
NeuralNetFastAI	5.659	0.900	7.630	0.838
NeuralNetTorch	5.468	0.897	7.274	0.841
WeightedEnsemble	**5.340**	**0.910**	**7.013**	**0.860**

D. Experiment Settings and Evaluation Metrics

We utilized the AutoML library AutoGluon [38] for training our machine learning models, applying the default preset *"medium_quality"* for hyperparameter optimization. The models were trained using root mean squared error (RMSE) as the loss function to optimize prediction accuracy. Our hardware setup included an Intel(R) Core(TM) i9-10900X CPU @ 3.70GHz paired with an NVIDIA RTX 4090 24GB GPU, running on Ubuntu 20.04.6 LTS.

The model performance was evaluated based on mean absolute error (MAE) and Pearson's correlation coefficient (R) between predicted brain age and chronological age, ensuring robust assessment of predictive accuracy and consistency.

III. RESULTS

A. Effect of Data Harmonization

Table II shows that models trained on non-harmonized data (MAE = 5.340 – 10.573; R = 0.643 – 0.910) consistently outperformed those trained on harmonized data (MAE = 7.013 – 12.821; R = 0.482 – 0.860). This trend was observed across all models, with the WeightedEnsemble model exhibiting a notable decline in performance after harmonization (MAE increased by 1.673, R decreased by 0.050). Our results demonstrate that site harmonization did not improve model accuracy, as evidenced by higher MAEs and slightly lower R values (Table II). This suggests that harmonization may have inadvertently disrupted some of the valuable variability present in non-harmonized data. These findings suggest the importance of carefully considering the impact of data harmonization on model performance. While harmonization is often seen as a necessary step to mitigate site-specific biases, our results suggest that it can also reduce the model's ability to capture important information. Future research should explore alternative approaches to addressing site-specific biases that may preserve more of the valuable variability in the data.

B. Feature Importance

We computed SHAP values using the best-performing model to assess and compare feature importance between harmonized and non-harmonized datasets (Table III). In the non-harmonized data, SHAP analysis revealed substantial variability in feature importance across sites, indicating the presence of site-specific biases. For example, the left accumbens ranked as the most important feature with a SHAP value of 1.318, potentially reflecting site-related artifacts rather than true biological relevance in brain age prediction. Similarly, the right putamen volume, while consistently identified as a key feature, showed varying importance (SHAP value of 1.230) across different sites, highlighting the influence of site-specific factors on the model's interpretation. Several cortical features, such as cortical thickness in the left PFC and surface area in the left precentral, also ranked highly in the non-harmonized dataset, but may have been influenced by site-specific effects rather than true predictive power. The left thalamus volume, which had a lower SHAP value (0.556) before harmonization, became highly important after harmonization, suggesting that site-related factor might have obscured its true significance. These findings suggest the critical role of site harmonization, as site-specific biases can distort the model's interpretation of feature importance, leaning to less reliable and generalizable predictions.

TABLE III
TOP 10 ABSOLUTE MEAN SHAP VALUES FOR HARMONIZED AND NON-HARMONIZED SAMPLES (L = LEFT; R = RIGHT)

Harmonized	
Feature	**Mean(|SHAP value|)**
Intracranial volume	2.446
L thalamus volume	1.498
R thalamus volume	0.876
R amygdala volume	0.870
R putamen volume	0.848
R precentral CT	0.754
L occipital CT	0.656
L PFC CT	0.606
L fronto-opercular/insula SA	0.599
L temporal SA	0.505
Non-harmonized	
Feature	**Mean(|SHAP value|)**
L accumbens SA	1.318
R putamen volume	1.230
L PFC CT	0.709
L precentral SA	0.675
L occipital CT	0.657
R pallidum volume	0.628
L thalamus volume	0.556
L medial CT	0.550
L posterior CT	0.544
L medial CT	0.526

IV. CONCLUSION

This study demonstrates the effectiveness of AutoML for brain age prediction in multi-site MRI datasets, particularly when combined with a weighted ensemble approach. While AutoML simplifies the model development process, our findings highlight the importance of careful consideration of site harmonization. Site harmonization, while intended to reduce biases, can have varying effects on model performance. In our study, it often led to decreased accuracy, suggesting that site-specific factors may contain valuable information for

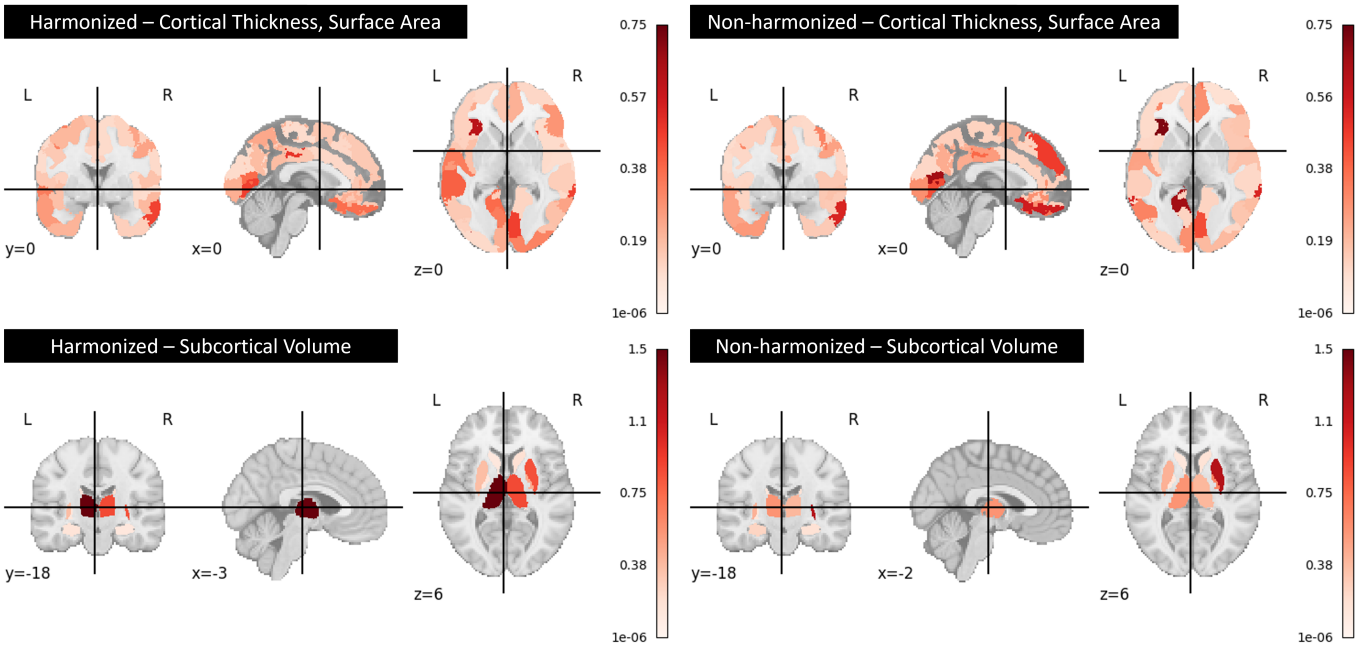

Fig. 2. Comparison of absoulte mean SHAP values across brain regional features (cortical thicknes, surface area, and subcortical volume) between harmonized and non-harmonized dataset.

prediction. SHAP value analysis further revealed the influence of site-specific biases on feature importance, emphasizing the need for targeted harmonization techniques. Future research should focus on developing more effective site harmonization methods tailored to specific datasets and imaging modalities. By addressing site-specific biases, we can improve the accuracy and generalizability of brain age prediction models, enabling more reliable insights into brain health and disease. Furthermore, applying these findings to specific populations, such as individuals with schizophrenia, can provide valuable insights into disease-related brain changes and inform the development of targeted interventions.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP), funded by the Korea government (MSIT) under grants RS-2024-00509257 (Global AI Frontier Lab), IITP-2024-RS-2024-00438239 (ITRC, Information Technology Research Center), and RS-2022-00155911 (Artificial Intelligence Convergence Innovation Human Resources Development, Kyung Hee University).

References

[1] J. H. Cole and K. Franke, "Predicting age using neuroimaging: innovative brain ageing biomarkers," *Trends in neurosciences*, vol. 40, no. 12, pp. 681–690, 2017.

[2] C. Constantinides, L. K. Han, C. Alloza, L. A. Antonucci, C. Arango, R. Ayesa-Arriola, N. Banaj, A. Bertolino, S. Borgwardt, J. Bruggemann *et al.*, "Brain ageing in schizophrenia: evidence from 26 international cohorts via the enigma schizophrenia consortium," *Molecular psychiatry*, vol. 28, no. 3, pp. 1201–1209, 2023.

[3] S. Shahab, B. H. Mulsant, M. L. Levesque, N. Calarco, A. Nazeri, A. L. Wheeler, G. Foussias, T. K. Rajji, and A. N. Voineskos, "Brain structure, cognition, and brain age in schizophrenia, bipolar disorder, and healthy controls," *Neuropsychopharmacology*, vol. 44, no. 5, pp. 898–906, 2019.

[4] J. Seitz-Holland, S. S. Haas, N. Penzel, A. Reichenberg, and O. Pasternak, "Brainage, brain health, and mental disorders: A systematic review," *Neuroscience & Biobehavioral Reviews*, p. 105581, 2024.

[5] J. Chen, J. Liu, V. D. Calhoun, A. Arias-Vasquez, M. P. Zwiers, C. N. Gupta, B. Franke, and J. A. Turner, "Exploration of scanning effects in multi-site structural mri studies," *Journal of neuroscience methods*, vol. 230, pp. 37–50, 2014.

[6] J. Park, J. Carp, K. M. Kennedy, K. M. Rodrigue, G. N. Bischof, C.-M. Huang, J. R. Rieck, T. A. Polk, and D. C. Park, "Neural broadening or neural attenuation? investigating age-related dedifferentiation in the face network in a large lifespan sample," *Journal of Neuroscience*, vol. 32, no. 6, pp. 2154–2158, 2012.

[7] B. B. Biswal, M. Mennes, X.-N. Zuo, S. Gohel, C. Kelly, S. M. Smith, C. F. Beckmann, J. S. Adelstein, R. L. Buckner, S. Colcombe *et al.*, "Toward discovery science of human brain function," *Proceedings of the national academy of sciences*, vol. 107, no. 10, pp. 4734–4739, 2010.

[8] L. M. Alexander, J. Escalera, L. Ai, C. Andreotti, K. Febre, A. Mangone, N. Vega-Potler, N. Langer, A. Alexander, M. Kovacs *et al.*, "An open resource for transdiagnostic research in pediatric mental health and learning disorders," *Scientific data*, vol. 4, no. 1, pp. 1–26, 2017.

[9] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium *et al.*, "The wu-minn human connectome project: an overview," *Neuroimage*, vol. 80, pp. 62–79, 2013.

[10] "https://brain-development.org/."

[11] R. L. Gollub, J. M. Shoemaker, M. D. King, T. White, S. Ehrlich, S. R. Sponheim, V. P. Clark, J. A. Turner, B. A. Mueller, V. Magnotta *et al.*, "The mcic collection: a shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia," *Neuroinformatics*, vol. 11, pp. 367–388, 2013.

[12] L. Snoek, M. M. van der Miesen, T. Beemsterboer, A. Van Der Leij, A. Eigenhuis, and H. Steven Scholte, "The amsterdam open mri collection, a set of multimodal mri datasets for individual difference analyses," *Scientific data*, vol. 8, no. 1, p. 85, 2021.

[13] A. J. Holmes, M. O. Hollinshead, T. M. O'keefe, V. I. Petrov, G. R. Fariello, L. L. Wald, B. Fischl, B. R. Rosen, R. W. Mair, J. L. Roffman *et al.*, "Brain genomics superstruct project initial data release with

structural, functional, and behavioral measures," *Scientific data*, vol. 2, no. 1, pp. 1–16, 2015.
[14] L. Tian, J. Wang, C. Yan, and Y. He, "Hemisphere-and gender-related differences in small-world brain networks: a resting-state functional mri study," *Neuroimage*, vol. 54, no. 1, pp. 191–202, 2011.
[15] M. A. Shafto, L. K. Tyler, M. Dixon, J. R. Taylor, J. B. Rowe, R. Cusack, A. J. Calder, W. D. Marslen-Wilson, J. Duncan, T. Dalgleish *et al.*, "The cambridge centre for ageing and neuroscience (cam-can) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing," *BMC neurology*, vol. 14, pp. 1–25, 2014.
[16] C. Aine, H. J. Bockholt, J. R. Bustillo, J. M. Cañive, A. Caprihan, C. Gasparovic, F. M. Hanlon, J. M. Houck, R. E. Jung, J. Lauriello *et al.*, "Multimodal neuroimaging in schizophrenia: description and dissemination," *Neuroinformatics*, vol. 15, pp. 343–364, 2017.
[17] X.-N. Zuo, J. S. Anderson, P. Bellec, R. M. Birn, B. B. Biswal, J. Blautzik, J. Breitner, R. L. Buckner, V. D. Calhoun, F. X. Castellanos *et al.*, "An open science resource for establishing reliability and reproducibility in functional connectomics," *Scientific data*, vol. 1, no. 1, pp. 1–13, 2014.
[18] S. C. Tanaka, A. Yamashita, N. Yahata, T. Itahashi, G. Lisi, T. Yamada, N. Ichikawa, M. Takamura, Y. Yoshihara, A. Kunimatsu *et al.*, "A multi-site, multi-disorder resting-state magnetic resonance image database," *Scientific data*, vol. 8, no. 1, p. 227, 2021.
[19] R. Botvinik-Nezer, R. Iwanir, F. Holzmeister, J. Huber, M. Johannesson, M. Kirchler, A. Dreber, C. F. Camerer, R. A. Poldrack, and T. Schonberg, "fmri data of mixed gambles from the neuroimaging analysis replication and prediction study," *Scientific data*, vol. 6, no. 1, p. 106, 2019.
[20] A. Sunavsky and J. Poppenk, "Neuroimaging predictors of creativity in healthy adults," *Neuroimage*, vol. 206, p. 116292, 2020.
[21] A. Kogan, K. Alpert, J. L. Ambite, D. S. Marcus, and L. Wang, "Northwestern university schizophrenia data sharing for schizconnect: A longitudinal dataset for large-scale integration," *Neuroimage*, vol. 124, pp. 1196–1201, 2016.
[22] P. J. LaMontagne, T. L. Benzinger, J. C. Morris, S. Keefe, R. Hornbeck, C. Xiong, E. Grant, J. Hassenstab, K. Moulder, A. G. Vlassenko *et al.*, "Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease," *medrxiv*, pp. 2019–12, 2019.
[23] D. Wei, K. Zhuang, L. Ai, Q. Chen, W. Yang, W. Liu, K. Wang, J. Sun, and J. Qiu, "Structural and functional brain scans from the cross-sectional southwest university adult lifespan dataset," *Scientific data*, vol. 5, no. 1, pp. 1–10, 2018.
[24] W. Liu, D. Wei, Q. Chen, W. Yang, J. Meng, G. Wu, T. Bi, Q. Zhang, X.-N. Zuo, and J. Qiu, "Longitudinal test-retest neuroimaging data from healthy young adults in southwest china," *Scientific data*, vol. 4, no. 1, pp. 1–9, 2017.
[25] R. A. Poldrack, E. Congdon, W. Triplett, K. Gorgolewski, K. Karlsgodt, J. Mumford, F. Sabb, N. Freimer, E. London, T. Cannon *et al.*, "A phenome-wide examination of neural and cognitive function," *Scientific data*, vol. 3, no. 1, pp. 1–12, 2016.
[26] B. Fischl, "Freesurfer," *Neuroimage*, vol. 62, no. 2, pp. 774–781, 2012.
[27] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, and B. T. Yeo, "Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri," *Cerebral cortex*, vol. 28, no. 9, pp. 3095–3114, 2018.
[28] B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. Van Der Kouwe, R. Killiany, D. Kennedy, S. Klaveness *et al.*, "Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain," *Neuron*, vol. 33, no. 3, pp. 341–355, 2002.
[29] J.-P. Fortin, N. Cullen, Y. I. Sheline, W. D. Taylor, I. Aselcioglu, P. A. Cook, P. Adams, C. Cooper, M. Fava, P. J. McGrath *et al.*, "Harmonization of cortical thickness measurements across scanners and sites," *Neuroimage*, vol. 167, pp. 104–120, 2018.
[30] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.
[31] S. A. Dudani, "The distance-weighted k-nearest-neighbor rule," *IEEE Transactions on Systems, Man, and Cybernetics*, no. 4, pp. 325–327, 1976.
[32] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, pp. 3–42, 2006.
[33] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.
[34] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
[35] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.
[36] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018.
[37] J. Howard and S. Gugger, "Fastai: a layered api for deep learning," *Information*, vol. 11, no. 2, p. 108, 2020.
[38] N. Erickson, J. Mueller, A. Shirkov, H. Zhang, P. Larroy, M. Li, and A. Smola, "Autogluon-tabular: Robust and accurate automl for structured data," *arXiv preprint arXiv:2003.06505*, 2020.