Explainable Age Predictions from Electroencephalography Data

Zara Cook[1,4] **Grant Sinha**[1] **Chengzong Zhao**[1,2] **Jack Wang**[2,1]
Sam M. Doesburg[4] **George Medvedev**[3] **Urs Ribary**[4] **Vasily A. Vakorin**[4] **Nabil Belacel**[2] **Pengcheng Xi**[2]
[1]University of Waterloo, Canada
[2]National Research Council Canada
[3]Fraser Health Authority, Canada
[4]Simon Fraser University, Canada
{zncook,gsinha,c233zhao,j88wang}@uwaterloo.ca
gmedvedev@gmail.com, {sam_doesburg,urs_ribary,vasily_vakorin}@sfu.ca
{nabil.belacel,pengcheng.xi}@nrc-cnrc.gc.ca

Abstract

Electroencephalography (EEG) serves as a widely acceptable clinical tool for monitoring and assessing brain activities. In leveraging artificial intelligence, machine learning techniques have been utilized for neurophysiological age prediction from EEG data. This research aims to enhance the transparency of such models, using SHapley Additive exPlanations (SHAP) to evaluate EEG feature significance. We employ EEGNet for feature extraction, training predictive models like Random Forest, Support Vector Regressor, and a Recurrent Neural Network. Additionally, we incorporate a transformer model for improved clarity and performance evaluation. Our data, sourced from public hospitals, indicates that the Transformer model notably excels in age prediction. This finding underscores the potential for more transparent machine learning in clinical EEG analysis. Our study advances the search for more interpretable and accountable models in healthcare, addressing trust concerns and facilitating informed decision-making in brain health assessment.

1 Introduction

Hospitals routinely employ electroencephalogram (EEG) recordings to assess the neurological functions of patients. The copious amounts of data generated by these examinations present a valuable opportunity for artificial intelligence (AI) to expedite analysis and identify significant neuromarkers studied in cognitive and clinical neuroscience. One of such neuromarkers is 'brain-predicted age' or brain age for simplicity [1]. This brain age holds particular significance as it provides insights into the ageing process, aiding in the identification of individuals at heightened risk of age-related cognitive impairments or clinical alterations and ultimately, mortality [1]. A notable observation arises when there is a disparity between an individual's predicted age and their chronological age, with the predicted age exceeding the latter. This discrepancy suggests that the individual's brain may be aging at an accelerated rate, indicating an elevated risk of neurological disorders or cognitive decline [1].

Deep learning (DL) models have seen a growing application in clinically- and cognitively-relevant prediction tasks using EEG recordings. They have emerged as a compelling alternative to the conventional reliance on engineered feature extraction, applied together with classical (non-neural) machine learning (ML) approaches. DL models, through training, have showcased their capacity to improve prediction accuracy by uncovering intricate features within EEG recordings that might be overlooked by traditional EEG analysis methods [2]. However, a critical concern associated with these DL models lies in their inherent lack of transparency, a concern that gains particular significance within clinical environments. In healthcare settings, where trust and transparency are paramount, ensuring the reliability and interpretability of AI models is essential.

When it comes to enhancing the transparency of deep learning (DL) models, Explainable AI (XAI) techniques step in to provide insights into the inner workings of these models, revealing how they arrive at their predictions. Among the leading XAI techniques, Shapley Additive explanations (SHAP) stands out as a mathematical framework rooted in cooperative game theory. It assigns a unique value, known as the Shapley value, to each feature within a prediction, quantifying its contribution to that prediction [3]. SHAP's capabilities extend to offering visual interpretations at both local and global levels. At the local level, SHAP provides in-depth explanations for individual predictions, indicating why a specific outcome was predicted. On a global scale, it delivers a comprehensive overview of feature importance across the entire dataset. This dual functionality, enabling the interpretation of individual predictions and the comprehension of broader data trends, makes SHAP one of the preferred choices for investigating model explainability [4].

Another promising approach to achieving greater model transparency is by adopting inherently explainable models, such as Transformers. These models derive their explainability from unique components known as attention mechanisms, which allow them to assign varying levels of importance to different input features. While Transformers have mainly made their mark in revolutionizing prediction tasks related to computer vision and natural language processing, their applicability to decoding EEG signals is gaining traction [5]. These inherently explainable Transformers could bring EEG analysis to a higher level of transparency, a feature that we explore in our study, focusing on accuracy in predicting brain age from EEG data.

Our contributions include:

- Significant performance enhancement via Transformers,
- Deep insights into model-input feature relations, and
- Comparative analysis of explainability across ML models.

2 Literature Review

2.1 Machine Learning Models

Traditional machine learning requires feature extraction for large datasets, with prominent methods being Random Forest Regression (RFR) and Support Vector Regression (SVR) [6]. In contrast, deep learning (DL) models autonomously extract features, often outperforming classical models, especially in tasks like image classification [7]. EEGNet, a specialized Convolutional Neural Network (CNN), and Long Short-term Memory (LSTM), a Recurrent Neural Network (RNN), are primary DL models for EEG analysis. EEGNet employs techniques like batch normalization and dropout for performance [8], while LSTMs, with better memory, excel at sequential data, even though they may require more training time [9–11].

In recent years, the transformer architecture, which was originally proposed for natural language processing tasks, has revolutionized various domains in AI due to its unparalleled ability to capture global dependencies in data [12]. Unlike CNNs which sometimes struggle with perceiving a wide range of internal relationships in data without deep structures, or RNNs that are constrained by sequential processing, transformers leverage the attention mechanism, making them inherently more flexible and efficient [12, 13].

As outlined by the research work in [5], traditional methods based on CNNs, while efficient, have limitations in recognizing global EEG dependencies. This is a significant concern given that EEG paradigms often possess strong overall relationships. They introduced Spatial-Temporal Tiny Transformer (S3T) which taps into the power of transformers to emphasize and utilize both spatial and temporal features in EEG data. By employing attention mechanisms, it effectively distinguishes spatial features and perceives global temporal features, leading to an improved EEG decoding.

Another notable advantage of transformers is their inherent explainability. The attention mechanism can highlight which parts of the input data are being focused on for a particular output, allowing

for a degree of interpretability in the results. As EEG analysis often demands clarity on which brain signals or patterns lead to specific interpretations or predictions, this feature is invaluable. Given the promising results of the S3T model on public datasets, it is evident that transformers hold significant potential for advancing EEG-based Brain-Computer Interface (BCI) technologies [5].

2.2 Explainability Techniques

Explainable AI (XAI) focuses on clarifying machine learning decisions. Local Interpretable Model-Agnostic Explanation (LIME) provides local explanations for any classifier, elucidating individual predictions [14]. While powerful, its insights might not always generalize. DeepLIFT, more specific to neural networks, back-propagates contributions to understand feature importance [15]. SHAP, based on cooperative game theory, presents both global and local model explanations. It calculates each feature's significance to a prediction, ensuring balanced attributions [16, 17]. Introduced in 2017, SHAP assesses features by their presence versus absence, showcasing their impact on predictions [4]. Given its comprehensive approach, SHAP was chosen for our study.

SHAP has its theoretical foundation in Shapley values, which guarantees an equitable distribution of contribution values among the features. However, a challenge with SHAP is its computational demand, especially for certain model types [3]. The SHAP methodology uses Shapley values to explain the output of machine learning models. Each feature gets a value based on its contribution to a specific prediction, making it evident how each feature influences the model's decisions [16, 17]. Launched as a model-agnostic solution in 2017 [4], SHAP interprets the influence of features by evaluating the performance difference when a feature is present versus its absence. This establishes how each feature contributes, either positively or negatively, to the prediction. According to [2], SHAP values are often considered superior to traditional feature importance techniques. While feature importance measures the overall influence of features using metrics like Gini importance, it is specific to certain machine learning models and does not provide the cooperative context that SHAP offers. This holistic and equitable way of assessing feature impact led to our selection of SHAP for our study.

3 Materials and Methods

3.1 Dataset

We utilized the dataset from paper [18], comprising EEG recordings taken over six years from a public hospital in British Columbia, Canada. Approved by Simon Fraser University and the Fraser Health Authority (protocol H18-02728, April 1, 2022), the data includes participants aged 15-99. Recorded using Natus Xltek EEG32U amplifiers, sessions varied in duration. For our study's focus on machine learning model efficacy, minimal preprocessing in line with [18] was performed. Data are also bandpass filtered between 0.5 Hz and 55 Hz [19]. Of the 7001 EEG recordings, 5000 were for training, 1000 for testing, and 1001 for validation.

3.2 Machine Learning Models

Adopting the methodology in [18], our EEG brain age prediction involved:

- Preprocessing: Cleaning and resampling EEG data to 128 Hz.
- EEGNet Feature Extraction: Using EEGNet to obtain latent features.
- Regression Modeling: Applying RFR, SVR, and LSTM for brain age prediction.
- Training/Evaluation: Dataset partitioned into subsets, using Mean Absolute Error (MAE) for training and evaluation. Additionally, a multi-layer perceptron (MLP) was added to the EEGNet model for direct age prediction.

Consistent with [18], machine learning strategies were used:

- Classical Machine Learning: Employed EEGNet for feature extraction, utilizing outputs for RFR and SVR techniques.
- Deep Learning: EEGNet and LSTM were explored for autonomous feature extraction from EEG data, with EEGNet comprising convolutions and an added MLP for age prediction. LSTM, an advanced RNN, identifies data sequence patterns.
- Transformer: Applied the S3T [5] for EEG decoding. It includes preprocessing, spatial and temporal transformation, with the goal being classification loss through cross-entropy [5]. Attention mechanisms in S3T offer insights into EEG signal dependencies, highlighting pertinent model areas for deep learning explainability [5].

3.3 Explainability

SHAP (SHapley Additive exPlanations) values, rooted in the cooperative game theory's Shapley values, provide powerful tools for explaining machine learning models. The TreeExplainer [20], as its name suggests, is designed primarily for tree-based models, offering exact Shapley value computations. Its strength lies in capturing intricate feature interactions, giving a clearer perspective on a model's overall behavior through the lens of localized explanations [4].

Kernel SHAP, meanwhile, is model-agnostic, serving as a bridge between the linear explanations of LIME [14] and the Shapley values. The primary advantage of Kernel SHAP is its emphasis on ensuring explanations uphold key properties like local accuracy and consistency [4].

Diving deeper, the Gradient Explainer amalgamates the ideas underpinning Integrated Gradients [21], SHAP, and SmoothGrad [22]. It distinguishes itself by allowing the use of an entire dataset as its reference background. With its foundational linear assumptions and considerations on feature independence, this explainer calculates SHAP values, proving invaluable when dealing with neural networks.

In our research, we harnessed these explainers in alignment with the unique architectures of the specific models being analyzed. The TreeExplainer was chosen for the Random Forest model due to its inherent tree-based structure. By leveraging the TreeExplainer's adeptness in elucidating interactions and feature dependencies, we gleaned profound insights into feature significance and the intricate dynamics between predictors.

For the LSTM, a neural architecture, the Gradient Explainer was deemed the most suitable. Neural networks, marked by their intricate mesh of weights and activations, mandate an advanced explanation strategy. With the Gradient Explainer's confluence of ideas from Integrated Gradients, SmoothGrad, and SHAP, we could deconstruct these networks effectively. By referencing the entire dataset as the background and leaning on its foundational assumptions, the Gradient Explainer enabled us to extract approximate SHAP values. These values illuminated the LSTM model's decision-making process, allowing us to trace back through neuron layers to spotlight the most influential features.

In the context of SVR, we opted for the Kernel Explainer, a model-agnostic tool. It demystified the SVM model's decisions through approximated SHAP values, enriching our understanding of how specific feature values shaped predictions in certain instances. Furthermore, the Kernel Explainer was also applied to LSTM, Random Forest, and LSTM models, setting the stage for a comparative analysis against more specialized explainers.

Table 1: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Standard Deviation of Absolute Errors (std.), and Training Times of models evaluated on EEG data.

Model Name	MAE	RMSE	std.	Time
EEGNet + LSTM	18.89	22.50	12.28	0:09:06
EEGNet + RFR	18.67	21.95	11.57	0:02:32
EEGNet + SVR	18.70	22.01	11.61	0:00:02
EEGNet + MLP	18.59	21.79	11.30	12:01:29
Transformer	17.22	-	-	-

4 Results

Inspired by [18], we systematically examined various training methods for the evaluation of EEG data. We employed the EEGNet archi-

tecture complemented by an Adam optimizer, explored the potential of LSTM, delved into an ensemble approach with the RFR consisting of 200 decision trees, and also scrutinized the SVR fortified by a radial basis function kernel. Our observations largely coincided with the findings presented in [18] with regard to mean absolute errors (MAEs). In a significant stride forward, we introduced the Transformer model to our analyses. Remarkably, this model eclipsed its counterparts, registering a compelling MAE of 17.22, as depicted in Table 1. This underpins the capabilities of the Transformer architecture when applied to EEG data interpretation.

SHAP played a pivotal role in our research, serving as the lens through which we gained clarity on the decision-making intricacies of the machine learning models. It also illuminated the ripple effect each feature exerted on the model's outcomes. Diving deeper into the LSTM model, it was meticulously architected incorporating pre-trained weights. We then orchestrated our dataset to align seamlessly with the input shape prerequisites. The LSTM model's inner workings were unraveled using a dual strategy: a gradient-oriented method, leveraging a sample from the dataset as foundational data, and a kernel-centric approach, which entailed reshaping the dataset, reminiscent of the adaptations executed for the SVR model.

For the Random Forest model both the TreeExplainer and KernelExplainer tools from SHAP were harnessed, the former excelling in speedy calculations for tree-based models, and the latter, with its model-agnosticism, undertaking elaborate computations to distill the Shapley values.

Shifting focus to the SVR model, our dataset was meticulously processed and structured to meet the model's specifications. The KernelExplainer played a pivotal role in deconstructing the SVR model's decisions post its training phase.

The insights learned from these models were visually represented through SHAP summary plots, as exhibited in Fig. 1. These plots offer a insight into feature relevance across a diverse range of time steps. They illustrate not just the prominence of each feature, but also the extent and nature of their influence. The plots are structured with features vertically, and their influence spanning horizontally with the centering of values around zero indicating minimal influence. Displacements to the right or left revealed the positive or negative contributions of features to the model's output. This is visually represented by a color gradient in our plots, with blue signifying lower feature values and pink indicating higher values, thus providing a spectrum of feature influence at a glance. The vertical ordering of features in the plot corresponds to their importance, with the most influential features based on the SHAP values positioned at the top. This ranking offers a hierarchy of feature significance, allowing use to quickly identify which features have the most substantial overall impact on the model's output.

A recurrent observation across the SHAP analyses was the dominant influence of Feature 104, underscoring its cardinal role. Other features such as Feature 4, Feature 95, and Feature 7 also frequently emerged as major contributors. Yet, there were features like Feature 103, Feature 107, and Feature 38, which showcased their influence predominantly in specific contexts, emphasizing their conditional relevance. These multifaceted insights pave the way for future endeavors in refined feature engineering and potential avenues for model optimization.

The results from the SHAP analysis inform clinicians about significant EEG features that are crucial in predicting brain age. This knowledge assists in focusing EEG visual inspections on these key features, enhancing the understanding of age-related brain activity.

As we further analyze our findings, a side-by-side assessment of the SHAP plots in Fig. 1 offers valuable insights. The Random Forest visualized in Fig. 1 a) using the TreeExplainer showed clear and easily interpretable SHAP values for key features. On the other hand, Fig. 1b) which utilized the KernelExplainer for the same model, provided a broader understanding of how features interact. Moving to the LSTM results, Fig. 1 c) GradientExplainer clearly highlighted how primary features affected the outcomes. Meanwhile, Fig. 1 d) KernelExplainer presented a more detailed view, requiring a closer look to grasp its intricacies. In conclusion, while tools like the TreeExplainer and GradientExplainer offer direct insights, the KernelExplainer provides a more detailed understanding, highlighting the importance of selecting the right tool for interpretation. These findings pave the way for improved approaches in future EEG data studies.

5 Conclusion

In this study, we explored the potential of the Transformer model in decoding EEG signals. While various deep learning and classical machine learning techniques were evaluated, it was the Transformer that showcased a boost in performance. While performance is a critical aspect, model transparency is equally essential. As we delve into the complexities of deep learning, it becomes increasingly important to comprehend the underlying factors influencing a model's conclusions.

The intricate decision-making processes of our models were unraveled using SHAP. This powerful tool allowed us to delve deep into the models, shedding light on the influence of individual features, and elucidating their roles in predictions. Our SHAP analyses revealed not only the overarching importance of certain features but also the nuanced interplay between them, enhancing our understanding of the models' inner workings. While our initial findings have been enlightening, they also pave the way for future exploration. The immediate trajectory of our research will be a deep dive into the attention mechanisms inherent to the Transformer. We believe that by unraveling these mechanisms, we can glean richer insights into how the model interprets EEG signals, ultimately bridging the gap between raw performance and transparent decision-making.

Acknowledgments

The authors would like to acknowledge support from Digital Health and Geospatial Analytics program at National Research Council Canada (DHGA-116).

References

[1] J. H. Cole, S. J. Ritchie, M. E. Bastin, M. C. V. Hernandez, S. M. Maniega, N. Royle, J. Corley, A. Pattie, S. E. Harris, Q. Zhang, N. R. Wray, P. Redmond, R. E. Marioni, J. M. Starr, S. R. Cox, J. M. Wardlaw, D. J. Sharp, and I. J. Deary, "Brain age predicts mortality," *Molecular Psychiatry*, vol. 23, no. 5, pp. 1385–1392, 2018.

[2] W. E. Marcílio and D. M. Eler, "From explanations to feature selection: assessing shap values as feature selection mechanism," in *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, Porto de Galinhas, Brazil, 2020, pp. 340–347.

[3] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, and R. Ranjan, "Explainable ai (xai): Core ideas, techniques, and solutions," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–33, 2023.

[4] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–15.

[5] Y. Song, X. Jia, L. Yang, and L. Xie, "Transformer-based spatial-temporal feature learning for eeg decoding," 2021.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[7] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, Jan 2015, arXiv:1404.7828 [cs]. [Online]. Available: http://arxiv.org/abs/1404.7828.

[8] V. Lawhern, A. Solon, N. Waytowich, S. Gordon, C. Hung, and B. Lance, "Eegnet: A compact convolutional network for eeg-based brain-computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, p. 056013, Oct 2018, arXiv:1611.08024 [cs, q-bio, stat]. [Online]. Available: http://arxiv.org/abs/1611.08024.

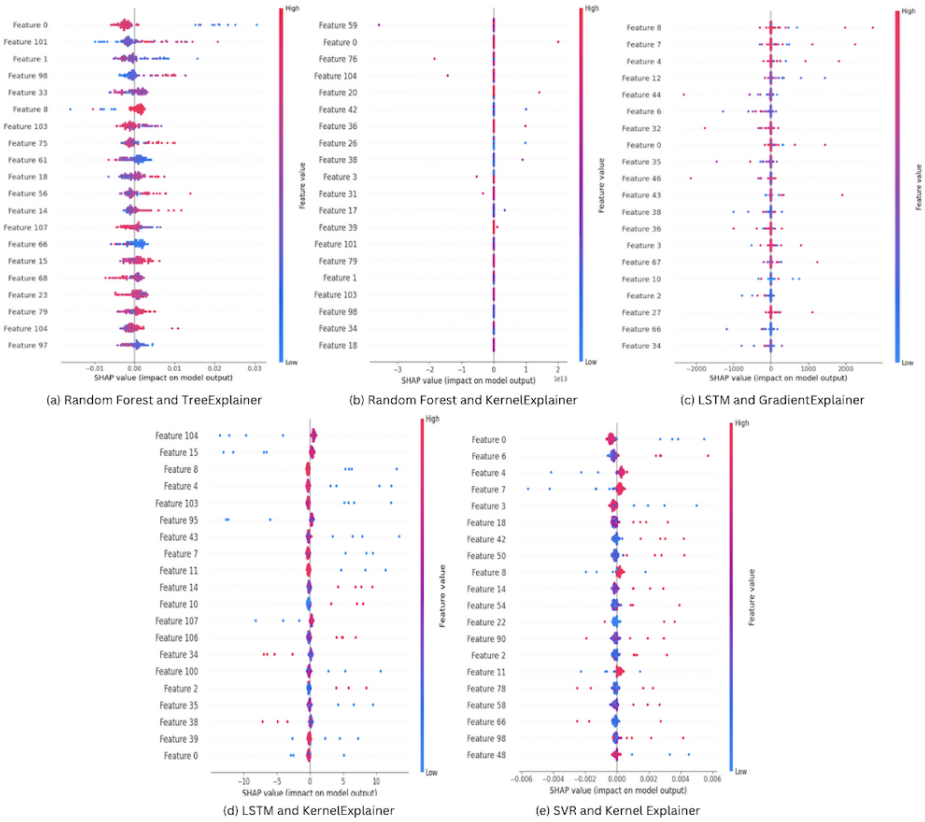

Fig. 1: Comparative visualization of SHAP values across different machine learning models and their respective explainers. The plots provide insights into feature importance and their impact on model output.

[9] H. Raza, A. Chowdhury, and S. Bhattacharyya, "Deep learning based prediction of eeg motor imagery of stroke patients for neuro-rehabilitation application," in *2020 International Joint Conference on Neural Networks (IJCNN)*. Glasgow, United Kingdom: IEEE, Jul 2020, pp. 1–8, [Online]. Available: https://ieeexplore.ieee.org/document/9206884/.

[10] K. M. Tsiouris, V. C. Pezoulas, M. Zervakis, S. Konitsiotis, D. D. Koutsouris, and D. I. Fotiadis, "A long short-term memory deep learning network for the prediction of epileptic seizures using eeg signals," *Computers in Biology and Medicine*, vol. 99, pp. 24–37, Aug 2018, [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S001048251830132X.

[11] P. Wang, A. Jiang, X. Liu, J. Shang, and L. Zhang, "Lstm-based eeg classification in motor imagery tasks," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 11, pp. 2086–2095, Nov 2018, [Online]. Available: https://ieeexplore.ieee.org/document/8496885/.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv:1706.03762 [cs]*, Dec 2017, arXiv: 1706.03762. [Online]. Available: http://arxiv.org/abs/1706.03762.

[13] N. Zhang, "Learning adversarial transformer for symbolic music generation," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–10, 2020.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, "'why should i trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug 2016, pp. 1135–1144.

[15] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," *arXiv preprint arXiv:1704.02685*, 2019, [Online]. Available: https://arxiv.org/abs/1704.02685.

[16] G. Marialuisa, B. G. Ilaria, G. P. Rudy, C. Federica, V. Nicola, P. Alessandro, S. S. Francesca, S. Nicola, and M. Gloria, "explainable ai allows predicting upper limb rehabilitation outcomes in sub-acute stroke patients," *IEEE Journal of Biomedical and Health Informatics*, Jan 2022.

[17] P. L. Ballester, J. S. Suh, N. C. W. Ho, L. Liang, S. Hassel, S. C. Strother, S. R. Arnott, L. Minuzzi, R. B. Sassi, R. W. Lam, R. Milev, D. J. Müller, V. H. Taylor, S. H. Kennedy, J. P. Reilly, L. Palaniyappan, K. Dunlop, and B. N. Frey, "Gray matter volume drives the brain age gap in schizophrenia: a SHAP study," *Schizophrenia*, vol. 9, no. 1, p. 3, Jan. 2023.

[18] G. Sinha, N. Belacel, Z. Gu, S. Doesburg, G. Medvedev, U. Ribary, V. Vakorin, and P. Xi, "Machine learning methods for electroencephalogram-based age prediction," *IEEE Sensors Conference 2023, Vienna, Austria*, October 2023.

[19] K. Jusseaume and I. Valova, "Brain age prediction/classification through recurrent deep learning with electroencephalogram recordings of seizure subjects," *Sensors*, vol. 22, no. 21, p. 8112, Oct 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/21/8112

[20] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 2522–5839, 2020.

[21] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *arXiv preprint arXiv:1703.01365*, 2017, [Online]. Available: http://arxiv.org/abs/1703.01365.

[22] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017, [Online]. Available: http://arxiv.org/abs/1706.03825.