

X-SiT: Inherently Interpretable Surface Vision Transformers for Dementia Diagnosis

Fabian Bongratz^{1,2*}[0009-0009-8879-1823], Tom Nuno Wolf^{1,2}[0000-0002-4728-9636], Jaume Gual Ramon¹, and Christian Wachinger^{1,2}[0000-0002-3652-1874]

¹ Lab for AI in Medical Imaging, Technical University of Munich, Munich, Germany
{fabi.bongratz,tn.wolf}@tum.de, *Corresponding author
² Munich Center for Machine Learning, Munich, Germany

Abstract. Interpretable models are crucial for supporting clinical decision-making, driving advances in their development and application for medical images. However, the nature of 3D volumetric data makes it inherently challenging to visualize and interpret intricate and complex structures like the cerebral cortex. Cortical surface renderings, on the other hand, provide a more accessible and understandable 3D representation of brain anatomy, facilitating visualization and interactive exploration. Motivated by this advantage and the widespread use of surface data for studying neurological disorders, we present the eXplainable Surface Vision Transformer (X-SiT). This is the first inherently interpretable neural network that offers human-understandable predictions based on interpretable cortical features. As part of X-SiT, we introduce a prototypical surface patch decoder for classifying surface patch embeddings, incorporating case-based reasoning with spatially corresponding cortical prototypes. The results demonstrate state-of-the-art performance in detecting Alzheimer’s disease and frontotemporal dementia while additionally providing informative prototypes that align with known disease patterns and reveal classification errors.

Keywords: Explainable AI · Cortical surface analysis · Diagnosis.

1 Introduction

Explainability, i.e., making an automated decision-making process understandable for humans, is a fundamental prerequisite for applications in clinical environments [34]. Yet, deep neural networks do not fulfill this property without further modification due to their complex information-processing architecture. Therefore, previous endeavors have focused on developing explainable methods for clinical reports, tabular measurements, video data, and medical images [12,21,22,33].

Cortical surface analysis, however, relies on non-Euclidean data for understanding various neurological and psychiatric conditions [2,5,27]. Specifically,

F. Bongratz and T.N. Wolf—Equal contribution.

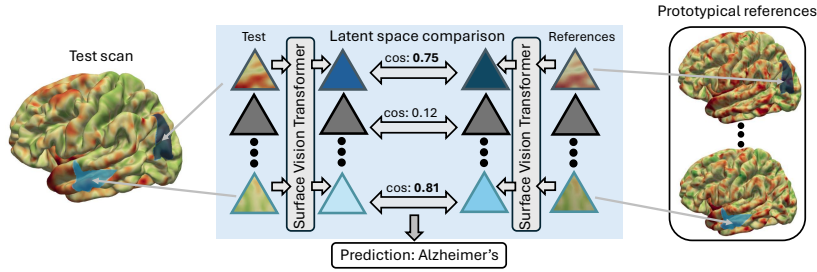


Fig. 1. Case-based reasoning with X-SiT. X-SiT learns discriminative prototypes of cortical surfaces; patches are colored in blue and surfaces represent cortical thickness.

triangular meshes are commonly used to represent the cortical sheet and morphological measurements thereof [7,11]. A central advantage of surface-based representations compared to 3D volumetric images is that they preserve spatial relations and topological properties of the cortical sheet, crucial for faithful explanations. Moreover, surfaces are well-suited for 3D visualization; they support interactive exploration, where specific areas of interest can be examined from different angles and depths, providing a more immersive user experience compared to traditional 2D images. Plotting and interpreting surface-based biomarkers, e.g., cortical thickness, is also straightforward. As these features are obtained after considerable processing, they also reduce potential sources of bias, e.g., inhomogeneities in image intensities [26].

Related work. Dedicated neural networks were previously developed for cortical surfaces, based on graph-convolutions [14,35] and transformers [4,6]. While the receptive field of graph-convolutional kernels is limited, transformer-based models can capture long-range dependencies across the cortical sheet. Unfortunately, the decision-making of transformers can only be understood vaguely by the attention weights [6], which are considered unreliable explanations [16,19]. Technically most related to our work, an extension of the Vision Transformer (ViT) [8] with a neural tree decoder (ViT-NeT) [17] was proposed to explain the decision-making of transformers in computer vision. However, ViT-Net has inherent flaws when it comes to processing registered cortical surface data. There are no spatial restrictions on the matching of prototypes; while a bird can be anywhere in an image, brain regions always align to a template. Moreover, the sequential traversal of local binary decisions prevents a global comparison of structural brain patterns. Most interpretable methods in neuroimaging are designed for images [23]. Notable exceptions use saliency maps [1,31], which do not explain the actual reasoning process of the neural network [28].

Contribution. We present the eXplainable Surface Vision Transformer (X-SiT), an *inherently interpretable* transformer-based neural network for the classification of cortical surface data. X-SiT leverages the cosine similarity to learned *corresponding cortical prototypes*, which represent a specific cohort, for the decision-making; see Figure 1 for a visualization. We evaluated X-SiT for detecting two

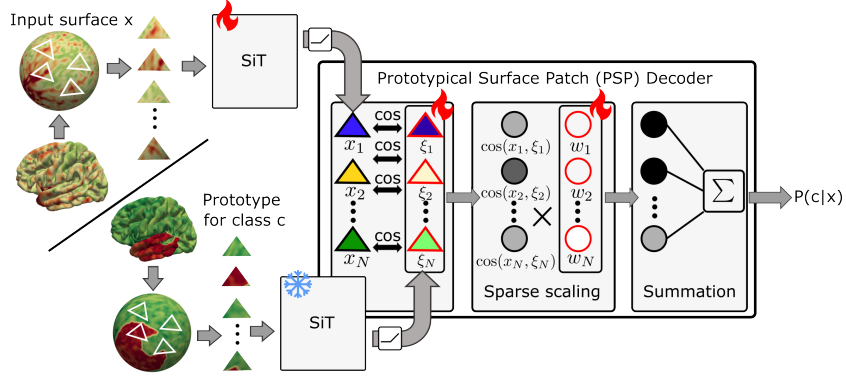


Fig. 2. Architecture of the eXplainable Surface Vision Transformer (X-SiT).

common forms of dementia, Alzheimer’s disease (AD) and frontotemporal dementia (FTD), demonstrating competitive performance to state-of-the-art models. We show that the prototypes learned by X-SiT closely match the discriminative disease patterns from the literature, and we demonstrate their benefit for comprehending the model’s predictions both locally and globally.

2 Methods

We illustrate the architecture of our X-SiT model in Figure 2. In the following, we describe its individual building blocks, namely the surface encoder and the prototypical surface patch (PSP) decoder. On a high level, the X-SiT takes a registered cortical surface \mathbf{x} as input and computes a probability $P(c | \mathbf{x})$ of the surface \mathbf{x} belonging to a certain target class c .

2.1 Surface Transformer Encoder

We use the Surface Vision Transformer (SiT) [6] to encode the input surface, given as triangular mesh with spherical topology. Throughout this work, we assume these surfaces were pre-registered to a template, therefore consisting of a fixed number of $|V|$ vertices. Each vertex comprises a set of F local features, such as cortical thickness, sulcal depth, and curvature. We partition the surface into a fixed set of N triangular, non-overlapping surface patches. Hence, we obtain a sequence of N patches, each comprising M vertices. These patches serve as the input to the SiT encoder, yielding patch-wise latent embeddings. Specifically, we convert each input patch of dimension $M \times F$ into a latent vector of dimension D , corresponding to a functional mapping $f_{\text{SiT}} : \mathbb{R}^{N \times M \times F} \mapsto \mathbb{R}^{N \times D}$. The resulting sequence of encoded patches, which we denote as $x = (x_1, \dots, x_N)$, serves as input to X-SiT’s decoder, which we will explain next.

2.2 PSP Decoder

To decode the patches into a probability $P(c \mid \mathbf{x})$ of the input \mathbf{x} belonging to a certain class c — in an interpretable manner, we propose a new prototypical surface patch (PSP) decoder. The core idea is to compare the encoded patches to learned prototypes, which consist of brain regions from exemplary cases representing the target class c , as illustrated in Figure 1. The underlying principle, case-based reasoning [3], allows us to reason as “this part in the input looks like that part from the training set”, the most human-comprehensible reasoning process [15,18,25]. However, the nature of cortical surfaces requires *corresponding prototypes*, i.e., matching brain regions, to allow for reasoning such as “this cortical region in the input looks like that corresponding cortical region from the training set”. This is in contrast to image-based prototypical networks [3,17,24], where the presence of prototypes is typically investigated across the entire image, but it ensures the comparisons made between cortical inputs and examples from the training set are clinically and anatomically meaningful. By allowing the model to take individual patches from different training cases, X-SiT can learn prototypical disease patterns that a single case could hardly cover.

Prototype similarity. Like the input surface, we encode the prototypes with our SiT encoder, cf. Figure 2. Thereby, we obtain a second, prototypical sequence of encoded patches, which we denote by $\xi = (\xi_1, \dots, \xi_N)$. Based on the correspondences, two encoded patches x_i and ξ_i with the same index i can be compared based on the cosine similarity of their latent representations:

$$\cos(x_i, \xi_i) = \frac{\sum_{j=1}^D x_{i,j} \xi_{i,j}}{\sqrt{\sum_{j=1}^D x_{i,j}^2} \sqrt{\sum_{j=1}^D \xi_{i,j}^2}}. \quad (1)$$

Since the prototypical patches represent the target class locally, the cosine similarity provides local evidence for or against characteristic structural patterns being present in the input surface. To further improve interpretability, we rectify all input embeddings to be positive via a ReLU function, prohibiting an increase of class probability in the absence of features. Therefore, only positive evidence of the target class is considered, as opposed to “no evidence” when $\cos(x_i, \xi_i) \approx 0$.

Sparse scaling. While the cosine similarity indicates the presence of class-related structural patterns, it does not distinguish between important and less important regions. To equip X-SiT with the capability to re-weight the importance of individual regions, we scale the cosine similarities with weights $w = \{w_1, \dots, w_N\}$, such that $\sum_{i=1}^N w_i = 1$. We enforce sparsity by suppressing the contribution of below-average regions, i.e., weights are set to zero iff $w_i < 1/N$. The rationale behind the sparsification is to encourage the model to focus primarily on discriminative brain regions. Moreover, sparse decision-making is generally known to improve model interpretability [29,32].

Summation and final prediction. Finally, we compute the class probability $P(c | \mathbf{x})$ as the weighted sum of cosine similarities:

$$P(c | \mathbf{x}) = \sum_{i=1}^N w_i \cos(x_i, \xi_i). \quad (2)$$

Given that the sum of the weights equals one (cf. definition above), a high probability $P(c | \mathbf{x}) \approx 1$, indicates that the input surface resembles the prototypes in all relevant regions, i.e., regions where $w_i > 0$.

2.3 Training

We train the SiT encoder, the sparse scaling weights, and the prototypes with binary cross-entropy end-to-end, as indicated by the flames in Figure 2. Therefore, the SiT maps the input surface \mathbf{x} to an optimal latent space for comparison with the prototypes. Every five epochs, we replace each prototypical patch — specifically, its latent representation ξ_i — with the most similar patch from all training instances at the prototype’s location. Thus, the prototypes correspond to actual training samples without impairing the stability of the training process.

3 Results and Discussion

3.1 Experimental Setting

We consider two clinical diagnostic tasks to evaluate the performance of the developed X-SiT model, namely Alzheimer’s disease (AD) diagnosis and detection of frontotemporal dementia (FTD). We compare our approach to non-interpretable state-of-the-art models for registered meshes, i.e., SpiralNet⁺⁺ [13], Spherical U-Net (S-UNet) [35], and the original SiT [6]. Additionally, we implemented an inherently interpretable surface transformer, SiT-NeT, following the architecture and design of ViT-NeT [17]. We tuned hyper-parameters³ for all models, based on the balanced accuracy (Bacc) on the validation set. We trained each model with five random initializations and AdamW [20], using inversely proportional class-frequency-based loss weights to mitigate class imbalance; all reported values are the mean and standard deviation across these runs.

3.2 Data and Preprocessing

We used public data from the Alzheimer’s disease neuroimaging initiative (ADNI)⁴ and the frontotemporal lobar degeneration neuroimaging initiative (NIFD)⁵. We

³ Search space included learning rate, weight decay, number of conv/transformer blocks, latent channels, and model-specific parameters such as sequence length for SpiralNet⁺⁺ and number of attention heads for SiT/SiT-NeT/X-SiT.

⁴ <https://adni.loni.usc.edu>

⁵ <https://memory.ucsf.edu/research-trials/research/allftd>

Table 1. Statistics of the datasets used for our experiments. We apply our model to diagnosing Alzheimer’s disease (AD) and frontotemporal dementia (FTD). We report the size of each cohort, the percentage of female subjects (%F), and the age (mean \pm std.).

Split	Task 1: AD diagnosis					Task 2: FTD diagnosis				
	Dataset	CN	AD	%F	Age	Dataset	CN	FTD	%F	Age
Train	ADNI	334	192	49.2	74.2 \pm 6.6	ADNI/ NIFD	392	44	50.7	71.3 \pm 8.1
Val		56	32	51.1	73.8 \pm 6.7		61	10	49.3	70.4 \pm 8.1
Test		55	32	49.4	74.3 \pm 6.5		64	13	50.6	69.6 \pm 8.1

Table 2. Classification results based on validation (V) and test (T) sets for the two considered diagnostic tasks, i.e., Alzheimer’s disease (AD) and frontotemporal dementia (FTD) vs. cognitively normal, respectively. We report mean \pm std. of five models with different weight initialization. We highlight **best** and *second* test results, and we indicate the model’s interpretability.

Model	Interp.	AD diagnosis			FTD diagnosis		
		Bacc (V)	Bacc (T)	F1 (T)	Bacc (V)	Bacc (T)	F1 (T)
SpiralNet ⁺⁺ [13]	✗	73.3 \pm 1.3	70.3 \pm 0.7	69.0 \pm 0.8	88.5 \pm 3.4	79.5 \pm 0.8	72.3 \pm 2.5
S-UNet [35] ¹	✗	65.0 \pm 2.6	57.6 \pm 3.5	57.2 \pm 3.5	76.9 \pm 5.2	73.3 \pm 5.3	73.0 \pm 7.0
SiT [6]	✗	85.3 \pm 2.8	80.5\pm0.5	<i>80.5\pm0.9</i>	85.3 \pm 2.9	80.6\pm1.8	74.3\pm4.4
SiT-NeT ²	✓	75.0 \pm 3.6	68.5 \pm 3.6	68.6 \pm 3.6	85.3 \pm 4.6	72.6 \pm 5.0	65.3 \pm 6.8
X-SiT (ours)	✓	84.5 \pm 1.4	<i>80.2\pm2.0</i>	80.6\pm3.1	87.5 \pm 3.0	<i>79.6\pm1.6</i>	<i>74.0\pm2.6</i>

¹ We used only the encoder of S-UNet, with adaptive average pooling and a classification head.

² We adapted the ViT-Net model [17] for cortical surface data.

used only baseline scans to avoid subject bias and data leakage. We split the data into training, validation, and test sets, stratified according to age, sex, and diagnosis; see Table 1 for an overview. We focused on subjects diagnosed with AD from ADNI and FTD patients (behavioral variant; bvFTD) from NIFD, comparing them to cognitively normal (CN) reference cohorts. Due to the low number of CN samples in the NIFD dataset ($n = 72$), we combined the CN groups from both studies for the FTD diagnosis task. We processed all T1w MR images with FreeSurfer (v7.2) [10], obtaining vertex-wise cortical thickness, curvature, and sulcal depth maps. Subsequently, we registered the data to the FsAverage template, represented as 6th-order icosphere (40,962 vertices). We concatenated the vertices of left and right hemispheres as inputs for all models.

3.3 Classification Performance

We compare the classification performance of X-SiT to the implemented baseline methods in Table 2. Notably, for AD detection, X-SiT and SiT outperform the graph convolution-based models (S-UNet and SpiralNet⁺⁺) by a large margin (+10% Bacc on the test set). This is likely due to the ability of transformers to capture the widespread patterns of cortical degeneration in AD [30], whereas

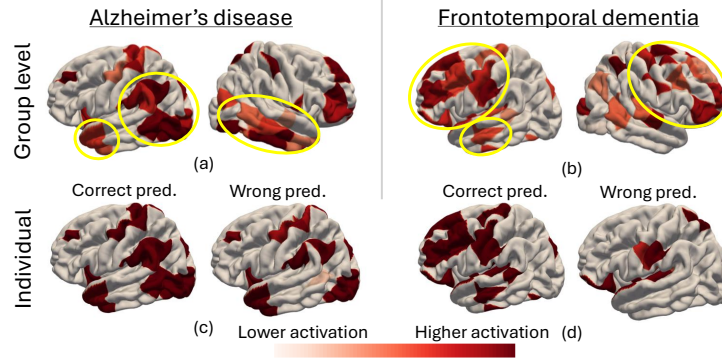


Fig. 3. (a) and (b) show mean activation (weighted similarity to prototypes) in red across correct test set classifications of the dementia class. Higher activation implies a higher contribution to the detection of the respective disease. Yellow ellipses indicate the hallmark regions of the respective dementia type: (a) temporal lobe and temporoparietal junction for Alzheimer’s disease and (b) frontal and temporal regions for frontotemporal dementia. (c) and (d) show model activation for individual predictions.

graph convolutions are inherently limited to aggregating structural information from local neighborhoods. For the diagnosis of FTD, this advantage seems less significant; SpiralNet⁺⁺ achieves the highest Bacc of 88.5% on the validation set. However, X-SiT and SiT are again at the forefront on the test set (Bacc of 80.6 and 79.6, respectively), indicating better generalization and less sensitivity to hyper-parameters tuned on the validation set. Overall, the results presented in Table 2 demonstrate that X-SiT maintains competitive performance compared to its direct non-interpretable baseline, SiT, while also providing valuable explanations (see Section 3.4 in the following). Notably, the proposed PSP decoder in X-SiT consistently outperformed the neural tree decoder in SiT-NeT.

3.4 Interpretability

Group level. X-SiT provides *global* explanations on the group level that help understanding the model’s general decision-making. In Figure 3 (a) and (b), we show the weighted patch-wise similarity, $w_i \cos(x_i, \xi_i)$, of input patches to real-instance prototypes, averaged across the respective disease group in the test set. Specifically, Figure 3 (a) indicates that X-SiT, trained for AD detection, focuses on regions that are actually characteristic of AD: the temporal lobe and the temporoparietal junction [30]. Additionally, we found contributions from the frontal and occipital lobes, both of which were previously associated with AD [9]. Similarly, when trained to detect FTD, X-SiT primarily considers the hallmark regions of this dementia type, the frontal and temporal lobe, see Figure 3 (b). Additionally, we found activation in the inferior parietal lobe of the right hemisphere, which is further in line with findings about FTD [9]. However, the temporal lobe received a comparably low weight in our FTD experiments.

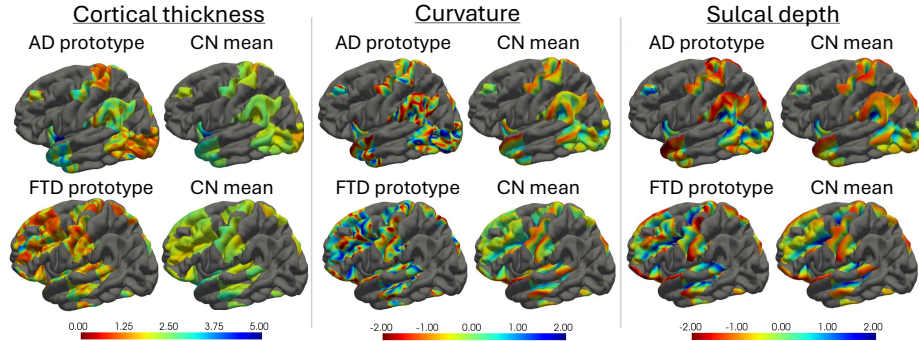


Fig. 4. Prototypical disease-specific input features, i.e., prototypes, learned by X-SiT. We also show the mean features for the CN group in the training set for comparison. Gray areas indicate regions that are ignored by the model ($w_i = 0$, cf. Section 2.2). The visualization is based on the FsAverage template.

A reason could be that we do not explicitly enforce the model to consider *all* discriminative regions; the model can learn to focus on the most relevant ones, therefore fostering easy interpretability.

Individual level. On the individual level, X-SiT’s explanations can aid in assessing the trustworthiness of a prediction. We depict patch-wise similarities between individual samples from the test set and the prototypes in Figure 3 (c) and Figure 3 (d). We observe that individual explanations for correctly predicted dementia cases predominantly follow the pattern observed at the group level. However, the individual explanation deviates considerably from the expected pattern in the case of wrong prediction. Consequently, users are enabled to question the model’s predictions and conduct a manual review of the diagnosis when discrepancies arise. This rigorous assessment is essential for selecting optimal treatment strategies, especially in ambiguous or borderline cases.

Prototypes. By design, the predictions of X-SiT refer to the learned prototypes presented in Figure 4. Prototypical patches are “stitched together” from different individuals; hence, they represent prototypical disease patterns that a single case could hardly cover. Compared to the average cortical thickness in CN cases, we observe a clear thinning in the learned prototypes. This is expected since atrophy in cortical gray matter is characteristic of the neurodegenerative process in both types of dementia [9]. The association of curvature and sulcal depth with dementia is less extensively studied in the literature [9,30] and we did not find considerable differences apart from smoothing effects from taking the mean of the CN group. Yet, they provide a comprehensive picture of the prototypical anatomy used for the decision-making in our model. Importantly, we found the prototypes to be relatively robust to random weight initialization, achieving an overlap of 76.3% and 71.7% for the five AD and FTD models, respectively.

4 Conclusion

In conclusion, we introduced X-SiT, an interpretable model for surface-based detection of dementia based on high-resolution cortical features derived from structural MRI. Our results demonstrate that X-SiT matches the accuracy of state-of-the-art non-interpretable models in the field. Moreover, the explanations provided by X-SiT elucidate its decision-making process and help identify failure cases, thereby enhancing its usability and trustworthiness in clinical settings.

Acknowledgments. This research was partially supported by the German Research Foundation (DFG, No. 460880779). We gratefully acknowledge the computational resources provided by the Leibniz Supercomputing Centre (www.lrz.de).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Azcona, E.A., Besson, P., Wu, Y., Punjabi, A., Martersteck, A., Dravid, A., Parish, T.B., Bandt, S.K., Katsaggelos, A.K.: Interpretation of brain morphology in association to alzheimer’s disease dementia classification using graph convolutional networks on triangulated meshes. In: *Shape in Medical Imaging*. p. 95–107 (2020)
2. de Chastelaine, M., Srokova, S., Hou, M., Kidwai, A., Kafafi, S.S., Racenstein, M.L., Rugg, M.D.: Cortical thickness, gray matter volume, and cognitive performance: a crosssectional study of the moderating effects of age on their interrelationships. *Cerebral Cortex* **33**(10), 6474–6485 (2023)
3. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems* **32** (2019)
4. Cheng, J., Zhang, X., Zhao, F., Wu, Z., Yuan, X., Gilmore, J.H., Wang, L., Lin, W., Li, G.: Spherical transformer on cortical surfaces. In: *Machine Learning in Medical Imaging*. p. 406–415 (2022)
5. Chouliaras, L., O’Brien, J.T.: The use of neuroimaging techniques in the early and differential diagnosis of dementia. *Molecular Psychiatry* **28**(10), 4084–4097 (2023)
6. Dahan, S., Fawaz, A., Williams, L.Z.J., Yang, C., Coalson, T.S., Glasser, M.F., Edwards, A.D., Rueckert, D., Robinson, E.C.: Surface vision transformers: Attention-based modelling applied to cortical analysis. In: *Medical Imaging with Deep Learning. Proceedings of Machine Learning Research*, vol. 172, pp. 282–303 (2022)
7. Dale, A.M., Fischl, B., Sereno, M.I.: Cortical surface-based analysis i. *NeuroImage* **9**(2), 179–194 (1999)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations* (2021)
9. Du, A.T., Schuff, N., Kramer, J.H., Rosen, H.J., Gorno-Tempini, M.L., Rankin, K., Miller, B.L., Weiner, M.W.: Different regional patterns of cortical thinning in alzheimer’s disease and frontotemporal dementia. *Brain* **130**(4), 1159–1166 (2006)
10. Fischl, B.: FreeSurfer. *NeuroImage* **62**(2), 774–781 (2012)
11. Fischl, B., Sereno, M.I., Dale, A.M.: Cortical surface-based analysis ii. *NeuroImage* **9**(2), 195–207 (1999)

12. Gallée, L., Beer, M., Götz, M.: Interpretable medical image classification using prototype learning and privileged information. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 435–445 (2023)
13. Gong, S., Chen, L., Bronstein, M., Zafeiriou, S.: Spiralnet++: A fast and highly efficient mesh convolution operator. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). p. 4141–4148 (2019)
14. Gopinath, K., Desrosiers, C., Lombaert, H.: Learnable pooling in graph convolutional networks for brain surface analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(2), 864–876 (2022)
15. Jeyakumar, J.V., Noor, J., Cheng, Y.H., Garcia, L., Srivastava, M.: How can i explain this to you? an empirical study of deep neural network explanation methods. *Advances in Neural Information Processing Systems* **33**, 4211–4222 (2020)
16. Kashefi, R., Barekatain, L., Sabokrou, M., Aghaeipoor, F.: Explainability of vision transformers: A comprehensive review and new perspectives. *arxiv. arXiv preprint arXiv:2311.06786* (2023)
17. Kim, S., Nam, J., Ko, B.C.: ViT-NeT: Interpretable vision transformers with neural tree decoder. In: International Conference on Machine Learning. *Proceedings of Machine Learning Research*, vol. 162, pp. 11162–11172 (2022)
18. Kim, S.S., Watkins, E.A., Russakovsky, O., Fong, R., Monroy-Hernández, A.: "help me help the ai": Understanding how explainability can support human-ai interaction. In: Conference on Human Factors in Computing Systems. pp. 1–17 (2023)
19. Komorowski, P., Baniecki, H., Biecek, P.: Towards evaluating explanations of vision transformers for medical imaging. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 3726–3732 (2023)
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: *International Conference on Learning Representations* (2019)
21. Mohammadjafari, S., Cevik, M., Thanabalasingam, M., Basar, A.: Using protopnet for interpretable alzheimer’s disease classification. In: *Canadian AI* (2021)
22. Pahud de Mortanges, A., Luo, H., Shu, S.Z., Kamath, A., Suter, Y., Shelan, M., Pöllinger, A., Reyes, M.: Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging. *npj Digital Medicine* **7**(1) (2024)
23. Munroe, L., da Silva, M., Heidari, F., Grigorescu, I., Dahan, S., Robinson, E.C., Deprez, M., So, P.W.: Applications of interpretable deep learning in neuroimaging: A comprehensive review. *Imaging Neuroscience* **2**, 1–37 (2024)
24. Nauta, M., Schlötterer, J., van Keulen, M., Seifert, C.: Pip-net: Patch-based intuitive prototypes for interpretable image classification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023)
25. Nguyen, G., Kim, D., Nguyen, A.: The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *Advances in Neural Information Processing Systems* **34**, 26422–26436 (2021)
26. Nyul, L., Udupa, J., Zhang, X.: New variants of a method of mri scale standardization. *IEEE Transactions on Medical Imaging* **19**(2), 143–150 (2000)
27. Pang, J.C., Aquino, K.M., Oldehinkel, M., Robinson, P.A., Fulcher, B.D., Breakspear, M., Fornito, A.: Geometric constraints on human brain function. *Nature* **618**(7965), 566–574 (2023)
28. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* **1**(5), 206–215 (2019)
29. Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., Zhong, C.: Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys* **16**, 1–85 (2022)

30. Singh, V., Chertkow, H., Lerch, J.P., Evans, A.C., Dorr, A.E., Kabani, N.J.: Spatial patterns of cortical thinning in mild cognitive impairment and alzheimer’s disease. *Brain* **129**(11), 2885–2893 (2006)
31. Spitzer, H., Ripart, M., Whitaker, K., et al.: Interpretable surface-based detection of focal cortical dysplasias: a multi-centre epilepsy lesion detection study. *Brain* **145**(11), 3859–3871 (2022)
32. Tibshirani, R.: Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **58**(1), 267–288 (1996)
33. Wolf, T.N., Pölsterl, S., Wachinger, C.: Don’t panic: Prototypical additive neural network for interpretable classification of alzheimer’s disease. In: *International Conference on Information Processing in Medical Imaging*. pp. 82–94 (2023)
34. Yang, G., Ye, Q., Xia, J.: Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion* **77**, 29–52 (2022)
35. Zhao, F., Xia, S., Wu, Z., Duan, D., Wang, L., Lin, W., Gilmore, J.H., Shen, D., Li, G.: *Spherical U-Net on Cortical Surfaces: Methods and Applications*, p. 855–866. Springer International Publishing (2019)