Identifying Alzheimer's Disease Prediction Strategies of Convolutional Neural Network Classifiers using R2* Maps and Spectral Clustering

Christian Tinauer
Department of Neurology
Medical University of Graz
Graz, Austria
christian.tinauer@medunigraz.at

Maximilian Sackl
Department of Neurology
Medical University of Graz
Graz, Austria
maximilian.sackl@medunigraz.at

Stefan Ropele
Department of Neurology
Medical University of Graz
Graz, Austria
stefan.ropele@medunigraz.at

Christian Langkammer
Department of Neurology
Medical University of Graz
Graz, Austria
christian.langkammer@medunigraz.at

***Abstract*—Deep learning models have shown strong performance in classifying Alzheimer's disease (AD) from R2* maps, but their decision-making remains opaque, raising concerns about interpretability. Previous studies suggest biases in model decisions, necessitating further analysis. This study uses Layer-wise Relevance Propagation (LRP) and spectral clustering to explore classifier decision strategies across preprocessing and training configurations using R2* maps. We trained a 3D convolutional neural network on R2* maps, generating relevance heatmaps via LRP and applied spectral clustering to identify dominant patterns. t-Stochastic Neighbor Embedding (t-SNE) visualization was used to assess clustering structure. Spectral clustering revealed distinct decision patterns, with the relevance-guided model showing the clearest separation between AD and normal control (NC) cases. The t-SNE visualization confirmed that this model aligned heatmap groupings with the underlying subject groups. Our findings highlight the significant impact of preprocessing and training choices on deep learning models trained on R2* maps, even with similar performance metrics. Spectral clustering offers a structured method to identify classification strategy differences, emphasizing the importance of explainability in medical AI.**

***Index Terms*—mri, alzheimer's disease, heatmapping, spectral clustering, chemical validation**

I. Introduction

Deep neural networks have demonstrated strong performance in Alzheimer's disease (AD) classification using magnetic resonance imaging (MRI) [1], but their decision-making processes remain largely opaque [2]. Ensuring that spurious data artifacts do not drive model accuracy is crucial for medical applications. While explainability methods such as Integrated Gradients [3], LIME [4], and Layer-wise Relevance Propagation (LRP) [5] have been used to highlight relevant regions in MRI-based classification, it remains unclear whether these networks primarily rely on disease-related biomarkers or unintended image characteristics [6].

In this work, we utilized the effective relaxation rate R2* as a quantitative MRI parameter for AD classification with a relevance-regularized convolutional neural network (CNN). R2* is defined as the inverse of the effective transverse relaxation time T2* (R2*=1/T2*), which reflects the decay of transverse magnetization in each voxel. Notably, R2* is highly correlated with iron concentration in gray matter [7], and increased iron levels in the basal ganglia are frequently observed in AD [8]. We hypothesize that CNNs implicitly learn such patterns and, with recent advances in explainability, we can now disentangle and visualize these learned features.

By applying Spectral Relevance Analysis (SpRAy) [9] to LRP-based heatmaps, we systematically investigate the spatial clustering of relevance in a deep learning model trained on R2* maps. This approach allows us to identify dominant feature clusters that drive classification decisions and assess their consistency across preprocessing variations. Our findings provide deeper insights into how CNNs utilize structural brain information, further refining the understanding of preprocessing influences and potential biases in deep learning-driven AD classification.

II. Methods

A. Dataset

We retrospectively selected 226 MRI datasets from 117 patients with probable AD (mean age=71.1±8.2 years, male/female=93/133) from our outpatient clinic and 226 MRIs from 219 propensity-logit-matched (covariates: age, sex) [10], [11] normal controls (NCs) (mean age=69.6±9.3 years, m/f=101/125) from an ongoing community-dwelling study. MRI data were acquired longitudinally over multiple sessions using a consistent MRI protocol at 3 Tesla (Siemens

This study was funded by the Austrian Science Fund (FWF grant numbers: P30134, P35887).

TimTrio), including a structural T1-weighted MPRAGE sequence with 1mm isotropic resolution (TR/TE/TI/FA = 1900 ms/2.19 ms/900 ms/9°, matrix = $176\times224\times256$) and a spoiled FLASH sequence ($0.9 \times 0.9 \times 2$mm³, TR/TE=35/4.92ms, 6 echoes, 4.92ms echo spacing, matrix = $208 \times 256 \times 64 \times 6$).

B. Preprocessing

Brain masks for each subject were obtained using FSL-SIENAX [12], and the structural T1-weighted image, and were subsequently used to perform skull-stripping. Using the data acquired from the spoiled FLASH sequence, we solved the inverse problem given as

$$S_{xy}[i] = S_{xy}[0]e^{-t[i]R_2^*}, \quad (1)$$

for $S_{xy}[0]$ and R_2^*, where $S_{xy}[i]$ the measured voxel signal intensity at echo i with echo time $t[i]$. The computations were executed for all voxels in the image volumes, yielding the R2* maps (matrix = $208 \times 256 \times 64$). The obtained R2* maps were affinely registered to the subject's MPRAGE sequence using FSL-flirt and subsequently nonlinearly registered to the MNI152 standard-space brain template using FSL-fnirt.

C. Classifier Network and Training

We employed a 3D subject-level classifier network based on [1], reducing the number and size of convolutional and fully connected layers to mitigate overfitting while maintaining validation accuracy. Batch normalization had no impact and was omitted, while max pooling was replaced with strided convolutions for improved interpretability [13], [14]. To enhance sparsity, all biases were constrained to be negative [14].

The final architecture consists of four blocks, each containing a $3 \times 3 \times 3$ convolutional layer (8 channels) and a down-convolutional layer (strided 2). This is followed by two fully connected layers (16 and 2 units, respectively), totaling 0.3 million trainable parameters. ReLU activations were used throughout, except for the Softmax output layer.

To focus the network on relevant features, we implemented a relevance-guided architecture, $Graz^+$, which extends the classifier with a relevance map generator. This approach incorporates an additional loss term that encourages the model to assign higher relevance to predefined focus regions while suppressing irrelevant areas. The binary attention masks used for this guidance were derived from the FSL-SIENAX brain masks during preprocessing. Full methodological details can be found in [6].

We trained models on R2* maps in subject space using the Adam optimizer [15] for 60 epochs with a batch size of 6. Data was split into training, validation, and test sets (70:15:15) while ensuring that all scans from the same subject remained in the same set. To maintain class balance, final sets were constructed by combining subsets from each cohort, and the process was repeated 10 times for random sampling analysis [16]. The learning rate was initially set to 10^{-3}, reduced by 0.3 after five consecutive epochs without validation loss improvement, and had a lower bound of 10^{-6}. Model weights were reset to their state at the start of a plateau phase.

D. Model Configurations

We evaluated three model configurations to assess the impact of skull-stripping and relevance-guided training [6] on classification performance. This design enabled us to isolate the effects of each component and analyze the learned features using heatmapping.

E. Heatmapping

Heatmaps were created using the LRP method with $\alpha = 1.0$ and $\beta = 0.0$, as described in [5]. Each voxel is attributed a relevance score (R). To analyze the relevance heatmaps, we grouped them and calculated mean heatmaps for each group.

F. Spectral Relevance Analysis

Spectral relevance analysis (SpRAy) enables efficient exploration of classifier behavior across large datasets by applying spectral clustering to inputs and heatmaps. This technique identifies common and atypical decision-making patterns, highlighting image features that may or may not reflect clinically relevant concepts. SpRAy is helpful in uncovering unexpected or artifact-driven classifier behaviors, similar to the "Clever Hans" effects found in [9], [17].

The SpRAy process implemented for this study involves six steps:

1) Compute relevance maps using LRP to identify focus areas for classification.
2) Warp the heatmaps to MNI152 image space.
3) Downsample the native and warped heatmaps to 2 mm isotropic resolution for efficient analysis.
4) Perform spectral clustering to group similar image- or relevance patterns.
5) Identify clusters with highest eigenvalue gap, indicating well-separated heatmap groups and computing mean heatmaps for groups.
6) Visualize the clusters using t-Stochastic Neighbor Embedding (t-SNE) [18], which aids in interpreting the results and understanding the relationship between clusters.

III. Results

Table I summarizes the performance metrics (accuracy, sensitivity, specificity, and AUC) for all configurations in the random sampling setup, evaluated in nonexcluded training sessions. Model A uses native R2* maps, Model B applies the brain mask to R2* maps for skull-stripping before classification, and Model C combines native R2* maps and relevance-guided training with brain masks.

We visualized the clustering of the heatmaps in native subject space and in MNI152 space using t-SNE, initialized with the normalized, symmetric, and positive semi-definite Laplacian matrix derived from the spectral clustering affinity matrix. Fig. 1 illustrates the grouping of heatmaps and corresponding warped heatmaps for all models. Group mean heatmaps, based on spectral clustering groupings, are presented in Fig. 2.

TABLE I
PERFORMANCE METRICS FOR ALL MODEL CONFIGURATIONS ON THE AD VS NC CLASSIFICATION TASK. MODELS ARE IDENTIFIED BY ID, WITH COLUMNS INDICATING USE OF SKULL-STRIPPING AND RELEVANCE-GUIDED TRAINING. BRACKETS [ ] SHOW 95% CONFIDENCE INTERVALS. NC: NORMAL CONTROL; AD: ALZHEIMER'S DISEASE; AUC: AREA UNDER THE RECEIVER OPERATING CHARACTERISTIC CURVE

Id	**Skull-stripping**	**Relevance-guided**	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
A	No	No	64.2±6.5% [53.5%, 76.9%]	61.3±9.6% [39.5%, 74.7%]	67.0±9.4% [51.2%, 83.7%]	64.12±0.06 [0.53, 0.77]
B	Yes	No	77.0±5.8% [64.9%, 85.7%]	75.1±7.8% [62.2%, 86.4%]	78.8±7.9% [62.9%, 90.1%]	76.94±0.06 [0.65, 0.85]
C	No	Yes	75.9±5.1% [67.9%, 85.8%]	69.7±9.7% [52.9%, 86.5%]	81.6±5.5% [72.2%, 92.3%]	75.64±0.05 [0.68, 0.86]

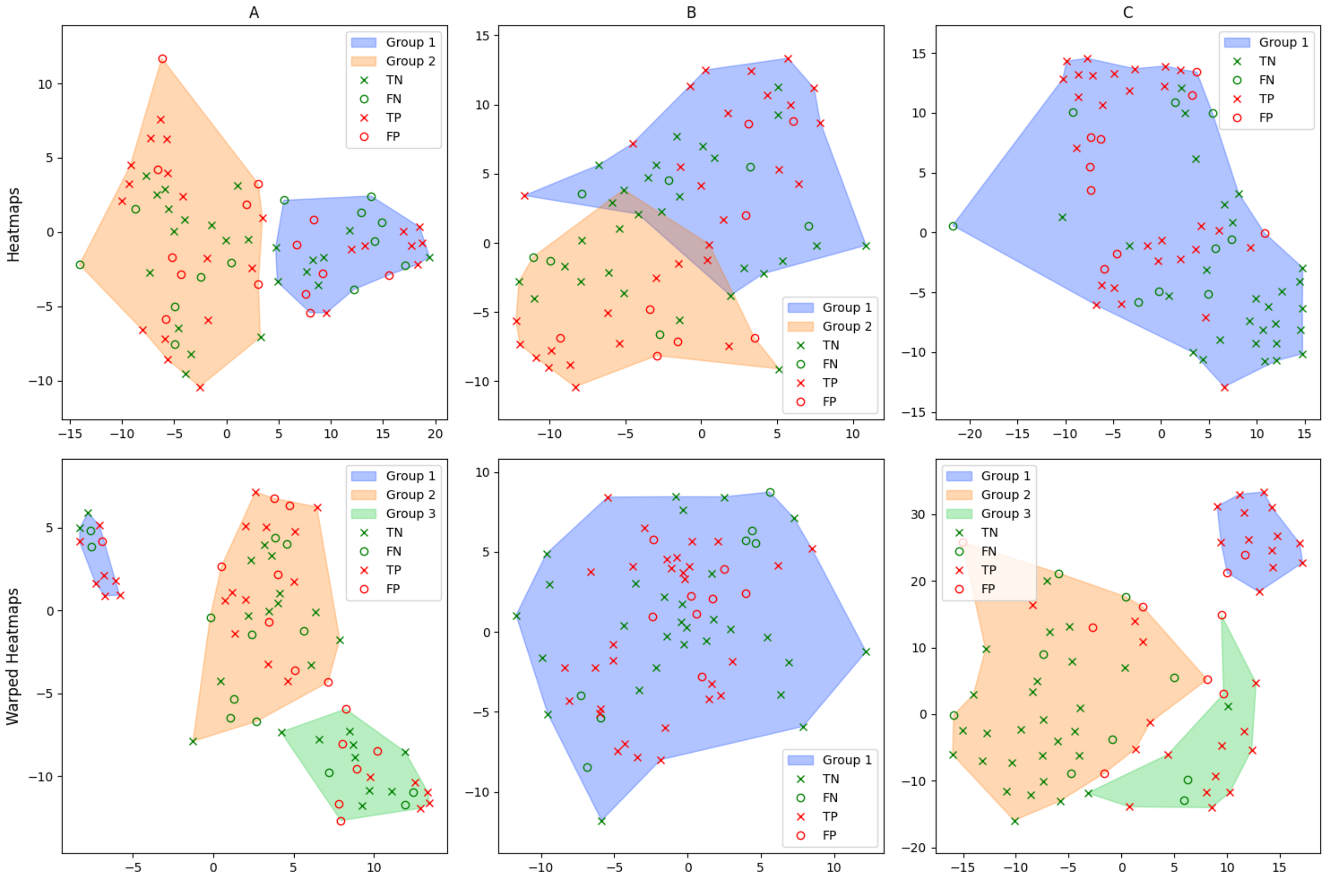

Fig. 1. t-SNE visualization of relevance heatmaps (row 1) and warped heatmaps (row 2) for models trained on native R2* maps (A), skull-stripped R2* maps (B), and with relevance-guided training (C). Spectral clustering was used to group the heatmaps, with the number of clusters determined by eigenvalue analysis. Points indicate individual heatmaps and are colored by classification outcome (TN, FN, TP, FP). Only the warped heatmaps from model C show clustering that aligns with subject groups (NC vs. AD), indicating spatially distinct feature patterns. NC = normal control; AD = Alzheimer's disease; TN = true negative; FN = false negative; TP = true positive; FP = false positive

IV. DISCUSSION

Our study extends previous research on the interpretability of deep learning models for AD classification using R2* maps. Earlier analyses [19] showed that CNNs primarily focus on relaxation rate changes in the basal ganglia. Here, we advance this understanding by applying spectral clustering to LRP-derived heatmaps, enabling a more structured assessment of decision patterns and uncovering systematic classification strategies beyond single-instance explanations.

The t-SNE visualization in Fig. 1 shows that heatmap warping before spectral clustering influences the grouping (row 1 vs. row 2). Notably, only model C, which applies relevance-guided training and warping, exhibited heatmap clustering aligning with subject groups (NC vs. AD). This suggests that models trained without explicit spatial constraints (models A and B) may rely on less structured features, while relevance-guided training helps capture more biologically meaningful patterns.

Mean heatmaps in Fig. 2, derived from spectral clustering groupings, reveal consistent relevance patterns that offer in-

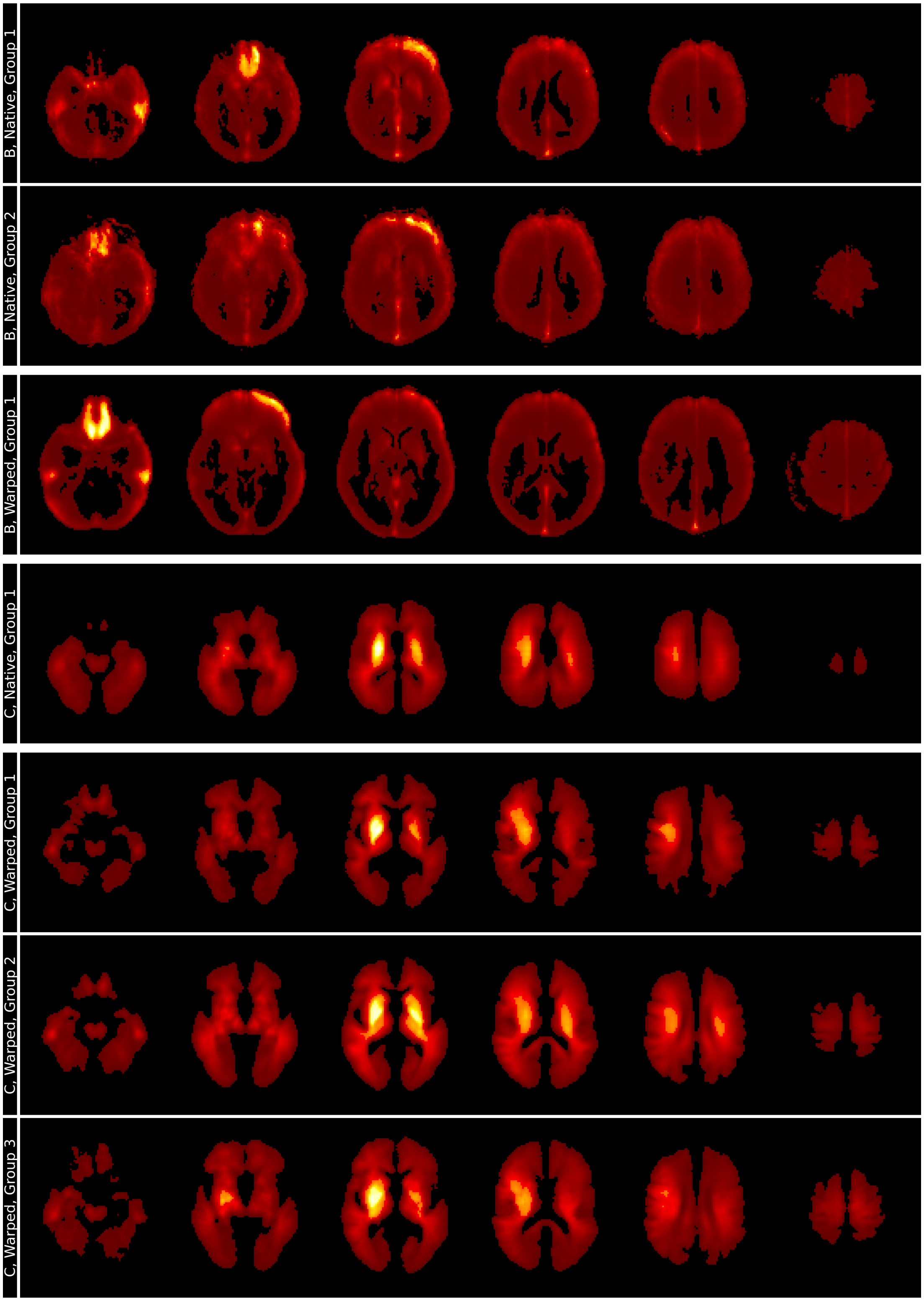

Fig. 2. Mean heatmaps for the groups identified using spectral clustering and eigenvalue analysis for models B and C. Model C shows a clear separation for warped heatmaps between Group 1 (row 5, all heatmaps are from AD) and Group 2 (row 6, most heatmaps are from NC). Model C highlights the left and the right basal ganglia, suggesting a more structured relevance pattern compared to Model B. Images are shown in standard-radiological view, causing the left and right side of the brain to be flipped.

sight into the classifier's decision strategies for models B and C. In the best-performing model (model C), which applies the relevance-guided approach, the separation between AD and NC predictions is more pronounced compared to models A and B. The mean warped heatmap for Group 1 (all AD) in model C shows greater relevance in the right basal ganglia, while the Group 2 (mostly NC) heatmap attributes equal relevance to both the left and right basal ganglia. This suggests that the relevance-guided approach enhances the model's ability to focus on meaningful features, providing a more interpretable decision strategy. Interestingly, Group 3 also highlights the left basal ganglia and surrounding tissue, but the group contains heatmaps from both NC and AD subjects, potentially reflecting structural differences and nonlinear registration. This shows the need for further exploration of these learned representations. In contrast, model B's heatmaps in native space show more reliance on brain masks and volume influences, with varying highlighted regions. After warping to the MNI152 space, these groups merge, emphasizing that postprocessing may obscure position-driven features.

The absence of significant performance differences between models B and C indicates that classification accuracy alone is insufficient to assess model robustness. Despite comparable performances across models, spectral clustering revealed distinct decision patterns, highlighting the importance of explainability techniques for evaluating model reliability. Importantly, our findings emphasize that seemingly minor training choices -such as skull-stripping or relevance-guided regularization- can substantially impact learned representations.

V. Conclusion

This study utilized quantitative MRI data (R2*) for deep learning classification and layer-wise relevance propagation (LRP) in a clinical cohort of Alzheimer's disease patients. By extending our previous research on heatmapping validation [19] by integrating chemical [20] and in-vivo [7] iron mapping studies, our findings confirm that heatmapping approaches can serve as valuable tools to identify areas of tissue changes and provide deeper insights into the internal mechanisms of deep learning-based classification networks. Spectral clustering applied to LRP-based heatmaps allowed us to evaluate classifier decision strategies systematically. Future studies are needed to further explore the influence of preprocessing artifacts on model decisions.

References

[1] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, O. Colliot, Alzheimer's Disease Neuroimaging Initiative, and Australian Imaging Biomarkers and Lifestyle flagship study of ageing, "Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation," *Medical Image Analysis*, vol. 63, p. 101694, Jul. 2020.

[2] C. Davatzikos, "Machine learning in neuroimaging: Progress and challenges," *NeuroImage*, vol. 197, pp. 652–656, Aug. 2019.

[3] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," *arXiv:1703.01365 [cs]*, Jun. 2017, arXiv: 1703.01365. [Online]. Available: http://arxiv.org/abs/1703.01365

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco California USA: ACM, Aug. 2016, pp. 1135–1144. [Online]. Available: https://dl.acm.org/doi/10.1145/2939672.2939778

[5] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLOS ONE*, vol. 10, no. 7, p. e0130140, Jul. 2015. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0130140

[6] C. Tinauer, S. Heber, L. Pirpamer, A. Damulina, R. Schmidt, R. Stollberger, S. Ropele, and C. Langkammer, "Interpretable brain disease classification and relevance-guided deep learning," *Scientific Reports*, vol. 12, no. 1, p. 20254, Nov. 2022.

[7] A. Damulina, L. Pirpamer, M. Soellradl, M. Sackl, C. Tinauer, E. Hofer, C. Enzinger, B. Gesierich, M. Duering, S. Ropele, R. Schmidt, and C. Langkammer, "Cross-sectional and Longitudinal Assessment of Brain Iron Level in Alzheimer Disease Using 3-T MRI," *Radiology*, vol. 296, no. 3, pp. 619–626, Sep. 2020.

[8] B. P. Drayer, "Imaging of the aging brain. Part II. Pathologic conditions," *Radiology*, vol. 166, no. 3, pp. 797–806, Mar. 1988.

[9] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking Clever Hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, no. 1, p. 1096, Mar. 2019.

[10] A. Kline and Y. Luo, "PsmPy: A Package for Retrospective Cohort Matching in Python," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2022, pp. 1354–1357, Jul. 2022.

[11] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, Apr. 1983. [Online]. Available: https://doi.org/10.1093/biomet/70.1.41

[12] S. M. Smith, Y. Zhang, M. Jenkinson, J. Chen, P. M. Matthews, A. Federico, and N. De Stefano, "Accurate, robust, and automated longitudinal and cross-sectional brain change analysis," *NeuroImage*, vol. 17, no. 1, pp. 479–489, Sep. 2002.

[13] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for Simplicity: The All Convolutional Net," Apr. 2015, arXiv:1412.6806 [cs]. [Online]. Available: http://arxiv.org/abs/1412.6806

[14] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognition*, vol. 65, pp. 211–222, May 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320316303582

[15] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Jan. 2017, arXiv:1412.6980 [cs]. [Online]. Available: http://arxiv.org/abs/1412.6980

[16] T. J. Bradshaw, Z. Huemann, J. Hu, and A. Rahmim, "A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging," *Radiology. Artificial Intelligence*, vol. 5, no. 4, p. e220232, Jul. 2023.

[17] C. Tinauer, M. Sackl, R. Stollberger, S. Ropele, and C. Langkammer, "Pfungst and Clever Hans: Identifying the unintended cues in a widely used Alzheimer's disease MRI dataset using explainable deep learning," Jan. 2025, arXiv:2501.15831 [eess]. [Online]. Available: http://arxiv.org/abs/2501.15831

[18] L. v. d. Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[19] C. Tinauer, A. Damulina, M. Sackl, M. Soellradl, R. Achtibat, M. Dreyer, F. Pahde, S. Lapuschkin, R. Schmidt, S. Ropele, W. Samek, and C. Langkammer, "Explainable Concept Mappings of MRI: Revealing the Mechanisms Underlying Deep Learning-Based Brain Disease Classification," in *Explainable Artificial Intelligence*, L. Longo, S. Lapuschkin, and C. Seifert, Eds. Cham: Springer Nature Switzerland, 2024, pp. 202–216.

[20] C. Langkammer, N. Krebs, W. Goessler, E. Scheurer, F. Ebner, K. Yen, F. Fazekas, and S. Ropele, "Quantitative MR imaging of brain iron: a postmortem validation study," *Radiology*, vol. 257, no. 2, pp. 455–462, Nov. 2010.