

Patch-based Intuitive Multimodal Prototypes Network (PIMPNet) for Alzheimer's Disease classification

Lisa Anita De Santi^{1,2,*}, Jörg Schlötterer^{3,4}, Meike Nauta⁵, Vincenzo Positano² and Christin Seifert³

¹Department of Information Engineering, University of Pisa, Pisa, Italy

²Fondazione Toscana G Monasterio - Bioengineering Unit, Pisa, Italy

³University of Marburg, Marburg, Germany

⁴University of Mannheim, Mannheim, Germany

⁵Datacation, Eindhoven, Netherlands

Abstract

Volumetric neuroimaging examinations like structural Magnetic Resonance Imaging (sMRI) are routinely applied to support the clinical diagnosis of dementia like Alzheimer's Disease (AD). Neuroradiologists examine 3D sMRI to detect and monitor abnormalities in brain morphology due to AD, like global and/or local brain atrophy and shape alteration of characteristic structures. There is a strong research interest in developing diagnostic systems based on Deep Learning (DL) models to analyse sMRI for AD. However, anatomical information extracted from an sMRI examination needs to be interpreted together with patient's age to distinguish AD patterns from the regular alteration due to a normal ageing process. In this context, part-prototype neural networks integrate the computational advantages of DL in an interpretable-by-design architecture and showed promising results in medical imaging applications. We present PIMPNet, the first interpretable multimodal model for 3D images and demographics applied to the binary classification of AD from 3D sMRI and patient's age. Despite age prototypes do not improve predictive performance compared to the single modality model, this lays the foundation for future work in the direction of the model's design and multimodal prototype training process.

Keywords

Interpretability-by-design, Prototype, Prototype-network, Multimodal Deep Learning, Alzheimer, MRI, Age

1. Introduction

There is a significant research interest in supporting Alzheimer's Disease (AD) diagnosis with Deep Learning (DL) models [1]. Existing diagnostic guidelines often integrate the clinical evaluation of the patient with structural Magnetic Resonance Imaging (sMRI), to detect pathological brain patterns like gray matter atrophy.

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

*Corresponding author.

✉ lisa.desanti@pdh.unipi.it (L. A. De Santi); joerg.schloetterer@uni-marburg.de (J. Schlötterer);

m.nauta@datacation.nl (M. Nauta); positano@ftgm.it (V. Positano); christin.seifert@uni-marburg.de (C. Seifert)

🆔 0000-0001-7239-4270 (L. A. De Santi); 0000-0002-3678-0390 (J. Schlötterer); 0000-0002-0558-3810 (M. Nauta);

0000-0001-6955-9572 (V. Positano); 0000-0002-6776-3868 (C. Seifert)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Brain alterations in sMRI might support the early and differential diagnosis and the prediction of disease’s progression. There are sets of common practices used for analysing sMRI acquisition, but there are still no universally accepted methods [2, 3, 4]. In addition, information collected from sMRI should be interpreted together with the patient’s age, as there are anatomical brain changes due to the physiological ageing process [5, 6].

DL architectures can facilitate the analysis of neuroimaging data, and might be able to identify unconventional AD subtypes and extract yet unknown image-based biomarkers [7, 8]. Prototypical-Part (PP) networks combine the advantages of DL models in an interpretable-by-design architecture, and are collecting interesting results in medical imaging applications where the black-box nature of standard DL models poses controversy [9].

There are currently different variants of PP networks, including PIPNet [10], originally applied to 2D images and then extended to handle 3D scans [11]. PIPNet showed appealing properties in the medical imaging domain [12], including a reduced number of part-prototypes, semantic significance of learned prototypes, and ability to cope with Out-of-Distribution data (which might be particularly useful in dementia diagnosis, where unusual neurodegeneration pattern are reported [4]). However, sMRI data should be interpreted together with patients’ demographics to discern age-related image alteration from pathological alteration, and existing PP models cannot be directly applied to perform this task. Adding non-image prototypes to the standard PP architecture is a non-trivial task, and there are no unique strategies available. There are some works which integrate the concept of learning prototypes from multiple modalities which are based on the concatenation (*deterministic prototypes*) or on the multimodal feature extraction (*shifted prototypes*). However, available models cannot be applied to our task, as specifically designed for images and textual data [13].

We present Patch-based Intuitive Multimodal Prototypes Network (PIMPNet), the first multimodal prototype classifier which learns 3D image part-prototypes and prototypical values from structured data, to predict patient’s cognitive level in AD from sMRI and age values.

2. Method

This section introduces the architecture (cf. Sect. 2.1 and Fig. 1) and the training process (Sect. 2.2), of PIMPNet.

2.1. Proposed Model: PIMPNet

We propose an *age-prototypes layer* integrated into the original PIPNet 3D model [11] to create our multimodal architecture. In contrast to “ordinary” age-binning for the inclusion of age information, the *age-prototypes layer* has the advantages of: (i) being able to learn important age values for the diagnostic task performed (which might not be equally distributed, and might not be easily identifiable in priors); (ii) not to assign different age bins to two patients of similar age close to the bin boundary.

Our PIMPNet has an **input layer** which takes the 3D image $\mathbf{x}_{img} \in \mathbb{R}^{ch \times S \times R \times C}$ and the age $\mathbf{x}_{age} \in \mathbb{R}^1$ as input, where ch, S, R, C respectively represents the number of channels, slices, rows and columns of the input image volume. Image \mathbf{x}_{img} and age \mathbf{x}_{age} are processed in parallel. A **CNN backbone** processes $\mathbf{x}_{img}, \mathbf{z} = f(\mathbf{x}_{img}; \mathbf{w}_f)$ extracting M 3-dimensional

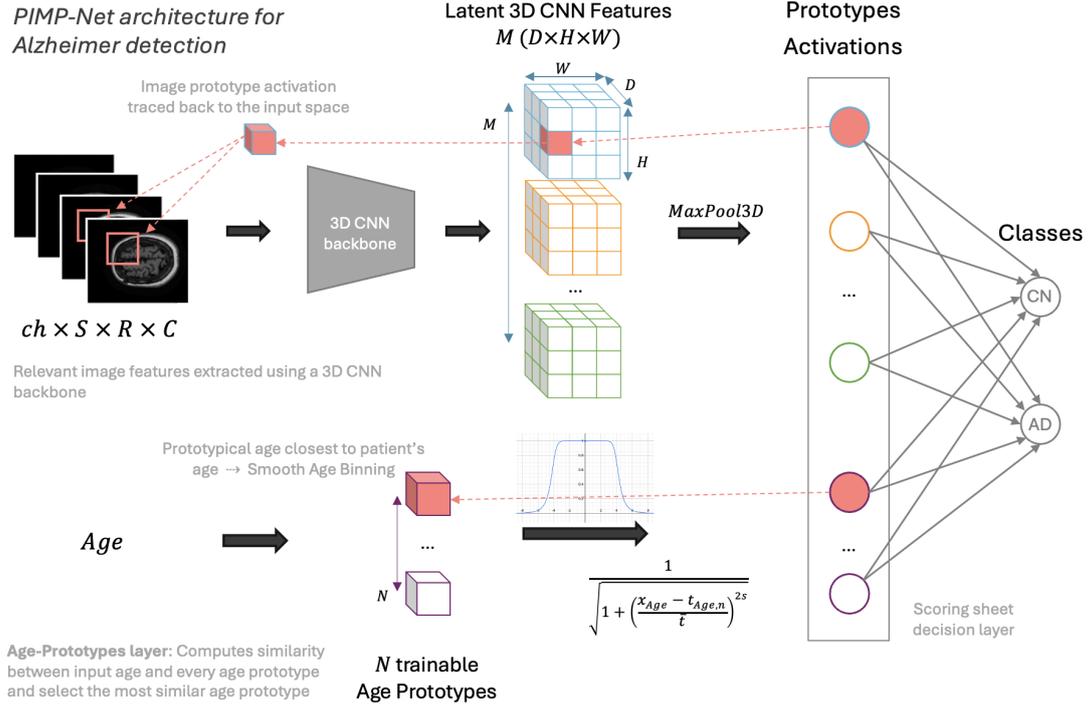


Figure 1: PIMPNet architecture

$(D \times H \times W)$ feature maps, where $\mathbf{z}_{m,d,h,w}$ represents the activation of image-prototype m at patch location d, h, w . Next, a **3D max-pooling** applied to every feature map extracts M image prototype presence scores $\mathbf{p}_{img} \in [0., 1.]^M$, where $\mathbf{p}_{img,m}$ measures the presence of the image prototype m in the input image. This defines the **image-prototypes layer**.

In parallel, we have the **age-prototypes layer**, constituted by N trainable 1-dimensional tensors $\mathbf{t}_{age,n} \in \mathbb{R}^{1 \times N}$, which aims to learn prototypical age values for the classification task. This layer computes age prototype presence scores $\mathbf{p}_{age} \in [0., 1.]^N$, a similarity measurement between the input age to every age prototype, defining a smooth age binning¹:

$$\mathbf{p}_{age,n} = \frac{1}{\sqrt{1 + \left(\frac{x_{age} - t_{age,n}}{\bar{t}}\right)^{2s}}}, \quad (1)$$

where \mathbf{t}_{age} are trainable parameters and \bar{t} and s are hyper-parameters which regulate the band and the slope of the similarity function. We then have a **prototypes layer** which concatenates the image and age prototype presence scores obtaining a layer of $L = M + N$ prototypes $\mathbf{p} \in [0., 1.]^L$ $\mathbf{p} = \text{concat}(\mathbf{p}_{img}, \mathbf{p}_{age})$. The final classification is performed by a **sparse linear**

¹The similarity function employed is inspired by the magnitude of a Butterworth filter [14]. In preliminary experiments, we used an exponential similarity function as in ProtoTree: $\mathbf{p}_{age,n} = \exp(-||x_{age} - t_{age,n}||)$, but as $\exp(-2) \approx 0.13$, a 2 years age difference would already result in little similarity, which is not in line with domain knowledge about age relevance for Alzheimer's disease.

positive layer $\mathbf{w}_c \in \mathbb{R}_{\geq 0}^{L \times K}$ which connects image and age prototypes to the K classes acting as a scoring sheet system. The K **class output scores** are given by the sum of the prototypes' presence score weighted for the contribution of prototype l to class k $\mathbf{w}_c^{l,k}$, i.e., $\mathbf{o} = \mathbf{p}\mathbf{w}_c$, where \mathbf{o} is $1 \times K$ and $\mathbf{o}_k = \sum_{l=1}^L \mathbf{p}_l \mathbf{w}_c^{l,k}$. PIMPNet returns the output class only using the most activated age-prototype, i.e., closest to the patient's age according to the similarity metric².

2.2. PIMPNet Training

We optimize PIMPNet's parameters by integrating the training of age prototypes into the original PIPNet training process [10]. This includes two main stages: (1) Self-Supervised pre-training of Image-Prototypes, and (2) PIMPNet training.

As the original PIPNet [10], the 1st stage generates positive pairs $\mathbf{x}'_{img}, \mathbf{x}''_{img}$ by applying data augmentation transformation to \mathbf{x}_{img} selected so that humans consider the two views similar. These are used to minimize the loss function $\lambda_A \mathcal{L}_A + \lambda_T \mathcal{L}_T$ by updating \mathbf{w}_f , where $\mathcal{L}_A = -\frac{1}{DHW} \sum_{(d,h,w) \in D \times H \times W} \log(\mathbf{z}'_{:,d,h,w} \cdot \mathbf{z}''_{:,d,h,w})$ is an *Alignment Loss* which optimizes positive pairs to activate the same prototype. Together with a softmax over $\mathbf{z}_{:,d,h,w}$, the alignment results in near-binary encodings where an image patch corresponds to exactly one prototype. $\mathcal{L}_T = -\frac{1}{M} \sum \log(\tanh(\sum \mathbf{p}_{img,b}) + \epsilon)$ is a *Tanh-Loss* used to prevent the trivial solution that one prototype node is activated on all image patches in each image in the dataset and instead activates multiple distinct prototypes per batch b . Only during training, output scores are calculated as $\mathbf{o} = \log((\mathbf{p}\mathbf{w}_c)^2 + 1)$, acting as regularization for sparsity.

The 2nd training stage includes the training of age prototypes, optimization of classification performance and image-prototypes fine-tuning for the downstream classification task. The optimization minimizes $\lambda_A \mathcal{L}_A + \lambda_T \mathcal{L}_T + \lambda_C \mathcal{L}_C$ by updating $\mathbf{w}_f, \mathbf{t}_{age}, \mathbf{w}_c$, where \mathcal{L}_C is the Log-likelihood classification loss

3. Evaluation

We used the multimodal dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database³. We selected "ADNI1 Standardized Screening Data Collection for 1.5T scans" processed with Gradwarp, B1 non-uniformity, and N3 correction, obtaining 307 CN and 243 AD sMRI brain scans and the corresponding patients' age. We report statistics on patients' demographic of the selected ADNI cohort in Table 1. We preprocessed sMRI data inspired by the pre-processing

Table 1

Patients' demographics of the selected ADNI cohort, further divided according to the clinical labels.

Class	N° subjects	Mean \pm SD Age	Age Range
<i>Both</i>	550	76 \pm 6	55-91
<i>CN</i>	307	76 \pm 5	60-90
<i>AD</i>	243	75 \pm 8	55-91

²We selected only the most activated age prototype during inference (not during the optimization process).

³<https://adni.loni.usc.edu>

Table 2

Performances comparison between PIPNet trained on 3D sMRI and PIMPNet trained on 3D sMRI + Age averaged over 5 folds.

Model	Acc	Bal Acc	SENS	SPEC	F1
PIPNet					
<i>ResNet-18 3D</i>	83 ± 04	83 ± 04	86 ± 06	79 ± 07	81 ± 05
<i>ConvNeXt-Tiny 3D</i>	65 ± 12	66 ± 09	56 ± 32	76 ± 15	66 ± 05
PIMPNet					
<i>ResNet-18 3D</i>	84 ± 04	83 ± 04	89 ± 03	77 ± 08	81 ± 05
<i>ConvNeXt-Tiny 3D</i>	72 ± 04	70 ± 04	86 ± 10	55 ± 14	63 ± 09

pipeline applied in previous works [15]. We transformed all the images to the common ICBM152 Non-Linear Symmetric 2009c standard space [16] with affine registration. We selected the grey matter structures by applying the ICBM152 Non-Linear Symmetric 2009c brain mask and kept a margin of 3 from its first and last non-empty slices. We applied an image downsampling of 2 and we scaled all image intensities to the range [0-1] with a min-max normalization. We implemented PIMPNet using PyTorch and MONAI⁴, training our models on an Intel Core i7 5.1 MHz PC, 32 GB RAM, equipped with an NVIDIA RTX3090 GPU with 24 GB of embedded RAM. As CNN backbones we used ResNet-18 3D pretrained on Kinetics400 [17] and ConvNeXt-Tiny 3D pretrained on the STOIC medical dataset (Study of Thoracic CT in COVID-19) [18]. We finetuned PIMPNet with Adam optimizer using the same hyperparameter settings of the original PIPNet [10]. We only reduced the batch size to 12 to adapt it to our computational capabilities and we set the learning rate of age prototypes to 0.1⁵. We arbitrarily set the number of age prototypes = 5 evenly spaced between 40 and 90 to cover the patients’ age range of our dataset. For the age similarity function, we respectively set $\bar{t} = 4$ and $s = 8$ ⁶. We performed 5-fold cross-validation with patient-wise splits. 20% of training images are used for validation.

We evaluated the models in terms of classification performance and with functionally grounded metrics of explainability. Results are reported in Tables 2 and 3. We compared PIMPNet performance (sMRI + age) with PIPNet-3D (sMRI only) [11], to evaluate if including age information improves diagnostic performance. We measured performance using Accuracy (Acc), Balanced Accuracy (Bal Acc), Sensitivity (SENS, Acc of Cognitive Normal class), Specificity (SPEC, Acc of Alzheimer’s Disease class), and F1 score (F1). We measured the Global size (GS) of the model as the total number of prototypes, and the Local size (LS) of explanations as the number of detected prototypes in a single 3D sMRI, averaged over all the images in the test set. Additionally, we report the Sparsity (Sp) of the decision layer as the percentage of zero-weights in the linear classification layer [10] to assess the compactness of the prototypes-classes layer. We further assessed whether prototypes are consistently located in the same brain region, and the purity of the prototypes in terms of the anatomical regions included based on the Cerabra atlas annotation [19]. More specifically, the Prototypes Localization Consistency (LC_p) evaluates

⁴<https://monai.io>

⁵Using the same learning rate of the original PIPNet used to train the image-prototypes (0.05) results in irrelevant updates of the Age Prototypes

⁶We leave an extensive hyperparameter search for learning the age prototypes for future work

Table 3

Functionally-grounded evaluation of PIPNet trained on 3D sMRI and PIMPNet trained on 3D sMRI + Age averaged over 5 folds. \uparrow and \downarrow : tendency for better values.

Model	GS \downarrow	LS \downarrow	Sp \uparrow	LC _p \downarrow	H _p \downarrow
ResNet-18 3D					
PIPNet	149 \pm 18	73 \pm 10	0.855 \pm 0.018	0.008 \pm 0.006	2.474 \pm 0.249
PIMPNet	143 \pm 35	74 \pm 20	0.861 \pm 0.033	0.006 \pm 0.006	2.424 \pm 0.162
ConvNeXt-Tiny 3D					
PIPNet	4 \pm 2	2 \pm 1	0.997 \pm 0.001	0.000 \pm 0.000	1.803 \pm 0.999
PIMPNet	10 \pm 9	4 \pm 4	0.993 \pm 0.002	0.000 \pm 0.000	1.543 \pm 0.626

Table 4

Prototypical Age Values $t_{Age,i}$ learned for folds M1, ..., M5 trained with different backbones.

Fold	$t_{Age,1}$	$t_{Age,2}$	$t_{Age,3}$	$t_{Age,4}$	$t_{Age,5}$	$t_{Age,1}$	$t_{Age,2}$	$t_{Age,3}$	$t_{Age,4}$	$t_{Age,5}$
	<i>ResNet-18 3D</i>					<i>ConvNeXt-Tiny 3D</i>				
M1	65.77	65.81	66.14	76.81	80.99	56.81	65.00	64.96	74.13	85.80
M2	68.46	69.40	70.38	77.04	82.38	55.75	58.39	64.96	74.32	85.59
M3	66.37	67.27	67.91	75.87	81.96	54.86	56.63	65.21	74.40	85.11
M4	66.72	66.72	67.07	77.07	79.75	58.22	58.59	66.50	75.88	89.09
M5	66.51	66.52	67.23	77.37	80.00	57.79	66.94	65.44	72.55	84.58

the differences in the coordinate centre of the prototypical part in the input image, while the Prototype Brain Entropy (H_p) as a measure of purity computes the Shannon Entropy of the brain regions included in the prototypical part [11]. We show the learned age prototypes t_{age} from five different folds (denoted as Mx where x indicates the current fold) in Table 4.

4. Discussion and Conclusion

Both PIPNet and PIMPNet with the ResNet-18 3D backbone achieve higher classification performance than with the ConvNext-Tiny backbone. Our preliminary results also show that the proposed Age-Prototype layer can learn prototypical age values; however, these do not improve classification performances compared to the baseline model. Our functionally-grounded evaluation of prototypes shows that all models learn prototypes consistently located in the same anatomical brain regions (low LC_p values). We also observe that the models trained with the ConvNeXt-Tiny 3D backbone have higher compactness. This might partially explain the lower performance scores (the number of prototypes learned is not enough for performing the diagnosis), but is an interesting observation for future research as such a highly compact model can be considered more interpretable than larger ones and can be easily evaluated by domain experts. We also observe that the image prototypes of the ConvNeXt-Tiny 3D backbone are generally purer⁷. Despite purity being a desirable property for prototypes [20], because of

⁷Purity is measured w.r.t. to the annotation provided by the CerebrA atlas

the design of the purity metric, also a prototype which only includes the background, i.e., a clinically irrelevant prototype, will have high purity⁸.

In summary, we proposed PIMPNet, an interpretable multimodal prototype-based classifier. The proposed architecture is the first prototypes-based network which performs an interpretable classification based on the detection of prototypes learned from different data modalities (3D images and age information). We applied PIMPNet to the binary classification of Alzheimer's Disease from 3D sMRI images together with the patient's age. Despite the usage of age prototypes do not improve predictive performance compared to the model trained with only images, we identified different potential reasons which define the future directions of our work. First, as the original PIPNet training paradigm includes a pretraining stage [10] of image prototypes, we plan to include an age prototypes pretraining step w.r.t. the log-likelihood classification loss. Second, we also plan to work on the model's design. As the simple concatenation of the prototype presence score might not be able to properly represent the relationship between age and image prototypes for the downstream task, we plan to combine image and age prototypes using a different (but still interpretable) classifier than a scoring-sheet system.

Acknowledgments

Data used in the preparation of this article was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The ADNI was launched in 2003 as a public-private partnership with the primary goal to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

References

- [1] M. A. Ebrahimighahnavieh, S. Luo, R. Chiong, Deep learning to detect alzheimer's disease from neuroimaging: A systematic literature review, *Computer Methods and Programs in Biomedicine* 187 (2020). doi:10.1016/j.cmpb.2019.105242.
- [2] L. De Santi, E. Pasini, M. Santarelli, D. Genovesi, V. Positano, An Explainable Convolutional Neural Network for the Early Diagnosis of Alzheimer's Disease from 18F-FDG PET, *Journal of Digital Imaging* 36 (2023). doi:10.1007/s10278-022-00719-3.
- [3] A. Chandra, G. Dervenoulas, M. Politis, Magnetic resonance imaging in Alzheimer's disease and mild cognitive impairment, 2019. doi:10.1007/s00415-018-9016-3.
- [4] P. Vemuri, C. Jack, Role of structural MRI in Alzheimer's disease, *Alzheimer's Research and Therapy* 2 (2010). doi:10.1186/alzrt47.
- [5] L. Zhao, W. Matloff, K. Ning, H. Kim, I. D. Dinov, A. W. Toga, Age-related differences in brain morphology and the modifiers in middle-aged and older adults, *Cerebral Cortex* 29 (2019) 4169–4193. doi:10.1093/cercor/bhy300.

⁸Posterior quantitative evaluation performed w.r.t. the CerebrA atlas revealed that the test set image prototypes (averaged over the 5-folds) obtained with the ConvNeXt-Tiny backbone have a higher percentage of background voxels included compared to the ones obtained with ResNet-18 (76.6% vs 59.2%)

- [6] R. Sivera, H. Delingette, M. Lorenzi, X. Pennec, N. Ayache, A model of brain morphological changes related to aging and alzheimer's disease from cross-sectional assessments, *NeuroImage* 198 (2019) 255–270. doi:10.1016/j.neuroimage.2019.05.040.
- [7] M. Böhle, F. Eitel, M. Weygandt, K. Ritter, Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification, *Frontiers in Aging Neuroscience* 10 (2019). doi:10.3389/fnagi.2019.00194.
- [8] M. Khojaste-Sarakhsi, S. S. Haghighi, S. F. Ghomi, E. Marchiori, Deep learning for Alzheimer's disease diagnosis: A survey, *Artificial Intelligence in Medicine* 130 (2022) 102332. doi:https://doi.org/10.1016/j.artmed.2022.102332.
- [9] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. D. Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith, S. Stumpf, Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions, *Information Fusion* 106 (2024) 102301. URL: <http://creativecommons.org/licenses/by/4.0/>. doi:10.1016/j.inffus.2024.102301.
- [10] M. Nauta, J. Schlötterer, M. van Keulen, C. Seifert, PIP-Net: Patch-Based Intuitive Prototypes for Interpretable Image Classification, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. doi:10.1109/CVPR52729.2023.00269.
- [11] L. A. D. Santi, J. Schlötterer, M. Scheschenja, J. Wessendorf, M. Nauta, V. Positano, C. Seifert, Pipnet3d: Interpretable detection of alzheimer in mri scans, 2024. arXiv:2403.18328.
- [12] M. Nauta, J. H. Hegeman, J. Geerdink, J. Schlötterer, M. v. Keulen, C. Seifert, Interpreting and correcting medical image classification with pip-net, in: *Artificial Intelligence. ECAI 2023 International Workshops*, 2024, pp. 198–215.
- [13] Y. Ma, S. Zhao, W. Wang, Y. Li, I. King, Multimodality in meta-learning: A comprehensive survey, *Knowledge-Based Systems* 250 (2022). doi:10.1016/j.knosys.2022.108976.
- [14] S. Butterworth, et al., On the theory of filter amplifiers, *Wireless Engineer* 7 (1930) 536–541.
- [15] A. W. Mulyadi, W. Jung, K. Oh, J. S. Yoon, K. H. Lee, H.-I. Suk, Estimating explainable Alzheimer's disease likelihood map via clinically-guided prototype learning, *NeuroImage* 273 (2023). doi:10.1016/j.neuroimage.2023.120073.
- [16] V. Fonov, A. Evans, R. McKinsty, C. Almli, D. Collins, Unbiased nonlinear average age-appropriate brain templates from birth to adulthood, *NeuroImage* 47 (2009) S102. URL: <https://www.sciencedirect.com/science/article/pii/S1053811909708845>. doi:https://doi.org/10.1016/S1053-8119(09)70884-5, organization for Human Brain Mapping 2009 Annual Meeting.
- [17] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, A closer look at spatiotemporal convolutions for action recognition (2017). URL: <http://arxiv.org/abs/1711.11248>.
- [18] D. Kienzle, J. Lorenz, R. Schön, K. Ludwig, R. Lienhart, Covid detection and severity prediction with 3d-convnext and custom pretrainings (2022). URL: <http://arxiv.org/abs/2206.15073>.
- [19] A. L. Manera, M. Dadar, V. Fonov, D. L. Collins, Cerebra, registration and manual label correction of mindboggle-101 atlas for mni-icbm152 template, *Scientific Data* 7 (2020). doi:10.1038/s41597-020-0557-9.
- [20] M. Nauta, C. Seifert, The co-12 recipe for evaluating interpretable part-prototype image classifiers, in: *Explainable Artificial Intelligence*, 2023, pp. 397–420.