Back to the basics with inclusion of clinical domain knowledge - A simple, scalable and effective model of Alzheimer's Disease classification

Sarah C. Brüningk[†] SARAH.BRUENINGK@BSSE.ETHZ.CH
Machine Learning & Computational Biology, Department of Biosystems Science and Engineering
ETH Zurich,Basel, Switzerland
Swiss Institute for Bioinformatics (SIB), Lausanne, Switzerland

Felix Hensel[†] FELIX.HENSEL@BSSE.ETHZ.CH
Machine Learning & Computational Biology, Department of Biosystems Science and Engineering
ETH Zurich, Basel, Switzerland
Swiss Institute for Bioinformatics (SIB), Lausanne, Switzerland

Louis P. Lukas LUKASL@STUDENT.ETHZ.CH
Machine Learning & Computational Biology, Department of Biosystems Science and Engineering
ETH Zurich,Basel, Switzerland

Merel Kuijs KUIJSM@STUDENT.ETHZ.CH
Machine Learning & Computational Biology, Department of Biosystems Science and Engineering
ETH Zurich,Basel, Switzerland

Catherine R. Jutzeler[**] CATHERINE.JUTZELER@BSSE.ETHZ.CH
Machine Learning & Computational Biology, Department of Biosystems Science and Engineering
ETH Zurich,Basel, Switzerland
Swiss Institute for Bioinformatics (SIB), Lausanne, Switzerland

Bastian Rieck[**] BASTIAN.RIECK@BSSE.ETHZ.CH
Machine Learning & Computational Biology, Department of Biosystems Science and Engineering
ETH Zurich, Basel, Switzerland
Swiss Institute for Bioinformatics (SIB), Lausanne, Switzerland

and Alzheimer's Disease Neuroimaging Initiative* †: These authors contributed equally. **: These authors share last authorship. *: Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Abstract

On high-resolution structural magnetic resonance (MR) images Alzheimer's disease (AD) is pathologically characterised by brain atrophy and an overall loss of brain tissue connectivity. In this study, we harness such prior clinical domain knowledge to evaluate MR image-based classification of AD patients from healthy controls using deliberately simple convolutional neural network (CNN) architectures. In addition to evaluating CNN performance on high resolution structural MR imaging data, we consider topological feature representations thereof to evaluate structural connectivity. We perform an ablation study, combined with model interpretability analysis, to evaluate the relevance of the specific image region used for classification. Notably, we find that by choosing a meaningful data representation comprising the left hippocampus, we achieve competitive performance (accuracy $84 \pm 7\%$) comparable to far more complex, heavily parameterised machine learning architectures. This implies that clinical domain knowledge may overrule the importance of model architecture design in the case of AD classification. This opens up new possibilities for interpretable architectures and simplifies model training in terms of computational cost and hardware requirements.

1. Introduction

Alzheimer's disease (AD) is the primary cause of dementia and the fifth leading cause of death in people over the age of 65 (Winblad et al., 2016). The societal and economic costs associated with AD are enormous and projected to significantly rise with increasing longevity (Wong, 2020). Despite immense research efforts, there is currently no curative treatment for AD. Pathologically, AD manifests as an accumulation of intracellular neurofibrillary tangles, neuritic plaques, as well as neuronal and synaptic loss (Scheff et al., 2006; Serrano-Pozo et al., 2011; Suemoto et al., 2017) progressing to large-scale changes to the brain morphology. These changes comprise atrophy, global loss of cerebral connectivity, volumetric shrinkage of distinct brain areas (e.g. amygdala and hippocampus), and enlargement of the ventricles (Barnes et al., 2009). Non-invasive neuroimaging, such as high-resolution magnetic resonance imaging (MRI), plays a major role in visualizing local atrophy and alterations to the global tissue connectivity (Rosenbloom and Pfefferbaum, 2008). Key clinical questions to address with these images include the early detection of disease onset (e.g. cerebral changes) and the prediction of disease progression. Solving these questions will ultimately allow to optimise available treatment strategies. Machine learning based image classification is well suited to facilitate the detection of imaging biomarkers. A first step towards this goal is the distinction of AD patients from cognitively normal subjects (CN) of comparable age.

A broad variety of machine learning approaches have previously been suggested for this task. Generally speaking, the main challenge for imaging-based classification is the efficient handling of large imaging data as well as facilitating interpretable solutions to build trust in the model prediction. The recent trend of model architectures has been moving towards more complex approaches, employing a combination of deep convolutional neural networks and unsupervised learning, or ensemble models combining multiple architectures as described in detail below. While such approaches are seen to perform well and reach state-of-the-art classification performance, they are also often hampered with respect to their (clinical) applicability; with increasing MRI resolution, memory requirements may quickly become prohibitive. Hence, several approaches prefer *down-sampling* input data to achieve good performance (Jin et al., 2019; Korolev et al., 2017; Oh et al., 2019). Downsampling data leads to the paradoxical situation that high-resolution data, which could be useful for an early detection of AD, is discarded. This raises the question of the necessity of complex models. Would simpler model architectures, which effectively select features and build on prior knowledge of the biological hallmarks of AD, be sufficient? In addition to significantly lower computational costs, a major advantage of simpler models is their interpretability. For instance, using interpretability aware model training can improve classification robustness, which in turn will promote trust in chosen models (Boopathy et al., 2020). In particular in the healthcare domain, model interpretability and explainability are essential to build trust in the suggested predictions and to allow the physician to base a decision on models that are "right for the right reason". As such, less complex models based on clinical domain knowledge could be more readily translatable to clinical application.

In this paper, we investigate two conceptually simple approaches embracing the pathological hallmarks of AD at both the local (brain tissue atrophy, particularly in the hippocampal region) and global (loss of whole brain tissue connectivity) scale. We first evaluate the performance of convolutional neural networks (CNNs) on subsets of full-resolution 3D MR images and perform an ablation study to compare different levels of biological scale and anatomical brain areas. Secondly, we employ topological data analysis (TDA) to assess the global changes in brain connectivity. Briefly, TDA leverages the mathematical theory of algebraic topology in order to detect topological features capturing the structural brain connectivity or changes thereof. We hypothesize that using simple model architectures that include prior domain knowledge of morphological hallmarks of AD and topological features thereof, will result in classification performance comparable to more complex model (e.g. Oh et al. (2019); Pan et al. (2020)). Moreover, our models will facilitate fast, scalable, and interpretable solutions through meaningful representations.

Generalizable Insights about Machine Learning in the Context of Healthcare

The hypothesis addressed in this contribution is translatable to a variety of machine learning problems covering:

- the optimised handling of large-scale medical imaging data in terms of computational cost
- inclusion of prior biological insights into model design, and data preselection
- topological data analysis for the application to structural MRI data.

For machine learning, medical imaging data is challenging owing to its large dimensions, which impose restrictions on hardware and thus exacerbate the optimisation of model hyperparameters in light of computationally expensive model training. We investigate the use of patch-based models and subsequent combination through graph neural networks or logistic regression to provide a scalable solution with minimal hardware requirements. In addition, we observe that machine learning research is heading towards increasingly complex model architectures and ensemble models to boost predictive performance. Such models are not only computationally expensive, but often inherently difficult to interpret. We hypothesise that restricting the image to brain functional units predominantly affected by AD (for example the hippocampus) may be more effective to guide classification while potentially also allowing for more scalable and hence computationally less expensive solutions. Finally, the application of TDA is a growing field in machine learning research (Carrière et al., 2020; Hofer et al., 2017). It provides a means of feature extraction based on intrinsic connectivity information and has previously shown success for the analysis of functional MRI data (Rieck et al., 2020). To our knowledge, no application to structural MRI data has been reported to date and we investigate its utility in this context. There is a clear clinical motivation for why the analysis of tissue connectivity may be suitable for AD classification. Moreover, TDA provides an effective means of data compression which would align with the aim to provide more scalable and computationally effective solutions for machine learning on medical imaging data.

2. Related Work

	(1)*	(2)*	(3)*	(4)*	(5a)*	(8a)*	(5b)†	(6)†,¶	(7)‡	(8b1§)††	(9)††,¶
ACC	0.79 ±0.08	0.77 ±0.06	0.90 **	0.67§ ±0.01	0.76 ±0.02	0.74 **	0.87 ±0.04	0.99	0.84 ±0.05	0.81 **	0.96 ±0.01
ACC‡‡	0.77 ±0.07	0.74 ±0.06						0.80 ±0.05			
AUC	0.88 ±0.08	0.78 ±0.04	**	0.79 ±0.01	**	**	0.91 **	0.99 ±0.02	0.92 ±0.08	**	**
Data	ADNI	ADNI	ADNI	ADNI, AIBL	ADNI	ADNI	ADNI	ADNI, CAD- Dementia	ADNI	ADNI	OASIS

Table 1: Previously reported performance estimates for CNNs for classification of AD and CN subjects (left: simple 3D CNNs, right: more complex models). Accuracy (ACC), and area under the receiver operating curve (AUC) are reported. Additionally we included re-evaluations by Oh et al. (2019) since this publication used a comparable data split and test set to our work. (1) Korolev et al. (2017); (2) Rieke et al. (2018); (3) Bäckström et al. (2018); (4) Liu et al. (2020); (5a,b) Oh et al. (2019); (6) Hosseini-Asl et al. (2018); (7) Pan et al. (2020); (8) Valliani and Soni (2017); (9) Hon and Khan (2017). *3D CNN architecture; †3D Convolutional Autoencoder and 3D CNN; ‡Ensemble based on 2D CNNs; †† transfer learning based on fine tuning existing model architectures with 2D slices. **Missing information not reported in original publication; §Three class classification (AD vs. MCI vs. CN); ¶Unclear data split used in the original publication (see also Wen et al. (2020)); ‡‡ additional values reported are performance estimates provided by Oh et al. (2019) who reproduced the original architecture. Data bases: AIBL (Ellis et al., 2009); CADDementia (Bron et al., 2015); OASIS (Marcus et al., 2010).

Classification of CN and AD subjects has been attempted using a wide variety of machine learning methods, ranging from linear support vector machines to deep 3D convolutional neural network (CNN) architectures (Rathore et al., 2017; Wen et al., 2020). Data types used for this task included genetic markers, clinical scores, cerebrospinal fluid markers, and a variety of MRI modalities. As the present study is based on structural MRI, the remainder of this section is focusing to previous work based on this neuroimaging modality. Table 1 summarises performance estimates for a selection of studies highlighted by Wen et al. (2020), predominantly using data from the ADNI database as used in our work. Early approaches typically warranted extensive preprocessing of scans to extract features, such as cortical thickness measurements or tissue density (Rathore et al., 2017). Recent advances in the AD classification task have been focusing on the application of deep CNNs for end-to-end learning approaches on naive images (Wen et al., 2020). Both 2D (Aderghal et al., 2017; Pan et al., 2020) and 3D convolutional architectures have been explored previously (Hosseini-Asl et al., 2018; Korolev et al., 2017). Approaches building on 2D brain image slices might provide an appealing solution in a setting where computational resources are constrained. However, in order to achieve state-of-the-art performance, large ensemble models combining multiple architectures were required. For example, Pan et al. (2020) selected a total of 15 base classifiers from an initial set of 123 options, each comprising a separate CNN trained on a distinct 2D image slice. Hence, despite the use of simple model architectures underpinning the ensemble, there was only a limited reduction in terms of computational resources. Earlier work also considered transfer learning from natural image classification, such as ResNet (He et al., 2016) or Inception Net (Szegedy et al., 2015) and fine tuning these architectures' output layers on either single (Valliani and Soni, 2017) or multiple 2D MRI slices (Hon and Khan, 2017).

Restricting models to 2D image slices implies a loss of information, especially with respect to spatial relationships. Consequently, many approaches have favoured full 3D MRIs evaluated by means of 3D CNNs. Simple feed forward architectures with varying numbers of convolutional layers were presented by Bäckström et al. (2018); Korolev et al. (2017); Rieke et al. (2018) and Liu et al. (2020); Korolev et al. (2017) also assessed residual convolutional networks; and Hosseini-Asl et al. (2018) and Oh et al. (2019) described combinations of 3D convolutional autoencoders for feature extraction with subsequent classification by a 3D CNN. Studies reporting results from standard feed forward architectures included multiple layers, each consisting of at least one set of convolutions (of varying size), activation and maxpooling (Bäckström et al., 2018; Korolev et al., 2017; Liu et al., 2020; Rieke et al., 2018). While no results have been reported for smaller/shallower architectures, Liu et al. (2020) observed that wider architectures provided greater performance gains, while additional layers only provided marginal improvements. Liu et al. (2020) further assessed the impact of a number of modifications such as instance instead of batch normalisation; magnitude of spatial downsampling in early layers; and dataset size. They reported that subsampling, i.e. including fewer instances of each diagnostic group, performed worse. This may be expected given the relatively limited number of instances available and the large number of parameters to train for deep models. Interestingly, limiting spatial downsampling by using a larger number of small kernels in the early layers of the network improved performance. As diagnosis of AD, especially at early stages, requires the detection of subtle differences in the MRI scan, this observation suggests that aggregating a large collection of (highly) localised features could improve the classification performance in more challenging cases.

Many previous studies did not consider the interpretability of their approach (Bäckström et al., 2018; Hosseini-Asl et al., 2018) or only provided limited interpretability analysis as part of their model evaluation (Korolev et al., 2017; Liu et al., 2020; Oh et al., 2019).

An exception is the publication by Rieke et al. (2018), who designed their study with the explicit goal of assessing a variety of interpretability methods in addition to classification performance. Rieke et al. (2018) assessed two occlusion and two gradient based methods to identify the most discriminative brain regions. Results matched previous findings from AD research, as all methods assessed highlighted the hippocampus and adjacent regions. Given the pathological variability of AD (Ferreira et al., 2020), however, these patterns might vary between individual subjects. Importantly, current interpretability analysis exclusively focused on large scale brain changes, as manifested in distinct brain functional subunits, but did not evaluate the local imaging features driving the identification of these subregions.

3. Cohort

3.1. Cohort Selection

This study was performed on T1-weighted MRIs obtained from the Alzheimer's Disease Neuroimaging Initiative[1] (ADNI) and comprised a collection of 358 subjects, including AD patients and healthy controls of matched age (mean age 77 ± 7 years). For each subject scans were acquired at multiple time points (yearly up to ten years) using a designated imaging protocol[2].

3.2. Feature Choices

All images were preprocessed using established pipelines: Following conversion to BIDS format using `clinica`[3], `fmriprep`[4] was used to perform registration to MNI reference space, bias field correction, and brain extraction for all images. The obtained $193 \times 229 \times 193$ pixel images were intensity normalised. All images were initially cropped to the inner brain volume of $120 \times 144 \times 120$ voxels to exclude the majority of the acquired image outside the brain. This crop was further subdivided into 64 non-overlapping image patches of size $30 \times 36 \times 30$. These patch dimensions provided a trade-off between image detail captured and overall patch size. Finally, the left and right hippocampus were segmented based on the Cerebrum Atlas (Manera et al., 2020), with a uniform three voxel dilation applied to the atlas mask to ensure full coverage. See Figure 1 for an overview of all image subsets used for classification.

3.3. Feature Choices for TDA

Persistent homology (Barannikov, 1994; Edelsbrunner and Harer, 2010) is the flagship tool in TDA for the extraction of topological information from a data set. We will briefly provide some intuition of the concept; for an in-depth background on TDA, specifically in the context of machine learning, we refer to a recent survey (Hensel et al., 2021). In order to apply TDA in our setting, we first need a *filtered simplicial* (i.e. *cubical–*) *complex*, which can be thought of as a higher dimensional analogue of a graph). In our case, as we are dealing with 3D images, this cubical complex structure is naturally given by the voxel grid together with

[1] `http://adni.loni.usc.edu/`

[2] In depth description of the protocol can be found here: `http://adni.loni.usc.edu/methods/documents/`

[3] `https://github.com/aramis-lab/clinica`

[4] `https://fmriprep.org/en/stable/`

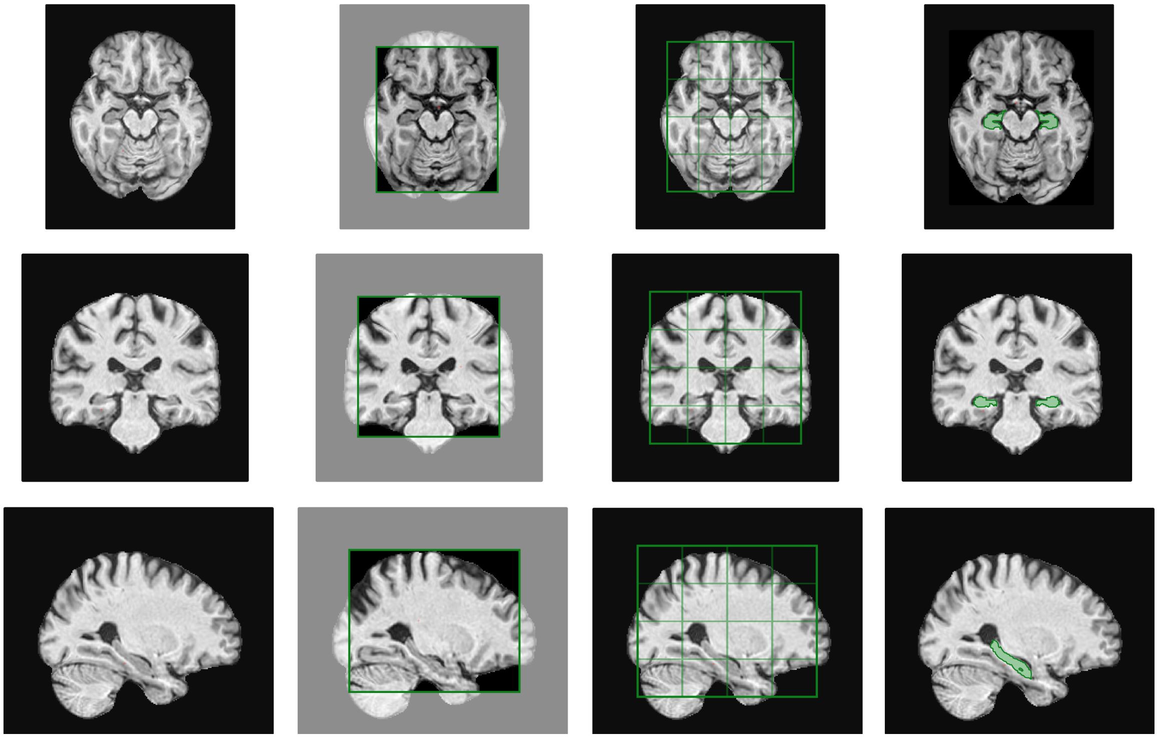

Figure 1: Schematic illustration of the T1-weighted MR image subvolumina used in this analysis (top: horizontal, middle: coronal, bottom: sagittal planes). Columns from left to right represent: full image (193 × 229 × 193); inner brain volume (120 × 144 × 120); image patches (each 30 × 36 × 30); hippocampus (green), $\sim 33 \times 46 \times 48$).

a filtration over *intensity* (i.e. greyscale). By applying a threshold value $\tau \in [0, 1]$ to a 3D image, we obtain a cubical complex, the τ-superlevel-set, by only retaining voxels whose intensity is at least τ. The topological features of the τ-superlevel-set are computed for τ ranging in $[0, 1]$. In this case, topological features can occur in dimensions $d \in \{0, 1, 2\}$ and represent *connected components* ($d = 0$), *cycles* or *tunnels* ($d = 1$) and *voids* ($d = 2$). Persistent homology tracks the *changes* of these topological features as τ is varied and stores this information in a so-called persistence diagram for each of the dimensions $d = 0, 1, 2$. We performed the persistence homology calculation using the tool DIPHA[5] on each of the image subsets described above. In order to use persistent homology features for machine learning frameworks, we applied the *persistence image* (Adams et al., 2017) vectorisation method, which is a weighted discretisation of persistence diagrams via a Gaussian distribution. For this step we used persim[6], with standard parameters, a quadratic weighting function and an output resolution of 50 × 50 of the resulting persistence images (PIs) in each of the

[5] `https://github.com/DIPHA/dipha`

[6] `https://persim.scikit-tda.org/en/latest/index.html`

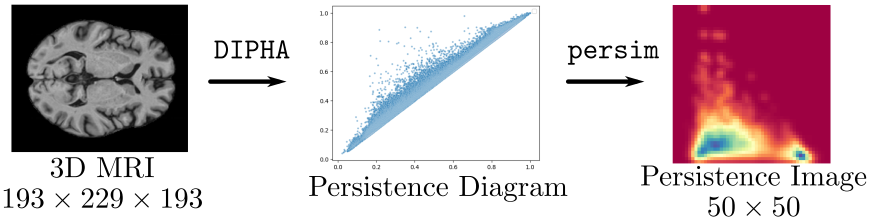

Figure 2: A schematic representation of the topological feature generation; from a 3D MRI image to a persistence image. Persistence homology is calculated via `DIPHA` and the resulting persistence diagram is then vectorised via `persim` to obtain a persistence image. This pipeline is repeated for each (homological) dimension $d \in \{0, 1, 2\}$.

dimensions $d \in \{0, 1, 2\}$. See Figure 2 for a schematic representation of the topological feature generation pipeline.

4. Methods

The relevant code and used subject IDs have been published on `https://github.com/BorgwardtLab/ADNI_3DCNNvsTDA.git`.

4.1. Model architecture

The underlying hypothesis of this study was that inclusion of clinical domain knowledge could boost classification performance of AD and CN subjects to allow for simpler, computationally less demanding architectures feasible for large scale application. We performed an ablation study through subsequential focusing of the model input data to more clinically relevant brain regions. We initially used the full-resolution inner brain volume as input, then evaluated geometric brain subregions (image patches), and finally, brain functional subunits (left and right hippocampus). For each of these image subsets both a TDA and purely image domain-based model was evaluated.

All model architectures with relevant hyperparameters are given in the supplementary material. For image-based classification, we used simple multilayer 3D CNNs with batch normalisation, dropout, and ReLU activation. Convolutional layers were followed by global average pooling and two dense layers prior to the output layer. For persistence image inputs, four-layer 2D CNNs (with comparable layers as described above) were trained separately for each of the three homological dimensions. The obtained pre-classification layer encodings from these three models were combined using a multilayer perceptron comprising three fully-connected layers, with optimised L_1 regularisation and sigmoid activation function, to arrive at the final classification based on topological features.

For each model, hyperparameters were optimised by random sampling, which, where applicable, was performed on a patch within the hippocampus. The relevant best hyperparameters were subsequently applied to all image patches. Models using optimal hyperparameters were trained within up to 2500 epochs on a NVIDIA TITAN RTX, 24 GiB RAM GPU. We implemented early stopping based on validation loss (binary cross entropy loss) and used the Adam optimiser with He uniform initialisation and balanced class weights. This method was implemented in Python using Keras/TensorFlow. Table 2 gives an overview of the relevant model complexity and training times for all investigated approaches.

Model	I-3D	P*-3D	HC-3D	I-TDA	P*-TDA	HC-TDA
Input dimension [voxels]	120x144x120	30x36x30	33x45x48	50x50	50x50	50x50
Input type	3D inner brain image	3D image patch	3D HC	2D PI	2D PI	2D PI
n	270k	72k	140k	54k	54k	54k
Training time/epoch	25 s	1 s	2 s	10 ms	10 ms	10 ms

Table 2: Overview of the implemented model architecture with relevant input data dimensions, number of trainable parameters (n), and training time per epoch. Abbreviations: TDA: Topological data analysis, PI: Persistence image, 2D: two dimensional, CNN: Convolutional neural network, P*: Best single patch $30 \times 36 \times 30$ voxels, I: Inner image subset of dimension $120 \times 144 \times 120$ voxels, HC: Hippocampus.

4.2. Patch ensemble models

We compared two strategies for combining the information of the 64 image patches: (i) logistic regression (LR) as a means of unstructured data combination, and (ii) graph neural networks (GNN) to evaluate the importance of geometrical proximity of image patches. Logistic regression was implemented using hyperparameters optimised via grid search, balanced class weights, and the `liblinear` solver Pedregosa et al. (2011). Here, class probabilities obtained from 2D/3D CNNs for each subject and patch (normalised, and centred around a decision boundary of zero) were used as inputs.

A simple GNN comprising four graph convolutional layers, with batch normalisation, dropout and ReLU activation function, was implemented using the `pytorch-lightning` library [7]. The graph structure underlying our GNN was chosen as a simple adjacency graph, in which nodes correspond to the image patches where edges connect patches that are direct (including diagonal) neighbours. This architecture provides a means to incorporate *between* patch connectivity information. We calculated the binary cross entropy loss and used the Adam optimizer with early stopping on validation loss. The CNN preclassification layer encodings were used as GNN node features for each image patch. Encodings were multiplied by the relevant output layer weights, and sorted with increasing variance between AD and CN groups to ensure most discriminative features being located in comparable positions for inter node operations. Figure 3 gives an overview of the GNN graph and node feature generation. In addition to combining different image patches, also TDA and 3D CNNs were combined in the same way by providing both, topological and structural model outputs as features to the relevant models.

4.3. Performance evaluation

We performed 5-fold cross validation (CV) splitting patients into training (80%) and test (20%) sets, stratified by class labels. Within the training samples a further 75-25% split

[7] `https://pytorch-lightning.readthedocs.io/en/latest/`

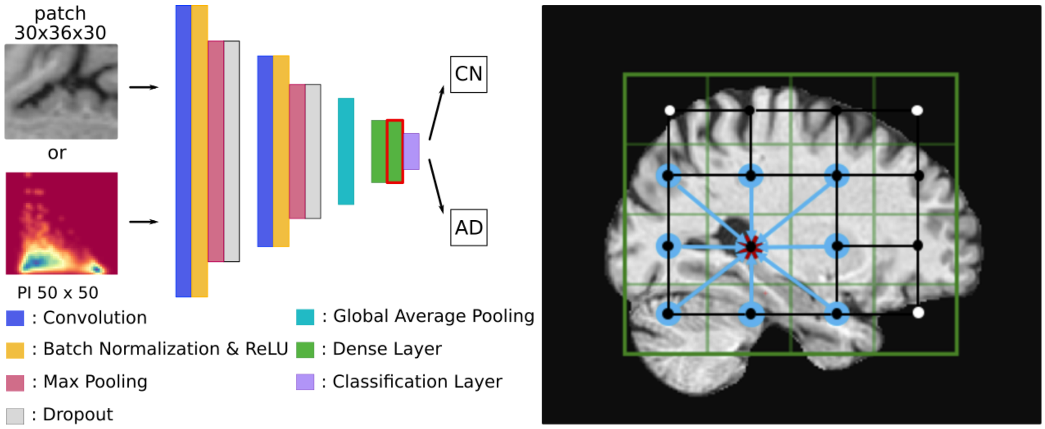

Figure 3: Overview of our model architecture using either 3D image patches or 2D persistence images (PI) as input. The preclassification layer encodings (highlighted in red) were used as node features for a GNN. In case of TDA features for all homological dimensions were included. The graph structure used to combine image patches is illustrated on the right. Here, light blue indicates neighbouring nodes to the central red node, with the relevant connections indicated.

was performed to obtain the final training and validation (used for hyperparameter tuning) sets. All longitudinal images per subject were used for training, whereas a single image, randomly selected from the longitudinal set, was selected for each validation and test patient. Depending on the split, 490–771/189-222 train, 74/74 validation, and 73–74/73–74 test images/subjects were used with an average train/validation/test prevalence of AD of $54 \pm 5\%/57 \pm 5\%/55 \pm 1\%$ for training, validation and test sets. Differences in included training number of images arise due to differences in the total number of scans acquired for each randomly selected subject.

Independent of the model architecture, training was repeated three times (runs) for each CV fold and we evaluated test set prediction accuracy, area under the receiver-operating curve (AUC), average precision score (APS), recall and precision. We averaged each metric over individual runs, then over folds, reporting mean values with standard deviations.

4.4. Interpretability analysis

For any healthcare machine learning application it is essential to evaluate the recognised image features that drive classification. The aim of interpretability analysis is to understand the evidence in the given data for the assigned class, as well as evidence that points to different classes. An understanding of the classification mechanism, which is crucial to safety- and trust-critical applications, starts with an assessment of the quality of the evidence with respect to the classification task. We use gradient-weighted class activation mapping Selvaraju et al. (2016, grad-CAM) to obtain a post-hoc evaluation of model interpretability. This method does not impose constraints on the CNN architecture and generates a class activation map that highlights the contributions of image features to the class prediction.

Grad-CAM is calculated using the n output feature maps of the final convolutional layer, which are indexed by $m \in \{1, 2, \ldots, n\}$. A feature map weight a_m^c is computed based on the average gradient of the score for class c (y^c) with respect to the activations along the x-, y-, and z-axis in the three-dimensional feature map A^m.

$$a_m^c = \frac{1}{|xyz|} \sum_x \sum_y \sum_z \frac{\partial y^c}{\partial A_{xyz}^m} \tag{1}$$

Each feature map is multiplied by its corresponding weight and summed up to obtain the full class activation map. A ReLU function is applied to the map to preserve those image features that support the classification decision.

$$\text{map}^c = \text{ReLu}\left(\sum_m a_m^c A^m\right) \tag{2}$$

The resulting class activation map is converted to a heat map, highlighting the image features most likely guiding the classification decision when changed. Grad-CAM was originally implemented in a 2D setting. Apart from small modification to allow for 3D evaluation, we use the originally-published code (Nguyen, 2020) to perform grad-CAM interpretability analysis on all of our 3D CNNs.

5. Results

5.1. Image-based classification improves with clinical domain knowledge

Table 3 provides an overview of the classification performance of models using 3D image subsets with increasing clinical domain knowledge upon selection of the input image region. We started with classification on the full resolution inner brain region I. This model had $\sim 270k$ parameters and trained at a speed of $\sim 25s$ per epoch. It obtained an accuracy, and AUC ($\text{ACC}_{I,3D} = 0.79 \pm 0.05$, $\text{AUC}_{I,3D} = 0.88 \pm 0.05$, $\text{APS}_{I,3D} = 0.91 \pm 0.05$) that were comparable to previous results with similar architectures (see Table 1). In a second step we analyzed 3D image patches. The relevant 3D CNNs comprised $\sim 70k$ trainable parameters and training time per epoch dropped to $\sim 1s$. Depending on the anatomical location of the patch, model performance varied as shown in Figure 4A. We observed maximum performance if the image patch was located in proximity to the hippocampus (best performing patch: $\text{ACC}_{P*,3D} = 0.81 \pm 0.05$, $\text{AUC}_{P*,3D} = 0.89 \pm 0.05$, $\text{APS}_{P*,3D} = 0.92 \pm 0.03$). There was no further improvement if image patch models were combined, neither using LR ($\text{ACC}_{LR,3D} = 0.81 \pm 0.04$, $\text{AUC}_{LR,3D} = 0.88 \pm 0.03$, $\text{APS}_{LR,3D} = 0.90 \pm 0.03$), nor with a GNN ($\text{ACC}_{\text{GNN},3D} = 0.79 \pm 0.03$, $\text{AUC}_{\text{GNN},3D} = 0.86 \pm 0.04$, $\text{APS}_{\text{GNN},3D} = 0.88 \pm 0.04$). We finally evaluated a 3D CNN on the left and right hippocampus separately ($\sim 140k$ trainable parameters, $\sim 2s$ per epoch). Whereas results in the right hippocampus fell in the same range as the best image patch network ($\text{ACC}_{HCr,3D} = 0.80 \pm 0.08$, $\text{AUC}_{HCr,3D} = 0.88 \pm 0.06$, $\text{APS}_{HCr,3D} = 0.91 \pm 0.04$), the model trained on the left hippocampus yielded the best results ($\text{ACC}_{HCl,3D} = 0.84 \pm 0.07$, $\text{AUC}_{HCl,3D} = 0.91 \pm 0.05$, $\text{APS}_{HCl,3D} = 0.93 \pm 0.03$).

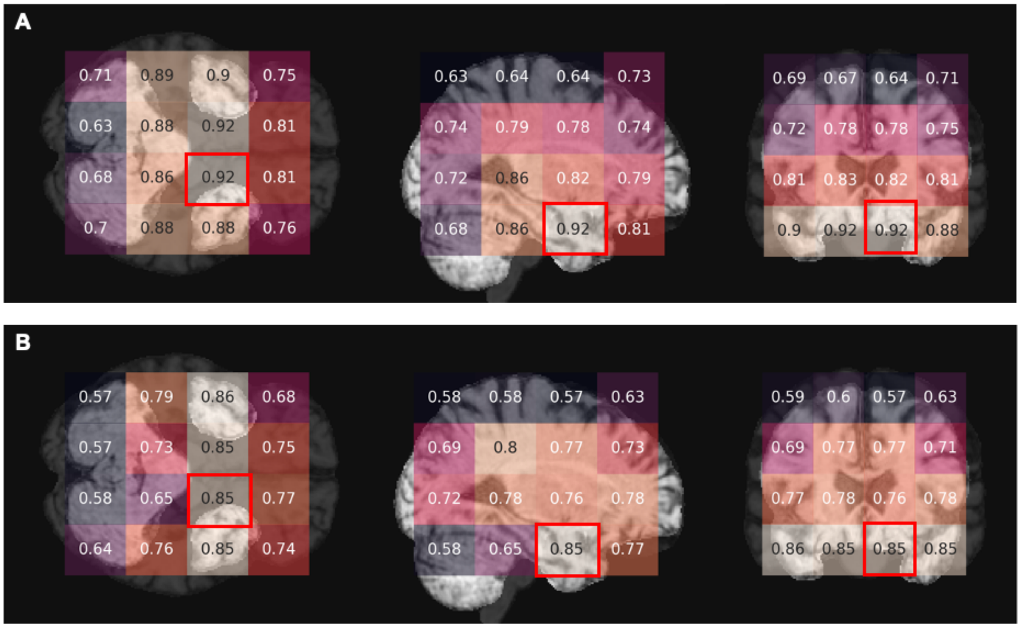

Figure 4: Classification APS of image patches either directly using 3D CNNs (A) or following conversion to persistence images (B). In the background the relevant top and bottom MR image slices of the 3D volume are shown. It can be observed that patches within the hippocampal brain region perform best. The best performing patch referenced in Table 3 and Table 4 is highlighted.

Model	I-3D	LR-3D	GNN-3D	P*-3D	HC left (3D)	HC right (3D)
ACC	0.79 ± 0.05	0.81 ± 0.04	0.82 ± 0.05	0.81 ± 0.05	$\mathbf{0.84 \pm 0.07}$	0.80 ± 0.08
AUC	0.88 ± 0.05	0.88 ± 0.03	0.90 ± 0.03	0.89 ± 0.05	$\mathbf{0.91 \pm 0.05}$	0.88 ± 0.06
APS	0.91 ± 0.05	0.90 ± 0.03	0.91 ± 0.03	0.92 ± 0.03	$\mathbf{0.93 \pm 0.03}$	0.91 ± 0.04
Recall	0.85 ± 0.07	0.85 ± 0.05	0.86 ± 0.03	0.85 ± 0.06	$\mathbf{0.87 \pm 0.08}$	0.85 ± 0.07
Precision	0.79 ± 0.05	0.82 ± 0.04	0.83 ± 0.07	0.82 ± 0.05	$\mathbf{0.84 \pm 0.06}$	0.80 ± 0.00

Table 3: Classification performance image subsets with 3D CNNs. Results are shown as test set ($n = 74$) mean values and standard deviations taken over three repeat runs and five cross validation folds. Abbreviations: 3D: three dimensional, P*: Best single patch, I: inner image subset of dimension $120 \times 144 \times 120$ voxels, HC: Hippocampus, LR: Logistic regression comprising all patches, GNN: Graph neural network comprising all patches, ACC: Accuracy, AUC: Area under the receiver operator curve, APS: Average precision score.

5.2. Connectivity information does not improve prediction

TDA provides an effective means of extracting connectivity features (and its higher analogues; homology) of data. This yields a substantial compression of the the original 3D image data to 50×50 pixel persistence images. Thus, despite comparable numbers of model parameters in TDA approaches (independent of original input image size, $\sim 50k$ trainable parameters) training was much faster than the relevant 3D CNNs (~ 10 ms per epoch) but came at the cost of loss of predictive power. To what degree the task-relevant information is preserved in our topological features can be seen in Table 4, which provides an overview of the classification performance of models using persistence images. As above, we report results for models building on different image subsets. We observed that overall TDA-based models were inferior to image-based classification. However, general trends between models using different image subset were preserved as discussed above: Classification based on the inner brain region, I, achieved an average accuracy of 71% ($\text{ACC}_{I,\text{TDA}} = 0.71 \pm 0.05$, $\text{AUC}_{I,\text{TDA}} = 0.78 \pm 0.03$, $\text{APS}_{I,\text{TDA}} = 0.79 \pm 0.03$). Within the range of uncertainties, this was comparable to the performance of the 3D image-based CNN but marginally inferior to the results obtained for the best performing image patch in the hippocampal region ($\text{ACC}_{P*,\text{TDA}} = 0.74 \pm 0.03$, $\text{AUC}_{P*,\text{TDA}} = 0.83 \pm 0.03$, $\text{APS}_{P*,\text{TDA}} = 0.85 \pm 0.03$). Figure 4B gives an overview of the APS within image patches of different brain regions. Again, patches within the medial temporal lobe areas, comprising parts of the hippocampus, performed best. LR-based combination of image patches did not improve performance ($\text{ACC}_{LR,\text{TDA}} = 0.81 \pm 0.04$, $\text{AUC}_{LR,\text{TDA}} = 0.88 \pm 0.05$, $\text{APS}_{LR,\text{TDA}} = 0.84 \pm 0.06$). However, depending on the evaluation metric, both the left hippocampus image subset alone or patch combination through a GNN were the best approach using persistence images ($\text{ACC}_{HCl,\text{TDA}} = 0.74 \pm 0.03$, $\text{AUC}_{HCl,\text{TDA}} = 0.84 \pm 0.04$, $\text{APS}_{HCl,\text{TDA}} = 0.87 \pm 0.03$ vs. $\text{ACC}_{\text{GNN,TDA}} = 0.77 \pm 0.02$, $\text{AUC}_{\text{GNN,TDA}} = 0.84 \pm 0.04$, $\text{APS}_{\text{GNN,TDA}} = 0.85 \pm 0.04$). As in the case of the 3D image-based models, the right hippocampus was less informative of AD status than the left ($\text{ACC}_{HCr,\text{TDA}} = 0.73 \pm 0.05$, $\text{AUC}_{HCr,\text{TDA}} = 0.81 \pm 0.05$, $\text{APS}_{HCr,\text{TDA}} = 0.84 \pm 0.04$).

Finally, when we combined TDA with 3D CNN models for the left HC through LR we observed no further performance improvement beyond what was achieved using the 3D CNN alone: $\text{ACC}_{HCl,\text{TDA+3DCNN}} = 0.83 \pm 0.06$, $\text{AUC}_{HCl,\text{TDA+3DCNN}} = 0.90 \pm 0.04$, $\text{APS}_{HCl,\text{TDA+3DCNN}} = 0.92 \pm 0.03$.

5.3. Interpretability analysis

In Figure 5 we show selected 2D image slices with the relevant class activation maps for different image subsets. For models considering a large fraction of the image (Figure 5A), typical brain regions affected by AD are highlighted. These include the ventricles and the hippocampus region at the base of the brain. In smaller image subsets, such as patches (Figure 5B) and the left hippocampus (Figure 5C), groves between the brain mass are emphasised indicating the presence of brain atrophy. In the case of the left hippocampus it remained open for interpretation if the highlighted low intensity image regions were associated with brain tissue atrophy or the result of a change in shape of the structure not captured by the applied atlas-based segmentation.

6. Discussion

Informed choice of image subset allows for less-complex model architecture Despite choosing a deliberately simple model architecture we have demonstrated competitive

Model	I-TDA	LR-TDA	GNN-TDA	P*-TDA	HC left-TDA	HC right-TDA
ACC	0.71 ± 0.05	0.75 ± 0.02	**0.77 ± 0.02**	0.74 ± 0.03	0.74 ± 0.04	0.73 ± 0.05
AUC	0.78 ± 0.03	0.82 ± 0.03	**0.84 ± 0.04**	0.83 ± 0.03	**0.84 ± 0.04**	0.81 ± 0.05
APS	0.79 ± 0.03	0.84 ± 0.03	0.85 ± 0.04	0.85 ± 0.03	**0.87 ± 0.03**	0.84 ± 0.04
Recall	0.77 ± 0.10	0.80 ± 0.08	0.81 ± 0.05	**0.82 ± 0.08**	0.80 ± 0.08	0.80 ± 0.09
Precision	0.73 ± 0.05	0.76 ± 0.04	**0.78 ± 0.05**	0.75 ± 0.06	0.75 ± 0.05	0.74 ± 0.05

Table 4: Classification performance of models using 2D PIs calculated from different image subsets as inputs. Results are shown as test set ($n = 74$) mean values and standard deviations taken over three repeat runs and five cross validation folds. Results represent an agglomeration of the PI dimensions 0, 1, and 2. Abbreviations: TDA: Topological data analysis, PI: Persistence image, 2D: two dimensional, CNN: Convolutional neural network, P*: Best single patch $30 \times 36 \times 30$ voxels, I: Inner image subset of dimension $120 \times 144 \times 120$ voxels, HC: Hippocampus, LR: Logistic regression comprising all patches, GNN: Graph neural network comprising all patches, ACC: Accuracy, AUC: Area under the receiver operator curve, APS: Average precision score.

classification performance for a subset of the models evaluated, particularly when restricting the image data to clinically relevant brain subregions. A model considering exclusively the left hippocampus reached a comparable performance to far more complex model architectures previously reported (Oh et al., 2019) on the same data (see Table 1). It has been suggested both by clinical researchers (Delacourte et al., 1999; Kälin et al., 2017; Mueller et al., 2010), as well as machine learning studies (Oh et al., 2019; Rieke et al., 2018), that the hippocampus is highly informative of AD status. Interestingly, we observed a performance asymmetry—for both image features, as well as topological features—between the left and right hippocampus. This observation is indeed in agreement with previous clinical insight. In general, it has been observed that the left hemisphere is more strongly affected by atrophy in AD (Laakso et al., 1995; Shi et al., 2009; Thompson et al., 2003). Recent results based on the ADNI data further indicated noticeable left-right asymmetry in total hippocampal volume in AD patients (Sarica et al., 2018). Importantly, cognitive performance of patients worsened with increasing hippocampal volume asymmetry (Sarica et al., 2018). Wachinger et al. (2016) also observed increasing deviations in the shape of the hippocampal regions for AD patients. They suggested that shape features were even more informative of disease state and progression than volume asymmetry. Here, despite using an atlas based segmentation of the hippocampus, which may not yield absolute accurate outlines, we allowed for visualisation of shape deviations through the dilation of the field of included voxels. The strong performance suggests that our approach is able to provide class predictions that align with clinical knowledge of AD, emphasising that a meaningful image representation was obtained. Interpretability analysis further pointed towards local brain atrophy, or shape deformation in the relevant volume, highlighting low image intensities. Despite predominant AD imaging biomarkers being related to the hippocampus and amygdala, it has been reported, that other structural brain changes and global loss of brain tissue connectivity are further hallmarks of AD. Such features were available for the patch-based and inner brain image-based classification tasks. Our interpretability analysis showed that, in addition to

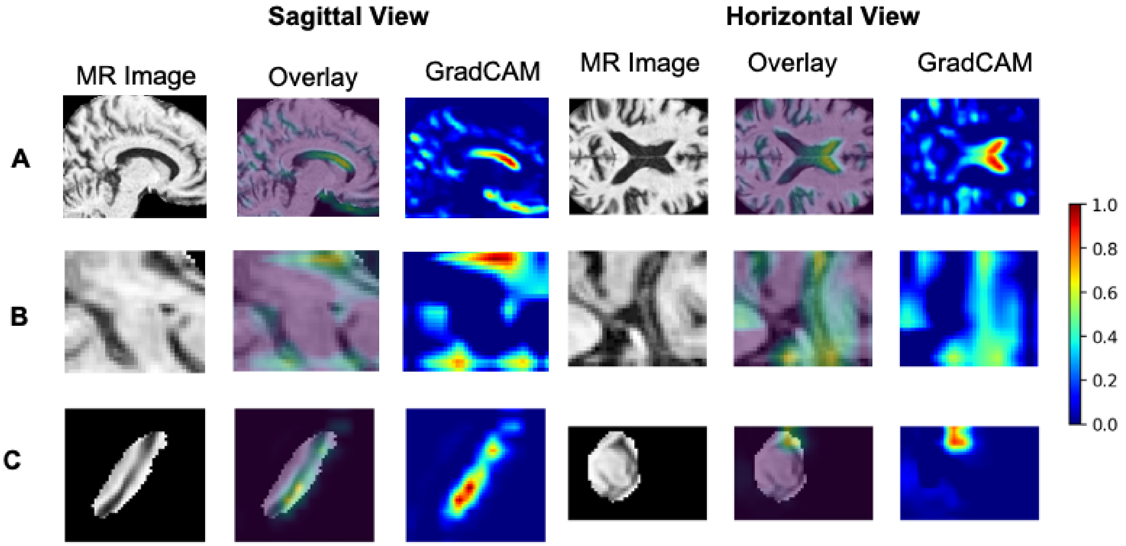

Figure 5: Grad-CAM interpretability analysis of the 3D CNNs at different scales. A selected 2D slice of the 3D MR image volume is shown together with the relevant class activation maps for a correctly classified AD patient with respect to the AD class. A) Inner brain image subset. B) Best performing patch in the left hippocampus region. C) Left hippocampus. For visualisation purposes, images are not drawn to scale, but preserve the original dimensions. The colour-bar indicates the normalised gradient-weighted class activation.

the hippocampus, the ventricular volume drove image classification. However, this did not translate to a performance benefit. Our ablation study showed, that classification based on a large image subset was comparable to previous work with similar models (Oh et al., 2019), demonstrating the generalizable improvement of classification performance by focusing on smaller image subsets containing brain structures affected by AD. The discrepancy in classification performance on the inner brain image and the hippocampus alone may be explained by the more favourable ratio of model parameters to image features in the latter case, as well as an effective preconditioning of the model, i.e. focusing on the most informative image subset. This may overrule potential other imaging features present outside the hippocampus. Interestingly, a patch-based ensemble model based on either LR or a GNN was unable to reproduce the classification results obtained based on the hippocampus data. This may be due to potentially confounding features present in other brain regions.

Model interpretability analysis Previous work mostly focused on high level interpretability analysis of the full MRI to reveal brain subregions, as the most discriminative image features between AD and CN subjects. Here, we go a step further by providing interpretability analysis at both the global and local scale of imaging features. We first demonstrated the importance of the hippocampus for AD classification through classification performance analysis of different image patches. In a second step, by means of grad-CAM, we identified that the extent of hippocampal brain tissue atrophy, or deformation of this structure, was driving classification at the local scale. This is in line with the biological hallmarks of AD in terms of progressive brain atrophy and implies that our methodology was successful in providing image classification based on clinically relevant features.

Topological data analysis did not further improve performance There was a strong motivation to investigate the use of TDA for the classification of AD from MRI data as AD manifests through a global loss of brain tissue connectivity. It was expected to provide an ideal setting for TDA as a tool capturing overall structural connectivity. At the same time, TDA may be regarded as a means of feature extraction, allowing for fast model training (∼10 ms per epoch) with minimal hardware requirements. We investigated whether topological features, represented as persistence images, hold sufficient information to allow for robust classification of AD from structural MRIs. Despite our best efforts, we were unable to obtain results competitive to our 3D CNN models using the TDA approach neither at the patch, hippocampus, or inner brain image level. However, it is essential to note that the classification performance of our TDA GNN model combining 64 image patches yielded accuracy and AUC levels that were in the range of previous, simple model architectures (see Table 1) making it suitable as a computationally efficient baseline model. Despite performing a substantial compression of 3D image data, topological representations are capable of encoding pertinent information for the anticipated classification task. These observations support our original hypothesis that topological features are relevant and can be leveraged for AD prediction to some extent. However, due to the intensity-based evaluation approach, image noise may impact on the persistence and number of homological features (connected components) detected. Previous work (Brüningk et al., 2020) investigated the use of image preprocessing techniques, such as smoothing, to mitigate this challenge but observed no significant boost in performance. This was likely due to the loss of small, local image features in the presence of smoothing. We consider our investigation a first step in the direction of using TDA in connection with structural MRI data analysis, but suggest that there remains room for exploration of optimised topological feature extraction and representation.

Limitations The diagnosis of AD per se is not an open clinical question given that other indication, such as cognitive performance or motor skills, are readily informative of AD vs. CN. The presented analysis should be understood as the foundation towards evaluation of more sophisticated clinical questions in the realm of AD research, including the prediction of the transition time of mild cognitively impaired to AD patients.

A further limitation of our approach's focus on specific image subregions (i.e. the left hippocampus) might arise due to the presence of heterogeneity in AD pathology (Ferreira et al., 2020; Habes et al., 2020). While a majority of AD diagnoses can be assigned to subtypes affecting predominantly the hippocampus (typical AD, limbic-predominant AD; pooled frequency 76% of cases (Ferreira et al., 2020)), two further subtypes of AD have been identified. One of these is the hippocampal sparing subtype (15% of cases (Ferreira et al., 2020)) which would be expected to be difficult to detect using our proposed architecture and may account for the failure to correctly stratify AD and CN subjects in all cases. Interestingly, the hippocampal sparing AD subtype shows an earlier age of onset and occurs more frequently in males. Using the augmentation of a 3D CNN with an age encoding, as successfully demonstrated in Liu et al. (2020), could potentially improve our approach's performance in these rarer cases. The fourth subtype described in the literature is characterised by minimal atrophy. It is unclear how restricting the input to particular brain regions would affect a classifier's ability to correctly predict a diagnostic label for patients affected by this subtype.

Finally, we took first steps into the investigation of topological analysis of structural MRI images and their relevance in the context of AD prediction. Despite our best efforts, there remains room for improvement in classification performance based on topological features. In addition to the limited classification performance, persistence images are inherently less intuitive to interpret than imaging information. A cornerstone of any medical ML application is to build trust in the suggested predictions—interpretability and explainability analysis is key towards this goal. This makes it essential to address the interpretation of topological differences for future application of this technique in the realm of medical data. Here we provide a solution that remains interpretable at the global level through the means of ensemble pooling of individual image patches. However, it remains unclear which anatomical image features drive classification performance. Such findings would not only be informative in terms of interpretability of the model but also yield a deeper insight into how (the progression of) AD affects brain connectivity.

Summary and outlook In conclusion, we demonstrated that a well-informed choice of the image subregion, derived from clinical domain knowledge, may overrule the need for complex model architectures and ensemble models. Despite a strong clinical and conceptual motivation for the use of topological features for AD classification, models based on connectivity features were outperformed by image-based 3D CNNs on the same data. Including clinical priors in 3D image-based models facilitates the use of comparably simple architectures. Our model was post-hoc interpretable, fast and computationally less expensive than previously reported architectures while maintaining competitive performance. We plan to investigate this approach further for interpretablity-aware model training (Boopathy et al., 2020) that requires a computationally efficient architecture. The grayscale intensity distributions of MR images are subject to the characteristics of the specific machine and the operating mode Lee et al. (2020). This implies that, without additional intensity correction, for example based on phantom scans, these differences in intensity distribution may significantly affect feature extraction and algorithm classification performance. Interpretability-aware models could improve the generalizability of the model to be applicable to different clinics and scanners. Interpretability-aware training is suggested to yield more robust predictions, which could improve model generalizability to out-of-domain imaging data acquired at multiple institutions and with different MR machines. While generalizability may also be improved using techniques from domain adaptation, we consider interpretability to be a crucial aspect for methods in this area. Moreover, we plan to investigate AD subtypes with the presented architecture to evaluate the importance of other functional subunits of the brain for AD subtyping. Such information is expected to impact on patient progression profiles and could truly advance the clinical understanding and treatment of AD.

We suggest that the proposed method of focusing analysis to clinically relevant image subsets translates also to other domain-intensive approaches. A hypothetical examples could be the treatment response prediction of brain tumours. Here, it is well known that tumour perfusion as well as anatomical location correlate with outcome. By including this clinical domain knowledge, e.g. through performing image classification based on only the brain region containing the tumour in addition to features encoding the anatomical location of the selected subset, may be superior to simply performing classification on the full brain image. This may be particularly important given a limited sample size and heterogeneity in tumours. A second example could be the evaluation of cognitive performance of former victims of traumatic brain injuries. Rather than including the full brain imaging data, specific brain subregions, such as the hippocampus, could be considered on their own.

Acknowledgments

This work was partially funded and supported by the Swiss National Science Foundation (Spark grants 190466 (F.H., B.R.) and 190647 (S.C.B., C.R.J); Ambizione Grant #PZ00P3186101, C.R.J). The content provided here is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. The funders had no role in study design, data collection & analysis, decision to publish, or preparation of the manuscript. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.;Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.;Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: a stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18, 2017. ISSN 1532-4435.

Karim Aderghal, Jenny Benois-Pineau, Karim Afdel, and Catheline Gwenaëlle. FuseMe: Classification of SMRI Images by Fusion of Deep CNNs in 2D+ϵ Projections. In *Pro-*

ceedings of the 15th International Workshop on Content-Based Multimedia Indexing, CBMI '17, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450353335. doi: 10.1145/3095713.3095749.

Serguei A. Barannikov. The framed Morse complex and its invariants. *Advances in Soviet Mathematics*, 21:93–115, 1994. doi: 10.1090/ADVSOV/021/03.

Josephine Barnes, Jonathan W. Bartlett, Laura A. van de Pol, Clement T. Loy, Rachael I. Scahill, Chris Frost, Paul Thompson, and Nick C. Fox. A meta-analysis of hippocampal atrophy rates in Alzheimer's disease. *Neurobiology of Aging*, 30(11):1711–1723, 2009. ISSN 0197-4580. doi: 10.1016/j.neurobiolaging.2008.01.010.

Akhilan Boopathy, Sijia Liu, Gaoyuan Zhang, Cynthia Liu, Pin-Yu Chen, Shiyu Chang, and Luca Daniel. Proper Network Interpretability Helps Adversarial Robustness in Classification, 2020.

Esther E. Bron, Marion Smits, Wiesje M. van der Flier, Hugo Vrenken, Frederik Barkhof, Philip Scheltens, Janne M. Papma, Rebecca M.E. Steketee, Carolina Méndez Orellana, Rozanna Meijboom, Madalena Pinto, Joana R. Meireles, Carolina Garrett, António J. Bastos-Leite, Ahmed Abdulkadir, Olaf Ronneberger, Nicola Amoroso, Roberto Bellotti, David Cárdenas-Peña, Andrés M. Álvarez Meza, Chester V. Dolph, Khan M. Iftekharuddin, Simon F. Eskildsen, Pierrick Coupé, Vladimir S. Fonov, Katja Franke, Christian Gaser, Christian Ledig, Ricardo Guerrero, Tong Tong, Katherine R. Gray, Elaheh Moradi, Jussi Tohka, Alexandre Routier, Stanley Durrleman, Alessia Sarica, Giuseppe Di Fatta, Francesco Sensi, Andrea Chincarini, Garry M. Smith, Zhivko V. Stoyanov, Lauge Sørensen, Mads Nielsen, Sabina Tangaro, Paolo Inglese, Christian Wachinger, Martin Reuter, John C. van Swieten, Wiro J. Niessen, and Stefan Klein. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge. *NeuroImage*, 111:562–579, 2015. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2015.01.048.

Sarah C. Brüningk, Felix Hensel, Catherine R. Jutzeler, and Bastian Rieck. Image analysis for Alzheimer's disease prediction: Embracing pathological hallmarks for model architecture design, 2020.

Karl Bäckström, Mahmood Nazari, Irene Yu-Hua Gu, and Asgeir Store Jakola. An efficient 3d deep convolutional network for alzheimer's disease diagnosis using mr images. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 149–153, 2018. doi: 10.1109/ISBI.2018.8363543.

Mathieu Carrière, Frederic Chazal, Yuichi Ike, Theo Lacombe, Martin Royer, and Yuhei Umeda. PersLay: A neural network layer for persistence diagrams and new graph topological signatures. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2786–2796. PMLR, 26–28 Aug 2020.

A. Delacourte, J.P. David, N. Sergeant, L. Buée, A. Wattez, P. Vermersch, F. Ghozali, C. Fallet-Bianco, F. Pasquier, F. Lebert, H. Petit, and C. Di Menza. The biochemical pathway of neurofibrillary degeneration in aging and Alzheimer's disease. *Neurology*, 52 (6):1158–1158, 1999. ISSN 0028-3878. doi: 10.1212/WNL.52.6.1158.

Herbert Edelsbrunner and John L. Harer. *Computational topology*. American Mathematical Society, Providence, RI, 2010. ISBN 978-0-8218-4925-5. doi: 10.1090/mbk/069. An introduction.

Kathryn Ellis, Ashley Bush, David Darby, Daniel De Fazio, Jonathan Foster, Peter Hudson, Nicola Lautenschlager, Nat Lenzo, Ralph Martins, Paul Maruff, et al. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *International Psychogeriatrics*, 21(4):672–687, 2009. doi: 10.1017/S1041610209009405.

Daniel Ferreira, Agneta Nordberg, and Eric Westman. Biological subtypes of Alzheimer disease: A systematic review and meta-analysis. *Neurology*, 94(10):436–448, 2020. doi: 10.1212/WNL.0000000000009058.

Mohamad Habes, Michel J. Grothe, Birkan Tunc, Corey McMillan, David A. Wolk, and Christos Davatzikos. Disentangling Heterogeneity in Alzheimer's Disease and Related Dementias Using Data-Driven Methods. *Biological Psychiatry*, 88(1):70–82, 2020. ISSN 0006-3223. doi: https://doi.org/10.1016/j.biopsych.2020.01.016. URL https://www.sciencedirect.com/science/article/pii/S0006322320300500. Convergence and Heterogeneity in Psychopathology.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. doi: 10.1109/cvpr.2016.90.

Felix Hensel, Michael Moor, and Bastian Rieck. A survey of topological machine learning methods. *Frontiers in Artificial Intelligence*, 4, 2021. ISSN 2624-8212. doi: 10.3389/frai.2021.681108.

Christoph D. Hofer, Roland Kwitt, Marc Niethammer, and Andreas Uhl. Deep Learning with Topological Signatures. *CoRR*, abs/1707.04041, 2017. URL http://arxiv.org/abs/1707.04041.

Marcia Hon and Naimul Mefraz Khan. Towards Alzheimer's disease classification through transfer learning. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1166–1169, 2017. doi: 10.1109/BIBM.2017.8217822.

Ehsan Hosseini-Asl, Mohammed Ghazal, Ali Mahmoud, Ali Aslantas, Ahmed Shalaby, Manual Casanova, Gregory Barnes, Georgy Gimel'farb, Robert Keynton, and Ayman El-Baz. Alzheimer's disease diagnostics by a 3D deeply supervised adaptable convolutional network. *Frontiers in Bioscience-Landmark*, 23:584–596, 01 2018. doi: 10.2741/4606.

Dan Jin, Jian Xu, Kun Zhao, Fangzhou Hu, Zhengyi Yang, Bing Liu, Tianzi Jiang, and Yong Liu. Attention-based 3D Convolutional Network for Alzheimer's Disease Diagnosis and Biomarkers Exploration. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1047–1051. IEEE, 2019. ISBN 978-1-5386-3641-1. doi: 10.1109/ISBI.2019.8759455.

Sergey Korolev, Amir Safiullin, Mikhail Belyaev, and Yulia Dodonova. Residual and Plain Convolutional Neural Networks for 3D Brain MRI Classification. *CoRR*, abs/1701.06643, 2017. URL http://arxiv.org/abs/1701.06643.

Andrea M. Kälin, Min T. M. Park, M. Mallar Chakravarty, Jason P. Lerch, Lars Michels, Clemens Schroeder, Sarah D. Broicher, Spyros Kollias, Roger M. Nitsch, Anton F. Gietl, Paul G. Unschuld, Christoph Hock, and Sandra E. Leh. Subcortical shape changes, hippocampal atrophy and cortical thinning in future Alzheimer's disease patients. *Frontiers in Aging Neuroscience*, 9:38, 2017. ISSN 1663-4365. doi: 10.3389/fnagi.2017.00038.

MP Laakso, H Soininen, K Partanen, E-L Helkala, P Hartikainen, P Vainio, M Hallikainen, T Hänninen, and PJ Riekkinen Sr. Volumes of hippocampus, amygdala and frontal lobes in the MRI-based diagnosis of early Alzheimer's disease: correlation with memory functions. *Journal of neural transmission-Parkinson's disease and dementia section*, 9 (1):73–86, 1995. doi: 10.1007/BF02252964.

Dong-Ho Lee, Yan Li, and Byeong-Seok Shin. Generalization of intensity distribution of medical images using GANs. *Human-centric Computing and Information Sciences*, 10 (1):17, dec 2020. ISSN 2192-1962. doi: 10.1186/s13673-020-00220-2.

Sheng Liu, Chhavi Yadav, Carlos Fernandez-Granda, and Narges Razavian. On the design of convolutional neural networks for automatic detection of Alzheimer's disease. In Adrian V. Dalca, Matthew B.A. McDermott, Emily Alsentzer, Samuel G. Finlayson, Michael Oberst, Fabian Falck, and Brett Beaulieu-Jones, editors, *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 116 of *Proceedings of Machine Learning Research*, pages 184–201. PMLR, 13 Dec 2020. URL http://proceedings.mlr.press/v116/liu20a.html.

Ana Manera, Mahsa Dadar, Vladimir Fonov, and Louis Collins. CerebrA, registration and manual label correction of Mindboggle-101 atlas for MNI-ICBM152 template. *Scientific Data*, 7(237), 2020. doi: https://doi.org/10.1038/s41597-020-0557-9.

Daniel S Marcus, Anthony F Fotenos, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *Journal of Cognitive Neuroscience*, 22(12):2677–2684, 2010. doi: https://doi.org/10.1162/jocn.2009.21407.

Susanne G. Mueller, Norbert Schuff, Kristine Yaffe, Catherine Madison, Bruce Miller, and Michael W. Weiner. Hippocampal atrophy patterns in mild cognitive impairment and Alzheimer's disease. *Human Brain Mapping*, 31(9):1339–1347, 2010. doi: https://doi.org/10.1002/hbm.20934.

Hoa Nguyen. Github repository by nguyenhoa93. GradCam & Guided GradCAM. https://github.com/nguyenhoa93/GradCAM_and_GuidedGradCAM_tf2, 2020. Accessed: 2021-03-19.

Kanghan Oh, Young Chul Chung, Ko Woon Kim, Woo Sung Kim, and Il Seok Oh. Classification and Visualization of Alzheimer's Disease using Volumetric Convolutional Neural Network and Transfer Learning. *Scientific Reports*, 9(1):1–16, 2019. ISSN 20452322. doi: 10.1038/s41598-019-54548-6.

Dan Pan, An Zeng, Longfei Jia, Yin Huang, Tory Frizzell, and Xiaowei Song. Early Detection of Alzheimer's Disease Using Magnetic Resonance Imaging: A Novel Approach Combining Convolutional Neural Networks and Ensemble Learning. *Frontiers in Neuroscience*, 14:259, 2020. ISSN 1662-453X. doi: 10.3389/fnins.2020.00259.

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*, 12:2825–2830, 2011.

Saima Rathore, Mohamad Habes, Muhammad Aksam Iftikhar, Amanda Shacklett, and Christos Davatzikos. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *Neuroimage*, 155:530–548, 2017. doi: 10.1016/j.neuroimage.2017.03.057.

Bastian Rieck, Tristan Yates, Christian Bock, Karsten Borgwardt, Guy Wolf, Nick Turk-Browne, and Smita Krishnaswamy. Uncovering the Topology of Time-Varying fMRI Data using Cubical Persistence. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6900–6912. Curran Associates, Inc., 2020.

Johannes Rieke, Fabian Eitel, Martin Weygandt, John-Dylan Haynes, and Kerstin Ritter. Visualizing Convolutional Networks for MRI-based Diagnosis of Alzheimer's Disease. *CoRR*, abs/1808.02874, 2018. URL http://arxiv.org/abs/1808.02874.

Margaret J Rosenbloom and Adolf Pfefferbaum. Magnetic resonance imaging of the living brain: evidence for brain degeneration among alcoholics and recovery with abstinence. *Alcohol Research & Health*, 31(4):362–376, 2008.

Alessia Sarica, Roberta Vasta, Fabiana Novellino, Maria Grazia Vaccaro, Antonio Cerasa, Aldo Quattrone, and The Alzheimer's Disease Neuroimaging Initiative . MRI asymmetry index of hippocampal subfields increases through the continuum from the mild cognitive impairment to the Alzheimer's disease. *Frontiers in Neuroscience*, 12:576, 2018. ISSN 1662-453X. doi: 10.3389/fnins.2018.00576.

Stephen W Scheff, Douglas A Price, Frederick A Schmitt, and Elliott J Mufson. Hippocampal synaptic loss in early Alzheimer's disease and mild cognitive impairment. *Neurobiology of aging*, 27(10):1372–1384, 2006. doi: 10.1016/j.neurobiolaging.2005.09.012.

Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization. *CoRR*, abs/1610.02391, 2016. doi: 10.1007/s11263-019-01228-7.

Alberto Serrano-Pozo, Matthew P Frosch, Eliezer Masliah, and Bradley T Hyman. Neuropathological alterations in Alzheimer disease. *Cold Spring Harbor perspectives in medicine*, 1(1):a006189, 2011. doi: 10.1101/cshperspect.a006189.

Feng Shi, Bing Liu, Yuan Zhou, Chunshui Yu, and Tianzi Jiang. Hippocampal volume and asymmetry in mild cognitive impairment and Alzheimer's disease: Meta-analyses of mri studies. *Hippocampus*, 19(11):1055–1064, 2009. doi: https://doi.org/10.1002/hipo.20573.

Claudia K Suemoto, Renata EL Ferretti-Rebustini, Roberta D Rodriguez, Renata EP Leite, Luciana Soterio, Sonia MD Brucki, Raphael R Spera, Tarcila M Cippiciani, Jose M Farfel, Alexandre Chiavegatto Filho, et al. Neuropathological diagnoses and clinical correlates in older adults in Brazil: A cross-sectional study. *PLoS Medicine*, 14(3):e1002267, 2017. doi: 10.1371/journal.pmed.1002267.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. doi: 10.1109/CVPR.2015.7298594.

Paul M Thompson, Kiralee M Hayashi, Greig De Zubicaray, Andrew L Janke, Stephen E Rose, James Semple, David Herman, Michael S Hong, Stephanie S Dittmer, David M Doddrell, et al. Dynamics of gray matter loss in Alzheimer's disease. *Journal of neuroscience*, 23(3):994–1005, 2003. doi: 10.1523/JNEUROSCI.23-03-00994.2003.

Aly Valliani and Ameet Soni. Deep Residual Nets for Improved Alzheimer's Diagnosis. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics*, ACM-BCB '17, page 615, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450347228. doi: 10.1145/3107411.3108224.

Christian Wachinger, David H Salat, Michael Weiner, Martin Reuter, and Alzheimer's Disease Neuroimaging Initiative. Whole-brain analysis reveals increased neuroanatomical asymmetries in dementia for hippocampus and amygdala. *Brain*, 139(12):3253–3266, 2016. doi: 10.1093/brain/aww243.

Junhao Wen, Elina Thibeau-Sutre, Mauricio Diaz-Melo, Jorge Samper-González, Alexandre Routier, Simona Bottani, Didier Dormont, Stanley Durrleman, Ninon Burgos, Olivier Colliot, et al. Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis*, 63, 2020. doi: https://doi.org/10.1016/j.media.2020.101694.

Bengt Winblad, Philippe Amouyel, Sandrine Andrieu, Clive Ballard, Carol Brayne, Henry Brodaty, Angel Cedazo-Minguez, Bruno Dubois, David Edvardsson, Howard Feldman, et al. Defeating Alzheimer's disease and other dementias: a priority for European science and society. *The Lancet Neurology*, 15(5):455–532, 2016. doi: 10.1016/S1474-4422(16)00062-4.

Winston Wong. Economic burden of Alzheimer disease and managed care considerations. *The American journal of managed care*, 26(8 Suppl):S177—S183, August 2020. ISSN 1088-0224. doi: 10.37765/ajmc.2020.88482.