Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

ACCEPTED MANUSCRIPT

Development of a deep learning network for Alzheimer's disease classification with evaluation of imaging modality and longitudinal data

To cite this article before publication: Alison Deatsch *et al* 2022 *Phys. Med. Biol.* in press https://doi.org/10.1088/1361-6560/ac8f10

Development of a Deep Learning Network for Alzheimer's Disease Classification with Evaluation of Imaging Modality and Longitudinal Data

***Alison Deatsch*[1+], Matej Perovnik[2], Mauro Namías[3], Maja Trošt[2], Robert Jeraj[1,4]**
for the Alzheimer's Disease Neuroimaging Initiative*

1 University of Wisconsin–Madison; 1111 Highland Ave, Madison, WI, USA 53705
2 University Medical Centre Ljubljana; Zaloška cesta 2, 1000 Ljubljana, Slovenia
3 Fundación Centro Diagnóstico Nuclear; Av Nazca 3449, Buenos Aires, Argentina C1417CVE
4 University of Ljubljana; Jadranska ulica 19, 1000 Ljubljana, Slovenia
*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
+ Corresponding author; email address: deatsch@wisc.edu

Abstract

Objective: Neuroimaging uncovers important information about disease in the brain. Yet in Alzheimer's disease (AD), there remains a clear clinical need for reliable tools to extract diagnoses from neuroimages. Significant work has been done to develop deep learning (DL) networks using neuroimaging for AD diagnosis. However, no particular model has emerged as optimal. Due to a lack of direct comparisons and evaluations on independent data, there is no consensus on which modality is best for diagnostic models or whether longitudinal information enhances performance. The purpose of this work was (1) to develop a generalizable DL model to distinguish neuroimaging scans of AD patients from controls and (2) to evaluate the influence of imaging modality and longitudinal data on performance.

Approach: We trained a 2-class convolutional neural network (CNN) with and without a cascaded recurrent neural network (RNN). We used datasets of 772 (N_{AD}=364, $N_{control}$=408) 3D ^{18}F-FDG PET scans and 780 (N_{AD}=280, $N_{control}$=500) T1-weighted volumetric-3D MR images (containing 131 and 144 patients with multiple timepoints) from the Alzheimer's Disease Neuroimaging Initiative (ADNI), plus an independent set of 104 (N_{AD}=63, N_{NC}=41) ^{18}F-FDG PET scans (one per patient) for validation.

Main Results: ROC analysis showed that PET-trained models outperformed MRI-trained, achieving maximum AUC with the CNN+RNN model of 0.93 ± 0.08, with accuracy 82.5±8.9%. Adding longitudinal information offered significant improvement to performance on ^{18}F-FDG PET, but not on T1-MRI. CNN model validation with an independent ^{18}F-FDG PET dataset achieved AUC of 0.99. Layer-wise Relevance Propagation heatmaps added CNN interpretability.

Significance: The development of a high-performing tool for AD diagnosis, with the direct evaluation of key influences, reveals the advantage of using ^{18}F-FDG PET and longitudinal data over MRI and single timepoint analysis. This has significant implications for the potential of

neuroimaging for future research on AD diagnosis and clinical management of suspected AD patients.

Keywords:

Alzheimer's disease, deep learning, ^{18}F-FDG PET, MRI, longitudinal, neuroimaging

Funding:

This work was supported by University of Wisconsin Carbone Cancer Center Support Grant P30 CA014520.

1. Introduction

Alzheimer's disease (AD) is the sixth leading cause of death in the United States, affecting an estimated 5.5 million people age 65 and older (Association, 2018; Oldan et al., 2021). Correct diagnosis of AD is essential for proper care of patients and for the development of interventions and therapies. Models estimate that early and accurate AD diagnosis could save up to $7.9 trillion in medical and care costs (Association, 2018). At present, as many as 12% to 23% of AD diagnoses are not confirmed at autopsy (Świetlik & Białowąs, 2019). Diagnosing AD at an early stage is hampered by the variability of the clinical symptoms and the subtleties of early brain changes, resulting in low sensitivity (71%-87%) and specificity (44%-71%) (Beach et al., 2012; Rathore et al., 2017; Świetlik & Białowąs, 2019). Current methods for early diagnosis are improving, but are often costly (e.g. amyloid PET scans) and invasive (lumbar puncture for cerebrospinal fluid (CSF) analysis). Despite a clear clinical need and significant efforts, there is currently a lack of a reliable, generalizable, affordable, and non-invasive tool for the diagnosis of AD.

Neuroimaging can reveal important information for AD diagnosis, but there is no consensus on which imaging modality is best. Both structural MRI and ^{18}F-FDG PET are widely accessible, non-invasive, and less expensive than more complex imaging. In neuroimaging-based classification studies of AD, structural MRI is the most frequently used, is included in most standard of care for AD, and has widespread availability (Ding et al., 2019; Martí-Juan et al., 2020; Rathore et al., 2017). However, ^{18}F-FDG PET is increasingly popular as it illustrates metabolic changes related to AD occurring prior to the onset of structural changes, allowing for earlier diagnosis relative to MRI and clinical symptoms (Femminella et al., 2018; Reiman & Jagust, 2012; Smailagic et al., 2015). MRI is still recommended by clinical guidelines as the first choice for initial AD diagnosis (Moonis et al., 2020), yet in some cases ^{18}F-FDG PET may be more appropriate. Other more complex imaging modalities such as amyloid PET are also widely used. While a negative amyloid scan rules out AD, the presence of amyloid alone does not confirm AD. In addition, ^{18}F-FDG PET correlates better with patients' cognitive performance (Khosravi et al., 2019) and reveals more about the stage of the neurodegeneration. Amyloid and other complex imaging modalities are also much more expensive and not available to everyone. Thus, ^{18}F-FDG PET and MRI remain the most commonly used neuroimaging modalities in diagnosis/follow up of AD.

In general, the sensitivity of AD diagnosis by visual inspection of images is found to be slightly better (7-11%) for ^{18}F-FDG PET over MR (Bloudek et al., 2011; Femminella et al., 2018; Johnson et al., 2013). However, a 2017 review across various methods of predicting AD found discordant results regarding which imaging modality is superior (A Sanchez-Catasus et al., 2017).

Significant work has been done to develop deep learning (DL) neural networks with neuroimaging inputs for AD diagnosis and staging. Convolutional neural networks (CNNs) have been frequently employed, often with very promising classification accuracy (Billones et al., 2016; Choi et al., 2019; Choi et al., 2018; Ding et al., 2019; Jo et al., 2019; Kazemi & Houghten, 2018; Spasov et al., 2019; Wang et al., 2019; Zhang et al., 2019; Świetlik & Białowąs, 2019). Deep neural networks with feature extraction(Lu et al., 2018), visual-based classifiers (Wood et al., 2019), and recurrent neural networks (RNNs) (Cui et al., 2019; Liu et al., 2018) have also been explored.

Despite a wealth of attempts, no particular model has yet emerged as the optimal diagnostic method. The performance of various DL approaches is difficult to compare objectively due to the large number of factors affecting model performance. It is still not well understood which features or inputs make one model more advantageous than another. In particular, neuroimaging modalities are rarely isolated and compared directly as inputs to the same DL model. Thus their influence on model performance is not well understood. In machine learning classification specifically, ^{18}F-FDG PET has been shown to outperform MR in a few cases, though most state-of-the-art DL models do not directly compare their performance. Those that have objectively compared these modalities are mainly simpler ML classifiers such as Support vector machines (SVM) (Dukart et al., 2011; Samper-González et al., 2018), networks that require feature extraction (Lu et al., 2018), and recently a straightforward CNN (Huang et al., 2019). The influence of imaging modality on complex DL model performance remains an active and underexplored factor in the use of DL for AD diagnosis and is the focus of this work.

Another important feature to consider in order to determine the advantages of a particular model is the inclusion of longitudinal information. Most current studies in AD diagnosis, and particularly DL models, are limited to single timepoint images (Martí-Juan et al., 2020). However, the incorporation of longitudinal data has been shown to improve classification performance in several statistical analyses and classical image analysis methods (Gray et al., 2012; Rodrigues & Silveira, 2014; Sun et al., 2017; Zhang et al., 2012; Zhang et al., 2019). In the clinic, it has been shown that repeating an FDG PET scan can greatly clarify equivocal diagnoses and improve disease management. For example, in a retrospective study, Bergeron et al. demonstrated that conducting a second FDG PET scan reduced the number of unclear diagnoses from 80% to 34%, and led to diagnostic change in 24% of cases and treatment modification in 22% of patients (Bergeron et al., 2016). Recent DL analyses have demonstrated success using RNNs to incorporate longitudinal data from neuroimages using varying levels of image pre-processing, from limited feature extraction (Lee et al., 2019) to CNN feature maps (Cui et al., 2019; Gao et al., 2018).

An additional important consideration for DL models, especially for clinical translation, is generalizability, or the ability of the model to perform well on independent, unseen patient data (Rathore et al., 2017). Despite a growing consensus in the field of DL for medical imaging that models should include external validation from an independent dataset wherever possible, models with state-of-the-art performance are rarely validated in such a manner to demonstrate their generalizability.

The purpose of this work was to develop and evaluate a novel deep learning model with and without longitudinal imaging data to distinguish patients with AD from normal controls based on their metabolic and structural neuroimaging scans. For this purpose, we developed a model consisting of a convolutional neural network (CNN) run either with or without a cascaded recurrent neural network (RNN) for binary classification. Our primary objective was to investigate the influence of imaging modality on model performance. Our secondary objectives were (1) to explore the impact of adding longitudinal data to the model and (2) to determine how well the model generalizes to new, external data.

2. Data

2.1 ADNI Dataset

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. For up-to-date information, see www.adni-info.org. Data was obtained from the ADNI database in November of 2019 (^{18}F-FDG PET dataset) and October of 2020 (MRI dataset).

In this work, "patients with AD" were considered to be all those with a clinical AD diagnosis as long as no CSF biomarker evidence contradicted this status. Specifically, all patients with CSF biomarkers reported in ADNI were evaluated to confirm biological AD pathology. The CSF biomarkers were evaluated according to the NIA-AA research framework and classified into categories according to their amyloid (A), tau (T) and neurodegeneration (N) status, in a so called AT(N) classification (Jack Jr et al., 2018). Cutoff values of 980 pg/ml for Aβ, 21.6 pg/mL for p-tau, and 0.077 for $A\beta_{42/40}$ were used. All patients with contradictory clinical and AT(N) diagnoses were excluded from the dataset. For instance, any patient with a clinical diagnosis of AD but belonging to an AT(N) biomarker category of normal or non-AD pathology were excluded, i.e. A-/T±/N±. Likewise, any patient with a clinical diagnosis of normal control (NC) belonging to an AT(N) biomarker category in the AD continuum, i.e. A+/T±/N±, was also excluded. While A+ is common in cognitively normal elderly subjects, these patients were excluded for two reasons. One, studies have shown that even cognitively unimpaired subjects with positive amyloid biomarkers are at greater risk of subsequent development of cognitive impairment (Ebenau et al., 2020). Two, studies have also shown that hypometabolic changes can be observed prior to the appearance of the first symptoms (Gordon et al., 2018). Therefore, to exclude the possibility that presence of amyloid in the brain of clinically healthy subject

would cause changes in brain metabolism, normal control subjects with brain amyloidosis were not included in our reference group.

A total of 77 scans were excluded by this process. Of the AD diagnosed patients in our datasets, 190 ^{18}F-FDG PET images and 266 T1-MR images did not have CSF measurements at or within 4 years of the time of the scan (Mattsson et al., 2012). In accordance with most standard practice for the ADNI dataset, these patients were still included with the assumption of correct clinical diagnosis in order to maintain a large enough dataset for the DL model.

^{18}F-FDG PET scans were downloaded from ADNI with the minimum pre-processing available. For dynamically-acquired images, the final five minute frame was obtained. For the static PET/CT-acquired images, the whole 30 minute frame was obtained. Each ^{18}F-FDG PET scan was then pre-processed in 4 stages using SPM12: (1) brain extraction, (2) rigid registration to a custom ^{18}F-FDG template (Della Rosa et al., 2014), (3) spatial normalization to Montreal Neurological Institute (MNI) space (voxel size = 2x2x2mm), and (4) intensity normalization by the global mean. Pre-processing with fewer steps (e.g. no normalization) and more steps (i.e. adding smoothing with a 3D Gaussian kernel) was also performed. Performance of the DL model was evaluated for each variation of pre-processed data using receiver operating curve (ROC) analysis, while the amount of overfitting was evaluated using learning curves. The dataset with the four steps described was found to give the best performance based on area under the ROC curve (AUC) while maintaining minimal overfitting. Examples of pre-processed ^{18}F-FDG PET scans for an AD and NC subject are shown in the top row of Figure 1.

All T1-weighted 3D MR scans with Magnetization Prepared Rapid Gradient-Echo (MP-RAGE) sequencing (Mugler III & Brookeman, 1990) were downloaded from ADNI. These scans have undergone grad-warping, intensity correction, and scaling for gradient drift using phantom data. The MR scans were then additionally pre-processed in two steps: (1) brain extraction, (2) rigid registration to a custom MRI template, so that both the MR and the ^{18}F-FDG PET scans all had similar levels of pre-processing, namely, each had brain extraction, rigid registration, and spatial and intensity normalization. Examples of pre-processed MR scans for an AD and NC subject are shown in the bottom row of Figure 1.

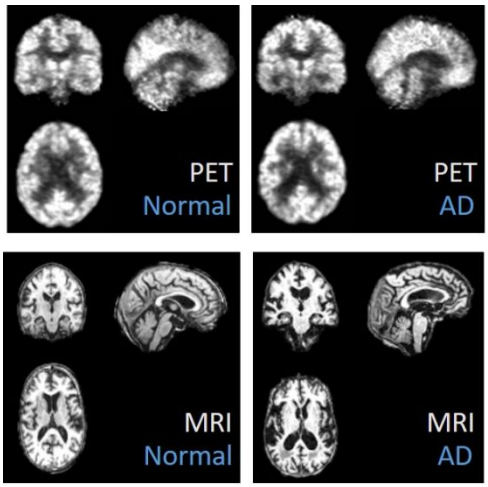

Figure 1. Example ^{18}F-FDG PET and MRI scans from the ADNI database after pre-processing for both AD and NC subjects.

After pre-processing, data quality assurance (QA) checks were performed. All PET scans which used 3D Filtered Back Projection reconstruction were removed from the dataset due its distinct noise pattern (N=99). In addition, any scans with poor segmentation, artifacts, motion, or other major visual issues were removed (N=83). To further maximize the reliability of the diagnosis, all patients with fluctuating disease status were eliminated (e.g. a patient who is classified as AD, then NC, then AD again at sequential timepoints).

A total of 772 (N_{AD}=364, $N_{control}$=408) 3D ^{18}F-FDG PET scans and 780 (N_{AD}=280, $N_{control}$=500) T1-weighted volumetric-3D MR images met all above criteria. See Table 1 for summaries. Data was shuffled randomly for each run and classes were balanced before input to the network. Balancing was done in an effort to give equal priority to each class during model training. Balancing classes in DL problems has been shown to improve accuracy and prevent the penalizing of minority samples. It also adds reliability to ROC analysis (Johnson & Khoshgoftaar, 2019). Thus for the CNN, 728 ^{18}F-FDG PET scans from 436 patients split evenly between AD patients and NC subjects (364 each) were used. Of these, 122 patients (61 each) had scans at multiple timepoints within a 2-year range (with time lags of one or two years between scans) and were thus used to evaluate the RNN. Similarly, performance was evaluated with 560 MRI scans (280 from each class) from 193 patients, 130 of whom (65 each) have scans at multiple timepoints within a 2-year range (with time lags of one or two years between scans).

Table 1. Description of the ^{18}F-FDG PET and T1-MRI datasets obtained from the ADNI database.

	ADNI ^{18}F-FDG PET		**ADNI T1-MRI**	
	AD	**NC**	**AD**	**NC**
TOTAL NUMBER OF SCANS	364	408	280	500
NUMBER OF PATIENTS WITH 2-3 SCANS	61	70	65	79
AVERAGE AGE	76.1	77.7	75.6	77.0
GENDER	59.6% male	60.8% male	57.0% male	54.9% male

2.2 Independent Validation Dataset

Performance of the CNN was also evaluated on an independent dataset obtained from University Medical Centre Ljubljana (UMCL) (Perovnik et al., 2022). This dataset consisted of ^{18}F-FDG PET scans from 104 patients (N_{AD} = 63, N_{NC} = 41), with only one scan per patient. All AD patients had a clinical AD diagnosis and CSF-confirmed AT(N) biological diagnosis. More specifically, all AD patients in this dataset fulfilled diagnostic criteria for amnestic type dementia (McKhann et al., 2011) and had positive Alzheimer CSF biomarkers, defined as Aβ42 < 650 pg/mL or Aβ42/Aβ40 < 0.077 for amyloid (A+) and p-tau > 60 pg/mL for tau positivity (T+). The detailed description of the sample is provided elsewhere (Perovnik et al., 2022). All normal controls in this dataset were healthy volunteers recruited from the local community with no specific inclusion/exclusion criteria. Each control subject completed a clinical neurological and neuropsychological examination for purposes of an earlier research project (Tomše et al., 2017).

All images were acquired as PET/CT scans with a 30-minute frame duration. In contrast to the ADNI dataset, this data was acquired using Time-of-Flight (TOF) PET scanners and Point Spread Function (PSF) reconstruction was applied. See Table 2 for a more detailed summary. The data was pre-processed in the same way as the ADNI data and pixel dimensions were matched to that of the ADNI scans. Image QA was again performed.

Table 2. Description of the ^{18}F-FDG PET dataset obtained from UMCL.

LJU ^{18}F-FDG PET DATASET	AD	NC
TOTAL NUMBER OF SCANS	63	41
NUMBER OF PATIENTS WITH 2-3 SCANS	0	0
AVERAGE AGE	73.0	65.3
GENDER	57.1% male	31.7% male

2.3 Consideration of Bias in the Longitudinal Cohort

For the use of longitudinal data, we considered the potential for bias due to selective dropout, or non-random loss of participants from a study (Chatfield et al., 2005). Comparisons of age, Mini-Mental State Exam (MMSE) score, and gender proportion were evaluated between the cohort of patients with multiple scans (used to evaluate the RNN) and the cohort of patients used for only single timepoint analysis (used to evaluate the CNN) to determine the level of bias in our datasets. Note that there is a significant overlap of patients across the two cohorts, since all patients with multiple scans were included in the single timepoint analysis as well.

Table 3 summarizes key demographics of our datasets. The average of each metric for the single timepoint cohort was compared to that of both the average value at the time of the initial scan for the longitudinal cohort as well as the overall average values from all three timepoints in the longitudinal cohort. This comparison was performed within each of the class

cohorts (AD and NC) and for both classes overall. The full dataset was considered in addition to each class separately in order to mimic the type of bias analysis performed in most of the studies cited in a large review of selective dropout bias (Chatfield et al., 2005), where bias analysis was performed over an entire dataset, without separating by any specific disease classes or subject type. Two sample t-tests were used with a p-value cutoff of 0.05. Histograms were examined to evaluate the normality of the data. The distributions of age are clearly approximately normally distributed. While the MMSE score distributions are less normal, the large sample size and conservative p-value cutoff allow the assumption of t-test validity. The only significant differences (denoted by the red text in Table 3) were found for overall MMSE score in the ^{18}F-FDG PET data and for starting AD MMSE score in the MR data. Thus it was concluded that our cohorts were not so different as to cause a significant bias in the analyses.

Table 3. Average values for age and cognitive score and percentage of male subjects in each sub-cohort for the ^{18}F-FDG PET (left) and T1-MRI (right) datasets. Red numbers indicate those that are significantly different (p-value < 0.05) from the single timepoint values.

	^{18}F-FDG PET			**T1-MRI**		
	AD	**NC**	**Overall**	**AD**	**NC**	**Overall**
AGE						
SINGLE TIMEPOINT	75.7	78.0	77.0	74.6	76.5	75.7
LONGITUDINAL STARTING	75.0	77.9	76.6	74.5	76.4	75.5
LONGITUDINAL OVERALL	76.0	78.5	77.3	74.6	76.6	75.7
MMSE SCORE						
SINGLE TIMEPOINT	20.1	29.1	26.9	22.1	29.0	26.5
LONGITUDINAL STARTING	21.2	29.1	28.1	23.6	29.2	26.6
LONGITUDINAL OVERALL	19.6	29.2	25.7	21.8	29.1	25.9
% MALE						
SINGLE TIMEPOINT	54.9	59.5	57.6	52.1	47.2	49.0
LONGITUDINAL STARTING	56.5	57.7	57.1	51.4	48.9	50.0

3. Methods

3.1 DL Model Overview

A binary classifier 3D CNN was trained both with and without a cascaded RNN. A schematic overview of the model is shown in Figure 2. The CNN inputs included brain images and clinical data (age, gender). The CNN evaluates spatial information from single timepoint scans and outputs class predictions. The feature maps from the second-to-last layer of the CNN are then input into the RNN for all patients with multiple scans. The RNN evaluates temporal information across 3 timepoints with 1-year gaps.

The DL model was built with the Keras library using TensorFlow as backend. All experiments were run on Python version 3.6. The model was trained in a workstation with an NVIDIA RTX Titan GPU with 24GB of memory. A single epoch of the CNN takes ~20 seconds to train over the entire training set, while each epoch of the RNN takes ~5 seconds.

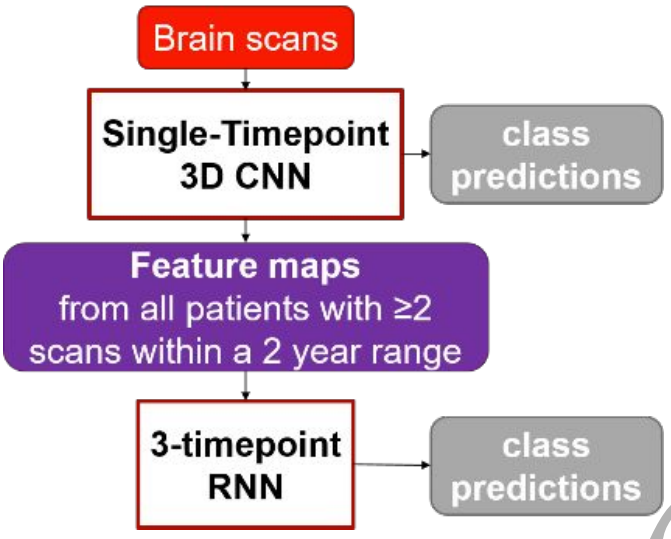

Figure 2. Schematic representation of the deep learning model setup.

The ADNI datasets were shuffled randomly and divided into training (60%), validation (20%), and test set (20%) splits. No data augmentation was performed. The resulting balanced number of scans in each set for the CNN can be found in Table 4 for the ^{18}F-FDG PET and T1-MRI datasets. The scans were then arranged by patient and by timepoint. If a patient had at least two scans spaced one to two years apart, they were included in the dataset for the RNN. Up to three scans with 1-year gaps between each scan were included for each patient. The resulting number of patients, each now with 2-3 scans, can be found in Table 4 for the ^{18}F-FDG PET and MRI datasets.

The models were run five times each with shuffled ADNI data. Performance was assessed using ROC analysis. Area under the curve (AUC) along with accuracy, sensitivity, and specificity at the optimal threshold (upper leftmost point, determined during training) were assessed for model comparison. Note that this threshold was determined by the ROC curve of the ADNI validation dataset (not the test set) in order to best preserve the generalizability of the model. Also note that this threshold (and the corresponding accuracy, sensitivity, and specificity) can be adjusted to prioritize true positive rate or false positive rate for optimized translation to the clinic, where the AUC provides a metric for how well the classes can be separated across all thresholds. Thus AUC values were used to optimize model parameters and to compare performance across models.

Table 4. Number of single timepoint scans used in the CNN and number of patients with 2-3 scans at multiple timepoints used in the RNN in the train/validation/test datasets for each imaging modality

	^{18}F-FDG PET						T1-MRI					
	CNN Scans			RNN Patients			CNN Scans			RNN Patients		
	AD	**NC**	**Total**	**AD**	**NC**	**Total**	**AD**	**NC**	**Total**	**AD**	**NC**	**Total**
TRAINING	234	234	468	40	40	80	180	180	360	42	42	84
VALIDATION	58	58	116	9	9	18	44	44	88	10	10	20
TESTING	72	72	144	12	12	24	56	56	112	13	13	26
TOTAL	364	364	728	61	61	122	280	280	560	65	65	130

3.2 Convolutional Neural Network (CNN)

The 3D-CNN used here was adapted from previous work by Spasov *et al.* (Spasov et al., 2019) The CNN model architecture, shown in Figure 3, employs 3D separable convolutional layers along with several fully connected layers. The formulation of this model architecture is explained in depth in Spasov *et al.* The version employed here is slightly simplified from its original use in that it only performs a single task and takes in a single imaging modality as input along with the clinical data. The number of parameters was also increased in each layer to improve performance based on a sensitivity study of model width using the ^{18}F-FDG PET dataset. In this sensitivity study, the number of parameters in each convolutional layer was tested systematically by doubling the number of features until overfitting was clear on the learning curves. The size of the final dense layer before the output was also varied from 2 to 100 and the value resulting in optimal performance in ROC analysis was chosen. The final model width still allows for a highly parameter-efficient model due to the use of separable and grouped convolutions.

Generally, the model uses three types of operational blocks (described in the inset of Figure 3), each following a similar pattern of reused layers: a convolutional or dense layer followed by batch normalization, an Exponential Linear Unit (ELU) activation function, and dropout. The convolutional blocks also utilize 3D max pooling to decrease the input image dimensionality.

Both the separable and grouped convolutions allow for more parameter-efficient image processing than traditional convolutions. The separable convolutional layers reformulate standard convolutions by separately performing depth-wise and then point-wise operations. This significantly reduces the number of parameters required. The model also employs grouped convolution. After the separable convolutional blocks, the feature maps are split into two groups along the channel axis and evaluated separately, requiring only half the parameters as only half of the channels are used to produce a single output feature map. This allows a reduction in the dimensionality of the activation maps and thus a more parameter-efficient model.

A skip, or residual, connection was also added such that the output from the last separable convolutional block is summed element-wise with the activation maps from the second

convolutional block, indicated by the circled plus sign in Figure 3. Residual connections have been shown to facilitate training as the neural net depth increases (Chollet, 2017; He et al., 2015).

The clinical data were processed in two sequential fully connected blocks with 64 and 20 units, respectively. The clinical features and flattened image embeddings were concatenated and processed by two dense layers, with the final layer providing two outputs for the binary prediction. To consider the influence of the inclusion of this clinical data, the model was also run without this clinical data included. ROC analysis indicates that the CNN performed consistently both with and without the clinical data included. The AUC and accuracy are well within the standard deviation for both the ^{18}F-FDG PET dataset and the MRI dataset. ROC curves, AUC values, and accuracy values of this supplementary experiment can be found in Figure A.1, Table A.1, and

Table A.2 in the Appendix.

A training batch size of 10 was used, the dropout rate was set to 0.3, and the L2 regularization penalty coefficient was set at $2\text{x}10^{-2}$. The convolutional kernel weight initialization follows the procedure described by He *et al.* (He et al., 2015). The loss was determined by sparse categorical cross entropy and minimized using the Adam optimizer with softmax activation and a step decay learning rate: $\mathrm{lr} = (1 \times 10^{-4}) \times 0.3^{\mathrm{epoch}/10}$. These hyperparameters were chosen to optimize the performance of the validation dataset while minimizing overfitting. The model was trained for 30 epochs. A detailed breakdown of the shape and number of parameters at each layer in the CNN can be found in Table A.3in the Appendix. Example learning curves shown in Figure 4 demonstrate the model's minimal overfitting during training.

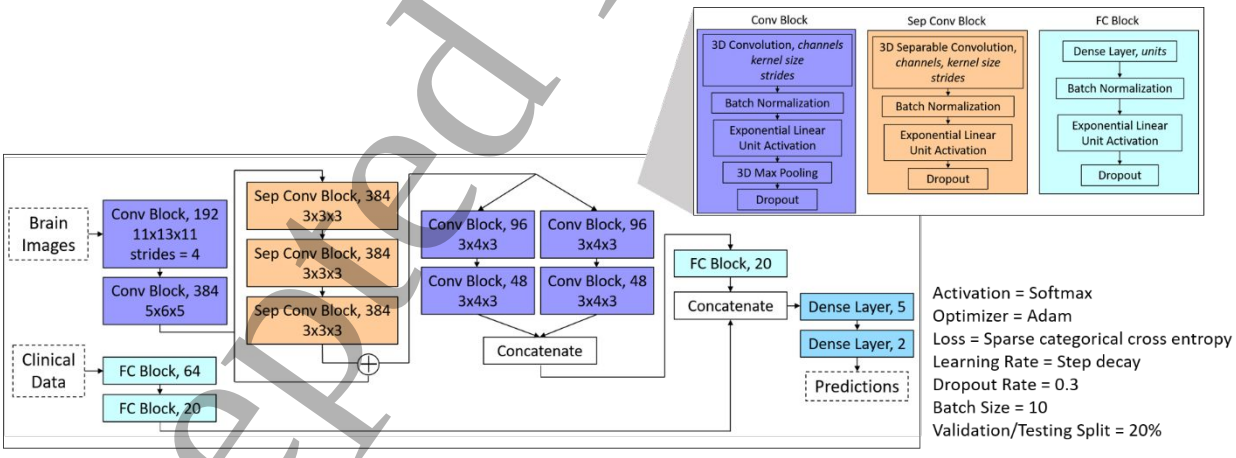

Figure 3. Model architecture of the 3D-CNN used in this work. Adapted from previous work by Spasov et al. (Spasov et al., 2019). The inset provides a description of each of the 3 types of blocks used in the model architecture. The parameters required for each type of layer are listed in italics in the inset, and the corresponding values used in particular blocks are provided for each block. If the strides are equal to the default value of 1, they are not listed. The circled plus sign indicates a residual, or skip, connection. On the bottom right are listed some key hyperparameters of the final trained model.

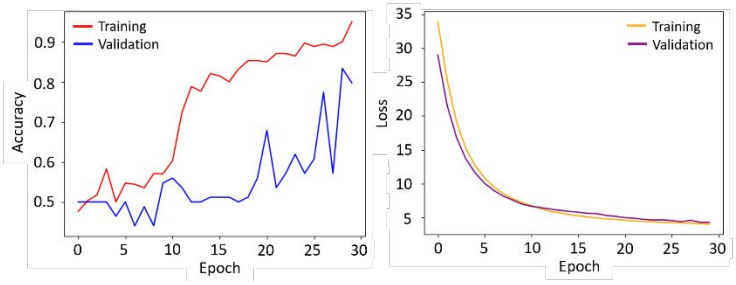

Figure 4. Example learning curves for the CNN training. Both accuracy (left) and loss (right) curves show minimal overfitting over the 30 training epochs.

3.3 Recurrent Neural Network (RNN)

The 3-layer RNN used in this work was designed following the example of Cui *et al.* (Cui et al., 2019) and is shown in Figure 5. Generally, the RNN relates the feature map from the second-to-last layer of the CNN at one timepoint to the maps at the timepoints both before and after. The CNN and RNN were trained end-to-end in tandem with a shuffled dataset input for each run.

The network is composed of three bidirectional gated recurrent unit (BGRU) blocks, each containing six gated recurrent units (GRU), described by the inset in Figure 5. The GRU is a specific variant of the long short-term memory (LSTM), which has shown slightly better performance and fewer parameters than the traditional LSTM (Chung et al., 2014). GRU layers use an update gate to determine what new information to add and a reset gate to determine which past information to discard (Cho et al., 2014). Each BGRU block consists of one forward and one backward sequence of GRUs, with one GRU per timepoint. In this way, the information at each timepoint is related to both its preceding and following timepoint. The output of each forward and backward GRU is then concatenated for each timepoint, and this information is fed into the next BGRU block. See the inset of Figure 5 for a visual explanation.

The input to the GRU is produced by the last dense layer of the CNN before the final softmax classification. A 5-dimensional feature map output is created for each image in the longitudinal dataset using the CNN. These maps are then organized sequentially by patient and fed into the corresponding GRU. The GRUs handle consistently spaced scans across all patients, so each GRU corresponds to one year of time between scans. The setup of three timepoints with 1-year gaps was chosen in order to maximize the availability of longitudinal data in the ADNI dataset. However, since not all patients have three scans exactly one year apart, a masking layer was added to handle missing inputs. The masking layer allows the GRU to recognize a placeholder where no scan was input and simply skip that layer, changing no information at either gate.

Three BGRU blocks are stacked to form a deep network, enhancing the longitudinal information flow. The outputs of the final block are concatenated to a dense layer with dropout and a softmax layer for the binary classification.

A training batch size of 5 was used, the dropout rate was set to 0.1, and the L2 regularization penalty coefficient was set at $1x10^{-6}$. The loss was determined by sparse categorical cross entropy and minimized using the Adam optimizer with softmax activation and a step decay

learning rate: $\mathrm{lr} = 0.002 \times 0.3^{\mathrm{epoch}/10}$. These hyperparameters were chosen to optimize the performance of the validation dataset while minimizing overfitting. The model was trained for 30 epochs. Example learning curves shown in Figure 6 demonstrate the model's minimal overfitting during training.

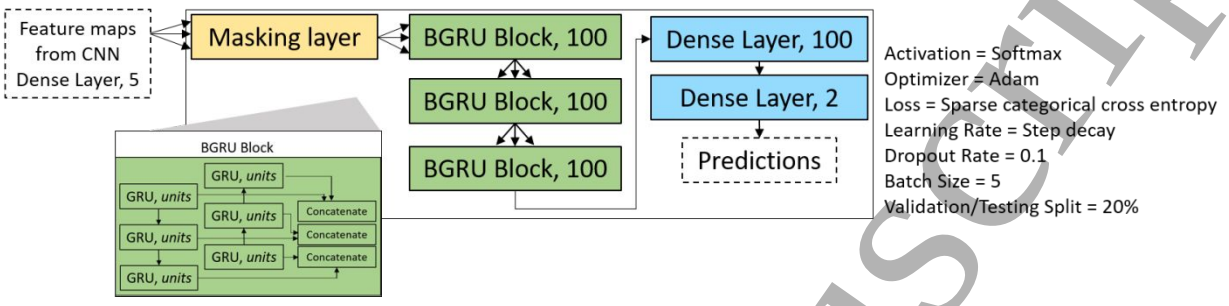

Figure 5. Model architecture of the RNN used in this work. The inset provides a description of each of the BGRU blocks. The same number of units (100) was used in each GRU layer. On the right are listed some key hyperparameters of the final trained model.

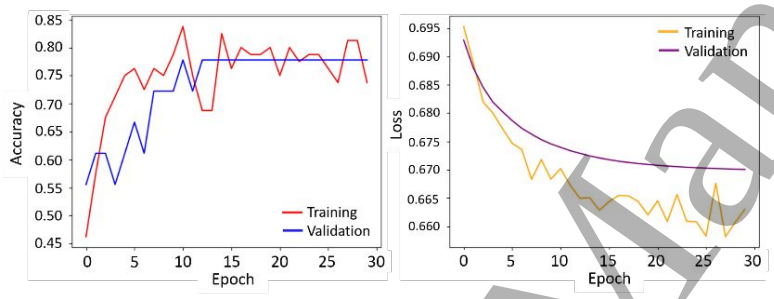

Figure 6. Example learning curves for the RNN training. Both accuracy (left) and loss (right) curves show minimal overfitting over the 30 training epochs.

3.4 Imaging Modality Comparison

The primary objective was to investigate the influence of imaging modality on model performance. The ADNI datasets described in Table 4, which produced the plots in Figure 7**Error! Reference source not found.**, contain unequal numbers of scans for each modality. To directly compare the performance of each imaging modality, the size of the ^{18}F-FDG PET dataset was restricted in the CNN and the size of the MRI dataset was restricted in the RNN, then the models were retrained and compared. This restriction was done by simply dropping scans from the end of the larger dataset before splitting into training, validation, and test sets until the number of scans was equal between the two modalities. This preserves the class balance between AD and NC. Since data was shuffled for each run, a different set of scans is dropped from each run.

3.5 Longitudinal Data Evaluation

A secondary objective of this study was to investigate the influence of including longitudinal data on model performance, evaluating whether adding the RNN to the model makes a significant difference. The ADNI datasets described in Table 4 which produced the plots in Figure 7**Error! Reference source not found.** contain unequal numbers of scans between the two model types. To directly evaluate the influence of longitudinal data, the number of scans

used in the CNN only model was restricted, then the models were retrained and compared. This restriction was again done by simply dropping scans from the end of the CNN dataset before splitting into training, validation, and test sets until the number of scans was equal to the number available for the RNN. This preserves the class balance between AD and NC, but means that the same exact scans were not necessarily used for both the CNN and the RNN in each run. Since data was shuffled for each run, a different set of scans is dropped from each run.

3.6 External Validation

Performance of the CNN was also evaluated on a held-out, independent test set of 104 ^{18}F-FDG PET scans obtained from UMCL for an external validation test. The model weights for the best performing run of the ADNI ^{18}F-FDG PET data trained CNN model were re-loaded and class predictions were made for each scan in the UMCL dataset. Performance was assessed using ROC analysis, and accuracy was evaluated at the same threshold as was deemed to be optimal for the original ADNI validation set during training.

3.7 Interpretability

Layer-wise Relevance Propagation (LRP) was employed to visualize the CNN decision-making. LRP creates a class-discriminative attention heatmap which indicates the relevance of each voxel to the final classification outcome. It operates by propagating the prediction backward in the neural network, using a set of purposely designed propagation rules (Montavon et al., 2019). LRP decomposes the network's output score into the individual contributions of the input neurons while keeping the total amount of relevance constant across layers. The heatmap does not rely on gradients, as with many other attention heatmap methods, but considers model weights and neuron activations (Bach et al., 2015; Samek et al., 2016). Final heatmaps show the average relevance of each voxel for contributing to the AD score (Böhle et al., 2019).

LRP was implemented using code from Bohle *et al.* (Böhle et al., 2019). The LRP-$\alpha_1\beta_0$ propagation rule was employed (Bach et al., 2015), which is equivalent to LRP-γ where $\gamma \to \infty$ and Excitation-Backpropagation (Montavon et al., 2019; Zhang et al., 2018). Employing LRP with a β value of zero allows only positive contributions to be shown in the heatmap, and highlights all positive contributions, regardless of their surroundings (Montavon et al., 2019).

In order to identify the most relevant regions for AD classification, an average heatmap was computed across all the correctly classified AD patients in the test set. This interpretability testing was performed on the best performing run of the CNN model for both ^{18}F-FDG PET and MRI.

4. Results

ROC analysis of the four model variations is shown in Figure 7**Error! Reference source not found.**. Corresponding AUC, accuracy at the optimal threshold, and sensitivity and specificity at the optimal threshold, are listed in Table 5, Table 6, and Table 7. The CNN+RNN model achieved an AUC of 0.93 ± 0.08 with ^{18}F-FDG PET data and 0.65 ± 0.12 for T1-MRI data, while the CNN only model achieved an AUC of 0.92 ± 0.01 with ^{18}F-FDG PET and 0.72 ± 0.10 with T1-MRI.

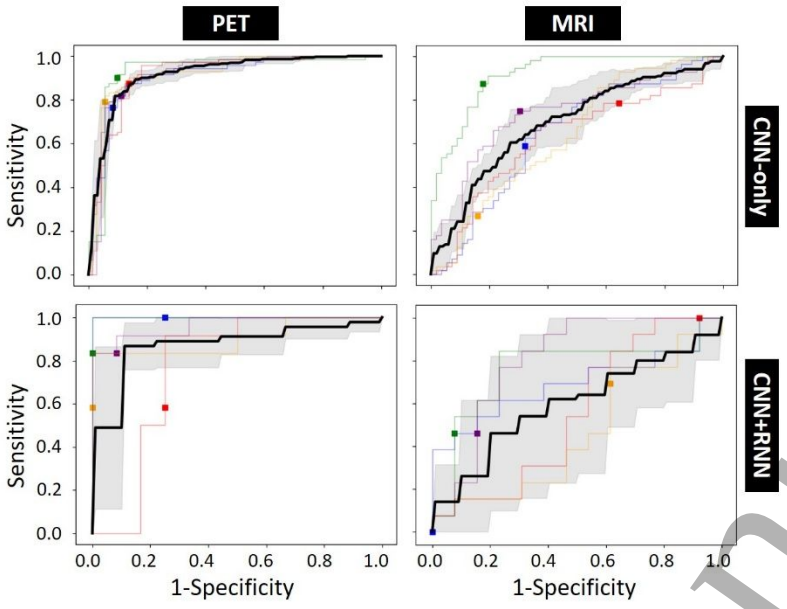

Figure 7. ROC curves for the CNN predictions (top row) and CNN+RNN predictions (bottom row) for both the ^{18}F-FDG PET (left) and MRI (right) datasets. The bold line indicates the average ROC curve over 5 runs and the shading indicates the uncertainty. The operating points corresponding to the optimal thresholds of the validation dataset are noted by the square markers.

*Table 5. AUC values for the ROC curves in Figure 7**Error! Reference source not found.** for the CNN model with and without the cascaded RNN when trained with ^{18}F-FDG PET data or T1-MR images from the ADNI database.*

AUC	^{18}F-FDG PET	T1-MRI
CNN ONLY	0.92 ± 0.01	0.72 ± 0.10
CNN + RNN	0.93 ± 0.08	0.65 ± 0.12

Table 6. Accuracy values for the CNN model with and without the cascaded RNN when trained with ^{18}F-FDG PET data or T1-MR images. Accuracy was calculated at the optimal threshold for the ADNI validation set.

ACCURACY	^{18}F-FDG PET	T1-MRI
CNN ONLY	86.7 ± 2.1%	66.6 ± 10.9%
CNN + RNN	82.5 ± 8.9%	58.4 ± 7.5%

Table 7. Sensitivity and specificity values for the CNN model with and without the cascaded RNN when trained with ^{18}F-FDG PET data or T1-MR images. Metrics were calculated at the optimal threshold for the ADNI validation set.

SENSITIVITY	^{18}F-FDG PET	T1-MRI
CNN ONLY	87.8 ± 4.5%	65.4 ± 21.4%
CNN + RNN	76.6 ± 16.2%	52.3 ± 32.8%

SPECIFICITY	^{18}F-FDG PET	T1-MRI
CNN Only	90.3 ± 2.8%	67.8 ± 17.3%
CNN + RNN	88.3 ± 11.3%	63.1 ± 38.1%

Following this successful development of a deep learning model for clinical AD diagnosis, the specific influences of various factors on the model were investigated.

4.1 Imaging Modality Comparison

The restricted ADNI datasets used to investigate the influence of imaging modality and the resulting ROC curves are shown in Figure 8 and the corresponding AUC values are listed in Table 8.

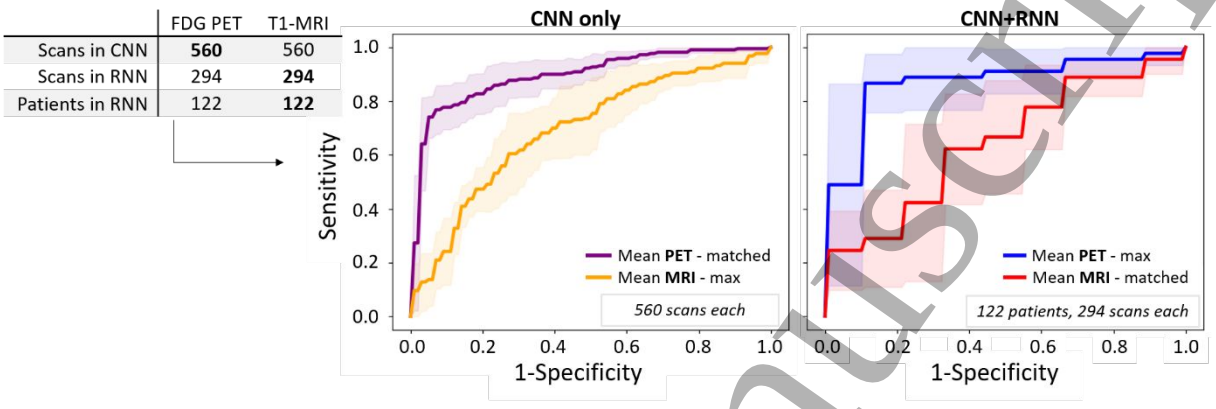

Figure 8. Roc curves comparing the performance of the two imaging modalities in both the CNN only (left) and the CNN+RNN models (right). The bold lines indicate the average ROC curve over 5 runs and the shading indicates the uncertainty. The dataset on the far left shows the total number of scans and patients used in each case, where bold indicates the datasets that were restricted to match the amount of data across the two modalities. "Max" indicates the dataset where the maximum number of scans was used, while "matched" indicates the dataset which was restricted to match the number of scans in the other.

Table 8. AUC values from the ROC curves in Figure 8 for the CNN model with and without the cascaded RNN when trained with equal numbers of ^{18}F-FDG PET scans and T1-MR images.

AUC	^{18}F-FDG PET	T1-MRI
CNN ONLY	0.88 ± 0.04	0.72 ± 0.10
CNN + RNN	0.93 ± 0.08	0.71 ± 0.09

4.2 Longitudinal Data Evaluation

The restricted ADNI datasets used to investigate the influence of longitudinal data and the resulting ROC curves are shown in Figure 9 and corresponding AUC values are listed in Table 9.

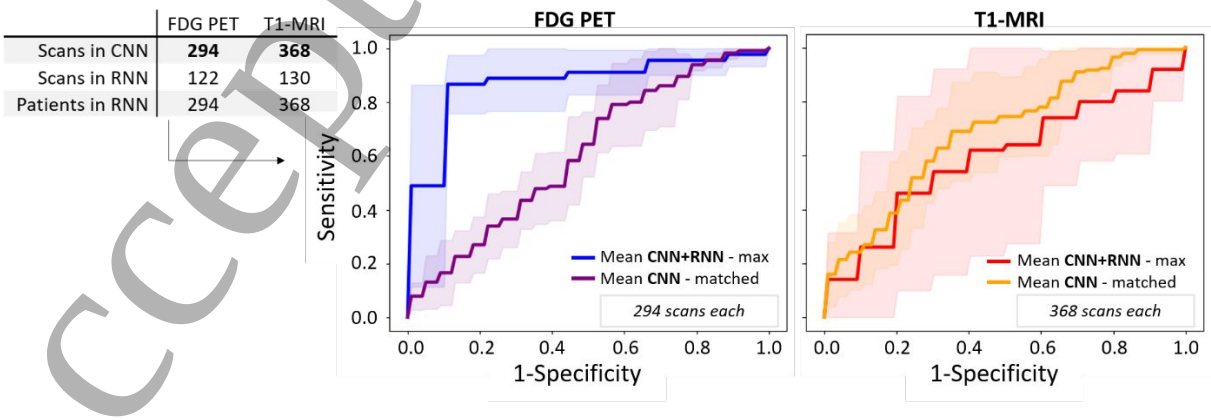

Figure 9. ROC curves comparing the performance of the model with and without the RNN and longitudinal data using both the ^{18}F-FDG PET data (left) and the MRI data (right). The bold lines indicate the average ROC curve over 5 runs and the shading

indicates the uncertainty. The dataset on the far left shows the total number of scans and patients used in each case, where bold indicates the datasets that were restricted to match the amount of data across the two model types. "Max" indicates the dataset where the maximum number of scans was used, while "matched" indicates the dataset which was restricted to match the number of scans in the other.

Table 9. AUC values from the ROC curves in Figure 9 for the CNN model with and without the cascaded RNN when trained with equal numbers of FDG PET scans and T1-MR images between the two model types.

AUC	^{18}F-FDG PET	T1-MRI
CNN ONLY	0.58 ± 0.08	0.63 ± 0.06
CNN + RNN	0.93 ± 0.08	0.65 ± 0.12

4.3 External Validation

ROC curves for the best performing run of the ADNI ^{18}F-FDG PET data trained CNN model tested with both ADNI data and the UMCL external validation dataset are shown in Figure 10**Error! Reference source not found.**. The resulting AUC and accuracy at the optimal threshold are shown in

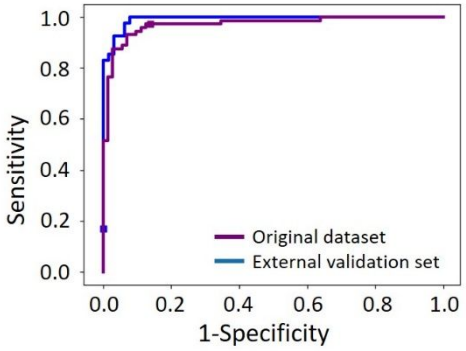

Figure 10. ROC curves for the ADNI ^{18}F-FDG PET test set (purple) and the external validation test set from LJU (blue) for the best performing run of the CNN model trained with the ADNI dataset. The operating points corresponding to the optimal thresholds of the ADNI validation dataset are noted by the square markers.

Table 10.

For this run, the model achieved an AUC on the ADNI test set of 0.97 and an accuracy at the optimal threshold of 91.7%. The external validation test set achieved an AUC of 0.99 with an accuracy (calculated at the same threshold) of 67.3%. At its own optimal threshold, the accuracy of the external validation set reaches 95.2%.

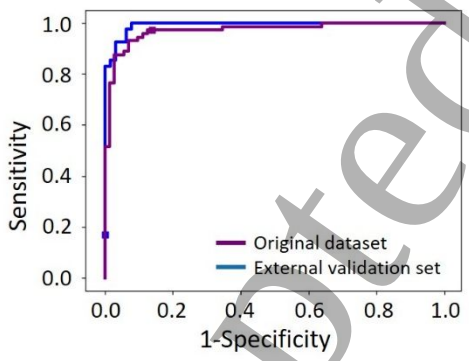

Figure 10. ROC curves for the ADNI ^{18}F-FDG PET test set (purple) and the external validation test set from LJU (blue) for the best performing run of the CNN model trained with the ADNI dataset. The operating points corresponding to the optimal thresholds of the ADNI validation dataset are noted by the square markers.

Table 10. Performance results for the ADNI test set and the external validation test set for the best performing run of the CNN model trained with the ADNI dataset. Accuracy is calculated at the optimal threshold of the ADNI validation set.

	AUC	ACCURACY
ORGINAL DATASET (ADNI TEST SET)	0.97	93.1%
EXTERNAL VALIDATION SET (LJU TEST SET)	0.99	67.3%

4.4 Interpretability

The LRP algorithm produced heatmaps highlighting the most important voxels for the CNN decision. Figure 11 shows the average LRP heatmaps across the correctly classified (true positive) AD scans and the correctly classified NC scans (true negative) for both ^{18}F-FDG PET and MRI from the ADNI dataset

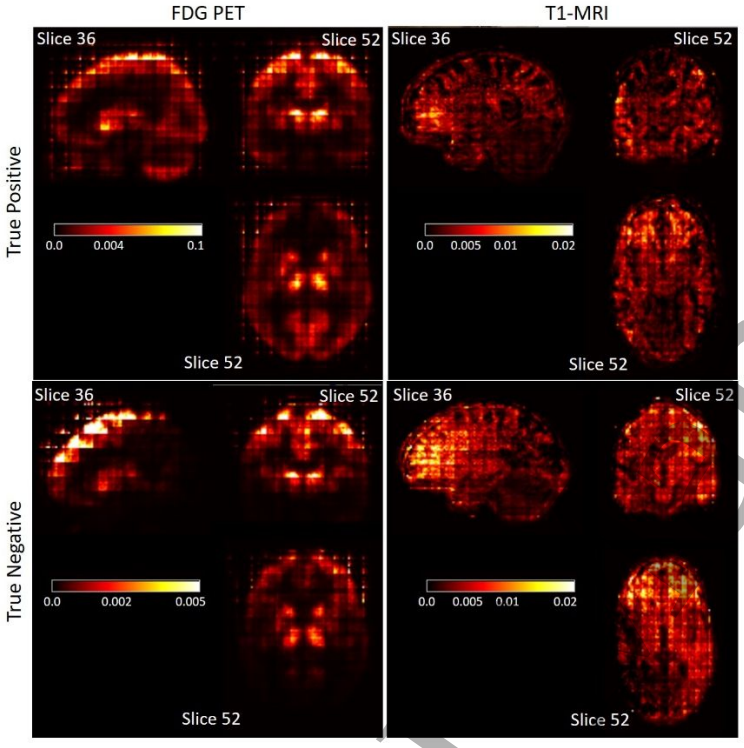

Figure 11. Average Layer-wise Relevance Propagation (LRP) heatmaps of the true positive (top row) scans of Alzheimer's patients and the true negative (bottom row) scans of normal controls for both ^{18}F-FDG PET (left) and T1-MRI (right).

5. Discussion

According to the metric of AUC, the best performing variation of the model presented here is the CNN+RNN model trained with ^{18}F-FDG PET data, producing an AUC of 0.93 ± 0.08 and an accuracy of 82.5 ± 8.9% with sensitivity and specificity of 76.6 ± 16.2% and 88.3 ± 11.3%, respectively. A higher accuracy of 86.7 ± 2.1% (with a similar AUC of 0.92 ± 0.08, sensitivity of 87.8 ± 4.5%, and specificity of 90.3 ± 2.8%) was achieved by the CNN only model also trained with ^{18}F-FDG PET data. Note that the choice of threshold (and corresponding accuracy) is only meant to give an idea of the potential of the model. This threshold can be optimized at translation based on allowable levels of false positive and false negative rates in the clinic. The AUC provides a metric for how well the classes can be separated across all thresholds.

The ROC curves demonstrate a high performing model with AUC approaching state of the art performance for an AD diagnostic tool, especially for models using ^{18}F-FDG PET data. While higher performing models have been presented, particularly using MR images, this model has specific advantages which allow it to act as the groundwork for further optimization as a

diagnostic tool. These advantages include that it does not rely on many clinical variables (such as neuropsychological test scores), is trained on a balanced dataset, requires minimal pre-processing of the input images, and generalizes well to an independent dataset. The dataset used for training also offers an advantage in that it has been scrubbed to exclude any contradictions between clinical AD diagnosis and biological AD pathology. This increases the accuracy and decreases the uncertainty in the ground truth diagnosis, allowing for more optimized model training.

5.1 Influence of Imaging Modality

When the ^{18}F-FDG PET data is limited to the equivalent number of scans as the number of MR images (560 total, balanced between classes), the PET-trained model significantly outperforms the MRI-trained model in both the CNN only and the CNN+RNN case. This outcome corroborates the findings of several other previously mentioned studies comparing the performance of ^{18}F-FDG PET and MRI on various AD diagnostic tasks using other types of ML classifiers and CNNs (Dukart et al., 2011; Huang et al., 2019; Lu et al., 2018; Samper-González et al., 2018). One hypothesis to explain this is that T1-MR images tend to vary more from one another than ^{18}F-FDG PET images. Thus using T1-MR images to diagnose AD with this model may require a larger or more harmonized dataset in order to see the same performance as ^{18}F-FDG PET.

This result also has potential implications in the clinic, suggesting that ^{18}F-FDG PET should be strongly considered for inclusion in early Standard-of-Care (SOC) practice for suspected AD patients. At present, ^{18}F-FDG PET is not typically ordered until after MRI and other laboratory and neuropsychological testing has been performed. Due to the retrospective nature of the ADNI dataset, this could have some influence on our results, though the timescales are likely too small to capture significant disease advancement. However, the high performance of the model with PET data suggests that its inclusion at the outset of suspected AD may be beneficial and should at minimum be explored in prospective studies.

5.2 Influence of Longitudinal Data

When the number of scans input to the CNN only is limited to the equivalent number of scans as are input to the CNN+RNN (294 total for ^{18}F-FDG PET and 368 total for T1-MRI, balanced between classes), CNN+RNN model significantly outperforms the CNN only when using the ^{18}F-FDG PET scans. Thus, for ^{18}F-FDG PET data, including longitudinal information appears to increase the information the model can learn, thus offering significant improvement to model performance. However, the same improvement is not seen in models trained with MR images. Since there is more data in the MRI dataset, the two modalities might be expected to show different performance overall, but that should not be expected to translate to a greater or lesser difference between the CNN only and CNN+RNN models. Following the same hypothesis as in the comparison of the imaging modalities above, it is possible that the MR scans are too varied from one another for the longitudinal change between scans to offer additional distinct information to be gleaned by the RNN portion of the model. It is also possible that changes in

the structure of the brain, captured by T1-MRI, occur more slowly than the metabolic changes captured by ^{18}F-FDG PET. Thus utilizing longitudinal information with T1-MRI may require monitoring over longer periods of time.

This result also has potential implications both for future research and in the clinic. Firstly, more AD diagnostic tools should explore the incorporation of longitudinal imaging data. This has been largely underexplored, particularly in deep learning models, but these results point toward its beneficial influence even over relatively short timescales. Secondly, regular ^{18}F-FDG PET follow-up scans should be considered as an addition to clinical SOC for suspected AD cases as the inclusion of this longitudinal information led to a more predictive model.

5.3 External Validation

When the CNN model was tested on an independent ^{18}F-FDG PET dataset, it performed extremely well achieving an AUC of 0.99. This is indeed highly promising for the generalizability of this model to other unseen data, particularly because (1) most of the ^{18}F-FDG PET scans obtained from ADNI were dynamically acquired while all scans in the UMCL independent dataset were acquired as static PET/CT and thus with higher signal-to-noise-ratio, (2) all patients in the UMCL dataset had their clinical diagnosis confirmed by CSF biomarker data for pathological diagnosis, (3) the UMCL dataset was acquired using TOF PET scanners and PSF reconstruction was applied, in contrast to the ADNI data.

The lower accuracy of 67.3% obtained when using the optimal threshold of the ADNI validation set indicates the limits of this generalizability, suggesting that the model could still benefit from a more robust training dataset. In the absence of more training data, this threshold can be tuned for specific clinical cases in one of two ways: (1) to optimize performance on a particular dataset (in this case achieving a maximum accuracy of 95.2%) or (2) to optimize performance to a particular outcome (i.e. desired sensitivity or specificity). The high AUC value indicates a wide range of possible generalizable thresholds for good clinical performance, even in the absence of a local dataset.

5.4 Interpretability

The LRP heatmaps shown in Figure 11 highlight regions of importance for the CNN decision-making towards each class. It appears from these maps that the CNN considers essentially the whole brain for its decision, with a focus on a few particular regions. The differences between the hot regions on the ^{18}F-FDG PET and T1-MRI suggest that when a patient is experiencing decline from AD, different brain regions may be affected structurally while others are affected metabolically.

5.5 Future Work

Future work should include obtaining datasets with increasing variability to explore the limitations of the model's generalizability. It should also include datasets where all patients have CSF data available, to increase ground truth diagnosis accuracy and create a tool for diagnosis of pathological AD. In addition, this work was limited to only testing the

generalizability of the CNN trained with ^{18}F-FDG PET. Future work should also aim to perform a similar analysis on an independent dataset of MR images as well as on longitudinal datasets of both ^{18}F-FDG PET and MRI to test the generalizability of the RNN. Extensions of the evaluation of this model to other imaging modalities, such as T2-MRI, fMRI, amyloid-PET, and tau-PET, would also make for interesting future explorations. A future study with multi-modality data would also be a very relevant and interesting extension of this work. This model could be easily adapted to accept multi-modality inputs, though the ADNI dataset will be significantly restricted to include only patients with multiple scan types at similar timepoints in order to incorporate the longitudinal aspects. Future studies should also compare the LRP results to other attention heatmap methods and to compare to the brain patterns identified using non-DL methods, such as PCA (Perovnik et al., 2022). In addition, application of this model to other clinical questions such as the differential diagnosis across various types of dementias (e.g. dementia with Lewy bodies or frontotemporal dementia), or the study of the preclinical stages of AD and their potential for progression/conversion to dementia is an important future expansion of this work.

6. Conclusion

We have successfully developed a deep learning model which can function as a generalizable, high-performance tool for AD diagnosis, achieving a maximum AUC with the CNN+RNN model of 0.93 ± 0.08. The direct comparison of imaging modality and evaluation of the inclusion of longitudinal data reveals the improved performance with ^{18}F-FDG PET data and its longitudinal information. These conclusions have significant implications both for future research on AD diagnostic models and management of suspected AD patients in the clinic. The CNN model generalizes well to a substantial external, independent dataset, showing promise for generalizability to other datasets and future translation. The CNN heatmaps move this work toward the identification of a quantitative imaging biomarker for AD.

The experimental results reveal a high-performing, generalizable, novel DL tool for the diagnosis of AD. The direct evaluation of these key influences offers insights about the potential of ^{18}F-FDG PET and T1-MRI and the changes they capture over time for AD diagnosis.

Acknowledgments

Special thanks to the Neurology Clinic and Department of Nuclear Medicine at University Medical Center Ljubljana for the use of their data for the validation dataset. Particular thanks to Luka Ležaić, M.D., Jan Jamšek, M.D., and Anka Cuderman, M.D.

Appendix

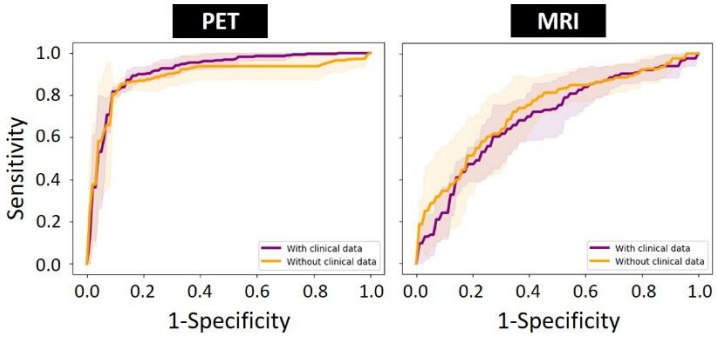

Figure A.1. ROC curves for the CNN only model run with (purple) and without (yellow) the clinical data included for the ADNI ^{18}F-FDG PET dataset (left) and the MRI dataset (right).

*Table A.1. AUC values for the ROC curves in Figure A.1**Error! Reference source not found.** for the CNN model with and without the clinical data when trained with ^{18}F-FDG PET data or T1-MR images from the ADNI database.*

AUC	^{18}F-FDG PET	T1-MRI
CNN WITH CLINICAL DATA	0.92 ± 0.01	0.72 ± 0.10
CNN WITHOUT CLINICAL DATA	0.92 ± 0.03	0.73 ± 0.06

*Table A.2. Accuracy values calculated at the optimal threshold of the ROC curves in Figure A.1**Error! Reference source not found.** for the CNN model with and without the clinical data when trained with ^{18}F-FDG PET data or T1-MR images from the ADNI database.*

ACCURACY	^{18}F-FDG PET	T1-MRI
CNN WITH CLINICAL DATA	86.7 ± 2.1%	66.6 ± 10.9%
CNN WITHOUT CLINICAL DATA	87.4 ± 2.6%	67.7 ± 5.1%

Table A.3. Detailed structure of the CNN.

Layer Name	**Details**	**Output Shape**	**Number of Parameters**
Input	Load imaging data	(91,109,91)	0
Conv3D_1	Start Conv Block	(23,28,23,1)	302,208
Batch_norm_1		(23,28,23,1)	768
elu_1		(23,28,23,1)	0
Max_Pooling3D_1		(12,14,12,1)	0
Dropout_1		(12,14,12,1)	0
Conv3D_2	Start Conv Block	(12,14,12,3)	11,059,584
Batch_norm_2		(12,14,12,3)	1,536
elu_2		(12,14,12,3)	0
Max_Pooling3D_2		(6,7,6,384)	0
Dropout_2		(6,7,6,384)	0

Sep_Conv3D_1	Start Sep Conv Block	(6,7,6,384)	158,208
Batch_norm_3		(6,7,6,384)	1,536
elu_3		(6,7,6,384)	0
Dropout_3		(6,7,6,384)	0
Sep_Conv3D_2	Start Sep Conv Block	(6,7,6,384)	158,208
Batch_norm_4		(6,7,6,384)	1,536
elu_4		(6,7,6,384)	0
Dropout_4		(6,7,6,384)	0
Sep_Conv3D_3	Start Sep Conv Block	(6,7,6,384)	158,208
Batch_norm_5		(6,7,6,384)	1,536
elu_5		(6,7,6,384)	0
Dropout_5		(6,7,6,384)	0
Add_1	Skip Connection	(6,7,6,384)	0
elu_6		(6,7,6,384)	0
Lambda_1		(6,7,6,192)	0
Lambda_2		(6,7,6,192)	0
Conv3D_3	Start Group 1 Conv Block	(6,7,6,96)	663,648
Conv3D_4	Start Group 2 Conv Block	(6,7,6,96)	663,648
Batch_norm_6		(6,7,6,96)	384
Batch_norm_7		(6,7,6,96)	384
elu_7		(6,7,6,96)	0
elu_8		(6,7,6,96)	0
Max_Pooling3D_3		(3,4,3,96)	0
Max_Pooling3D_4		(3,4,3,96)	0
Dropout_6		(3,4,3,96)	0
Dropout_7		(3,4,3,96)	0
Conv3D_5	Start Group 1 Conv Block	(3,4,3,48)	165,936
Conv3D_6	Start Group 2 Conv Block	(3,4,3,48)	165,936
Batch_norm_8		(3,4,3,48)	192
Batch_norm_9		(3,4,3,48)	192
elu_9		(3,4,3,48)	0
elu_10		(3,4,3,48)	0
Input_xls	Load clinical data	(2)	0
Max_Pooling3D_3		(2,2,2,48)	0
Max_Pooling3D_4		(2,2,2,48)	0
Dense_1	Start FC Block	(64)	192
Dropout_8		(2,2,2,48)	0
Dropout_9		(2,2,2,48)	0

Batch_norm_10		(64)	256
Concatenate_1	Recombine groups	(2,2,2,96)	0
elu_11		(64)	0
Reshape_1		(768)	0
Dropout_10		(64)	0
Dense_3	Start FC Block (imaging)	(20)	15,380
Dense_2	Start FC Block (clinical)	(20)	1,300
Batch_norm_11		(20)	80
Batch_norm_12		(20)	80
elu_12		(20)	0
elu_13		(20)	0
Dropout_11		(20)	0
Dropout_12		(20)	0
Concatenate_2	Combine imaging and clinical data	(40)	0
Dense		(5)	205
Class_Output (Dense)	Final Softmax Layer	(2)	12

Total parameters: 13,521,153
Trainable parameters: 13,516,913

References

A Sanchez-Catasus, C., N Stormezand, G., Jan van Laar, P., P De Deyn, P., Alvarez Sanchez, M., & AJO Dierckx, R. (2017). FDG-PET for prediction of AD dementia in mild cognitive impairment. A review of the state of the art with particular emphasis on the comparison with other neuroimaging modalities (MRI and perfusion SPECT). *Current Alzheimer Research*, *14*(2), 127-142.

Association, A. (2018). 2018 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, *14*(3), 367-429.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, *10*(7), e0130140.

Beach, T. G., Monsell, S. E., Phillips, L. E., & Kukull, W. (2012). Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer Disease Centers, 2005–2010. *Journal of neuropathology and experimental neurology*, *71*(4), 266-273.

Bergeron, D., Beauregard, J.-M., Guimond, J., Fortin, M.-P., Houde, M., Poulin, S., . . . Laforce Jr, R. (2016). Clinical Impact of a Second FDG-PET in Atypical/Unclear Dementia Syndromes. *Journal of Alzheimer's Disease*, *49*, 695-705. https://doi.org/10.3233/JAD-150302

Billones, C. D., Demetria, O. J. L. D., Hostallero, D. E. D., & Naval, P. C. (2016). DemNet: a convolutional neural network for the detection of Alzheimer's disease and mild cognitive impairment. 2016 IEEE region 10 conference (TENCON),

Bloudek, L. M., Spackman, D. E., Blankenburg, M., & Sullivan, S. D. (2011). Review and meta-analysis of biomarkers and diagnostic imaging in Alzheimer's disease. *Journal of Alzheimer's Disease*, *26*(4), 627-645.

Böhle, M., Eitel, F., Weygandt, M., & Ritter, K. (2019). Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Frontiers in aging neuroscience*, *11*, 194.

Chatfield, M. D., Brayne, C. E., & Matthews, F. E. (2005). A systematic literature review of attrition between waves in longitudinal studies in the elderly shows a consistent pattern of dropout between differing studies. *Journal of clinical epidemiology*, *58*(1), 13-19.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Choi, H., Ha, S., Kang, H., Lee, H., Lee, D. S., & Initiative, A. s. D. N. (2019). Deep learning only by normal brain PET identify unheralded brain anomalies. *EBioMedicine*, *43*, 447-453.

Choi, H., Jin, K. H., & Initiative, A. s. D. N. (2018). Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. *Behavioural brain research*, *344*, 103-109.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition,

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Cui, R., Liu, M., & Initiative, A. s. D. N. (2019). RNN-based longitudinal analysis for diagnosis of Alzheimer's disease. *Computerized Medical Imaging and Graphics*, *73*, 1-10.

Della Rosa, P. A., Cerami, C., Gallivanone, F., Prestia, A., Caroli, A., Castiglioni, I., . . . and the, E.-P. E. T. C. (2014). A Standardized [18F]-FDG-PET Template for Spatial Normalization in Statistical Parametric Mapping of Dementia. *Neuroinformatics*, *12*(4), 575-593. https://doi.org/10.1007/s12021-014-9235-4

Ding, Y., Sohn, J. H., Kawczynski, M. G., Trivedi, H., Harnish, R., Jenkins, N. W., . . . Mari Aparici, C. (2019). A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain. *Radiology*, *290*(2), 456-464.

Dukart, J., Mueller, K., Horstmann, A., Barthel, H., Möller, H. E., Villringer, A., . . . Schroeter, M. L. (2011). Combined evaluation of FDG-PET and MRI improves detection and differentiation of dementia. *PloS one*, *6*(3), e18111.

Ebenau, J. L., Timmers, T., Wesselman, L. M., Verberk, I. M., Verfaillie, S. C., Slot, R. E., . . . Van Den Bosch, K. A. (2020). ATN classification and clinical progression in subjective cognitive decline: The SCIENCe project. *Neurology*, *95*(1), e46-e58.

Femminella, G. D., Thayanandan, T., Calsolaro, V., Komici, K., Rengo, G., Corbi, G., & Ferrara, N. (2018). Imaging and molecular mechanisms of Alzheimer's disease: a review. *International journal of molecular sciences*, *19*(12), 3702.

Gao, L., Pan, H., Liu, F., Xie, X., Zhang, Z., Han, J., & Initiative, A. s. D. N. (2018). Brain disease diagnosis using deep learning features from longitudinal MR images. Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data,

Gordon, B. A., Blazey, T. M., Su, Y., Hari-Raj, A., Dincer, A., Flores, S., . . . Xiong, C. (2018). Spatial patterns of neuroimaging biomarker change in individuals from families with autosomal dominant Alzheimer's disease: a longitudinal study. *The Lancet Neurology*, *17*(3), 241-250.

Gray, K. R., Wolz, R., Heckemann, R. A., Aljabar, P., Hammers, A., Rueckert, D., & Initiative, A. s. D. N. (2012). Multi-region analysis of longitudinal FDG-PET for the classification of Alzheimer's disease. *NeuroImage*, *60*(1), 221-229.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE international conference on computer vision,

Huang, Y., Xu, J., Zhou, Y., Tong, T., Zhuang, X., & Initiative, A. s. D. N. (2019). Diagnosis of Alzheimer's disease via multi-modality 3D convolutional neural network. *Frontiers in Neuroscience*, *13*, 509.

Jack Jr, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., . . . Karlawish, J. (2018). NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*, *14*(4), 535-562.

Jo, T., Nho, K., & Saykin, A. J. (2019). Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. *Frontiers in aging neuroscience*, *11*, 220.

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 27. https://doi.org/10.1186/s40537-019-0192-5

Johnson, K. A., Sperling, R. A., Gidicsin, C. M., Carmasin, J. S., Maye, J. E., Coleman, R. E., . . . Fleisher, A. S. (2013). Florbetapir (F18-AV-45) PET to assess amyloid burden in Alzheimer's disease dementia, mild cognitive impairment, and normal aging. *Alzheimer's & Dementia*, *9*(5), S72-S83.

Kazemi, Y., & Houghten, S. (2018). A deep learning pipeline to classify different stages of Alzheimer's disease from fMRI data. 2018 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB),

Khosravi, M., Peter, J., Wintering, N. A., Serruya, M., Shamchi, S. P., Werner, T. J., . . . Newberg, A. B. (2019). 18F-FDG is a superior indicator of cognitive performance compared to 18F-florbetapir in Alzheimer's disease and mild cognitive impairment evaluation: a global quantitative analysis. *Journal of Alzheimer's Disease*, *70*(4), 1197-1207.

Lee, G., Nho, K., Kang, B., Sohn, K.-A., & Kim, D. (2019). Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Scientific reports*, *9*(1), 1-12.

Liu, M., Cheng, D., Yan, W., & Initiative, A. s. D. N. (2018). Classification of Alzheimer's disease by combination of convolutional and recurrent neural networks using FDG-PET images. *Frontiers in neuroinformatics*, *12*, 35.

Lu, D., Popuri, K., Ding, G. W., Balachandar, R., & Beg, M. F. (2018). Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images. *Scientific reports*, *8*(1), 1-13.

Martí-Juan, G., Sanroma-Guell, G., & Piella, G. (2020). A survey on machine and statistical learning for longitudinal analysis of neuroimaging data in Alzheimer's disease. *Computer methods and programs in biomedicine*, *189*, 105348.

Mattsson, N., Portelius, E., Rolstad, S., Gustavsson, M., Andreasson, U., Stridsberg, M., . . . Zetterberg, H. (2012). Longitudinal Cerebrospinal Fluid Biomarkers over Four Years in Mild Cognitive Impairment. *Journal of Alzheimer's Disease*, *30*, 767-778. https://doi.org/10.3233/JAD-2012-120019

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., . . . Phelps, C. H. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 263-269. https://doi.org/https://doi.org/10.1016/j.jalz.2011.03.005

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K.-R. (2019). Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, 193-209.

Moonis, G., Subramaniam, R. M., Trofimova, A., Burns, J., Bykowski, J., Chakraborty, S., . . . Pannell, J. S. (2020). ACR appropriateness criteria® dementia. *Journal of the American College of Radiology*, *17*(5), S100-S112.

Mugler III, J. P., & Brookeman, J. R. (1990). Three-dimensional magnetization-prepared rapid gradient-echo imaging (3D MP RAGE). *Magnetic resonance in medicine*, *15*(1), 152-157.

Oldan, J., Jewells, V., Pieper, B., & Wong, T. (2021). Complete Evaluation of Dementia: PET and MRI Correlation and Diagnosis for the Neuroradiologist. *American Journal of Neuroradiology*, *42*(6), 998-1007.

Perovnik, M., Tomše, P., Jamšek, J., Emeršič, A., Tang, C., Eidelberg, D., & Trošt, M. (2022). Identification and validation of Alzheimer's disease-related metabolic brain pattern in biomarker confirmed Alzheimer's dementia patients. *Scientific Reports*, *12*(1), 11752. https://doi.org/10.1038/s41598-022-15667-9

Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A., & Davatzikos, C. (2017). A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage*, *155*, 530-548.

Reiman, E. M., & Jagust, W. J. (2012). Brain imaging in the study of Alzheimer's disease. *Neuroimage*, *61*(2), 505-516.

Rodrigues, F., & Silveira, M. (2014). Longitudinal FDG-PET features for the classification of Alzheimer's disease. 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society,

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., & Müller, K.-R. (2016). Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, *28*(11), 2660-2673.

Samper-González, J., Burgos, N., Bottani, S., Fontanella, S., Lu, P., Marcoux, A., . . . Wen, J. (2018). Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data. *NeuroImage*, *183*, 504-521.

Smailagic, N., Vacante, M., Hyde, C., Martin, S., Ukoumunne, O., & Sachpekidis, C. (2015). 18 F-FDG PET for the early diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI). *Cochrane Database of Systematic Reviews*(1).

Spasov, S., Passamonti, L., Duggento, A., Lio, P., Toschi, N., & Initiative, A. s. D. N. (2019). A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. *Neuroimage*, *189*, 276-287.

Sun, Z., van de Giessen, M., Lelieveldt, B. P., & Staring, M. (2017). Detection of conversion from mild cognitive impairment to Alzheimer's disease using longitudinal brain MRI. *Frontiers in neuroinformatics*, *11*, 16.

Tomše, P., Jensterle, L., Grmek, M., Zaletel, K., Pirtošek, Z., Dhawan, V., . . . Trošt, M. (2017). Abnormal metabolic brain network associated with Parkinson's disease: replication on a new European sample. *Neuroradiology*, *59*(5), 507-515. https://doi.org/10.1007/s00234-017-1821-3

Wang, H., Shen, Y., Wang, S., Xiao, T., Deng, L., Wang, X., & Zhao, X. (2019). Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease. *Neurocomputing*, *333*, 145-156.

Wood, D., Cole, J., & Booth, T. (2019). NEURO-DRAM: a 3D recurrent visual attention model for interpretable neuroimaging classification. *arXiv preprint arXiv:1910.04721*.

Zhang, D., Shen, D., & Initiative, A. s. D. N. (2012). Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. *PloS one*, *7*(3), e33182.

Zhang, F., Li, Z., Zhang, B., Du, H., Wang, B., & Zhang, X. (2019). Multi-modal deep learning model for auxiliary diagnosis of Alzheimer's disease. *Neurocomputing*, *361*, 185-195.

Zhang, J., Bargal, S. A., Lin, Z., Brandt, J., Shen, X., & Sclaroff, S. (2018). Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, *126*(10), 1084-1102.

Świetlik, D., & Białowąs, J. (2019). Application of artificial neural networks to identify alzheimer's disease using cerebral perfusion SPECT data. *International journal of environmental research and public health*, *16*(7), 1303.