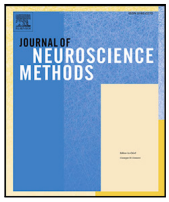
Journal of Neuroscience Methods 411 (2024) 110253

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth



A minimalistic approach to classifying Alzheimer's disease using simple and extremely small convolutional neural networks

Edvard O.S. Grødem [a,b,*], Esten Leonardsen [b,c], Bradley J. MacIntosh [a,d], Atle Bjørnerud [a], Till Schellhorn [a], Øystein Sørensen [b], Inge Amlien [b], Anders M. Fjell [b], for the Alzheimer's Disease Neuroimaging Initiative [1]

[a] Computational Radiology & Artificial Intelligence unit, Division of Radiology and Nuclear Medicine, Oslo University Hospital, 0372, Oslo, Norway
[b] Center for Lifespan Changes in Brain and Cognition, Department of Psychology, University of Oslo, 0373, Oslo, Norway
[c] Norwegian Centre for Mental Disorders Research, Oslo University Hospital & Institute of Clinical Medicine, University of Oslo, 0373, Oslo, Norway
[d] Department of Medical Biophysics, Sunnybrook Research Institute, University of Toronto, M5G 1L7, Toronto, Canada

ARTICLE INFO

Dataset link: adni.loni.usc.edu

Keywords:
Alzheimer's disease
Convolutional neural network
Mild cognitive impairment
Artificial neural network

ABSTRACT

Background: There is a broad interest in deploying deep learning-based classification algorithms to identify individuals with Alzheimer's disease (AD) from healthy controls (HC) based on neuroimaging data, such as T1-weighted Magnetic Resonance Imaging (MRI). The goal of the current study is to investigate whether modern, flexible architectures such as EfficientNet provide any performance boost over more standard architectures.
Methods: MRI data was sourced from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and processed with a minimal preprocessing pipeline. Among the various architectures tested, the minimal 3D convolutional neural network SFCN stood out, composed solely of 3x3x3 convolution, batch normalization, ReLU, and max-pooling. We also examined the influence of scale on performance, testing SFCN versions with trainable parameters ranging from 720 up to 2.9 million.
Results: SFCN achieves a test ROC AUC of 96.0% while EfficientNet got an ROC AUC of 94.9 %. SFCN retained high performance down to 720 trainable parameters, achieving an ROC AUC of 91.4%.
Comparison with existing methods: The SFCN is compared to DenseNet and EfficientNet as well as the results of other publications in the field.
Conclusions: The results indicate that using the minimal 3D convolutional neural network SFCN with a minimal preprocessing pipeline can achieve competitive performance in AD classification, challenging the necessity of employing more complex architectures with a larger number of parameters. This finding supports the efficiency of simpler deep learning models for neuroimaging-based AD diagnosis, potentially aiding in better understanding and diagnosing Alzheimer's disease.

1. Introduction

Alzheimer's disease (AD) is a neurodegenerative disease and the most common cause of dementia (Jack et al., 2018). Today 10.7% of the population over the age of 65 has dementia caused by Alzheimer's disease (Association, 2022). The cause of AD is not fully understood, and while there are multiple drugs approved by the FDA, their utility has been limited due to moderate symptom relief and severe side effects (Association, 2022; Athar et al., 2021; Loera-Valencia et al., 2019). There is a pressing need for high-precision diagnostic tools that can identify patients with a high risk of developing AD. Such a tool could be useful for selecting high-risk subjects for clinical trials. Ideally, the diagnostic should be performed before the onset of full AD, such that severe neurodegeneration has not taken place. Usually, an individual will experience mild cognitive impairment (MCI) prior to being diagnosed with AD. A large group of MCI patients remain stable and do not progress to AD (Nettiksimmons et al., 2014). These groups are often called stable MCI (sMCI) and progressive MCI (pMCI). In this study, we present a deep learning pipeline for classifying AD patients

* Corresponding author at: Computational Radiology & Artificial Intelligence unit, Division of Radiology and Nuclear Medicine, Oslo University Hospital, 0372, Oslo, Norway.
E-mail address: edvardgr@uio.no (E.O.S. Grødem).

[1] 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 in: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

https://doi.org/10.1016/j.jneumeth.2024.110253
Received 3 April 2024; Received in revised form 12 July 2024; Accepted 16 August 2024
Available online 20 August 2024

from healthy controls (HC) using T1-weighted magnetic resonance images (MRI) of the brain. AD vs HC classification provides a basis to evaluate model architecture for a classification task based on two cognitively distinct groups (Jack et al., 2008). By investigating the properties of the AD vs HC classification, we hope to gain insight that will enable sMCI vs pMCI classification.

The field of deep learning is evolving quickly. A major part of this research is conducted on 2D image classification tasks on datasets such as ImageNet (Russakovsky et al., 2015). New architectures and training regimes have greatly improved classification accuracy (Tan and Le, 2019; Canziani et al., 2016). However, MRI brain scans are 3D volumes. Often 2D architectures are expanded to 3D data by simply replacing 2D convolutions with their 3D counterparts, assuming that the techniques that improve performance in 2D natural images generalize across dimensions and domains (Uemura et al., 2020; Liang et al., 2018; Ruiz et al., 2020; Chen et al., 2019).

Peng et al. (2021) challenged this assumption by proposing the 3D CNN named Simple Fully Convolutional Network (SFCN) for brain-age prediction on T1w MRI. SFCN was specifically designed to be simple and "shallow" compared to modern architectures. In this paper, we rigorously test SFCN against the popular architectures DenseNet (Huang et al., 2017), and EfficientNet (Tan and Le, 2019) for AD vs HC classification. It is well established that, in general, the performance of a model is highly dependent on the number of parameters (He et al., 2016; Nakkiran et al., 2021). We investigate this claim explicitly by studying the performance of the SFCN as we shrink the feature width towards one.

Data leakage in machine learning refers to the phenomenon where information from the test data influences the training of the model. This issue can contribute to systematic errors and biases in reported results of AD classification as discussed in Wen et al. (2020). For example, data leakage can occur when a dataset is improperly split such that one participant with multiple assessments is part of both the training and test sets. Another typical example is using the test set for hyper-parameter selection (Kriegeskorte et al., 2009). Oversight in constructing the datasets can lead to overly optimistic classification accuracy (i.e. > 98%) (Wen et al., 2020). It is prudent to build on the AD classification literature to understand better the factors contributing to high performance. Some of the earlier findings may be inflated, in part due to sub-optimal data management. In the present study, we make use of a relatively large sample and rely on 5-fold cross-validation. We tune hyper-parameters using a validation set, avoiding prematurely exposing the models to the test set. Hyper-parameters for all architectures were found using the same search procedure, and the final test results were only generated once. We believe this minimizes the likelihood of data leakage and, subsequently, inflation of our reported results.

Our primary contribution is demonstrating that the simple and shallow SFCN architecture can achieve competitive performance with more complex architectures like DenseNet and EfficientNet in AD vs HC classification, using minimal preprocessing. Searching for deep learning architectures is computationally demanding and can consume significant research time. We show that our results are competitive with other efforts in the literature, which typically involve much more complex pipelines. Additionally, we demonstrate that good results can be obtained with very small architectures. Smaller models are less hardware-intensive, making research on AD classification, as well as clinical inference, more accessible. The models with trained weights and the dataset splits can be found at: https://github.com/CRAI-OUS/simple_ad

2. Methods

2.1. Dataset and preprocessing

We used structural T1-weighted MRIs of the brain from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (Jack et al., 2008). The ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD.

Table 1
Feature width for the configurations of SFCN. Model 7 is the "Base" model and is identical to the original SFCN model architecture.

Model	Parameters	Block 1	Block 2	Block 3	Block 4	Block 5	FcBlock 6
SFCN-0	712	2	2	2	3	3	2
SFCN-1	2677	4	4	4	6	6	4
SFCN-2	4883	4	6	6	8	8	6
SFCN-3	13.3k	4	8	8	16	16	8
SFCN-4	46.4k	4	8	16	32	32	8
SFCN-5	184.8k	8	16	32	64	64	16
SFCN-6	738.0k	16	32	64	128	128	32
SFCN-7-Base	2.95M	32	64	128	256	256	64

We used a minimal preprocessing pipeline to ensure that as much of the information in the raw data was available to the model, and to minimize the complexity of the total pipeline. Each brain scan was skullstripped with HD-BET (Isensee et al., 2019), a deep learning-driven skullstripping tool. HD-BET was chosen because it is fast ($t < 10$ s/scan) compared to other alternatives such as Freesurfer (Fischl, 2012). Rapid processing is important in making the models more accessible for clinical use. The scans were resampled to 1 mm isotropic resolution and cropped to a size of $160 \times 192 \times 160$. The top 5 percentile of the intensities were clipped and each scan was normalized to the interval [0, 1]. Clipping the top intensities ensures that noisy outliers do not shift the contrast when rescaling intensities to [0, 1].

Only HC and AD patients were considered in this paper, which amounted to 1597 subjects scanned for a total of 5054 MRI sessions. The data were divided into 5 folds such that all sessions of each subject were contained in one fold. Each fold was stratified so that the male-to-female ratio, average age, and diagnosis ratio were matched for each fold. We created 5 data splits where the train and validation set consisted of 4 folds, and the remaining fold was used for testing. 10% of the 4 train–validation folds were used for validation, which was also balanced with respect to the training set. For testing and validation, one random session was used for each subject, as opposed to training when all available sessions were used. This was to avoid a few subjects with a lot of sessions influencing the test results too much. We refer to split 0 as the combination of the training, validation, and test set that uses fold 0 as the test set. An overview of the folds and the stratified values can be seen in Table A.1.

2.2. Architectures

In this paper, we compare three architectures: DenseNet (Huang et al., 2017) was chosen since it is a popular architecture in medical classification (Liang et al., 2018; Uemura et al., 2020; Ruiz et al., 2020). EfficientNet (Tan and Le, 2019) was chosen since it can be scaled to a small size, which makes it practical for training on 3D T1 scans. While published in 2019, it is still competitive on ImageNet classification with a Top-1 accuracy of 84.3%.

For these two architectures, we used the 3D implementation provided by Monai (Cardoso et al., 2022). We used DenseNet121 and EfficientNet-B0 as these were the smallest versions of the models, and we therefore could train the models using a single GPU per model.

Our focus in this paper is the Simple Fully Convolutional Network (SFCN) (Peng et al., 2021) architecture. It was originally designed for brain-age estimation, for which it has achieved state-of-the-art performance on (Peng et al., 2021; Gong et al., 2021; Leonardsen et al., 2022). It has also been successfully applied to AD classification (Leonardsen et al., 2023; Gupta et al., 2023). As the name suggests, the architecture

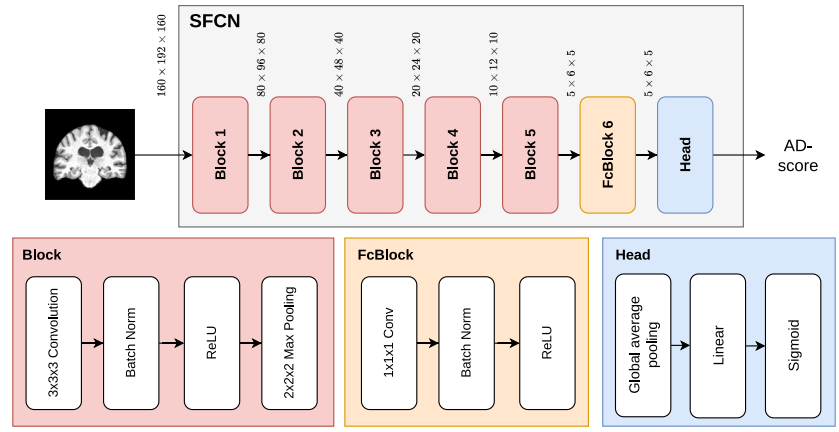


Fig. 1. Diagram of the SFCN architecture. The spatial dimension of the feature space is shown between each block. The feature width for each feature space can be found in Table 1.

was designed to be simple. SFCN consists of 6 blocks of $3 \times 3 \times 3$ convolution, batch normalization, ReLU activation, and max pooling. At the final layer, global average pooling is performed, followed by a single linear layer. A diagram of the architecture can be seen in Fig. 1.

We investigated how the number of trainable parameters affected the predictive performance. SFCN was down-scaled by keeping the depth constant and dividing the width of each layer by a multiple of 2. For the smallest models, we deviated from this rule to avoid layers of width 1. The models were ranked by their number of parameters from SFCN-0 to SFCN-7. SFCN-7 is the base model, with the original size as defined by Peng et al. (2021). The configuration of the down-scaled SFCN models can be seen in Table 1.

2.3. Hyper parameter selection and training the models

The performance of deep learning models is sensitive to the hyper-parameters of the models. Different architectures might benefit from different hyper-parameters. For a fair comparison between different architectures, we therefore performed a grid-based hyper-parameter search for all architectures on the training and validation set of split 0, searching for an optimal optimizer, learning rate, and weight decay. In order to fit the training on a single GPU, we fixed the batch size to 4. For each architecture, the hyper-parameters that gave the highest Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) on the validation set were selected. Each model was then trained with optimal hyper-parameters on the other 4 splits and tested on their respective test sets. As optimizers, we tested both AdamW (Loshchilov and Hutter, 2017) and Stochastic Gradient Descent (SGD). For AdamW we used $\beta_1 = 0.9, \beta_2 = 0.95$. No momentum was used with SGD. The models were trained for 50 epochs with binary cross-entropy loss. The learning rate was scheduled using linear warmup up to epoch 10, followed by cosine decay. No weight decay was used for normalizing layers and bias terms (Brock et al., 2021; He et al., 2022; Jia et al., 2018). Since a model might benefit from early stopping, two sets of the trainable parameters were saved for each training session, one from the last epoch and one for the epoch with the highest validation ROC AUC. We used the checkpoint of the last epoch as our default but investigated the properties of the best ROC AUC checkpoint. The hyper-parameters of the grid search can be found in Table A.2 and the optimal parameters for each model in Table A.3. A summary of the method for training SFCN-Base on AD vs HC classification can be found in Fig. 2.

When training smaller versions of SFCN, we used the best hyper-parameters found for SFCN-Base. However, we changed the batch size from 4 to 16, and we used a linear scaling law to adapt the learning rate to the new batch size (Goyal et al., 2017). Since the smaller configuration of SFCN could benefit from less weight decay, a hyper-parameter search for weight decay was performed. The hyper-parameters for the smaller versions of SFCN can be found in Table A.4.

2.4. Metrics

We choose to rely on ROC AUC as our metric for hyper-parameter selection as well as model evaluation. ROC AUC is in many ways considered a better performance metric than accuracy and related measures such as sensitivity and specificity (Dinga et al., 2019). This is partly due to ROC AUC being invariant to class imbalance. Furthermore, we consider it a clinical task to decide what rate of false positives and false negatives can be tolerated in clinical settings. We do, however, report the accuracies of our models, to facilitate an intuitive interpretation of model performance and enable comparisons with other methods.

3. Results

3.1. AD vs HC classification

We found that EfficientNet, DenseNet and SFCN-Base performed similar on the test sets, with an average ROC AUC of 94.9%, 94.9%, and 96.0%, respectively. SFCN had a slightly higher ROC AUC than the two other architectures in 4 out of 5 test folds. A comparison of the ROC AUCs and accuracies across the architectures can be seen in Fig. 3. Using the checkpoint with the highest validation ROC AUC yielded a similar result with a test ROC AUC of 95.2% for EfficientNet, 94.7% for DenseNet, and 95.7% for SFCN.

Next, we compare our results to the results from other publications. From the list of AD classification publications with “no data leakage” by Wen et al. (2020) we selected the 5 publications with the highest accuracy. In addition, we selected a few newer publications on AD classification. The complete comparison can be found in Table 2. Overall, our simple pipeline performed competitively with state-of-the-art methods.

3.2. Qualitative model characteristics

We investigated the correlation of the model predictions. Using the pre-sigmoid output of the models, we calculated the Pearson Correlation Coefficient (PCC) for pairs of the architectures. We found the models to be highly correlated, with PCC for DenseNet and EfficientNet of 0.88, SFCN and EfficientNet of 0.90, and SFCN and DenseNet of 0.91.

Simple AD classification

Skullstrip T1 brain scans and resample to 1mm isotropic resolution. Crop the top 5 percentile voxel intensities and scale intensities to [0, 1]. Train the SFCN architecture on AD vs HC:

- Optimizer: SGD
- Learning rate scheduler: Linear warmup to 20% of total epochs followed by cosine decay.
- Data augmentation: Random translation up to 6 voxels, random flip over the medial plane.
- Grid search: Epochs, learning rate, and weight decay.

Fig. 2. Summary of the deep learning pipeline for AD vs HC classification using a simple convolutional neural network.

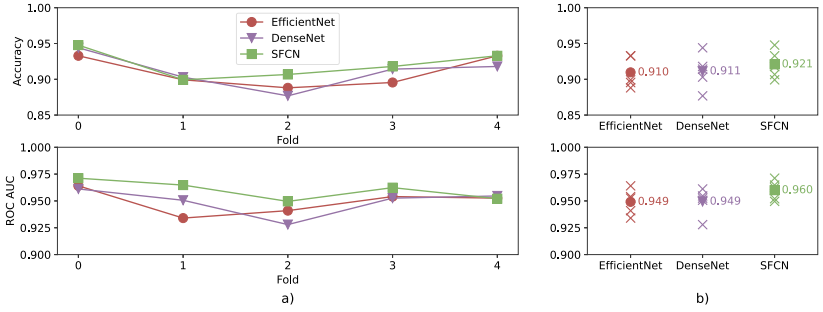


Fig. 3. (a) AD vs HC test accuracy and ROC AUC for EfficientNet, DenseNet, and SFCN-Base for each split. (b) Comparison of average test accuracy and ROC AUC for AD versus HC using EfficientNet, DenseNet, and SFCN-Base.

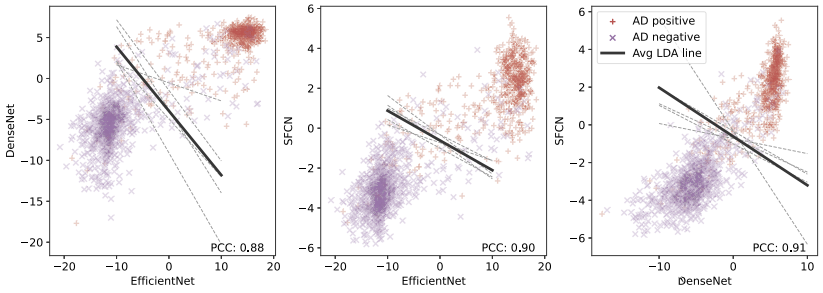


Fig. 4. The pre-sigmoid output of each model architecture plotted up against each other. The value on the axis represent the value of the output before the final sigmoid activation models. Higher values represent that the model given the input has higher confidence in the AD class. Lower values represent higher confidence in the HC class. The scatter plot is created by combining the 5 test sets. The pre-sigmoid outputs are from the 5 models that trained on the train set belonging to each test set. Pearson correlation coefficient (PCC) for each model pair is shown in the right corner. The black lines are the class separation lines from the linear discriminant analysis (LDA) models when fitted on the pre-sigmoid output of the 5 validation sets.

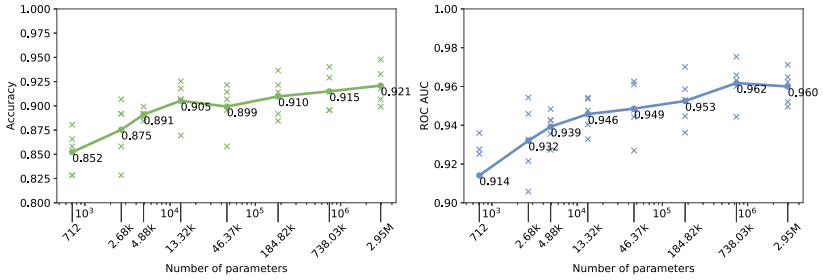


Fig. 5. AD vs HC classification test accuracy and ROC AUC for SFCN of different sizes. The lines mark the average ROC AUC and accuracy and the points mark the accuracy and ROC AUC for each of the split.

Table 2
Comparison of AD vs HC deep-learning classifiers on ADNI data. Methods marked with a '+' indicate "no data leakage" as per Wen et al. (2020). **Table abbreviations:** BA: Balanced Accuracy. **Modalities:** T1w: T1 weighted MRI, MD: Mean diffusivity MRI, FDG-PET: Fluorodeoxyglucose positron emission. **Preprocessing:** The main steps of the preprocessing pipelines of the methods. If given in the method, the software used in preprocessing is given in parentheses. R: Linear registration, NR: Nonlinear Registration, N: Normalization or bias field correction , SS: Skullstripping, Seg: Segmentation, LMD: Custom landmark detection, None: No preprocessing was performed. **Preprocessing software:** SMP: Statistical Parametric Mapping (Ashburner et al., 2012), FS: Freesurfer (Fischl, 2012), DPARSF: Data Processing Assistant for Resting-State fMRI (Yan and Zang, 2010), HD-BET: (Isensee et al., 2019), FSL: MRIB Software Library (Woolrich et al., 2009), ANTs: Advanced Normalizations Tools (Tustison et al., 2021), DARTEL: Diffeomorphic Anatomical Registration Through Exponentiated Lie algebra (Goto et al., 2013). **Models Types**: 2.5D CNN: CNN, which processes images as slices using 2D convolution and integrates the information across slices. 3D CNN: CNN with 3D convolution operating on the full image, 3D ROI CNN: 3D CNN that operates on subregions of the image. VAE: Variational Autoencoder, FEAT MLP: Multilayer perceptron that processes features extracted from the image.

	ROC AUC	Accuracy	Modalities	Preprocessing	Model Type
Gupta et al. (2023)	–	88.6% (BA)	T1w	R	2.5D CNN
Zhang et al. (2022)	96.1%	93.2%	T1w	N(ANTs)-SS(ANTS)-R(ANT)	2.5D CNN
Cobbinah et al. (2022)	94.9%	93.1%	T1w	None	VAE,3D CNN
Lu et al. (2022)	96.3%	90.0%	T1w	SEG(DPARSF)-NR[DARTEL]	3D CNN
Wen et al. (2020)	–	89.0% (BA)	T1w	N(ANTs)-SS(ANTs)-R(ANTs)	3D ROI CNN
Lu et al. (2018)	–	84.6%	T1w, FDG-PET	SEG(FS)-NR(LDDMM)	FEAT MLP
Liu et al. (2018a) +	95.0%	91.2%	T1w, FDG-PET	N-S (Wang et al., 2011), R(FSL)	3D ROI CNN
Liu et al. (2018b) +	95.9%	91.1%	T1w	LMD	3D ROI CNN
Aderghal et al. (2018)+	–	90.0%	T1w, MD	N(SPM)-R(SPM)-SS(SPM)	2.5 CNN
Bäckström et al. (2018)+	–	90.1%	T1w	N(FS)-SS(FS)	3D CNN
Li et al. (2018)+	92.4%	89.5%	T1w	SS-N-R(FSL)	2.5 CNN
SFCN-Base (Our implementation)	96.0%	92.1%	T1w	SS(HD-BET)	3D CNN

To further test if different models picked up different useful features we used Linear Discriminant Analysis (LDA) to construct new classifiers from the output of the three model architectures. We fitted 5 LDAs on the validation sets and did inference on the test data. In Fig. 4 the pre-sigmoid output of the models can be seen plotted against each other together with the LDA classification line. The LDA models that used all three architectures had an average ROC AUC of 96.19%, an increase of only 0.19% compared to SFCN.

3.3. Small architecture performance

Next, we investigated how the size of SFCN affects the test accuracy. Using the same hyper-parameters as SFCN-Base we tested 7 smaller architectures. We found that SFCN-3 with 13k parameters achieved a ROC AUC of 94.6%. SFCN-3 has only 0.44% of the parameters of SFCN-Base, but the relative reduction in ROC AUC is only 1.45%. We further observed that the extremely small SFCN-0 width 712 parameters and a feature width of [2, 2, 2, 3, 3, 2] still have a respectable 91.4% ROC AUC. The performance metrics as a function of model size can be seen in Fig. 5.

We visualized all the feature maps from SFCN-0 of 4 individuals. We selected the two subjects from the test set for which the model gave the highest and lowest AD scores. The feature maps are displayed in Fig. 6.

4. Discussion

We investigated whether modern architectures, proven effective on ImageNet, increase the AD vs HC classification performance over simpler architectures for structural MRI data. Our results show that the three architectures, SFCN-Base, DenseNet, and EfficientNet, achieved approximately the same ROC AUC, with a slight advantage to SFCN-Base. Considering the extensive efforts in designing these architectures, this finding is unexpected. SFCN represents a simplistic architecture and might be considered outdated. This raises questions about why DenseNet and EfficientNet, which are effective classifiers for 2D tasks, do not exhibit the same efficiency for AD classification using 3D brain MRI. Furthermore, why do our results remain competitive with other studies that employ pretraining or carefully designed pipelines and architectures?

Although we cannot present strong evidence for why SFCN is sufficient for AD classification, we can hypothesize. DenseNet and EfficientNet are architectures built to perform well on ImageNet. ImageNet has 1000 diverse labels, many of which have very distinct characteristics. In contrast, AD classification only has two classes, with relatively minor visual differences distinguishing the AD group from the HC group, especially compared to the visual diversity in ImageNet. The images are also centered and always facing the same direction, simplifying the classification task. We see that on the AD classification task the model capacity can be very small with SFCN-0 having only 712 parameters and still reaching an ROC AUC of 91.4%. On ImageNet classification, the drop in performance is observed in models with many orders of magnitude more parameters than the models in our experiments. The ImageNet top-1 accuracy of DenseNet drops from 77.85% to 74.8% when the model size changes from 33M to 7M parameters (Huang et al., 2017).

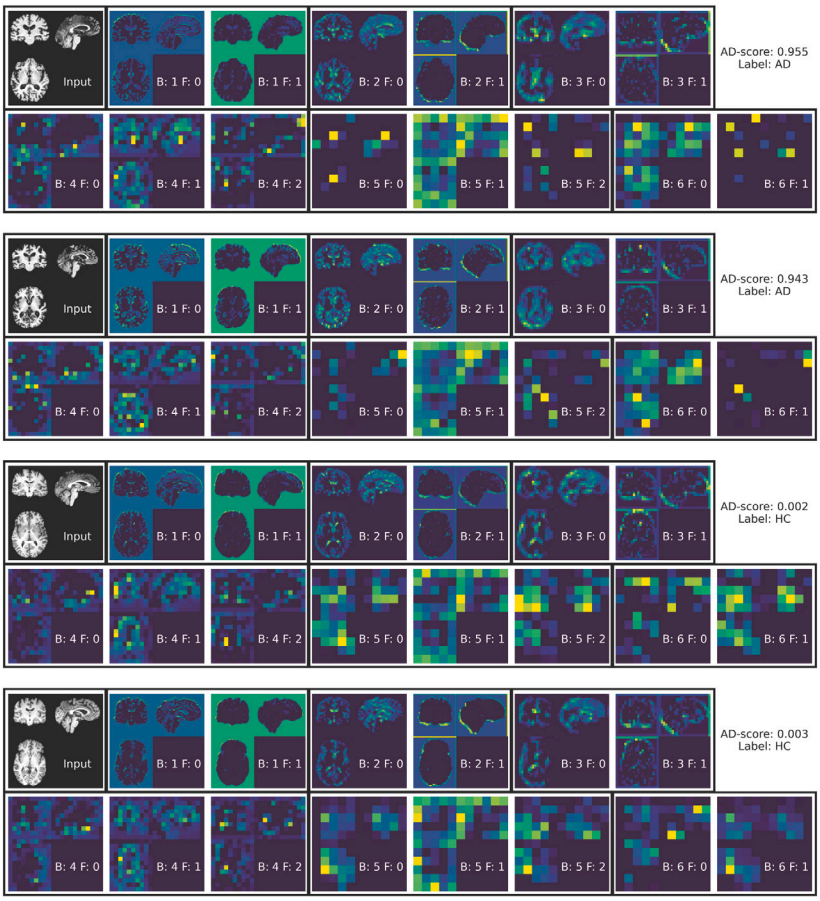


Fig. 6. All feature maps of SFCN-0. The two subjects with the highest and lowest AD-score in the test set are displayed. Each 3D feature map is displayed as 3 orthogonal slices along with the index of the Block(B) and Feature (F). Zoom in to see details.

A visual inspection of the feature maps of SFCN-0 in Fig. 6 shows that the network extracts regions of Cerebrospinal fluid (CSF) in the first layers while discarding the texture in the rest of the brain. This may be a hint as to why SFCN is able to achieve a high ROC AUC even with a tiny model. Since CSF is darker then brain tissue in T1w MR images, a simple threshold is sufficient for an estimate of brain atrophy. This threshold function can be easily implemented by a linear layer and a ReLU function, similar to the first layer of SFCN. Understanding how SFCN processes these regions further down the network is challenging, due to the shrinking spatial size and convoluted interactions between the features. However, since one can see all the features of the first layer, we believe that the CSF segmentation-like behavior is central to further processing.

We see a potential in utilizing the small version of SFCN, and similar simple models, for new applications. It could be possible to explicitly analyze the features of such a model to understand what it has learned. Feature analysis is usually not feasible when a model has hundreds or thousands of features, but with a model with no more than 10 features in each layer, visualizing and interpreting them is a realistic possibility.

Other applications can be in a clinical setting on a machine without a GPU. In this case, a slight decrease in performance may be acceptable if the model is small enough to enable rapid analysis on a single CPU.

5. Conclusion

In this paper, we have demonstrated that a simple preprocessing pipeline and the simple architecture SFCN yielded competitive results on AD vs HC classification relative to two other larger and more sophisticated model architectures. We found that the SFCN model architecture could be scaled to a surprisingly small size, with only a small deterioration in performance. This work suggests that a simple CNN with minimal preprocessing could serve as a viable baseline when testing new machine learning pipelines for AD-related classification.

Ethics

The clinical experiments of ADNI have been approved by the ethics board selected by the participating institutes of ADNI. The Office for Human Research Protections (OHRP) has reviewed and approved each ethics board. Informed consent from all participants in ADNI has been conducted in accordance with US 21 CFR 50.25, the Tri-Council Policy Statement: Ethical Conduct of Research Involving Humans and the Health Canada and ICH Good Clinical Practice. All methods in the current study were carried out following the guidelines and regulations of ADNI.

Data and code availability

The data used in the current study can be accessed on requested from ADNI at https://adni.loni.usc.edu/data-samples/access-data/. The code is available at https://github.com/CRAI-OUS/simple_ad.

Table A.1
The stratification of the splits. All sessions of ADNI with HC and AD were used. The ratio of AD to HC changes from training to validation and test since AD subjects on average have fewer sessions.

	Subjects	Sessions	AD/HC	M/F	Age
Train	[857, 857, 857, 857, 857]	[2868, 2882, 2885, 2873, 2839]	[0.34, 0.35, 0.33, 0.35, 0.35]	[0.5, 0.51, 0.51, 0.51, 0.5]	[75.78, 75.82, 75.71, 75.77, 75.74]
Validation	[215, 215, 215, 215, 215]	[215, 215, 215, 215, 215]	[0.4, 0.4, 0.4, 0.4, 0.4]	[0.49, 0.49, 0.48, 0.49, 0.49]	[74.64, 74.62, 74.77, 74.63, 74.6]
Test	[268, 268, 268, 268, 268]	[268, 268, 268, 268, 268]	[0.4, 0.4, 0.4, 0.4, 0.4]	[0.49, 0.49, 0.49, 0.49, 0.49]	[74.67, 74.78, 74.75, 74.59, 74.71]

CRediT authorship contribution statement

Edvard O.S. Grødem: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Esten Leonardsen:** Writing – review & editing, Methodology, Conceptualization. **Bradley J. MacIntosh:** Writing – review & editing, Supervision, Conceptualization. **Atle Bjørnerud:** Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition, Conceptualization. **Till Schellhorn:** Writing – review & editing, Methodology, Conceptualization. **Øystein Sørensen:** Writing – review & editing, Methodology, Formal analysis. **Inge Amlien:** Writing – review & editing, Resources, Data curation. **Anders M. Fjell:** Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare no competing interests.

Data availability

The data is available at from ADNI at adni.loni.usc.edu.

Acknowledgments

The project was supported by a grant from the South-Eastern Norway Regional Health Authority (HSØ- 2021079). 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, United States, the National Institute of Biomedical Imaging and Bioengineering, United States, 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.

Appendix A

See Tables A.1–A.4.

Table A.2
Values used during the grid search over optimizer, learning rate, and weight decay for each of the main architectures, EfficientNet, DenseNet, and SFCN. Different learning rates were used for SDG and AdamW.

Parameters	Values	
Optimizer	SDG	AdamW
Learning rate	[0.005, 0.1, 0.05]	[0.001, 0.0003, 0.0001]
Weight Decay	[0.001, 0.01, 0.1]	[0.001, 0.01, 0.1]

Table A.3
The best hyper-parameters found after the grid search for SFCN, DenseNet, and EfficientNet.

Model type	Learning rate	Weight decay	Optimizer	Batch size
EfficientNet	0.001	0.01	AdamW	4
DenseNet	0.005	0.01	SGD	4
SFCN	0.005	0.1	SGD	4

Table A.4
The best parameters for SFCN after grid search for the different number of parameters. Model 0 used a batch size of 8 in order to avoid an unknown CUDA error with a batch size of 16.

Model	Learning rate	Weight decay	Optimizer	Batch size
SFCN-0	0.05	0.01	SGD	8
SFCN-1	0.05	0.01	SGD	16
SFCN-2	0.05	0.0001	SGD	16
SFCN-3	0.02	0.01	SGD	16
SFCN-4	0.05	0.01	SGD	16
SFCN-5	0.05	0.01	SGD	16
SFCN-6	0.02	0.1	SGD	8
SFCN-7-Base	0.005	0.1	SGD	4

References

Aderghal, K., Khvostikov, A., Krylov, A., Benois-Pineau, J., Afdel, K., Catheline, G., 2018. Classification of Alzheimer disease on imaging modalities with deep CNNs using cross-modal transfer learning. In: 2018 IEEE 31st International Symposium on Computer-Based Medical Systems. CBMS, IEEE, pp. 345–350.

Ashburner, J., Barnes, G., Chen, C., Daunizeau, J., Flandin, G., Friston, K., Gitelman, D., Kiebel, S., Kilner, J., Litvak, V., et al., 2012. SPM8 manual. Funct. Imaging Lab. Inst. Neurol..

Association, A., 2022. 2022 Alzheimer's disease facts and figures. Alzheimer's & Dementia 18 (4), 700–789. http://dx.doi.org/10.1002/alz.12638.

Athar, T., Al Balushi, K., Khan, S.A., 2021. Recent advances on drug development and emerging therapeutic agents for Alzheimer's disease. Mol. Biol. Rep. 48 (7), 5629–5645.

Bäckström, K., Nazari, M., Gu, I.Y.-H., Jakola, A.S., 2018. 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, IEEE, pp. 149–153.

Brock, A., De, S., Smith, S.L., Simonyan, K., 2021. High-performance large-scale image recognition without normalization. In: International Conference on Machine Learning. PMLR, pp. 1059–1071.

Canziani, A., Paszke, A., Culurciello, E., 2016. An analysis of deep neural network models for practical applications. arXiv preprint arXiv:1605.07678.

Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al., 2022. MONAI: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701.

Chen, S., Ma, K., Zheng, Y., 2019. Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625.

Cobbinah, B.M., Sorg, C., Yang, Q., Ternblom, A., Zheng, C., Han, W., Che, L., Shao, J., 2022. Reducing variations in multi-center Alzheimer's disease classification with convolutional adversarial autoencoder. Med. Image Anal. 82, 102585.

Dinga, R., Penninx, B.W., Veltman, D.J., Schmaal, L., Marquand, A.F., 2019. Beyond accuracy: measures for assessing machine learning models, pitfalls and guidelines. BioRxiv 743138.

Fischl, B., 2012. FreeSurfer. Neuroimage 62 (2), 774–781.
Gong, W., Beckmann, C.F., Vedaldi, A., Smith, S.M., Peng, H., 2021. Optimising a simple fully convolutional network for accurate brain age prediction in the PAC 2019 challenge. Front. Psychiatry 12, 627996.
Goto, M., Abe, O., Aoki, S., Hayashi, N., Miyati, T., Takao, H., Iwatsubo, T., Yamashita, F., Matsuda, H., Mori, H., et al., 2013. Diffeomorphic anatomical registration through exponentiated lie algebra provides reduced effect of scanner for cortex volumetry with atlas-based method in healthy subjects. Neuroradiology 55, 869–875.
Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K., 2017. Accurate, large minibatch sgd: Training ImageNet in 1 hour. arXiv preprint arXiv:1706.02677.
Gupta, U., Chattopadhyay, T., Dhinagar, N., Thompson, P.M., Ver Steeg, G., 2023. Transferring models trained on natural images to 3D MRI via position encoded slice models. In: 2023 IEEE 20th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1–5.
He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.
Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708.
Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H.-P., Heiland, S., Wick, W., et al., 2019. Automated brain extraction of multisequence MRI using artificial neural networks. Hum. Brain Mapp. 40 (17), 4952–4964.
Jack, Jr., C.R., Bennett, D.A., Blennow, K., Carrillo, M.C., Dunn, B., Haeberlein, S.B., Holtzman, D.M., Jagust, W., Jessen, F., Karlawish, J., et al., 2018. NIA-AA research framework: toward a biological definition of Alzheimer's disease. Alzheimer's & Dementia 14 (4), 535–562.
Jack, Jr., C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L. Whitwell, J., Ward, C., et al., 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imaging: An Off. J. Int. Soc. Magn. Reson. Med. 27 (4), 685–691.
Jia, X., Song, S., He, W., Wang, Y., Rong, H., Zhou, F., Xie, L., Guo, Z., Yang, Y., Yu, L., et al., 2018. Highly scalable deep learning training system with mixed-precision: Training ImageNet in four minutes. arXiv preprint arXiv:1807.11205.
Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S., Baker, C.I., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nature Neurosci. 12 (5), 535–540.
Leonardsen, E.H., Peng, H., Kaufmann, T., Agartz, I., Andreassen, O.A., Celius, E.G., Espeseth, T., Harbo, H.F., Høgestøl, E.A., de Lange, A.-M., et al., 2022. Deep neural networks learn general and clinically relevant representations of the ageing brain. NeuroImage 256, 119210.
Leonardsen, E.H., Persson, K., Grødem, E., Dinsdale, N., Schellhorn, T., Roe, J.M., Vidal-Piñeiro, D., Sørensen, Ø., Kaufmann, T., Westman, E., et al., 2023. Characterizing personalized neuropathology in dementia and mild cognitive impairment with explainable artificial intelligence. medRxiv 2006–2023.
Li, F., Liu, M., Initiative, A.D.N., et al., 2018. Alzheimer's disease diagnosis based on multiple cluster dense convolutional networks. Comput. Med. Imaging Graph. 70, 101–110.
Liang, S., Zhang, R., Liang, D., Song, T., Ai, T., Xia, C., Xia, L., Wang, Y., 2018. Multimodal 3D DenseNet for IDH genotype prediction in gliomas. Genes 9 (8), 382.
Liu, M., Cheng, D., Wang, K., Wang, Y., 2018a. Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis. Neuroinformatics 16 (3), 295–308.
Liu, M., Zhang, J., Adeli, E., Shen, D., 2018b. Landmark-based deep multi-instance learning for brain disease diagnosis. Med. Image Anal. 43, 157–168.
Loera-Valencia, R., Cedazo-Minguez, A., Kenigsberg, P., Page, G., Duarte, A., Giusti, P., Zusso, M., Robert, P., Frisoni, G., Cattaneo, A., et al., 2019. Current and emerging avenues for Alzheimer's disease drug targets. Journal of Internal Medicine 286 (4), 398–437.
Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.
Lu, B., Li, H.-X., Chang, Z.-K., Li, L., Chen, N.-X., Zhu, Z.-C., Zhou, H.-X., Li, X.-Y., Wang, Y.-W., Cui, S.-X., et al., 2022. A practical Alzheimer's disease classifier via brain imaging-based deep learning on 85,721 samples. J. Big Data 9 (1), 1–22.
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. Sci. Rep. 8 (1), 5697.
Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., Sutskever, I., 2021. Deep double descent: Where bigger models and more data hurt. J. Stat. Mech. Theory Exp. 2021 (12), 124003.
Nettiksimmons, J., DeCarli, C., Landau, S., Beckett, L., Initiative, A.D.N., et al., 2014. Biological heterogeneity in ADNI amnestic mild cognitive impairment. Alzheimer's & Dementia 10 (5), 511–521.
Peng, H., Gong, W., Beckmann, C.F., Vedaldi, A., Smith, S.M., 2021. Accurate brain age prediction with lightweight deep neural networks. Med. Image Anal. 68, 101871.
Ruiz, J., Mahmud, M., Modasshir, M., Shamim Kaiser, M., Alzheimer's Disease Neuroimaging Initiative, f.t., et al., 2020. 3D DenseNet ensemble in 4-way classification of Alzheimer's disease. In: International Conference on Brain Informatics. Springer, pp. 85–96.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) 115 (3), 211–252. http://dx.doi.org/10.1007/s11263-015-0816-y.
Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 6105–6114.
Tustison, N.J., Cook, P.A., Holbrook, A.J., Johnson, H.J., Muschelli, J., Devenyi, G.A., Duda, J.T., Das, S.R., Cullen, N.C., Gillen, D.L., et al., 2021. The ANTsX ecosystem for quantitative biological and medical imaging. Sci. Rep. 11 (1), 9068.
Uemura, T., Näppi, J.J., Hironaka, T., Kim, H., Yoshida, H., 2020. Comparative performance of 3D-DenseNet, 3D-ResNet, and 3D-VGG models in polyp detection for CT colonography. In: Medical Imaging 2020: Computer-Aided Diagnosis. 11314, SPIE, pp. 736–741.
Wang, Y., Nie, J., Yap, P.-T., Shi, F., Guo, L., Shen, D., 2011. Robust deformable-surface-based skull-stripping for large-scale studies. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011: 14th International Conference, Toronto, Canada, September 18-22, 2011, Proceedings, Part III 14. Springer, pp. 635–642.
Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., Dormont, D., Durrleman, S., Burgos, N., Colliot, O., et al., 2020. Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. Med. Image Anal. 63, 101694.
Woolrich, M.W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., Beckmann, C., Jenkinson, M., Smith, S.M., 2009. Bayesian analysis of neuroimaging data in FSL. Neuroimage 45 (1), S173–S186.
Yan, C., Zang, Y., 2010. DPARSF: a MATLAB toolbox for "pipeline" data analysis of resting-state fMRI. Front. Syst. Neurosci. 4, 1377.
Zhang, F., Pan, B., Shao, P., Liu, P., Shen, S., Yao, P., Xu, R.X., Initiative, A.D.N., et al., 2022. A single model deep learning approach for Alzheimer's disease diagnosis. Neuroscience 491, 200–214.