From longitudinal measurements to image classification: Application to longitudinal MRI in Alzheimer's disease

Samaneh Abolpour Mofrad (✉ sam@hvl.no)
Western Norway University of Applied Sciences

Hauke Bartsch
University of Bergen

Alexander Selvikvåg Lundervold
Western Norway University of Applied Sciences

Article

Keywords:

Posted Date: April 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1541040/v1

From longitudinal measurements to image classification: Application to longitudinal MRI in Alzheimer's disease

Samaneh A. Mofrad[1,3,*], Hauke Bartsch[2,3], Alexander S. Lundervold[1,3], and for the Alzheimer's Disease Neuroimaging Initiative[**]

[1]Dept. of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, 5020, Norway
[2]Dept. of Informatics, University of Bergen, Bergen, Norway
[3]MMIV, Dept. of Radiology, Haukeland University Hospital, Bergen, Norway
[*]Corresponding author: Samaneh A. Mofrad, Pb. 7030, 5020 Bergen, Norway, sam@hvl.no
[**]Data included in this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

ABSTRACT

We propose a novel method of constructing representations of multiple one-dimensional longitudinal measurements as two-dimensional grey-scale images. This can be used to turn classification problems from longitudinal settings into simpler image classification problems, allowing for the application of newer deep learning methods on longitudinal measurements. Our approach is applicable to situations with balanced or imbalanced longitudinal data sets, and where there are missing data at some time points. To evaluate our approach, we apply it to an important and challenging task: the prediction of dementia from brain volume trajectories derived from longitudinal MRI. We construct an ensemble of convolutional neural network models to classify two groups of subjects: those diagnosed with mild cognitive impairment at all examinations (stable MCI) versus those starting out as MCI but later converting to Alzheimer's disease (converted AD). Models were trained on image representations derived from $N = 736$ subjects sourced from the ADNI database ($471/265$ sMCI/cAD). We obtained an accuracy of a resulting ensemble model of 76%, measured on an independent test set. Our approach is simple and easy to apply but competitive (in terms of accuracy) with results reported in other machine learning approaches with similar classification on comparable tasks. This indicates that our approach can lead to useful representations of longitudinal data.

Introduction

Deep neural networks form the basis for a wide range of state-of-the-art medical image analysis tasks and have drawn a lot of interest over the past years[1,2]. While deep learning techniques are behind successful applications in various fields, in many cases it is difficult to find an appropriate representation of the input data used to train deep learning models, that highlights the useful predictive features in the data.

Longitudinal data is one such case. Here measurements are taken repeatedly through time with multiple outcomes at each time point. Some of these difficulties are due to the inherent properties of longitudinal data, like inter-correlation between the set of observations of one subject[3] and the unbalanced observations for subjects[4].

Motivated by studies where time-series or speech recognition data were represented as images[5–7], we propose a pipeline for producing two-dimensional (2D) images from longitudinal data. This enables the use of well-studied techniques from deep learning for two-dimensional image classification.

First, we gathered all the data collected from each subject in a matrix so that one axis is associated with time points and the other with the corresponding values of those time points. Then, we scaled the columns' values separately to get a standard range for each variable. Next, we mapped each scaled matrix to a gray-scale image, so that the pixel intensity represents the matrix values. The 2D images can then be used to train a deep neural network classifier.

To evaluate our proposed pipeline in a concrete setting, we used a longitudinal data source with a large number of subjects, which contains ascending, descending, and categorical data, where the number of time points and the length between them varies significantly. We used data from subjects diagnosed with various levels of dementia: Alzheimer's Disease (AD), which is a common irreversible neurodegenerative disorder characterized by a cognitive impairment that gradually worsens over time[8,9], and Mild Cognitive Impairment (MCI), which is a transitional state from normal cognition to dementia [10]. We ran the experiment on two subgroups labeled as stable MCI (sMCI), who were diagnosed as MCI at all scans, and converged AD (cAD), who were diagnosed as MCI at the beginning but later developed AD.

After preparing 2D images for sMCI and cAD subjects, we investigated the effect of data augmentation techniques, model architectures, and hyper-parameter selection. We used the results from these investigations to construct an ensemble model that can classify conversion to AD versus stable MCI with an average accuracy of 76%. This is a competitive result when compared with other approaches, indicating the usefulness of the proposed pipeline also for other problems related to longitudinal measurement.

Methods

Data

Data 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, with an overall goal to validate biomarkers for use in clinical treatment trials for patients with AD. The study was approved by the Institutional Review Boards at each ADNI site (see the full list here: http://adni.loni.usc.edu). Informed consent was obtained from all subjects prior to enrollment. All methods were carried out in accordance with relevant guidelines and regulations. The present study was approved by the ADNI Publication Committee (ADNI DPC).

We constructed a longitudinal data set by selecting all the subjects from all the three ADNI phases (ADNI1, ADNI Go and ADNI3) that had at least three MRI scans. Our data set consists of 736 subjects (female/male: 299/437) with a total of 3956 MRI scans (see Table 1 for details). We considered two longitudinal labels based on the ADNI diagnoses, as defined in our previous work[11]. If subjects were defined with MCI at all visits, we labeled them as stable MCI (sMCI) and when subjects converting from MCI to AD we labeled them as converted AD (cAD). These groups can be used to uncover features associated with progressing from MCI to AD.

Group	Subjects	MRI	Gender (f/m)
sMCI	471	2424	195/276
cAD	265	1532	104/161
	736	3956	299/437

Table 1. Longitudinal subjects and MR images: total number of subjects, MR images, and gender distribution within two subgroups.

Availability of Data and Materials

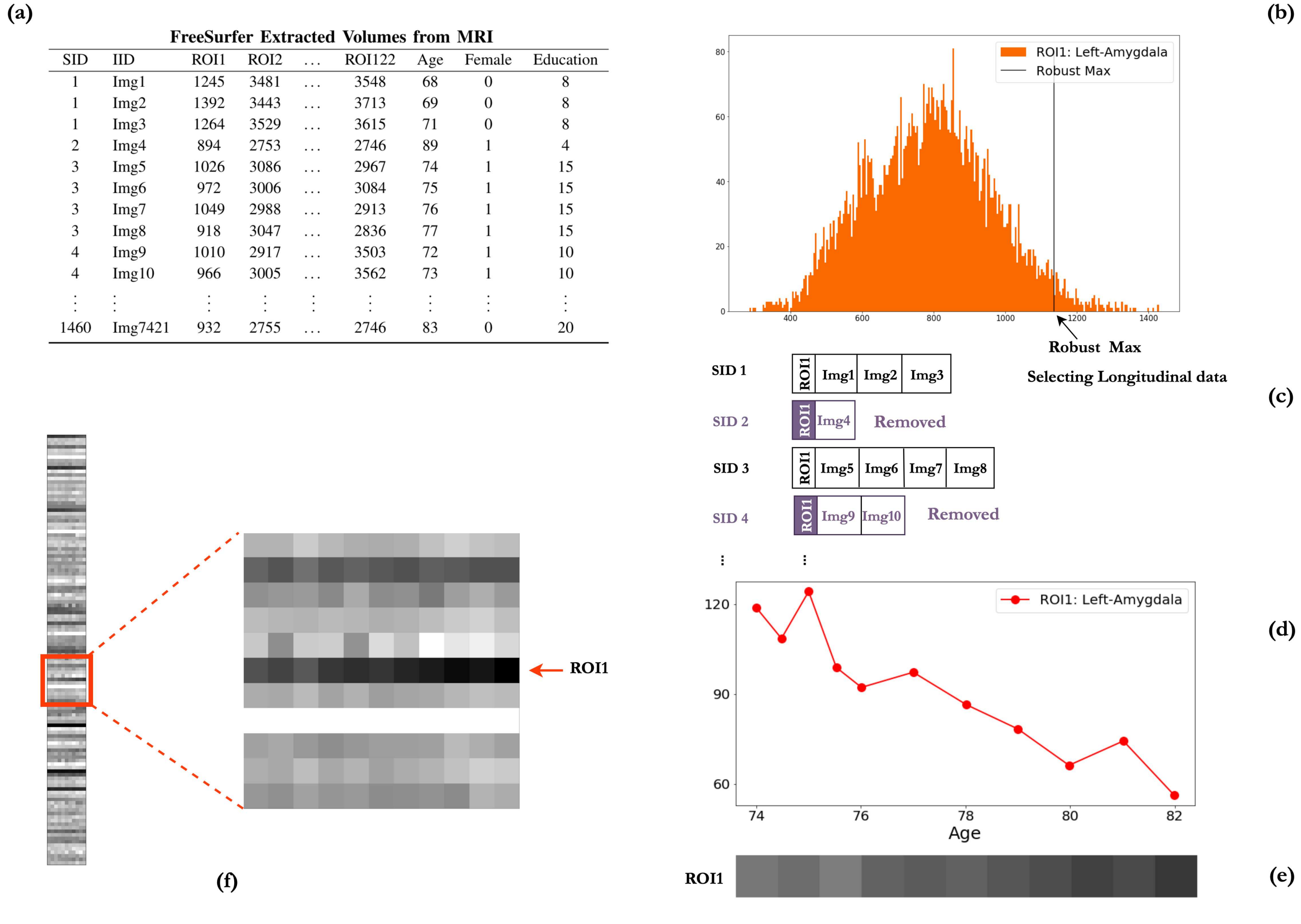
The data that support the findings of this study are available from ADNI database (adni.loni.usc.edu) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of ADNI.

Image preparation

We used the measured volumes of all the regions in the brain that are extracted by Freesurfer[12] v.6.0 from the T1-weighted MR images. For each individual, we obtained such measurements at all time points, and thereby a two-dimensional matrix for each subject containing the volumes of brain regions at the time points. Further, to include their possible influence in our study, we added three more rows to the matrix: gender (male $= 0$, female $= 1$), the level of education (varies between 4 years and 20 years), and age of subject at MRI examinations (between 54 to 96). Therefore, for subject i ($i = 1, \dots, 736$) we had a matrix x_i, so that $x_i \in \mathbf{R}^{m \times n_i}$, where $m = 125$ (number of ROIs + 3), and $3 \leq n_i \leq 11$ is the number of scans for subject i. The goal was then to construct a two-dimensional image for each subject based on its matrix.

We first selected 20% of subjects for the final test set at random, controlling for class labels (to have 20% of each label in the test set), gender (to have 20% of both male and female in the test set), and age (to have a similar range of age in both the training and test set). Then for the rest of the data (including subjects with less than three MRI scans), we assigned the ROIs and the three additional variables: age, gender, and education level as columns in a table, and inserted the extracted volumes from images into its rows (see Fig. 1a).

Next, we scaled the volumes using a min and max for each region, to get them into a similar range. For scaling we considered all the available data in ADNI (1460 subjects, female/male: 642/818, with a total of 7421 MRI scans.), not only sMCI and cAD subjects, to obtain a more general range for the volumes of regions in the brain. More specifically, the volumes of the considered regions of interest have positive values measured in cubic millimeters. For each region of interest, we



FreeSurfer Extracted Volumes from MRI

SID	IID	ROI1	ROI2	…	ROI122	Age	Female	Education
1	Img1	1245	3481	…	3548	68	0	8
1	Img2	1392	3443	…	3713	69	0	8
1	Img3	1264	3529	…	3615	71	0	8
2	Img4	894	2753	…	2746	89	1	4
3	Img5	1026	3086	…	2967	74	1	15
3	Img6	972	3006	…	3084	75	1	15
3	Img7	1049	2988	…	2913	76	1	15
3	Img8	918	3047	…	2836	77	1	15
4	Img9	1010	2917	…	3503	72	1	10
4	Img10	966	3005	…	3562	73	1	10
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1460	Img7421	932	2755	…	2746	83	0	20

Figure 1. Here we illustrate an example of preparing images from brain regions volumes extracted from MR images. The left part of the figure is for all ROIs, whereas on the right side, we explain a specific ROI. Note that we should first detach the test set. **a)** We assign the ROIs to the columns of a table where each row corresponds to the volumes of ROIs for one image. The number of MR images varies from one ID to another. **b)** For each ROI, we find a robust max and replace the upper outliers with this value. Then, we scale the volumes in the column by the min and robust max between 0 and 255. **c)** Next, we select the longitudinal subjects which have at least three images. The graph in **(d)** shows the left-Amygdala scaled volumes versus age. **e)** This graph maps to a gray-scale image so that pixel intensity represents the changes in the values. **f)** Finally, we attach the gray images of all ROIs on top of each other to get an image for each subject.

considered the real minimum value as the lower limit of that region. In the regions of interest data, we observe a trend that exhibits a lower mean to be skewed toward a few participants with very large volumes. In order to better utilize the limited resolution of the intensity values (8-12bit), we opted to perform a single-sided winsorizing operation where the largest 2.5% of all the volume values are replaced with less extreme values (Fig. 1b). We call these upper limits *robust max*. Then, we scaled each column of the table based on its minimum value and its robust max, $\frac{x_i - min}{robust\ max - min}$.

Note that the scaling of one subject is affected by all the other subjects in the table. To avoid data leakage, it was, therefore, important to separate a test set before scaling. The test set subjects, and potentially other new, previously unseen subjects, are scaled using the minimum and (robust) maximum computed using the training data.

After scaling the values in all columns, we selected the longitudinal subjects for which at least three MRI scans were available (Fig. 1c). Every subject has a volume-trajectory for each ROI (Fig. 1d) which we mapped to an image where the pixels' intensity in the image represents the ROI's scaled values at time points (see Fig. 1e, for one ROI). Then, we add the images of all ROIs on top of each other, plus the intensity images of age, education, and gender (Fig. 1f). This resulted in one image per participant, based on the volume extracted from the longitudinal MRI scans.

The images have different dimensions caused by subjects having missing data and the different number of visits among the subjects. Figure 2a shows images with different sizes for both sMCI and cAD subgroups. To determine if we can identify the differences between the prepared images in subgroups by their pixels' intensity we constructed the images in Fig. 2b. For all subjects, we linearly interpolated the values of ROIs to have the same image dimension for all subjects, and then we calculated the average of the matrices associated with all images in each subgroup. These average images (Fig. 2b) highlight the differences in the intensities of sMCI compared to cAD. Note that we only used interpolation in the construction of the average images (Fig. 2a). The below classification experiment was done using images of different sizes, as in Fig. 2a.

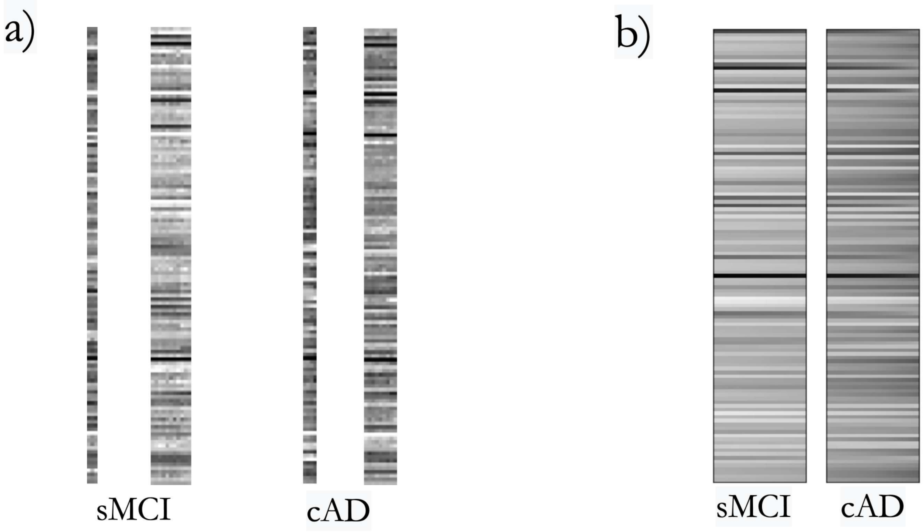


Figure 2. a) The images have different sizes as we have longitudinal data with different lengths. Here there are two examples of images for each subgroup with different sizes. b) The average image for all subjects in two longitudinal subgroups after normalization and interpolation.

Regularization techniques

During the training of our models, we used multiple regularization techniques. Both general explicit techniques such as dropout, batch normalization, and weight decay, and data augmentation that were tailored to our specific data set, as described here.

Aiming to balance the class sizes and to use a source of variance in our data set to boost our models' generalization ability, we augmented the data set by adding Gaussian noise to the existent images. We presume that the obtained volumes for ROIs contain noise (confer the instability in trajectory graph in Fig. 1d), which likely is related to the physical and biological situation during scanning, uncertainly concerning the quality of T1 weighted images, and our chosen segmentation tool (FreeSurfer). To estimate how the variability in the volumes of the ROIs produced by repeated scans in a short time affects our constructed 2D images, we identified 14 subjects in ADNI who had at least two MRI scans within a month or less. It's natural to assume that the volumes of one's brain regions change very little over one month, but comparing the extracted volumes of ROIs for these two repeated MRI scans showed differences in volume (from $\pm 5.3\ mm^3$ for Left-vessel to $\pm 5563.2\ mm^3$ for Cerebral White Matter volume). For each ROI, we averaged 14 standard deviations, measured separately for two collected volumes of each subject, and called it σ_{roi}. Then, we added Gaussian noise with zero mean and the measured standard deviation for each ROI ($P_{roi}(\mu = 0, \sigma_{roi})$) to the training set until we collected 600 subjects for each class. Then we incorporated the noisy data into one table. Afterward, we normalized the new table by using the min and robust max saved for each ROI (Fig. 1b) and then prepared images based on the noisy versions of existing subjects by following the steps in Fig. 1c to 1f.

Model selection

Since model performance is typically very sensitive to hyper-parameter tuning, we performed an extensive search over a wide set of hyper-parameters. To find the optimized values for learning rate, weight decay, dropout, and the CNN structures, we selected 10 different training-validation sets (hereafter 10 folds). For each label, we randomly selected 18% of the training set in order to keep the same percent of gender and the same range of age in both training and validation sets when possible. Note that we put the test set aside before this step. A grid search over model architectures and these hyper-parameters was conducted by varying the following:

- CNN model: we considered `ResNet18` and `ResNet34`, 18-layer and 34-layer residual networks[13], as implemented in the `Torchvision` library[14].

- Probability of dropouts on the hidden layers (ps): we passed eight values for ps: 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6 and 0.7.
- Weight decay (wd): for wd we passed four values: 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}.
- Maximum learning rate (max-lr): these five values were tested for max-lr: 3×10^{-2}, 3×10^{-3}, 3×10^{-4}, 1.5×10^{-4} and 1×10^{-4}.

To construct and train our binary CNN classifiers we used `fastai`[15] version 1.0.61, a deep learning library based on PyTorch. Instead of training all layers with a constant learning rate or decreasing the learning rate with a fixed or exponential value, we applied the cyclical learning rates method[16] as implemented in `fastai`, which varies the learning rate cyclically between a reasonable set of minimum and maximum boundaries[15]. The batch size was set to eight for all models.

We also compared the performance of the ResNet models with and without pretraining, using the pretrained ResNet18 and ResNet34 models available in `PyTorch`, fine-tuning them on our data using `fastai`. Further, we investigated whether batch normalization had a significant effect on the performance of models.

During model selection, we monitored the training and validation loss, error rates, accuracy, precision, recall, and F_1 score on the validation set. For each combination of parameters, we estimated the optimal number of epochs by finding the epochs associated with the smallest validation loss separately for all 10 folds and computing their average.

After the grid search, we selected the top-performing models in terms of accuracies over the 10 validation sets, and we got our final results by ensembling these models using both soft and hard voting strategies. In hard voting, the ensemble predicts the majority vote among the individual models, while soft voting is based on averaging the class probabilities of the models.

There are several sources of randomness in the PyTorch CNN models, leading to slightly different results every time a model is used. To limit the effect of such randomness, we fixed the random seed in both `Numpy` and `Pytorch`. Further, to also reduce the effect of other sources of randomness, such as dropout layers, we trained the ensemble models 20 times with both hard and soft voting and reported the average and standard deviation for the final results.

Finally, to investigate whether spatial relationships reflected in the ordering of the various ROI measurements influence the model, we randomly shuffled the order of ROIs in the images ten times. We then applied the same ensemble model to the identical pair of training and test sets for these ten different image sets.

Results

Our results are based on the steps described in Section "Model selection" applied to two subgroups of subjects, sMCI and cAD (see Fig. 2).

Model selection

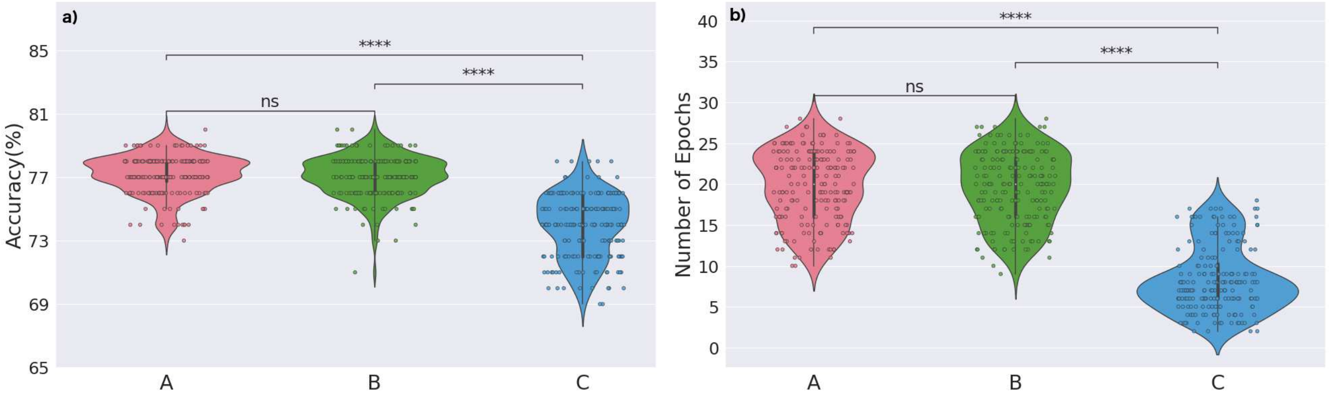


Figure 3. Comparing the effect of batch-normalization and transfer learning during grid-search; A: Batch-normalization and non-pretraining ResNet18/ResNet34) models, B: without batch-normalization and non-pretraining, C: batch-normalization and pretraining. The average accuracies **a)** and epochs associated with the lower validation loss **b)** over validation sets show the similarity between A and B, while in C are reduced significantly. P-value; ns: $p > 0.05$ and ****: $p < 0.0001$.

Fig. 3a shows the similarity in the performance (accuracy) of the models with and without batch normalization (A and B, respectively), and also the significant decline in the performance when using the pretrained ResNet18/ResNet34 models with batch normalization (C). As our images are quite different from the images used for pre-training, the low performance of this

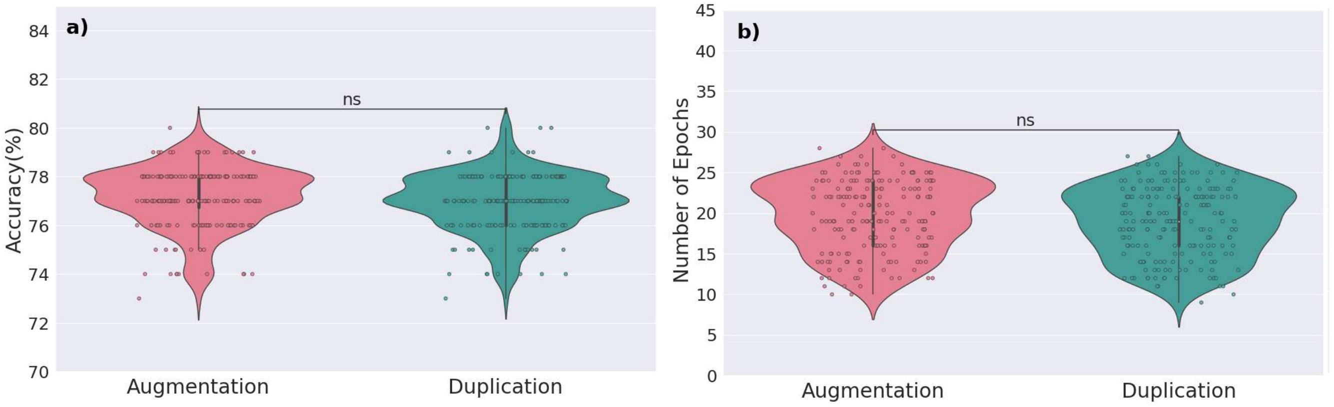


Figure 4. The comparison between grid-search over data with Gaussian noise and duplicating the existing images shows an insignificant difference between the accuracy **a)** and the number of epochs associated with the lowest validation loss **b)**. P-value; ns: $p > 0.05$.

model is perhaps to be expected. During the grid search, we saved the number of epochs associated with the average of lower validation loss for 10 validation sets. Fig. 3b shows the similarity in the number of epochs for models with and without batch normalization (A and B) while the number of epochs for the case with transfer learning (C) is significantly smaller. Thus, while transfer learning speeds up the training step, the accuracy is not as high as in cases A and B. We chose to use model A in the following, i.e. without pretraining and with batch normalization.

To explore the value of our data augmentation approach, we repeated the experiment once again by duplicating the existing images instead of adding Gaussian noise. Fig. 4 compares the performance of the models when augmentation is according to the Gaussian noise ($\mu = 0,\ \sigma_{roi}$) with the case of duplicating the available images. There is an insignificant difference between the accuracies (Fig. 4a) and the number of epochs associated with the average of lowest validation loss (Fig. 4b).

Ensemble model

Models	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
ResNet	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	34	34	34
epoch	13	15	21	14	15	14	13	16	17	16	16	19	16	17	17	20	16	13	17
ps	1e-1	2e-1	2e-1	2e-1	4e-1	4e-1	4e-1	5e-1	5e-1	5e-1	5e-1	6e-1	6e-1	6e-1	6e-1	7e-1	3e-1	3e-1	5e-1
wd	1e-4	1e-1	1e-2	1e-3	1e-1	1e-1	1e-4	1e-1	1e-2	1e-3	1e-1	1e-2	1e-1	1e-2	1e-3	1e-2	1e-1	1e-3	1e-4
max-lr	3e-4	3e-4	3e-3	3e-4	3e-4	1.5e-4	1.5e-4	3e-4	3e-4	3e-4	1e-4	3e-4	1e-4	1.5e-4	1e-4	3e-4	3e-4	1.5e-4	3e-4

Table 2. 19 selected models based on: network architectures (ResNet18 and ResNet34), hyper-parameters (ps: probability of dropout layers, wd: weight decay, max-lr: maximum learning rate), and the number of epochs.

Based on the results of the previous section we selected the top 19 models (see Table 2) to investigate whether constructing an ensemble model improves the results compared to the individual models in terms of accuracy and robustness. Fig. 5a shows the average of receiver operating characteristic (ROC) curves of all models for a specific validation set. The mean and standard deviation of areas under the ROC curves (ROC AUC) for each validation set was computed. These curves (Fig. 5a) highlight the differences in model performance over the validation sets.

For each model, we averaged the ROC curves across ten validation sets (Fig. 5b). Based on the curves in Fig. 5b, the differences between the ROC AUC mean and standard deviation of the 19 models are insignificant (between 0.79 ± 0.04 and 0.81 ± 0.04). Therefore, while the models differ in their performance on single validation sets (Fig. 5a), their averages over all the validation sets are quite close (Fig. 5b).

Further, the mean ROC AUCs for the ensemble model, applied separately to each validation set, are plotted in black in Fig. 5a and b. The ROC AUC mean value and standard deviation for the ensemble model are 0.83 ± 0.04, which is higher than the top individual ROC AUC on seven validation sets (Fig. 5a), and higher than all ROC AUC for individual models when averaged over ten validation set (Fig. 5b).

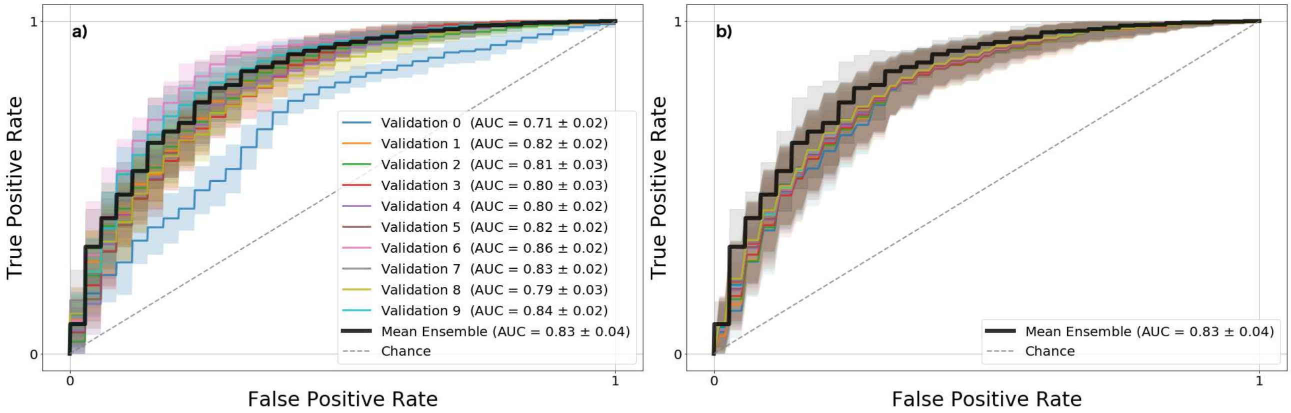


Figure 5. **a)** Averaged ROC curve and standard deviation of AUC based on a specific validation set (color) for top 19 selected models. **b)** Averaged ROC curve and standard deviation of AUC based on a specific model (color) for all validation sets. While the models have different performances on each specific validation set (**a**), the average performance on all the validation sets have similarities (**b**). The average ROC for the ensemble model is in black.

Classification of sMCI versus cAD

The final evaluation of our models was conducted by training the 19 selected models on the combined training and validation sets and forming two ensemble models, based on soft and hard voting. Their performance on the separate test set was then computed. To reduce the effect of randomness, we repeated the computations 20 times and averaged the results. Fig. 6 shows the final results for the soft and hard voting ensembles. The averages of accuracy, weighted precision, recall, and F_1score for hard voting are as follows: 75.9%, 77.1%, 75.9%, and 76.1% respectively (Fig. 6a). The averages for soft voting are as follows: 76.3%, 77.5%, 76.3%, and 76.6% respectively (Fig. 6b). Fig. 6c represents the accuracies of 20 runs for both hard and soft voting.

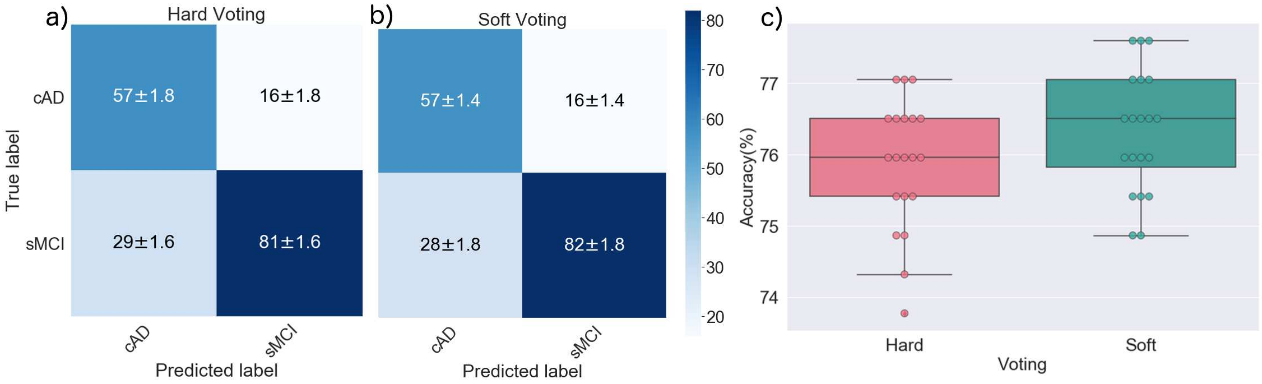


Figure 6. **a)** The confusion matrix after hard voting over 19 models obtained the average accuracy, precision, recall, and F_1score of 76%,77%, 76%, and 76% respectively. **b)** Soft voting over 19 models obtained accuracy, precision, recall, and F_1score of 76.%, 78%, 76%, and 77% respectively. **c)** Boxplot for accuracy of 20 runs of ensemble models.

Finally, we randomly shuffled the order of ROIs in the images 10 times, and evaluated the same 19 models on these images. The accuracies ranged from 72% to 80%, with an average of 75.5%.

Discussion

We proposed a method to represent the information of longitudinal metadata as two-dimensional images, enabling the construction of image-based machine learning classifiers. This approach makes it possible to use complex sets of longitudinal data together with standard image classification methods that are not designed specifically for longitudinal data. It can be used in settings with balanced or imbalanced longitudinal data, with missing data at some time points.

We evaluated the method in an experiment based on the ADNI data set, aiming to classify stable MCI versus converged AD subjects using convolutional neural network models. We achieved higher-than-chance results, with an average accuracy of

76%. Our results show that our proposed method is competitive with other CNN-based approaches in the literature[17], where the reported accuracies range from 62% to 82% in similar classification problems, based on varying machine learning methods and multi-modal data sources (see e.g.[11,17–21]).

For example, Cue et al.[18] classified sMCI vs. cAD subjects with an accuracy of 71.71%, sensitivity of 65.27% and a specificity of 75.27% based on a CNN and recurrent neural network and data sourced from ADNI. Shmulev et al[19] used a ResNet-based CNN model directly on MRI data from ADNI, predicting cAD with an accuracy of 62%, sensitivity of 75% and specificity of 54%. Li et al.[22] applied support vector machine on 36 sMCI and 39 cAD subjects obtaining an 80% accuracy for classification. They increased the accuracy to 82% by adding functional MRI data to their model. Lian et al.[20] classified sMCI vs. cAD with the accuracy of 81%, but with a sensitivity of 53%, and Wen et al.[23] in their systematic review of CNN studies, reported that their accuracy is on a severely imbalanced dataset (one class is less than half of the other).

Our approach can be seen as a relatively simple method to deal with longitudinal data that can lead to competitive classification results.

Our study has some limitations related to the chosen source of data. As shown in Fig. 1, the trajectories of brain volume exhibit some instability. This instability affects pixel intensity and therefore makes the classification more challenging. Another limitation of this study is the difficulty in diagnosing MCI and AD. Some studies have shown the establishment of AD in the brain years before cognitive impairments appear in the behavioral functionality of the brain[24]. In the data from ADNI, the time span between visits for subjects is half a year on average, and we potentially have some sMCI that are almost AD. A possibility to make the model more robust is to drop one or two last visits of the sMCI subjects since they may be showing AD symptoms in less than six months.

Our results indicate that the proposed approach to longitudinal data analysis can be suitable as a supplementary analysis method next to more established statistical and machine learning analysis streams. Further assessment requires applying the method to multiple different data sources with more diversity in the form of data, such as ascending, descending, categorical, cyclical, or heterogeneous trends over time. An interesting area to explore would be time-series of resting-state functional MRI or longitudinal data of other progressive diseases such as schizophrenia or Parkinson's disease, which have different patterns of changes in brain volumes[25–27].

We hope that this integration of newer machine learning methods will create additional avenues for researchers to work on longitudinal data.

References

1. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. image analysis* **42**, 60–88 (2017).

2. Lundervold, A. S. & Lundervold, A. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik* **29**, 102–127 (2019).

3. Diggle, P. *et al. Analysis of longitudinal data* (Oxford University Press, 2002).

4. Grimes, D. A. & Schulz, K. F. Cohort studies: marching towards outcomes. *The Lancet* **359**, 341–345 (2002).

5. Kalash, M. *et al.* Malware classification with deep convolutional neural networks. In *2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, 1–5 (IEEE, 2018).

6. Shulga, D. *et al.* Toward Explainable Automatic Classification of Children's Speech Disorders. In *International Conference on Speech and Computer*, 509–519 (Springer, 2020).

7. Qin, Z., Zhang, Y., Meng, S., Qin, Z. & Choo, K.-K. R. Imaging and fusing time series for wearable sensor-based human activity recognition. *Inf. Fusion* **53**, 80–87 (2020).

8. Association, A. P. *Diagnostic and statistical manual of mental disorders (DSM-5)* (Pilgrim Press, Washington, 2013).

9. Park, D. C. & Reuter-Lorenz, P. The adaptive brain: aging and neurocognitive scaffolding. *Annu. review psychology* **60**, 173–196 (2009).

10. Geda, Y. E. Mild cognitive impairment in older adults. *Curr. psychiatry reports* **14**, 320–327 (2012).

11. Mofrad, S. A., Lundervold, A. & Lundervold, A. S. A predictive framework based on brain volume trajectories enabling early detection of Alzheimer's disease. *Comput. Med. Imaging Graph.* **90**, 101910 (2021).

12. Fischl, B. FreeSurfer. *Neuroimage* **62**, 774–781 (2012).

13. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).

14. Contributors, T. Torchvision. models (2018).

15. Howard, J. & Gugger, S. Fastai: A layered API for deep learning. *Information* **11**, 108 (2020).

16. Smith, L. N. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 464–472 (IEEE, 2017).

17. Wen, J. *et al.* Convolutional Neural Networks for Classification of Alzheimer's Disease: Overview and Reproducible Evaluation. *Med. Image Analysis* 101694 (2020).

18. Cui, R., Liu, M., Initiative, A. D. N. *et al.* Rnn-based longitudinal analysis for diagnosis of alzheimer's disease. *Comput. Med. Imaging Graph.* **73**, 1–10 (2019).

19. Shmulev, Y., Belyaev, M., Initiative, A. D. N. *et al.* Predicting conversion of mild cognitive impairments to Alzheimer's disease and exploring impact of neuroimaging. In *Graphs in Biomedical Image Analysis and Integrating Medical Imaging and Non-Imaging Modalities*, 83–91 (Springer, 2018).

20. Lian, C., Liu, M., Zhang, J. & Shen, D. Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's Disease diagnosis using structural MRI. *IEEE transactions on pattern analysis machine intelligence* (2018).

21. Mofrad, S. A., Lundervold, A. J., Vik, A. & Lundervold, A. S. Cognitive and MRI trajectories for prediction of Alzheimer's disease. *Sci. Reports* **11**, 1–10 (2021).

22. Li, Y. *et al.* Discriminant analysis of longitudinal cortical thickness changes in Alzheimer's disease using dynamic and network features. *Neurobiol. aging* **33**, 427–e15 (2012).

23. Wen, J. *et al.* Convolutional Neural Networks for Classification of Alzheimer's Disease: Overview and Reproducible Evaluation. *arXiv preprint arXiv:1904.07773* (2019).

24. Jack Jr, C. R. & Holtzman, D. M. Biomarker modeling of Alzheimer's disease. *Neuron* **80**, 1347–1358 (2013).

25. Yilmaz, R., Hopfner, F., van Eimeren, T. & Berg, D. Biomarkers of Parkinson's disease: 20 years later. *J. Neural Transm.* **126**, 803–813 (2019).

26. Fei, X. *et al.* Impact of region of interest size on transcranial sonography based computer-aided diagnosis for Parkinson's disease. *Math. Biosci. Eng.* **16**, 5640–5651 (2019).

27. Rodrigues-Amorim, D. *et al.* Schizophrenia: a review of potential biomarkers. *J. psychiatric research* **93**, 37–49 (2017).

Acknowledgements

Our work was supported by the Trond Mohn Research Foundation, grant number BFS2018TMT07.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Ageing, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Author contributions statement

S.A.M. took part in conceiving the study and conducted all the experiments, contributed to the analysis of the results, and drafted the manuscript. H.B. and A.S.L. contributed to conceiving, designing and analyzing, interpreting results, and reviewing the manuscript.

Competing interests

We declare that the authors have no competing interests as defined by Nature Research, or other interests that might be perceived to influence the results and/or discussion reported in this paper.