How You Split Matters: Data Leakage and Subject Characteristics Studies in Longitudinal Brain MRI Analysis

Dewinda J. Rumala[0000−0002−1458−2238]

Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
dewinda.207022@mhs.its.ac.id

Abstract. Deep learning models have revolutionized the field of medical image analysis, offering significant promise for improved diagnostics and patient care. However, their performance can be misleadingly optimistic due to a hidden pitfall called 'data leakage'. In this study, we investigate data leakage in 3D medical imaging, specifically using 3D Convolutional Neural Networks (CNNs) for brain MRI analysis. While 3D CNNs appear less prone to leakage than 2D counterparts, improper data splitting during cross-validation (CV) can still pose issues, especially with longitudinal imaging data containing repeated scans from the same subject. We explore the impact of different data splitting strategies on model performance for longitudinal brain MRI analysis and identify potential data leakage concerns. GradCAM visualization helps reveal shortcuts in CNN models caused by identity confounding, where the model learns to identify subjects along with diagnostic features. Our findings, consistent with prior research, underscore the importance of subject-wise splitting and evaluating our model further on hold-out data from different subjects to ensure the integrity and reliability of deep learning models in medical image analysis.

Keywords: Data Leakage · Deep Learning · MRI · Alzheimer's Disease

1 Introduction

Medical image analysis has witnessed remarkable advancements with the integration of deep learning models, particularly Convolutional Neural Networks (CNNs), which have demonstrated great potential in enhancing diagnostics and patient care. Deep learning models have shown remarkable accuracy in classifying medical images, including brain MRI scans, and have become invaluable tools for medical professionals. However, the reliability of these models can be jeopardized by different biases [3], especially a hidden challenge known as 'data leakage'.

Data leakage occurs when information from the test set unintentionally leaks into the training process [10], leading to over-optimistic performance during model evaluation [1]. This misleading optimism can result in a false perception of the model's efficacy, potentially compromising clinical decision-making and

patient outcomes. Data leakage can arise from various sources, including improper data splitting strategies during cross-validation (CV) procedures [19,2].

In this study, we focus on investigating data leakage in the context of 3D medical imaging, specifically using 3D CNNs for longitudinal brain MRI analysis. Longitudinal brain MRI data is particularly challenging due to repeated scans from the same subjects over time, introducing complex temporal dependencies that can exacerbate data leakage. While 3D CNNs have been considered less prone to data leakage than their 2D counterparts [19], the impact of different data splitting strategies on model performance in this domain remains relatively unexplored.

Our primary objective is to examine the influence of data splitting strategies, specifically subject-wise, record-wise, and late-wise, on the performance of 3D CNN models in longitudinal brain MRI when employed for Alzheimer's Disaese (AD) classification. We aim to uncover potential issues related to data leakage and identity confounding as argued by [2], where the model learns to identify individual subjects rather than focusing solely on the diagnostic features of interest. To achieve this, we employ GradCAM visualization to gain insights into the attention patterns of the CNN models during classification.

This study contributes valuable insights into the impact of data leakage on deep learning models in the domain of medical image analysis, with a specific focus on longitudinal brain MRI data. Our findings are expected to shed light on the importance of proper data splitting strategies and the significance of subject-wise splitting in mitigating data leakage concerns. Ultimately, our research aims to enhance the integrity and reliability of deep learning models, promoting their seamless integration into clinical practice for longitudinal brain MRI classification.

Related work. When employing deep learning models for Alzheimer's disease (AD) diagnosis using MR images, many studies have traditionally relied on two-dimensional (2D)-based analysis using 2D CNNs. However, the slice-level analysis is prone to bias, including data leakage, as highlighted by Yagis et al. [19]. In response to these limitations, recent works have shifted towards using 3D CNNs for volume-based analysis, which has shown to be more superior compared to 2D-based analysis [13]. Several studies have demonstrated the advantages of 3D CNNs in capturing spatial information and improving the accuracy of medical image analysis [7,6,20,11,13].

Despite significant advancements and high accuracies achieved by deep learning models, many studies overlook the importance of proper data splitting procedures to avoid data leakage. Specifically, in longitudinal data, the occurrence of identity confounding should be avoided by using subject-wise split, as advocated by Neto et al. [2] and Saeb et al. [16]. Additionally, the commonly overlooked practice of late-wise split (where data splitting is performed after data transformation) has been identified as a potential source of data leakage, as suggested by Yagis et al. [19]. Thus, it should be avoided to ensure a reliable evaluation of deep learning models.

2 Methods

2.1 Data Collection and Processing

We obtained and used the preprocessed MRI data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [8] (Full details about the preprocessing step can be found at the ADNI website[1]), where we specifically selected T1-weighted and T2-weighted images acquired using a 3T scanner from the same demographic. To ensure data consistency, we applied strict inclusion and exclusion criteria and included only subjects with complete 3T MRI scans for both modalities. The collected dataset consists of longitudinal data from multiple visits for each subject, comprises 25 patients of AD, 41 patients of healthy controls (CN), and 45 patients with mild cognitive impairment (MCI). For each class, there are 150 scans in total, except for AD, where we collected only 50 scans.

The data was then further analyzed using the the Computational Anatomy Toolbox 12 (CAT12) [4] implemented in the Statistical Parametric Mapping 12 (SPM12) software. In CAT12, MRI images were processed using the usual VBM pipeline. As a result, all images have been rescaled to $113 \times 137 \times 113$ and the intensity values are in range between 0 and 1. And to increase the number of data, we performed 3D augmentation techniques using the libraries provided by [17]. The augmentation techniques performed include flipping and 5 degree rotation as previously practiced by [7]. Thus, the final size of our data collection is 300 volume image scans for every class. Further detail about data statistics can be found in Table S1.

2.2 Training Setup

We employed the widely recognized CNN architecture, DenseNet121, which was originally designed for 2D image analysis. To adapt it for 3D medical image analysis, we utilized the 3D architecture version of DenseNet121 as proposed by Soleovyel et al. [17]. Based on their extensive investigation, 3DDenseNet121 demonstrated superior performance compared to other architectures when applied for 3D medical image analaysis task. In addition, we made a few modifications to the architecture to better suit our needs.

Following the last 3D convolutional layer in the architecture, we incorporated a 3D global average pooling layer to reduce the spatial dimensions and obtain a fixed-length feature vector. This feature vector is then connected to fully connected layers with 2 units $\{128, 3\}$, enabling us to perform three-way classification for the target classes of CN, MCI, and AD.

Between the fully connected layers, we incorporated rectified linear unit (ReLU) activation layers to introduce non-linearity and enhance the network's expressiveness. The final layer is a softmax classification layer, which provides probability scores for each class. We utilized the categorical cross-entropy loss function to train the model for the multi-class classification task. More information about the training setup is summarized in the supplemental Table S2.

---

[1] http://adni.loni.usc.edu/methods/mri-analysis

2.3   Evaluation Scheme

Our main goal is to assess the impact of data leakage during CV by considering three different splitting strategies, each adjusted to our longitudinal data conditions, as detailed in Section 2.1. To better illustrate the implementation, we provide a toy example of the three splitting strategies in Fig. S1. In this study, the dataset is then divided into train/validation/test sets with a rough ratio of 70/10/20%, respectively. We conduct CV across 5-folds to thoroughly evaluate the performance and generalizability of our deep learning models under various splitting strategies.

To ensure an unbiased evaluation of our models and maintain a balanced representation of each class in each fold, we employ a stratified splitting strategy. This involves partitioning the images per class into k-folds, ensuring an equal distribution of images for each class across the folds. By doing so, we prevent any class-specific biases during training and evaluation, thus providing a fair assessment of our deep learning model's performance.

These three evaluation schemes allow us to comprehensively investigate and understand the potential impact of data leakage on the model's performance under different data splitting conditions. This provides a holistic assessment of the deep learning model's reliability and generalizability

Subject-Wise Split. In this scheme, all image scans of each subject are assigned as a group in a fold, regardless of their visit times before any data transformation is applied (early splitting). This approach ensures that the scans from the same subject are not spread across different folds but rather kept together, allowing the model to be evaluated on unseen subjects during each fold. This helps to avoid any potential data leakage and identity confounding between subjects as suggested by Neto, et al. [14].

Record-Wise Split. In this scheme, image scans from all subjects are grouped together into a fold based on the records or visits time in an early split manner. With this strategy, data from different visits of the same subject may appear in different folds, allowing the model to be trained and evaluated on the same subjects during each fold. While some studies [12] believe that this strategy might be conditionally appropriate and can lead to better results compared to subject-wise fashion data split, it may not be the best approach and can lead to overconfident performance due to data leakage, specifically caused by the occurrence of identity confounding as argued by Neto et al. [14].

Late Split. Data splitting is performed after the augmentation process in this scheme, by using a sequential numbering approach (more information described in Fig. S1). This split allows the augmented image scans of the same subjects from a certain visit to appear in different folds. According to previous studies by Varoquaux et al. [18] and Yagis et al. [19], this splitting strategy has been shown to cause data leakage and should be avoided.

3   Result

Evaluation Results on 5-Fold Data. We first investigated the impact of different data splitting strategies towards model performance on the 5-fold data. Table 1 reports the results for the experiments described in section 2.3 on different MRI sequences of T1-weighted and T2-weighted MRI. As seen in the table, the performance of the models varied across the different splitting strategies.

Upon comparing the performance across various data splitting strategies, a significant difference was observed (P=0.0389), with the record-wise strategy achieving the highest mean accuracy across all MRI sequences. Conversely, the subject-wise strategy obtained the lowest mean accuracy when evaluated on both sequences.

Furthermore, our investigation revealed no statistically significant difference between the T1-weighted and T2-weighted MRI sequences in terms of model performance (P=0.7921). This finding suggests that both sequences are equally suitable for the three-way classification task of CN, MCI, and AD. Consequently, the choice of MRI sequence does not significantly influence the overall classification performance.

Table 1. Training and testing on longitudinal T1-weigted and T2-weighted MRI. Values are presented as mean and standard deviation across folds, expressed in percentages.

MRI Sequence	Scheme	Acc	Prec	Rec	F1-score
	Subject-wise	67.11 ± 6.11	69.38 ± 6.02	67.11 ± 6.12	68.28 ± 5.63
T1-weighted	Record-wise	97.33 ± 1.86	97.54 ± 1.66	97.33 ± 1.86	97.34 ± 1.85
	Late-wise	81.33 ± 12.37	89.45 ± 8.31	79.31 ± 13.29	89.44 ± 77.64
	Subject-wise	61.56 ± 5.23	63.81 ± 9.56	62.56 ± 5.59	61.55 ± 5.77
T2-weighted	Record-wise	93.33 ± 1.55	$95.53 \pm 1,45$	95.33 ± 1.55	95.42 ± 1.77
	Late-wise	88.89 ± 6.30	89.32 ± 6.34	88.89 ± 6.30	88.75 ± 6.37

Further Evaluation on More Subjects. We evaluated the robustness of 3D CNN models trained with different splitting strategies using a separate hold-out dataset of longitudinal brain MRI. The hold-out data consists of image scans from subjects with different visits, including 8 CN, 11 MCI, and 7 AD subjects, totaling 30 image scans per class (see Table S1 for details). Due to data availability constraints, the evaluation focused on T1-weighted MRI images, but future research could benefit from including T2-weighted MRI images for a more comprehensive assessment.

From the previous training and evaluation of the model on 5-fold data, we obtained five different models for each data split strategy that were directly evaluated on the hold-out data. The results of this robustness evaluation are presented in Table 2. From this investigation, we discovered notable discrepancies between the 5-fold and hold-out data evaluation for all splitting strategies.

For instance, the record-wise strategy, which demonstrated an impressive mean accuracy of 97.33% during cross-validation, experienced a substantial drop to 38.71% when tested on the hold-out data. Similarly, the late-wise strategy, initially yielding a mean accuracy of 81.33% during cross-validation, exhibited a significant decline to 40.43% in the robustness evaluation. Conversely, subject-wise splitting displayed a least drastic drop in accuracy compared to the other splitting strategies, from 67.12% to 42.15%.

We observed no statistically significant difference between data split strategies on hold-out data ($P = 0.8235$). However, it is worth noting that subject-wise splitting strategy achieved the highest mean accuracy in this evaluation, while record-wise splitting obtained the lowest mean accuracy. These results contrast with the model's performance on 5-fold data.

Table 2. Robustness evaluation of trained models on hold-out data of T1-weighted MRI. Values are presented as mean and standard deviation obtained from models trained previously on different folds, expressed in percentages.

Splitting Scheme	Acc	Prec	Rec	F1-score
Subject-wise	42.15 ± 5.45	38.71 ± 7.54	42.12 ± 5.50	38.57 ± 4.99
Record-wise	38.71 ± 7.75	37.48 ± 9.20	38.63 ± 7.72	35.68 ± 7.37
Late-wise	40.43 ± 8.95	$37,62 \pm 13.31$	40.43 ± 8.95	39.92 ± 4.80

Grad-CAM Visualization. We present in Fig. 1 examples of the gradient class activation maps (Grad-CAM) for the CN and AD classes extracted and fused from all layers of the 3D CNN models. GradCAM visualization was performed on the hold-out data to gain insights into the attention patterns during classification for the 3D CNN models trained using different data split strategies.

The first row in Fig. 1 displays GradCAM examples of correctly classified CN and AD images from different splitting strategies. We observe that the patterns for achieving correct predictions on CN are relatively similar across the splitting strategies, where some activations are shown around middle area of the brain. However, there are notable differences between the attention patterns for correctly classified AD images under each splitting strategy.

The Grad-CAM results from the subject-wise splitting strategy show activations around the peripheral and middle regions of the brain, contributing to the correct identification of AD. However, when employing the record-wise splitting strategy, the activations become more scattered compared to using subject-wise splitting. While some activated areas remain shown in the left hemisphere, there are more irrelevant activated regions outside of the brain.

The same pattern is observed for the model trained using the late-wise splitting strategy, with even worse results. In this case, the activated areas are heavily concentrated in the corner of the right brain. The results from the record-wise and late-wise strategies may indicate that the models have learned hidden pat-
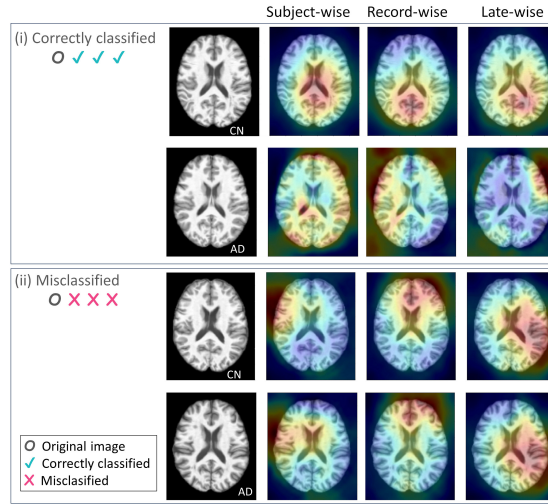
Fig. 1. GradCAM visualization examples for Normal Control (CN) and Alzheimer's Disease (AD) classes under different data split schemes on axial slices of T1-weighted MRI. Each pair of rows depicts (i) correctly classified images, and (ii) misclassified images across all splitting strategies

terns that may not be easily understood by radiologists, rather than focusing on the expected brain patterns for detecting AD.

Moving to the second row, which showcases misclassified images, we can observe similarities in the attention patterns across the different data split strategies. Notably, the heatmaps tend to drift off to the background and do not heavily highlight regions in the brain, which suggests challenges in accurately classifying these images.

4   Discussion and Conclusion

In this study, we conducted a comprehensive evaluation of the effect of data leakage on a 3D CNN model for classifying longitudinal brain MRI data into CN, MCI, and AD using three data splitting strategies: subject-wise, record-wise, and late-wise. Our investigation began with 5-fold data, revealing significant differences in accuracy across the splitting strategies.

The record-wise strategy performed impressively during cross-validation, while subject-wise showed the lowest accuracy. Surprisingly, when tested on hold-out data, all strategies experienced significant drops in accuracy. The record-wise strategy, initially the best, suffered the worst decline. Therefore, to gain further insights into the model's behavior, we performed GradCAM visualization of the model evaluated on the hold-out data.

The activation heatmaps for correctly classified images showed that there might be shortcut learning in record-wise and late-wise splitting to correctly

identify AD. This means that the models learned unintended patterns during cross-validation, which do not generalize well to new, unseen data [5,9]. This might explain why there is significant drop in record-wise and late-wise strategies. As argued by Neto, et al. [2], such record-wise strategy is prone for identity confounding, where the model try to learn identifying the subject rather than target class. Meanwhile, late-wise splitting was also found to cause data leakage [19]. Besides, it may might also experience identity confounding.

Meanwhile, in subject-wise splitting, the model shows more interpretable heatmaps, indicating that it is focusing on more relevant regions for AD diagnosis and not relying on shortcuts or irrelevant information. However, there is a decrease in performance when tested on hold-out data compared to cross-validation data, suggesting that the model might still be facing some challenges, which according to Little, et al. [12], subject-wise splitting strategies can lead to model under-fitting and larger classification errors. The occurence of under-fitting for this strategy makes sense, as the model showed unnecessary activations on irrelevant areas of misclassified images, which suggests that the model may not be learning all the relevant features needed for accurate diagnosis.

However, the less significant drop in performance compared to record-wise and late-wise splitting strategies indicates that subject-wise split might be more robust and less prone to data leakage. And as suggested by Neto, et al. [2], overcoming model under-fitting when applying subject-wise split can be addressed by having more subjects in the dataset. With a larger and more diverse dataset, the model can learn from a broader range of samples, enabling it to capture more complex patterns and improve its performance and generalization on both cross-validation and hold-out data.

We acknowledge certain limitations in this study, particularly the issue of under-fitting due to the limited number of subjects in the subject-wise approach. Additionally, we recognize the gender imbalance, as the study includes a higher number of female participants compared to male participants. For a less biased analysis, future studies should aim to include a larger and more balanced representation of participants in terms of gender and other demographic factors [15]. Moreover, exploring the effects of different splitting strategies on various demographic groups would be essential for ensuring a stronger fairness in the deep learning model's performance.

In summary, this study highlights the significant impact of data leakage on model performance, leading to over-optimistic results. Our findings align with previous research advocating for the use of subject-wise splitting approach [14,16,2] and early-split [19] to avoid data leakage and identity confounding. Moreover, it is advisable to incorporate hold-out data from different subjects whenever possible, in accordance with the recommendations of Varoquaux et al. [18]. Implementing these suggestions can improve deep learning model robustness and reliability in medical image analysis, particularly for longitudinal brain MRI classification.

References

1. Bussola, N., Marcolini, A., Maggio, V., Jurman, G., Furlanello, C.: AI Slipping on Tiles: Data Leakage in Digital Pathology. In: Del Bimbo, A., Cucchiara, R., Sclaroff, S., Farinella, G.M., Mei, T., Bertini, M., Escalante, H.J., Vezzani, R. (eds.) Pattern Recognition. ICPR International Workshops and Challenges, vol. 12661, pp. 167–182. Springer International Publishing, Cham (2021). `https://doi.org/10.1007/978-3-030-68763-2_13`

2. Chaibub Neto, E., Pratap, A., Perumal, T.M., Tummalacherla, M., Snyder, P., Bot, B.M., Trister, A.D., Friend, S.H., Mangravite, L., Omberg, L.: Detecting the impact of subject characteristics on machine learning-based diagnostic applications. npj Digital Medicine **2**(1), 99 (Oct 2019). `https://doi.org/10.1038/s41746-019-0178-x`

3. Drukker, K., Chen, W., Gichoya, J., Gruszauskas, N., Kalpathy-Cramer, J., Koyejo, S., Myers, K., Sá, R.C., Sahiner, B., Whitney, H., Zhang, Z., Giger, M.: Toward fairness in artificial intelligence for medical image analysis: Identification and mitigation of potential biases in the roadmap from data collection to model deployment. Journal of Medical Imaging **10**(06) (Apr 2023). `https://doi.org/10.1117/1.JMI.10.6.061104`

4. Gaser, C., Dahnke, R., Thompson, P.M., Kurth, F., Luders, E., Alzheimer's Disease Neuroimaging Initiative: CAT – A Computational Anatomy Toolbox for the Analysis of Structural MRI Data. Preprint, Neuroscience (Jun 2022). `https://doi.org/10.1101/2022.06.11.495736`

5. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (Nov 2020). `https://doi.org/10.1038/s42256-020-00257-z`

6. Ghazal, M.: Alzheimer rsquo s disease diagnostics by a 3D deeply supervised adaptable convolutional network. Frontiers in Bioscience **23**(2), 584–596 (2018). `https://doi.org/10.2741/4606`

7. Goenka, N., Tiwari, S.: AlzVNet: A volumetric convolutional neural network for multiclass classification of Alzheimer's disease through multiple neuroimaging computational approaches. Biomedical Signal Processing and Control **74**, 103500 (Apr 2022). `https://doi.org/10.1016/j.bspc.2022.103500`

8. Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L. Whitwell, J., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Gunnar Krueger, Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W., ADNI Study: The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. Journal

of Magnetic Resonance Imaging **27**(4), 685–691 (Apr 2008). `https://doi.org/10.1002/jmri.21049`

9. Jiménez-Sánchez, A., Juodelyte, D., Chamberlain, B., Cheplygina, V.: Detecting shortcuts in medical images – a case study in chest x-rays (2022)

10. Kaufman, S., Rosset, S., Perlich, C., Stitelman, O.: Leakage in data mining: Formulation, detection, and avoidance. ACM Trans. Knowl. Discov. Data **6**(4) (dec 2012). `https://doi.org/10.1145/2382577.2382579`, `https://doi.org/10.1145/2382577.2382579`

11. Korolev, S., Safiullin, A., Belyaev, M., Dodonova, Y.: Residual and plain convolutional neural networks for 3D brain MRI classification. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). pp. 835–838. IEEE, Melbourne, Australia (Apr 2017). `https://doi.org/10.1109/ISBI.2017.7950647`

12. Little, M.A., Varoquaux, G., Saeb, S., Lonini, L., Jayaraman, A., Mohr, D.C., Kording, K.P.: Using and understanding cross-validation strategies. Perspectives on Saeb et al. GigaScience **6**(5) (May 2017). `https://doi.org/10.1093/gigascience/gix020`

13. Narazani, M., Sarasua, I., Pölsterl, S., Lizarraga, A., Yakushev, I., Wachinger, C.: Is a PET all you need? A multi-modal study for alzheimer's disease using 3D CNNs. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 66–76. Springer Nature Switzerland, Cham (2022)

14. Neto, E.C., Perumal, T.M., Pratap, A., Bot, B.M., Mangravite, L., Omberg, L.: On the analysis of personalized medication response and classification of case vs control patients in mobile health studies: The mPower case study (2017). `https://doi.org/10.48550/ARXIV.1706.09574`

15. Ricci Lara, M.A., Echeveste, R., Ferrante, E.: Addressing fairness in artificial intelligence for medical imaging. Nature Communications **13**(1), 4581 (Aug 2022). `https://doi.org/10.1038/s41467-022-32186-3`

16. Saeb, S., Lonini, L., Jayaraman, A., Mohr, D.C., Kording, K.P.: The need to approximate the use-case in clinical machine learning. GigaScience **6**(5) (May 2017). `https://doi.org/10.1093/gigascience/gix019`

17. Solovyev, R., Kalinin, A.A., Gabruseva, T.: 3D convolutional neural networks for stalled brain capillary detection. Computers in Biology and Medicine **141**, 105089 (Feb 2022). `https://doi.org/10.1016/j.compbiomed.2021.105089`

18. Varoquaux, G., Cheplygina, V.: Machine learning for medical imaging: Methodological failures and recommendations for the future. npj Digital Medicine **5**(1), 48 (Apr 2022). `https://doi.org/10.1038/s41746-022-00592-y`

19. Yagis, E., Atnafu, S.W., García Seco de Herrera, A., Marzi, C., Scheda, R., Giannelli, M., Tessa, C., Citi, L., Diciotti, S.: Effect of data leakage in brain MRI classification using 2D convolutional neural networks. Scientific Reports **11**(1), 22544 (Nov 2021). `https://doi.org/10.1038/s41598-021-01681-w`

20. Zhang, J., Zheng, B., Gao, A., Feng, X., Liang, D., Long, X.: A 3D densely connected convolution neural network with connection-wise attention mechanism for Alzheimer's disease classification. Magnetic Resonance Imaging **78**, 119–126 (May 2021). `https://doi.org/10.1016/j.mri.2021.02.001`
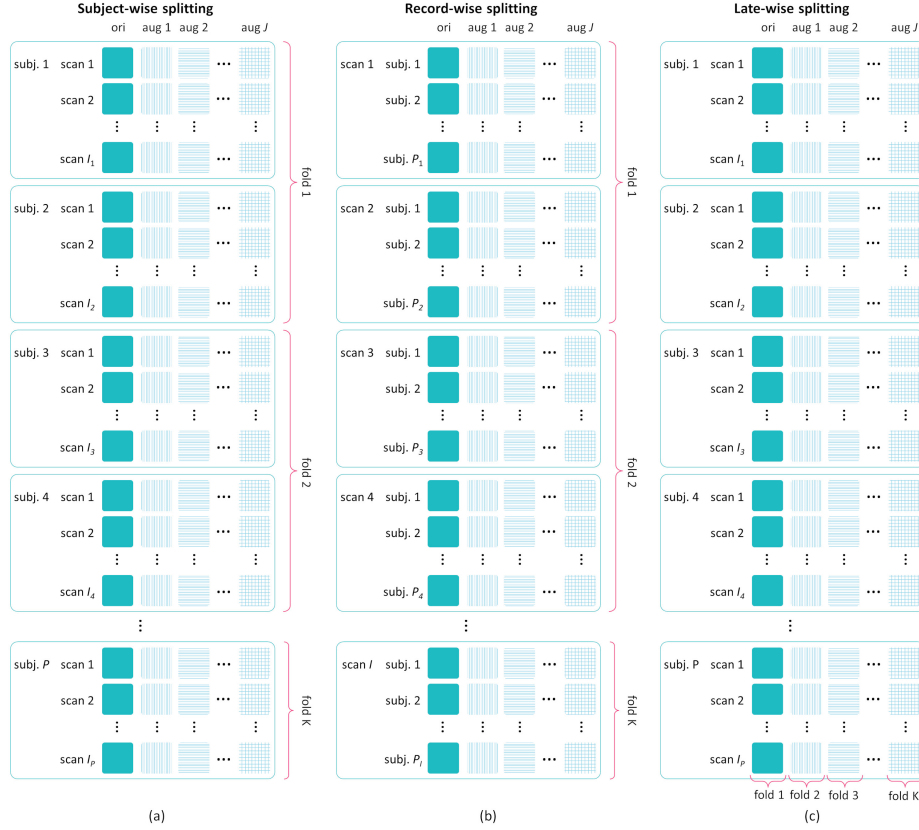
Supplementary Material

Fig. S1. A toy example of different data split strategies for longitudinal brain MRI. (a) Subject-wise splitting groups all image scans based on the subjects into k-folds. (b) Record-wise splitting groups image scans based on different visit times into k-folds. (c) Late-wise splitting groups image scans based on transformation technique into k-folds.

Table S1. Dataset Statistics. Ratio of female/male is based on number of images rather than number of subjects.

Collection	Groups	No of Subj.	Female/Male	Age	No of Scans	
					Before Aug	After Aug
5-fold data	CN	41	85/65	76.68 ± 4.15	150	300
	MCI	45	50/100	74.19 ± 8.57	150	300
	AD	25	30/20	74.22 ± 8.90	50	300
Hold-out data	CN	8	17/13	74.81 ± 3.13	30	-
	MCI	11	8/22	75.42 ± 7.26	30	-
	AD	7	16/1	78.26 ± 6.52	30	-

Table S2. Overview of the parameters used across all experiments: the learning rate was reduced using the ReduceLROnPlateau scheduler from TensorFlow by a factor of 0.1 when validation loss did not decrease after 10 epochs. Adam was used as the optimizer.

Parameter	Value
Learning rate	0.0001
Epsilon	0.0001
Beta 1	0.9
Beta 2	0.99
Epoch	100
Batch size	24