Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

When Near-Perfect Accuracy Is a Warning Sign: Frame-Level Data Leakage in 3D CNN Classification of the Alzheimer's Disease Spectrum from fMRI Volumes

ARIFUL Islam

sm.islam.ariful@gmail.com

East Delta University https://orcid.org/0009-0004-5285-3598

Research Article

Keywords: Alzheimer’s disease, mild cognitive impairment, fMRI, 3D convolutional neural network, data leakage, evaluation methodology, ADNI

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11694221/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

When Near-Perfect Accuracy Is a Warning Sign: Frame-Level Data Leakage in 3D CNN Classification of the Alzheimer's Disease Spectrum from fMRI Volumes

Md. Ariful Islam, for the Alzheimer's Disease Neuroimaging Initiative*
Department of Computer Science and Engineering, East Delta University
Chattogram, Bangladesh
sm.islam.ariful@gmail.com

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

***Abstract*—Deep learning studies on neuroimaging data frequently report accuracies close to 100%, yet such results can arise from data leakage rather than from learned disease signal. This paper is a case study of that failure in the author's own work. A 3D convolutional neural network (CNN) was trained to separate Alzheimer's disease (AD), mild cognitive impairment (MCI), early MCI (EMCI) and cognitively normal (CN) participants using 191,428 individual 3D volumes extracted from functional MRI (fMRI) acquisitions in the Alzheimer's Disease Neuroimaging Initiative (ADNI). Each time frame was treated as an independent image, and the pooled frames were split at random, at the image level, into training and validation sets. The network reached 100% validation accuracy by epoch 14 and made no errors on any of the 191,428 volumes. We show that this outcome is explained by frame-level leakage: frames of the same participant are nearly identical, so a model can recognise individuals instead of disease stages, and a random split places almost every participant on both sides. We document the training signatures that exposed the problem, relate them to leakage patterns reported in the Alzheimer's literature, quantify why image-level splitting of fMRI frames guarantees subject overlap, and propose a subject-level evaluation protocol. The corrected experiment is left to future work; the contribution of this paper is a transparent, reproducible account of how a plausible pipeline yields an invalid result, and how to detect it.**

***Index Terms*—Alzheimer's disease, mild cognitive impairment, fMRI, 3D convolutional neural network, data leakage, evaluation methodology, ADNI**

I. Introduction

Alzheimer's disease (AD) is the most common cause of dementia, and mild cognitive impairment (MCI) is often its prodromal stage [1]. Early detection matters for management and treatment [2], and progression risk differs sharply between stages: in a longitudinal cohort, dementia conversion rates were reported as 1.74% for controls, 4.33% for early MCI (EMCI) and 18.6% for late MCI [3]. Deep learning on neuroimaging is widely studied as a route to earlier and more scalable diagnosis [4], [5], [6], and deep learning models have shown high diagnostic accuracy across medical imaging tasks, although many studies have methodological weaknesses [7].

Reported accuracies in this field vary widely, and a substantial part of that variation comes from flawed evaluation. In a systematic review and reproducible re-evaluation of CNN-based AD classification, Wen *et al.* found that many published studies were at risk of data leakage, most often because images of the same participant appeared in both training and test sets, and showed that leakage can inflate accuracy dramatically [8]. Yagis *et al.* demonstrated the same effect for 2D slice-level splitting of brain MRI, where slice-level splits produced high accuracy even on data where subject-level splits did not [9]. More broadly, data leakage has been identified as a leading cause of irreproducible results in machine-learning-based science [10], and evaluation errors are among the most common methodological failures in medical imaging research [11].

Functional MRI adds a further axis along which leakage can occur. An fMRI acquisition is a 4D series in which the same brain volume is recorded every few seconds, often more than a hundred times, capturing changes in blood oxygenation that reflect neural activity [12]. If each time frame is treated as an independent image, a single participant contributes many near-identical samples.

This paper reports a case in which exactly this happened in the author's own research, and uses it to make the failure concrete. The contributions are:

1) a complete description of a 3D CNN pipeline on 191,428 ADNI fMRI frames that produced 100% validation accuracy;
2) an analysis of the training signatures that reveal frame-level leakage;
3) a simple quantitative argument showing that random image-level splitting of fMRI frames places essentially every participant in both partitions;
4) a subject-level evaluation protocol for frame-based fMRI classification.

II. Background and Related Work

A. Deep Learning for AD Classification

CNNs have been applied to structural MRI, PET and fMRI for AD staging. Huang *et al.* combined T1-weighted MRI and FDG-PET in a multi-modality 3D CNN and reported 90.10% accuracy for AD versus normal controls [13]. Goenka and Tiwari proposed AlzVNet, a volumetric CNN for multiclass AD classification [14]. Teng *et al.* reviewed

machine learning on fMRI data for brain disease research, noting both promise and methodological challenges such as small samples and heterogeneous preprocessing [4].

B. Representations of fMRI

fMRI is usually analysed through its temporal structure, for example as regional time series or as functional connectivity between regions. Matkovič *et al.* showed that static and dynamic functional connectomes derived from fMRI represent largely similar information [15]. The pipeline studied here used neither representation: it discarded temporal order and treated each 3D frame as a stand-alone image.

C. Data Leakage in Neuroimaging

Leakage occurs whenever information about the evaluation data influences training or model selection. Common forms in neuroimaging include (i) subject overlap, where data from one participant appear in more than one partition; (ii) splitting after augmentation or slicing; and (iii) using the test set for model selection or early stopping [8], [9]. All three are present in the case below.

III. Data and Pipeline as Originally Run

A. Data

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) [16], [17]. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The original goal of ADNI was to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). The current goals include validating biomarkers for clinical trials, improving the generalizability of ADNI data by increasing diversity in the participant cohort, and to provide data concerning the diagnosis and progression of Alzheimer's disease to the scientific community. For up-to-date information, see adni.loni.usc.edu.

The data were accessed after approval of the author's ADNI data use application. Participants belonged to four diagnostic groups: AD, MCI, EMCI and CN. Each participant's fMRI acquisition is a 4D blood-oxygen-level-dependent (BOLD) series. Every 3D time frame was exported from these series into a single folder per diagnostic class and labelled with the participant's diagnosis. Participant identifiers were not retained in the exported file names, and the number of participants per class was not recorded at the time. The resulting pool contained 191,428 volumes (Table I).

Ethics statement: The ADNI study was approved by the institutional review boards (IRBs) of all participating sites, and written informed consent was obtained from all participants or their authorized representatives at enrolment. This work is a secondary analysis of de-identified ADNI data obtained under the ADNI Data Use Agreement; no new data were collected and the author had no contact with participants, so no additional ethics approval was required.

TABLE I
Volumes per Diagnostic Class

Class	Volumes	Share
AD	27,382	14.3%
CN	62,511	32.7%
MCI	72,707	38.0%
EMCI	28,828	15.1%
Total	191,428	100%

B. Preprocessing and Split

Each volume was resampled to 64 × 64 × 64 voxels with one channel and normalised to [0, 1] by dividing by its maximum intensity. Random rotations and shifts were applied to training volumes during training. The pooled volumes were then split at random, at the image level, into 80% for training (153,142 volumes) and 20% for validation (38,286 volumes).

C. Model and Training

The network (Table II) has two 3D convolutional layers with 32 and 64 filters of size 3 × 3 × 3 and ReLU activation, each followed by 2 × 2 × 2 max pooling, then a 256-unit fully connected layer with ReLU, dropout with rate 0.5, and a four-way softmax. Because the flattened feature vector has 14 × 14 × 14 × 64 = 175,616 elements, the fully connected layer dominates the parameter count, giving about 45 M parameters in total. Training used the Adam optimiser at a learning rate of 10^{-3}, categorical cross-entropy loss and a batch size of 32, for up to 50 epochs with early stopping on validation loss (patience 12); training stopped at epoch 27.

TABLE II
3D CNN Architecture

Layer	Configuration	Output
Input	1 channel	64 × 64 × 64 × 1
Conv3D	32 filters, 3×3×3, ReLU	62 × 62 × 62 × 32
MaxPool3D	2×2×2	31 × 31 × 31 × 32
Conv3D	64 filters, 3×3×3, ReLU	29 × 29 × 29 × 64
MaxPool3D	2×2×2	14 × 14 × 14 × 64
Flatten	—	175,616
Dense	256, ReLU; dropout 0.5	256
Dense	4, softmax	4

IV. Observed Results

Fig. 1 shows training and validation curves from epoch 14 onward. Validation accuracy was 1.0000 at every epoch shown, and validation loss stayed near zero (at most 0.0076), while training accuracy was about 0.999 and training loss ranged from roughly 0.06 to 0.11.

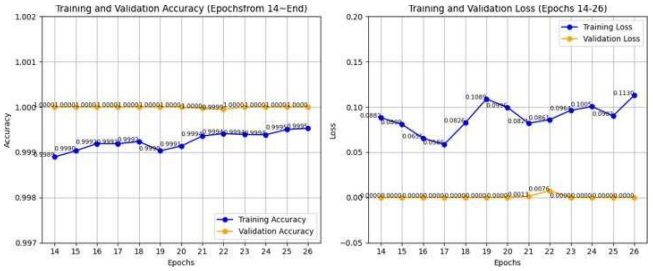


Fig. 1. Training and validation accuracy (left) and loss (right), epochs 14–26. Validation accuracy is 1.0000 throughout and validation loss is lower than training loss at every epoch.

Fig. 2 shows the confusion matrix produced at the end of the study. It contains no off-diagonal entries, and its counts (27,382 + 62,511 + 72,707 + 28,828) sum to 191,428, the size of the entire pool. The matrix was therefore computed on all volumes, including the 80% used for training, so it measures fit rather than generalisation.

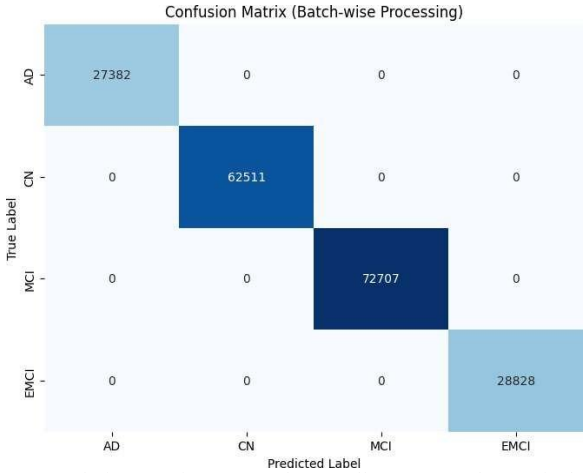


Fig. 2. Confusion matrix computed over all 191,428 volumes, including training data. Zero errors in every class.

V. Why the Result Is Invalid

A. Signature 1: Validation Better Than Training

Dropout and augmentation are active only during training, so validation loss is often slightly lower than training loss. Here, however, validation loss was at least an order of magnitude lower than training loss and validation accuracy was perfect from epoch 14 onward. That pattern indicates that validation samples are almost copies of training samples.

B. Signature 2: Perfect Separation of Subtle Classes

EMCI, MCI and CN differ only subtly in imaging, and subject-level evaluations in the literature report accuracies well below perfect for such distinctions [8]. A model that separates all four classes without a single error is far more likely to exploit a shortcut than to have solved the clinical problem.

C. Mechanism: Frames Identify Participants

Consecutive frames of one acquisition are recorded seconds apart. They share the participant's anatomy, head position, scanner and intensity profile, and differ mainly in small BOLD fluctuations. Because every frame of a participant carries the same diagnostic label, a network that can recognise the participant can predict the label without learning anything about disease.

A random image-level split makes such recognition possible for essentially every participant. If a participant contributes n frames and each frame is independently assigned to validation with probability $p = 0.2$, the probability that the participant appears in both partitions is

$$\Pr(\text{overlap}) = 1 - (1-p)^n - p^n. \quad (1)$$

For $n = 20$ this is already 0.988, and for typical acquisitions of more than 100 frames it is indistinguishable from 1. Under this split, the effective validation question becomes "which known person is this?", which a CNN answers easily. The effective sample size of the study is the number of participants, not the number of volumes.

D. Compounding Factors

Two further choices amplified the problem: the same leaked validation split was used for early stopping, and the final confusion matrix was computed on the training data as well. Augmentation by small rotations and shifts does not remove participant identity, so it offered no protection.

VI. Proposed Subject-Level Protocol

The following protocol removes the leakage described above and applies to any study that turns fMRI time series, or MRI slices, into multiple images per participant:

1) Keep the participant identifier in every exported file name (e.g., *subjectID*_frame012).

2) Split by participant, stratified by diagnosis, using grouped k-fold cross-validation; report the number of participants per class in every fold.

3) Hold out a separate set of validation participants for early stopping and model selection, and evaluate the test participants once.

4) Aggregate frame-level predictions per participant (e.g., mean class probability) and report subject-level metrics alongside frame-level ones.

5) Where possible, stratify by acquisition site and scanner, and report confidence intervals.

6) Consider representations that use the temporal information, such as functional connectivity or 4D models, rather than isolated frames.

Under this protocol the accuracy is expected to fall substantially below 100%. That lower number, not the near-perfect one, is the meaningful estimate of what the model has learned.

VII. Discussion

The pipeline in this study was not unusual: frame extraction, resampling, normalisation, augmentation and a standard 3D CNN. Each step is reasonable on its own, and the failure came from a single decision about how to split the data. That is why leakage of this kind is easy to miss and why it continues to appear in published work [8], [9].

A practical lesson is that near-perfect accuracy on a clinically subtle, multi-class neuroimaging task should trigger an audit rather than a write-up. Three quick checks

would have caught this case early: whether any participant appears in more than one partition, whether validation loss is implausibly far below training loss, and whether the evaluation set is disjoint from the training set.

VIII. Conclusion

A 3D CNN trained on 191,428 ADNI fMRI frames reached 100% validation accuracy for AD, MCI, EMCI and CN classification. This paper has shown that the result reflects frame-level data leakage rather than disease recognition, explained the mechanism quantitatively, and set out a subject-level protocol for evaluating frame-based fMRI models. Evaluating the model under this protocol is left to future work.

Acknowledgment

Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19AG024904). The grantee organization is the Northern California Institute for Research and Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including 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.

An AI-based writing assistant was used for language editing and formatting of this manuscript. All experiments, analyses and conclusions are the author's own, and the author takes full responsibility for the content.

References

[1] B. J. Kelley and R. C. Petersen, "Alzheimer's disease and mild cognitive impairment," *Neurol. Clin.*, vol. 25, no. 3, pp. 577–609, 2007.

[2] R. Hafiz *et al.*, "The latest advances in the diagnosis and treatment of dementia," *Cureus*, vol. 15, no. 12, 2023, Art. no. e50522.

[3] S.-Y. Lin *et al.*, "The clinical course of early and late mild cognitive impairment," *Front. Neurol.*, vol. 13, 2022.

[4] J. Teng, C. Mi, J. Shi, and N. Li, "Brain disease research based on functional magnetic resonance imaging data and machine learning: A review," *Front. Neurosci.*, vol. 17, 2023.

[5] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," *Insights Imaging*, vol. 9, no. 4, pp. 611–629, 2018.

[6] L. Alzubaidi *et al.*, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, 2021, Art. no. 53.

[7] R. Aggarwal *et al.*, "Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis," *npj Digit. Med.*, vol. 4, 2021, Art. no. 65.

[8] J. Wen *et al.*, "Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation," *Med. Image Anal.*, vol. 63, 2020, Art. no. 101694.

[9] E. Yagis *et al.*, "Effect of data leakage in brain MRI classification using 2D convolutional neural networks," *Sci. Rep.*, vol. 11, 2021, Art. no. 22544.

[10] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, 2023, Art. no. 100804.

[11] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: Methodological failures and recommendations for the future," *npj Digit. Med.*, vol. 5, 2022, Art. no. 48.

[12] G. H. Glover, "Overview of functional magnetic resonance imaging," *Neurosurg. Clin. N. Am.*, vol. 22, no. 2, pp. 133–139, 2011.

[13] Y. Huang, J. Xu, Y. Zhou, T. Tong, and X. Zhuang, "Diagnosis of Alzheimer's disease via multi-modality 3D convolutional neural network," *Front. Neurosci.*, vol. 13, 2019, Art. no. 509.

[14] N. Goenka and S. Tiwari, "AlzVNet: A volumetric convolutional neural network for multiclass classification of Alzheimer's disease through multiple neuroimaging computational approaches," *Biomed. Signal Process. Control*, vol. 74, 2022, Art. no. 103500.

[15] A. Matkovič, A. Anticevic, J. D. Murray, and G. Repovš, "Static and dynamic fMRI-derived functional connectomes represent largely similar information," *Netw. Neurosci.*, vol. 7, no. 4, p. 1266, 2023.

[16] Alzheimer's Disease Neuroimaging Initiative (ADNI). [Online]. Available: https://adni.loni.usc.edu

[17] R. C. Petersen *et al.*, "Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization," *Neurology*, vol. 74, no. 3, pp. 201–209, 2010.