How serious is data leakage in deep learning studies on Alzheimer's disease classification?

Junhao Wen, Elina Thibeau-Sutre, Jorge Samper-Gonzalez, Alexandre M Routier, Simona Bottani, Didier Dormont, Stanley Durrleman, Olivier Colliot, Ninon Burgos

How serious is data leakage in deep learning studies on Alzheimer's disease classification?

Junhao Wen[*,1], **Elina Thibeau--Sutre**[*,1], **Jorge Samper-González**[1], **Alexandre Routier**[1], **Simona Bottani**[1], **Didier Dormont**[1,2], **Stanley Durrleman**[1], **Olivier Colliot**[1,2,3], **Ninon Burgos**[1]

[*]The authors contributed equally to this study
[1]ARAMIS Lab, ICM, Inserm U1127, CNRS UMR 7225, Sorbonne University, Inria, Paris, France
[2]AP-HP, Department of Neuroradiology, Pitié-Salpêtrière Hospital, Paris, France
[3]AP-HP, Department of Neurology, Pitié-Salpêtrière Hospital, Paris, France

junhao.wen89@gmail.com elina.thibeausutre@icm-institute.org @AramisLabParis

In recent years, there has been a strong interest in the use of deep learning (DL) for assisting diagnosis of brain diseases from neuroimaging data. Unbiased evaluation of their performances is critical to assess their potential clinical value.

A major source of bias is data leakage, that can be difficult to detect for non-specialists.

In this study, focusing on the case of Alzheimer's disease (AD) diagnosis from T1 MRI using convolutional neural networks (CNN), we performed a rigorous literature search, assessed the prevalence of data leakage and analysed its causes. Additionally, we demonstrated the phenomenon of data leakage in a controlled setting by focusing on the impact of the data split strategy.

Methods

Literature Search

Search engines: PubMed & Scopus

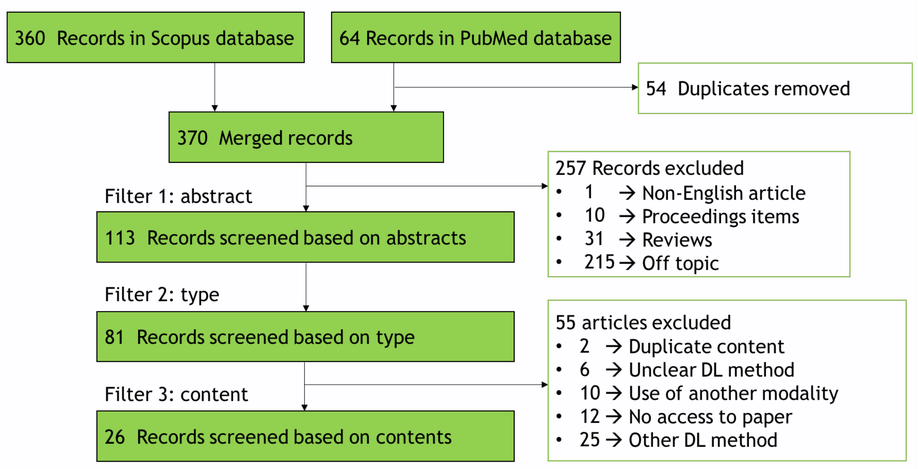


Diagram summarizing the bibliographic methodology.

Data Leakage

3 categories identified:

1. Biased split
Data extracted from the same individual is distributed in both the train and the test sets.

2. Late split
Test / train split is performed after another procedure (feature selection, pretraining...).

3. No independent test set
The performance is evaluated on the train and / or validation sets.

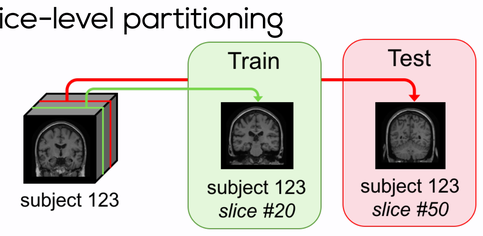
3 possible labels

- *Clear* when data leakage is explicitly witnessed
- *Unclear* when no sufficient explanation is offered
- *None detected* otherwise

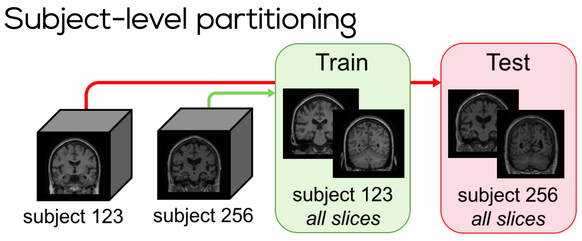
Application

Evaluation of the impact of biased split:
Two experiments were conducted with different data partitioning strategies for the AD vs CN classification on ADNI dataset.

A. Slice-level partitioning



B. Subject-level partitioning



Results

Summary of the studies performing classification of AD using CNNs on anatomical MRI

A. Studies without data leakage

Study	DOI	Accuracy AD vs CN	Data leakage
Aderghal et al, 2017	10.1007/978-3-319-51811-4_56	83,70%	None detected
Aderghal et al, 2018	10.1109/CBMS.2018.00067	90%	None detected
Backstrom et al, 2018 *	10.1109/ISBI.2018.8363543	90,11%	None detected
Cheng et al, 2017	10.1117/12.2281808	87,15%	None detected
Cheng and Liu, 2017	10.1109/CISP-BMEI.2017.8302281	85,47%	None detected
Islam and Zhang, 2018 **	10.1186/s40708-018-0080-3	(CN/mild/moderate /severe: 93,18%)	None detected
Korolev et al, 2017	10.1109/ISBI.2017.7950647	80,00%	None detected
Li et al, 2018	10.1109/IST.2017.8261566	88,31%	None detected
Li et al, 2018	10.1016/j.compmedimag.2018.09.009	89,50%	None detected
Liu et al, 2018	10.1007/s12021-018-9370-4	84,97%	None detected
Liu. et al, 2018	10.1016/j.media.2017.10.005	91,09%	None detected
Liu. et al, 2018	10.1109/JBHI.2018.2791863	90,56%	None detected
Senanayake et al, 2018	10.1109/ISBI.2018.8363832	76%	None detected
Shmulev et al, 2018	10.1007/978-3-030-00689-1_9	(sMCI/pMCI: 62%)	None detected
Valliani and Soni, 2017	10.1145/3107411.3108224	81,30%	None detected

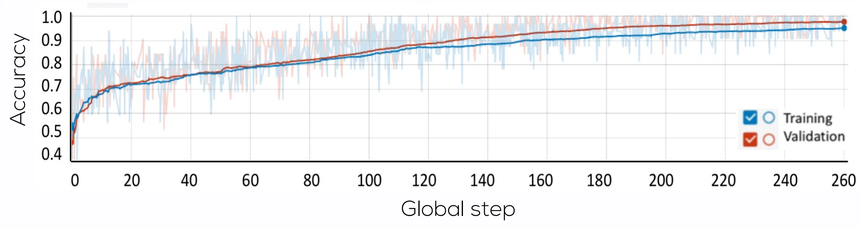
B. Studies with potential data leakage

Study	DOI	Accuracy AD vs CN	Data leakage	Categories 1	2	3
Aderghal et al, 2017	10.1145/3095713.3095749	91,41%	Unclear		X	X
Hon and Khan, 2017	10.1109/BIBM.2017.8217822	96,25%	Unclear	X	X	
Hosseini-Asl et al, 2018	10.2741/4606	99,30%	Unclear	X		
Islam and Zhang, 2017	10.1007/978-3-319-70772-3_20	(CN/mild/moderate /severe: 73,75%)	Unclear		X	X
Taqi et al, 2018	10.1109/MIPR.2018.00032	100%	Unclear			X
Vu et al, 2017	10.1109/BIGCOMP.2017.7881683	85,24%	Unclear	X		
Wang et al, 2018	10.1007/s10916-018-0932-7	97,65%	Unclear			X
Backstrom et al, 2018 *	10.1109/ISBI.2018.8363543	98,74%	Clear	X		
Farooq et al, 2017	10.1109/IST.2017.8261460	(AD/LMCI/EMCI/ CN: 98,88%)	Clear	X		
Gunawardena et al, 2017	10.1109/M2VIP.2017.8211486	(AD/MCI/CN: 96%)	Clear	X	X	
Vu et al, 2018	10.1007/s00500-018-3421-5	86,25%	Clear	X		X
Wang S. et al, 2017	10.1007/978-3-319-68600-4_43	(MCI/CN: 90,60%)	Clear	X	X	

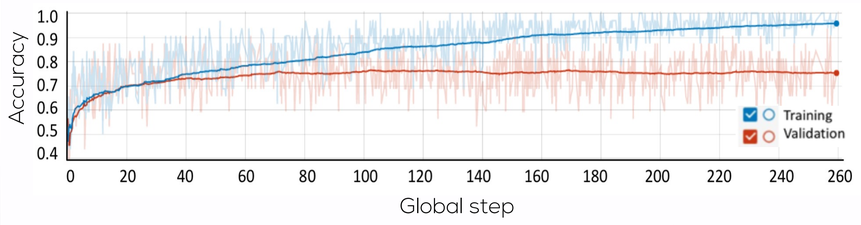
When different from AD vs CN, the classification task is specified in brackets.
* (Backstrom et al., 2018) studied the impact of a biased data split
** Use of imbalanced accuracy on an imbalanced dataset

Observation of performance bias

A. Slice-level partitioning



B. Subject-level partitioning



The training and validation accuracies (smoothed by a threshold of 0.99) are obtained during 150 epochs for both data split strategies over the same architecture.

Data leakage categories:
1. *Biased split*
2. *Late split*
3. *No independent test set*

Mean accuracy

None detected	*Clear* + *Unclear*
86.0 ± 4.5%	94.4 ± 5.6%

Test accuracy

subject-level split	*slice-level split*
75%	98%

Conclusion

Data leakage is a common problem in the literature (42% of surveyed papers). Moreover, it has a serious impact on performance evaluation, as demonstrated by the strong differences in accuracies in both the literature and our experiments. Thus the current literature of the domain may overestimate the performance of deep learning systems for automatic diagnosis of Alzheimer's disease.

Downloads

All the papers that were analyzed in the literature search may be found at www.zotero.org/groups/2337160/ad-dl