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

Junhao Wen, Elina Thibeau–Sutre, Jorge Samper-González, Alexandre 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]

[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

*denotes shared first authorship

Introduction

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, which refers to the use of test data in any part of the training process (Kriegeskorte et al., 2009; Rathore et al., 2017). Data leakage can be difficult to detect for non-specialists, in particular for DL approaches which are complex and very flexible. For instance, splitting slices or scans from the same patient into both training and test sets leads to a biased evaluation. 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 analyzed 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

A bibliographic search was systematically conducted on PubMed and Scopus for the classification of AD using CNNs from T1 MRI. We included only peer-reviewed papers either in journals or in recent conference proceedings (from 2017) up to the time of this search (6/11/18). The resulting articles were labeled into three categories: i) *Clear* when data leakage was explicitly witnessed; ii) *Unclear* when no sufficient explanation was offered and iii) *None detected*. They were further categorized according to the cause of data leakage.
In addition, we performed experiments for a particular case of biased data split. Two different strategies were studied: i) slice-level, where slice extraction was performed before data split, resulting in slices from the same patient being in both the training and test sets; ii) patient-level, where the data split was correctly done. T1 MRI were preprocessed using Clinica (Routier et al., 2018) and used as inputs of an adapted LeNet5 CNN (Lecun et al., 1998). We compared the classification performances obtained with the two data split strategies using baseline ADNI data (336 AD patients and 376 cognitively normal (CN) subjects).

Results

Among the 26 articles retrieved, 4 contained a *Clear* data leakage, 7 were labeled as *Unclear* and 14 as *None detected*. These proportions strongly differ depending on how the MRI is handled by the network: out of the 9 studies which dealt with 2D slices rather than the 3D volume, only two were labeled as *None detected*.
Accuracies obtained by studies labeled as *None detected* ($86{,}0 \pm 4{,}5\%$, for AD vs CN) strongly differed from studies labeled as *Unclear* or *Clear (*$94{,}4 \pm 5{,}6\%$). Three main causes of data leakage were identified (Table 1): i) *Biased split*, where data split was not done at the subject-level causing data from the same subject to appear in both the training and test sets (often in cases of slicing/patching of MRI volumes, using multiple visits or data augmentation); ii) *No independent test set*, where the test set was used to optimize and

fine-tune hyperparameters; iii) *Late split*, where other operations (e.g. pretraining and feature selection) were performed on the entire dataset before the data split. Note that we chose not to label as *Unclear* the studies that did not explain the origin of their architecture. The design or choice of network is often not detailed and thus may have been done by successive evaluations on the test set.
In the experiments, accuracy on the test set was 98% for (biased) slice-level data split and 75% for (unbiased) patient-level data split. The unbiased test accuracy (75%) was obtained at around the 60000th global step where the over-fitting occurred (Figure 1).

Conclusions

Data leakage is a common problem in the litterature (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.

References

Backstrom, K. et al. (2018), ‘An efficient 3D deep convolutional network for Alzheimer’s disease diagnosis using MR images’, in *IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 149–153.

Kriegeskorte, N. et al (2009), ‘Circular analysis in systems neuroscience: the dangers of double dipping’, Nature Neuroscience, vol. 12, pp. 535–540.

Lecun, Y. et al. (1998), ’Gradient-based learning applied to document recognition’, Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324.

Rathore, S. et al. (2017), ‘A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer’s disease and its prodromal stages’, NeuroImage, vol. 155, pp. 530–548.

Routier, A. et al. (2018), ‘Clinica: an open source software platform for reproducible clinical neuroscience studies’, In Annual Meeting of the Organization for Human Brain Mapping (OHBM 2018).

Figures

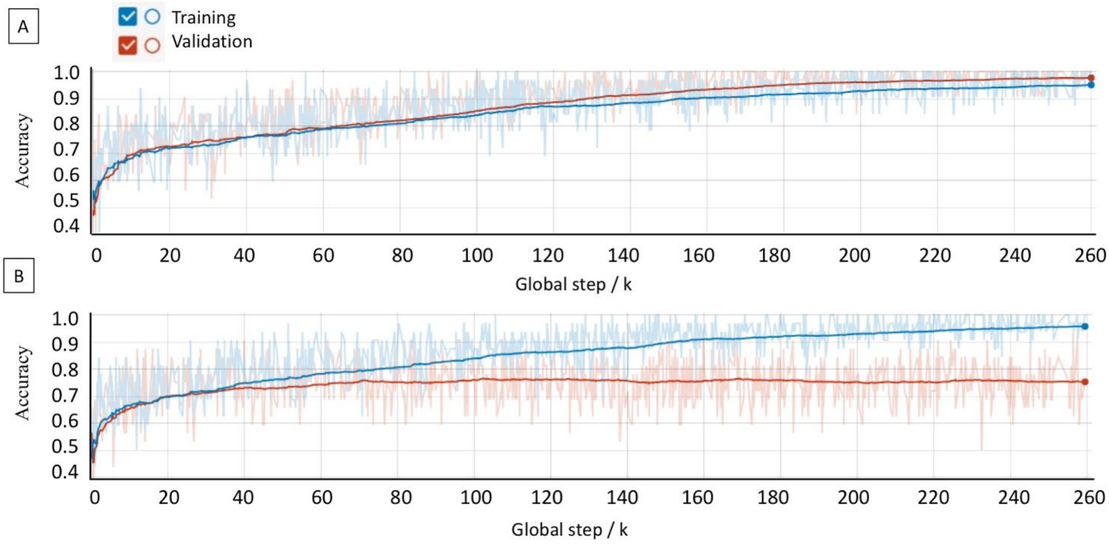

Figure 1. The training and validation accuracies (smoothed by a threshold of 0.99) are obtained during 150 epochs for both data split strategies over the same architectures. (A) slice-level data split; (B) patient-level data split.
The hyperparameters were fine-tuned on the training and validation dataset. Batch size, initial learning rate and dropout rate are 32, 0.01 and 0.5, respectively.

Table 1. Summary of the studies performing classification of AD using CNNs on anatomical MRI. When different from AD vs CN, the classification task is specified in brackets. (A) studies without data leakage; (B) studies with potential data leakage.
Abbreviations: 1: Biased split; 2: No independent test set ; 3: Late split.
* (Backstrom et al., 2018) experimented two data-partitioning strategies to study the consequences of a kind of biased data split and is thus linked to two different labels.
** Use of imbalanced accuracy on an imbalanced dataset, leading to an over-optimistic estimation of performance.

(A) None detected Table

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) Data leakage Table

Study	DOI	Accuracy AD vs CN	Data leakage	Categories 1	2	3
Aderghal et al, 2017	10.1145/3095713.3095749	91,41%	Unclear	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		