frontiers | Frontiers in Neuroscience

TYPE Opinion
PUBLISHED 02 December 2024
DOI 10.3389/fnins.2024.1509358

Check for updates

OPEN ACCESS

EDITED BY
Paul J. Laurienti,
Wake Forest University, United States

REVIEWED BY
Mohsen Bahrami,
Virginia Tech - Wake Forest University,
United States
Matthew Rosenblatt,
Yale University, United States

*CORRESPONDENCE
Federico Del Pup
✉ federico.delpup@studenti.unipd.it

RECEIVED 10 October 2024
ACCEPTED 18 November 2024
PUBLISHED 02 December 2024

CITATION
Del Pup F and Atzori M (2024) Toward improving reproducibility in neuroimaging deep learning studies.
Front. Neurosci. 18:1509358.
doi: 10.3389/fnins.2024.1509358

Toward improving reproducibility in neuroimaging deep learning studies

Federico Del Pup[1,2,3]* and Manfredo Atzori[2,3,4]

[1]Department of Information Engineering, University of Padua, Padua, Italy, [2]Department of Neuroscience, University of Padua, Padua, Italy, [3]Padova Neuroscience Center, University of Padua, Padua, Italy, [4]Information Systems Institute, University of Applied Sciences Western Switzerland (HES-SO Valais), Sierre, Switzerland

KEYWORDS

deep learning, reproducibility, replicability, repeatability, open source, BIDS

1 Introduction

After more than a decade since the Imagenet breakthrough (Krizhevsky et al., 2012), there is no doubt that Deep Learning (DL) has established itself as a powerful resource whose limits and risks remain difficult to assess (Bengio et al., 2024). This success, originated from the ability of deep neural networks to create representations of complex data with multiple levels of abstraction (LeCun et al., 2015), has inevitably attracted researchers from different domains, including multidisciplinary ones.

In computational neuroscience (Trappenberg, 2009), deep learning has offered novel insights into the functionalities of the brain, demonstrating a remarkable ability to exploit the intrinsic multimodality of the field (Saxe et al., 2021). This is also reflected in the increasing volume of publications encompassing diverse data types such as electroencephalographam (EEG), magnetoencephalographam (MEG), structural (MRI) and functional magnetic resonance imaging (fMRI) (Zhu et al., 2019; Zhang et al., 2021). Nevertheless, the potential of deep learning in neuroimaging data analysis is countered by several critical issues (Miotto et al., 2018), including the scarcity of large open datasets, the poor generalizability of DL models, their lack of interpretability, and the poor reproducibility of results, which is discussed in this work.

According to National Academies of Sciences, Engineering, and Medicine et al. (2019), reproducibility is defined as the ability to "obtain consistent results using the same input data; computational steps, methods, and code; and conditions of analysis." This definition differs from that of replicability, commonly mistaken as a synonym, which is instead defined as the ability to "obtain consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data." From the above definitions, it is possible to derive how reproducibility solely depends on how authors facilitate the emulation of the same computational environment. Improving reproducibility, especially in a deep learning scenario, is therefore crucial for ensuring methodological robustness and result trustworthiness. However, when analyzing neuroimaging deep learning studies, several researchers have expressed concerns on how authors rarely report the key elements that make a study reproducible (Ciobanu-Caraus et al., 2024; Colliot et al., 2023). A recent review of DL applications for medical image segmentation found that only 9% of the selected studies were reproducible (Renard et al., 2020), a conclusion also supported by other independent studies (Moassefi et al., 2024; Marrone et al., 2019; Ligneris et al., 2023). Furthermore, the same problem was discovered in DL-EEG applications, where, from a review of 154 selected papers, only 12 were found to be easily reproducible (Roy et al., 2019).

The above statistics highlight not only the severity of the reproducibility crisis, but also how often this important issue is overlooked by both authors and publishers. Insufficient reproducibility not only threatens the credibility of scientific findings, potentially hindering the discovery of new knowledge in the domain (e.g., treatments for neurological disorders), but also introduces inconsistencies in results due to factors such as dataset diversity, variability in preprocessing, and discrepancies in model implementation and evaluation. Such inconsistencies heighten the risk of misinterpreting data or drawing incorrect conclusions that support the validity of a framework over another, thereby posing a potential negative impact on clinical outcomes. For instance, various studies using deep learning to classify different types of dementia or to predict cognitive scores with extremely high accuracies have been questioned not only for their lack of reproducibility, but also for potential performance biases arising from inadequate data partitioning methods or ambiguous validation or model selection procedures (Brookshire et al., 2024; Wen et al., 2020). In contrast, other disciplines have achieved greater reproducibility through the use of standardized datasets and methodologies. For example, the ImageNet dataset (Krizhevsky et al., 2012) has significantly advanced computer vision by establishing benchmarks categorized by learning methods, while the General Language Understanding Evaluation (GLUE) benchmark provides a collection of resources for training, evaluating, and analyzing natural language understanding systems (Wang et al., 2018).

Consequently, there is a need for the neuroimaging field to adopt similar practices to improve the reliability of published research, which heavily depends on its reproducibility. To contribute in this direction, this paper outlines the key elements necessary for achieving reproducibility in neuroimaging deep learning studies and organize them in a table that can be also used as a checklist.

2 Improving reproducibility

Reproducibility can only be achieved if considered since the design of the study. Even in such cases, deep learning presents numerous sources of irreproducibility, making it difficult for researchers to account for them all. Various checklists have been proposed to guide neuroscientists in designing DL studies (Roy et al., 2019; Moassefi et al., 2024), yet they often focus on specific data types or do not effectively explain how certain features affect reproducibility. To provide a clear checklist for researchers, this section introduces an updated table listing 30 key elements that are essential to ensuring reproducibility. Features in Table 1 are organized into six categories and discussed concisely in the following sub-paragraphs, highlighting how the absence of such information can hinder reproducibility and affect the reliability of results. Additionally, specific locations (paper, supplementary material, repository) are suggested to help authors improve clarity and reduce density in the papers.

2.1 Software and hardware

Reproducibility is ensured only if authors share the source code within an open platform like GitHub.[1] However, in neuroscience this is done only 10% of the times, in contrast with other domains like computer vision, where the percentage can reach 70% (Pineau et al., 2021). Sharing a well organized code is crucial, as it provides valuable insight into the implementation correctness. Additionally, releasing an easily executable code can facilitate the design of fair comparisons between the original and new results, which helps building useful benchmarks.

On the paper side, the name and version of each main software library should be clearly reported. This helps to limit any randomness introduced by unnoticed changes in the library's code base, which can potentially compromise the reproducibility and repeatability of results (Alahmari et al., 2020). For the same reason, the entire environment should be also exported and uploaded within the repository. Furthermore, authors should report additional computational details to help tracking other known hardware non-determinism (Chen et al., 2022). These include: GPU model and number, CUDA version, training time, memory allocation, and parallelization across devices.

2.2 Dataset

Previously cited review studies agree that only a small portion of researchers share the data on dedicated open platforms. Sharing health data is indeed problematic due to strict privacy regulations. However, researchers should not feel discouraged, since there are many tools nowadays designed to facilitate data de-identification, reorganization in standardized formats (e.g., BIDS Gorgolewski et al., 2016), and sharing with a digital object identifier (e.g., Zenodo,[2] OpenNeuro[3] Markiewicz et al., 2021). Sharing raw data not only enhances the reliability of research, but also encourages other teams to try improving the original results. In addition, increasing the number of public data can facilitate the creation of multi-center datasets, which are crucial for training DL models being able to better generalize on unseen data. For example, the increasing availability large open-source datasets has encouraged researchers to explore promising deep learning architectures, such as transformers, improving the investigation of the spatiotemporal dynamics of the human brain (Kim et al., 2023; Tang et al., 2023).

Describing the dataset in detail is also important. The paper should clearly indicate the number of subjects, relevant demographic data, number of training samples, and data acquisition modalities. Data acquisition modalities may include the channel map, sampling rate, and reference for EEG data, or the scanner model, voxel size, and acquisition sequence for MRI data.

[1] https://github.com

[2] https://zenodo.org

[3] https://openneuro.org

TABLE 1 List of criteria a research study should fulfill to ensure reproducibility.

Category	Feature	Placement
Software and hardware		
1	Source code	Open repository (e.g., GitHub)
2	Software environment	Paper and source code
3	Computational resources	Paper
Dataset		
1	Raw or preprocessed data	Open platform (e.g., OpenNeuro)
2	Number of subjects	Paper and supplementary If multiple datasets are used, provide a summary table within the paper's main body
3	Demographic data	
4	Number of samples	
5	Acquisition modalities	
Data preprocessing		
1	Artifact handling algorithms	Paper, supplementary, source code Specify the parameters used for each preprocessing step and annotate manually performed operations
2	Normalization and standardization	
3	Harmonization and co-registration	
4	Resampling	
5	Data augmentation	
Model		
1	Architecture description	Paper, supplementary, source code Provide a detailed summary table within the supplementary material
2	Number of learnable parameters	
3	Input dimension	
Training hyperparameters		
1	Random seed	Paper Provide all the information in a dedicated subsection Consider uploading the hyperparameter search results within the main repository, or within the supplementary material
2	Parameter initialization	
3	Batch size	
4	Number of epochs	
5	Loss functions	
6	Optimization algorithms	
7	Learning rate and schedulers	
8	Stopping criteria	
9	Regularization	
10	Hyperparameters search method	
Model evaluation		
1	Data partition	Paper and supplementary Provide the result of each training within the main repository, or as a supplementary material
2	Validation scheme	
3	Performance metric	
4	Baseline comparison	

2.3 Data preprocessing

Medical data are usually processed with complex pipelines, not necessary fully automatic. It is therefore extremely important to make this step reproducible, as different implementation of the same pipeline can lead to different results. For example, minimal changes in the IClabel's thresholds (Pion-Tonachini et al., 2019) can drastically alter the EEG's spectral properties; similarly, the convergence threshold in the ANTs software (Avants et al., 2014) can heavily affects the MRI registration. Consequently, any customizable parameter (or manually performed operation) should be clearly described in the paper. Alternatively, researchers may upload preprocessed data alongside the raw data, enabling consistent input for analysis and assisting laboratories with limited computational resources in reproducing results. Data augmentations should also be clearly described, with particular regard to the range of random values used during their execution.

2.4 Model architecture

Deep learning models should be described on several fronts, especially if architectures are particularly complex. First, the paper should include a schematic representation of the model, and specify its input dimension and number of trainable parameters. (If the model comes from an external library, its name and version should be reported.) Second, supplementary material should provide a full summary table where the number of parameters, input and output dimensions, and other customizable features (e.g., weight constraints, grouped convolutions) are listed for each layer of the network. Finally, the actual model's implementation should be uploaded within the source code, as different implementations are characterized by different parameters initialization.

2.5 Training hyperparameters

Table 1 lists a set of 10 training hyperparameters that should be described within the paper. For each of them, it is important to specify the value of each customizable parameter, if necessary in a table. Authors are encouraged to pay special attention to the use of random seeds as a way to control randomness in the code. Simply setting the random seed at the beginning of a script might not be enough. Best practice is to check if model's parameters after initialization and at the end of the training are the same on multiple rerun. Additionally, libraries like Pytorch (Paszke et al., 2019) can be configured to use a set of deterministic algorithms which, although slower, can enhance reproducibility in DL experiments.

2.6 Model evaluation

In neuroimaging data analysis, it is best practice to evaluate DL models with subject-based cross-validation procedures, which require to partition the dataset multiple times in training and validation sets (Kunjan et al., 2021). Partitions should be reproducible, because different splits can dramatically change the results due to strong subject-based characteristics that can guide the learning process toward different loss minima. Furthermore, it is important to clearly state how the validation set is used during training. While switching to a train-validation-test split scenario as part of a nested cross-validation, as proposed in Kim et al. (2022), is preferable, it is unfortunately usual to find works that use the validation set to stop the training. This process introduces data leakage and affects the reliability of the results. Nevertheless, this information is often omitted (Pandey and Seeja, 2022), leaving the question of whether data leakage was introduced unanswered.

3 Discussion

Designing a reproducible deep learning experiment, especially in a complex and heterogeneous field like neuroscience, is extremely challenging. Researchers must consider several factors that can influence experimental outcomes; otherwise, reproducibility can be permanently lost. In computational neuroscience, DL applications to neuroimaging data require multiple training instances to be run as part of subject-based cross-validation procedures (e.g., Leave-N-Subject-Out, typically using 5 or 10 folds Poldrack et al., 2020). Even with modern GPUs capable of performing 10^{12} floating point operations per second, these strategies can be time-consuming and computationally intensive. It is natural that researchers might hesitate to re-run the analysis if they encounter reproducibility issues. Additionally, deep learning advances rapidly, with numerous papers published every month. While this rapidity fuels the research teams' need to quickly disseminate their work, it should also not undermine reproducibility, which is central to the scientific method and a necessary (though not sufficient) condition for a scientific statement to be accepted as new knowledge (Colliot et al., 2023). Focusing on reproducibility indirectly compels researchers to develop methodologically robust experiments that yield more reliable results.

Technical reproducibility in DL studies can only be assured if both the code and dataset are released by the study group. Although privacy regulations can complicate data sharing, teams can nonetheless make their source code publicly available (Ciobanu-Caraus et al., 2024). This will provide a complementary resource that is often necessary for clearly describing neuroimaging deep learning studies, especially when articles are subjected to word or page limits. Table 1 not only lists 30 key reproducibility elements, but also suggests specific places to present them without hindering the paper's readability. However, this alone might not be enough. Actions are required also from the publisher side; in fact, many of them have begun revising their policies to strongly encourage, and occasionally mandate, the sharing of code and research data.

In conclusion, while deep learning is a powerful tool, it still remains extremely sensitive to minimal variations in the experimental setup. It is essential to re-evaluate priorities, especially in sensitive applications like medical ones, and look beyond mere model performance. Without addressing critical issues such as reproducibility and generalizability, reported accuracies may be seen as mere numbers lacking practical meaning.

Author contributions

FD: Conceptualization, Writing – original draft, Writing – review & editing. MA: Funding acquisition, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the European Union's Horizon Europe research and innovation programme under grant agreement no 101137074 - HEREDITARY. This work has received funding from STARS@UNIPD funding program of the University of Padova, Italy, through the project: MEDMAX. Open Access funding provided by Università degli Studi di Padova | University of Padua, Open Science Committee.

Acknowledgments

FD would like to thank the other members of the Padova Neuroscience Center for their inspiring comments on the topic presented in this paper.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alahmari, S. S., Goldgof, D. B., Mouton, P. R., and Hall, L. O. (2020). Challenges for the repeatability of deep learning models. *IEEE Access* 8, 211860–211868. doi: 10.1109/ACCESS.2020.3039833

Avants, B. B., Tustison, N. J., Stauffer, M., Song, G., Wu, B., and Gee, J. C. (2014). The insight toolkit image registration framework. *Front. Neuroinform.* 8:44. doi: 10.3389/fninf.2014.00044

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., et al. (2024). Managing extreme AI risks amid rapid progress. *Science* 384, 842–845. doi: 10.1126/science.adn0117

Brookshire, G., Kasper, J., Blauch, N. M., Wu, Y. C., Glatt, R., Merrill, D. A., et al. (2024). Data leakage in deep learning studies of translational EEG. *Front. Neurosci.* 18:1373515. doi: 10.3389/fnins.2024.1373515

Chen, B., Wen, M., Shi, Y., Lin, D., Rajbahadur, G. K., and Jiang, Z. M. J. (2022). "Towards training reproducible deep learning models," in *Proceedings of the 44th International Conference on Software Engineering, ICSE '22* (New York, NY: Association for Computing Machinery), 2202–2214.

Ciobanu-Caraus, O., Aicher, A., Kernbach, J. M., Regli, L., Serra, C., and Staartjes, V. E. (2024). A critical moment in machine learning in medicine: on reproducible and interpretable learning. *Acta Neurochir.* 166:14. doi: 10.1007/s00701-024-05892-8

Colliot, O., Thibeau-Sutre, E., and Burgos, N. (2023). *Reproducibility in Machine Learning for Medical Imaging*. (New York, NY: Springer US), 631–653.

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scient. Data* 3, 1–9. doi: 10.1038/sdata.2016.44

Kim, J., Hwang, D.-U., Son, E. J., Oh, S. H., Kim, W., Kim, Y., et al. (2022). Emotion recognition while applying cosmetic cream using deep learning from EEG data; cross-subject analysis. *PLoS ONE* 17, 1–26. doi: 10.1371/journal.pone.0274203

Kim, P., Kwon, J., Joo, S., Bae, S., Lee, D., Jung, Y., et al. (2023). Swift: Swin 4d fmri transformer. *Adv. Neural Inf. Process. Syst*. 36, 42015–42037. doi: 10.48550/arXiv.2307.05916

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, eds. F. Pereira, C. Burges, L. Bottou, and K. Weinberger (New York: Curran Associates, Inc).

Kunjan, S., Grummett, T. S., Pope, K. J., Powers, D. M., Fitzgibbon, S. P., Bastiampillai, T., et al. (2021). "The necessity of leave one subject out (LOSO) cross validation for EEG disease diagnosis," in *Brain Informatics: 14th International Conference, BI 2021, Virtual Event* (Cham: Springer), 558–567.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Ligneris, M. D., Bonnet, A., Chatelain, Y., Glatard, T., Sdika, M., Vila, G., et al. (2023). "Reproducibility of tumor segmentation outcomes with a deep learning model," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)* (Cartagena: IEEE), 1–5.

Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E., et al. (2021). The OpenNeuro resource for sharing of neuroscience data. *Elife* 10:e71774. doi: 10.7554/eLife.71774

Marrone, S., Olivieri, S., Piantadosi, G., and Sansone, C. (2019). "Reproducibility of deep CNN for biomedical image processing across frameworks and architectures," in *2019 27th European Signal Processing Conference (EUSIPCO)* (A Coruna: IEEE), 1-5.

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Brief. Bioinform*. 19, 1236–1246. doi: 10.1093/bib/bbx044

Moassefi, M., Singh, Y., Conte, G. M., Khosravi, B., Rouzrokh, P., Vahdati, S., et al. (2024). Checklist for reproducibility of deep learning in medical imaging. *J. Imag. Inform. Med*. 4, 1–10. doi: 10.1007/s10278-024-01295-4

National Academies of Sciences, Engineering, and Medicine, Policy and Global Affairs, Committee on Science, Engineering, Medicine, and Public Policy, Board on Research Data and Information, Division on Engineering and Physical Sciences, Committee on Applied and Theoretical Statistics, Board on Mathematical Sciences and Analytics, Division on Earth and Life Studies, Nuclear and Radiation Studies Board, Division of Behavioral and Social Sciences and Education, Committee on National Statistics, Board on Behavioral, Cognitive, and Sensory Sciences, and Committee on Reproducibility and Replicability in Science (2019). *Reproducibility and Replicability in Science*. Washington, DC: National Academies Press.

Pandey, P., and Seeja, K. (2022). Subject independent emotion recognition from EEG using VMD and deep learning. *J. King Saud University – Comp. Inform. Sci*. 34, 1730–1738. doi: 10.1016/j.jksuci.2019.11.003

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "PyTorch: an imperative style, high-performance deep learning library," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (Red Hook, NY).

Pineau, J., Vincent-Lamarre, P., Sinha, K., Lariviere, V., Beygelzimer, A., d'Alche Buc, F., et al. (2021). Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *J.Mach. Learn. Res*. 22, 1–20. doi: 10.48550/arXiv.2003.12206

Pion-Tonachini, L., Kreutz-Delgado, K., and Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *Neuroimage* 198, 181–197. doi: 10.1016/j.neuroimage.2019.05.026

Poldrack, R. A., Huckins, G., and Varoquaux, G. (2020). Establishment of best practices for evidence for prediction: a review. *JAMA Psychiatry* 77, 534–540. doi: 10.1001/jamapsychiatry.2019.3671

Renard, F., Guedria, S., Palma, N. D., and Vuillerme, N. (2020). Variability and reproducibility in deep learning for medical image segmentation. *Sci. Rep*. 10:13724. doi: 10.1038/s41598-020-69920-0

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., and Faubert, J. (2019). Deep learning-based electroencephalography analysis: a systematic review. *J. Neural Eng.* 16:051001. doi: 10.1088/1741-2552/ab260c

Saxe, A., Nelli, S., and Summerfield, C. (2021). If deep learning is the answer, what is the question? *Nat. Rev. Neurosci.* 22, 55–67. doi: 10.1038/s41583-020-00395-8

Tang, C., Wei, M., Sun, J., Wang, S., and Zhang, Y. (2023). CsAGP: Detecting alzheimer's disease from multimodal images via dual-transformer with cross-attention and graph pooling. *J. King Saud Univers. – Comp. Inform. Sci.* 35:101618. doi: 10.1016/j.jksuci.2023.101618

Trappenberg, T. P. (2009). *Fundamentals of Computational Neuroscience*. Oxford: Oxford University Press.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, eds. T. Linzen, G. Chrupała, and A. Alishahi (Brussels: Association for Computational Linguistics), 353–355.

Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-Gonzlez, J., Routier, A., Bottani, S., et al. (2020). Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Med. Image Anal.* 63:101694. doi: 10.1016/j.media.2020.101694

Zhang, X., Yao, L., Wang, X., Monaghan, J., Mcalpine, D., and Zhang, Y. (2021). A survey on deep learning-based non-invasive brain signals: recent advances and new frontiers. *J. Neural Eng.* 18:031002. doi: 10.1088/1741-2552/abc902

Zhu, G., Jiang, B., Tong, L., Xie, Y., Zaharchuk, G., and Wintermark, M. (2019). Applications of deep learning to neuro-imaging techniques. *Front. Neurol.* 10:869. doi: 10.3389/fneur.2019.00869