



HAL
open science

Checklist Strategies to Improve the Reproducibility of Deep Learning Experiments with an Illustration

Ali Ben Abbess, Leonardo Meneguzzi, Pedro Pizzigatti Corrêa, David Mouillot, Romain David, Gérard Subsol, Danton Ferreira Vellenich, Rodolphe Devillers, Shelley Stall, Nicolas Mouquet, et al.

► **To cite this version:**

Ali Ben Abbess, Leonardo Meneguzzi, Pedro Pizzigatti Corrêa, David Mouillot, Romain David, et al.. Checklist Strategies to Improve the Reproducibility of Deep Learning Experiments with an Illustration. RDA 19th Plenary Meeting, Part Of International Data Week, Jun 2022, Seoul, South Korea. , 2022, 10.5281/zenodo.6587702 . hal-03738323

HAL Id: hal-03738323

<https://hal.science/hal-03738323>

Submitted on 26 Jul 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

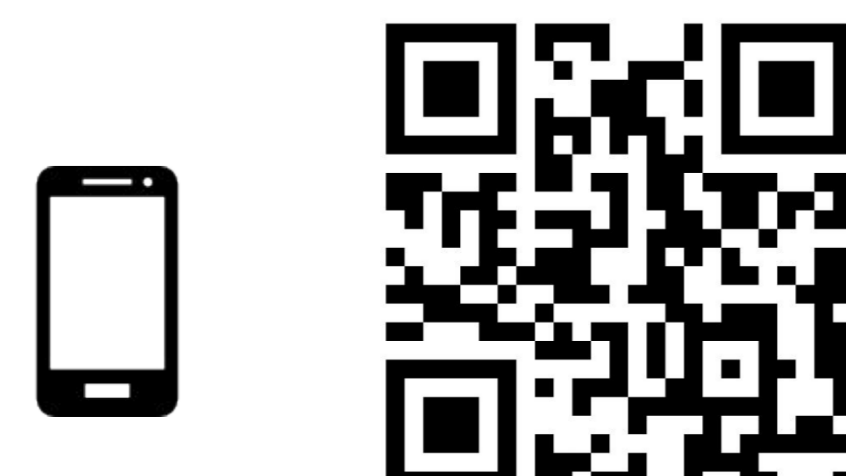
Reproducibility checklist: pre-screen your Deep Learning experiment prior to your submission or publication!



The description of the dataset
<input type="checkbox"/> Are there relevant statistics?
<input type="checkbox"/> Is the dataset open access?
<input type="checkbox"/> Are there clear details of training / validation / test splits?
<input type="checkbox"/> Is there an explanation of any data that was excluded, and all pre-processing steps?
<input type="checkbox"/> Is there a link to a downloadable version of the dataset or simulation environment?
<input type="checkbox"/> For new data collected, is there a complete description of the data collection process?
FAIR sub-principles (Findability, Accessibility, Interoperability, Reusability)
<input type="checkbox"/> Are (Meta)data assigned with a globally unique and persistent identifier? (F1)
<input type="checkbox"/> Are data described with rich metadata? (F2)
<input type="checkbox"/> Are (Meta)data retrievable by their identifier using a standardized communications protocol? (A1)
<input type="checkbox"/> Do (Meta)data include qualified references to other (meta)data? (I3)
<input type="checkbox"/> Are Metadata richly described with a plurality of accurate and relevant attributes? (R1)
<input type="checkbox"/> Are (Meta)data associated with detailed provenance? (R1.2)
Reported experimental results and theoretical claim
<input type="checkbox"/> Is there a clear measure or statistics used to report results?
<input type="checkbox"/> A description of results with central tendency & variation
<input type="checkbox"/> The average runtime for each result, or estimated energy cost
<input type="checkbox"/> Is there a clear statement of the claim?
<input type="checkbox"/> Is there a complete proof of the claim?

The description of the DL architecture and hyper-parameter optimisation process
<input type="checkbox"/> Is there a clear description of the mathematical and/or DL model?
<input type="checkbox"/> Does the paper use a Cross-Validation strategy?
<input type="checkbox"/> Is there a clear explanation of assumptions?
<input type="checkbox"/> Is there an analysis of the complexity of any algorithm?
<input type="checkbox"/> Does the paper use an Optimization procedure? Which one?
<input type="checkbox"/> Were the Hyper-Parameters handcrafted (selected manually)?
<input type="checkbox"/> Does the paper clearly mention the use of Learning rate?
<input type="checkbox"/> Does the paper clearly mention the use of Batch size?
<input type="checkbox"/> Does the paper use Dropout regularization?
<input type="checkbox"/> Are there a clear description of hyper-parameters?
<input type="checkbox"/> Is there an exact number of training and evaluation runs?
The infrastructure and implementation
<input type="checkbox"/> Does the paper detail the infrastructure adequately?
<input type="checkbox"/> Which framework was used?
The shared code
<input type="checkbox"/> Is the shared code Open source?
<input type="checkbox"/> Is there a specification of dependencies?
<input type="checkbox"/> Is there a training code?
<input type="checkbox"/> Is there an evaluation code?
<input type="checkbox"/> Is there a (Pre-)trained model(s)?
<input type="checkbox"/> Is there a README file?

Table 1: Compiled Checklist from Machine Learning checklist Pineau, (2020) [2], the recommendations from Renard et al. 2020 [3], and the FAIR sub-principles (following Hartley & Olsson 2020) [4] aiming to pre-screening Deep Learning experiments to achieve Reproducibility.



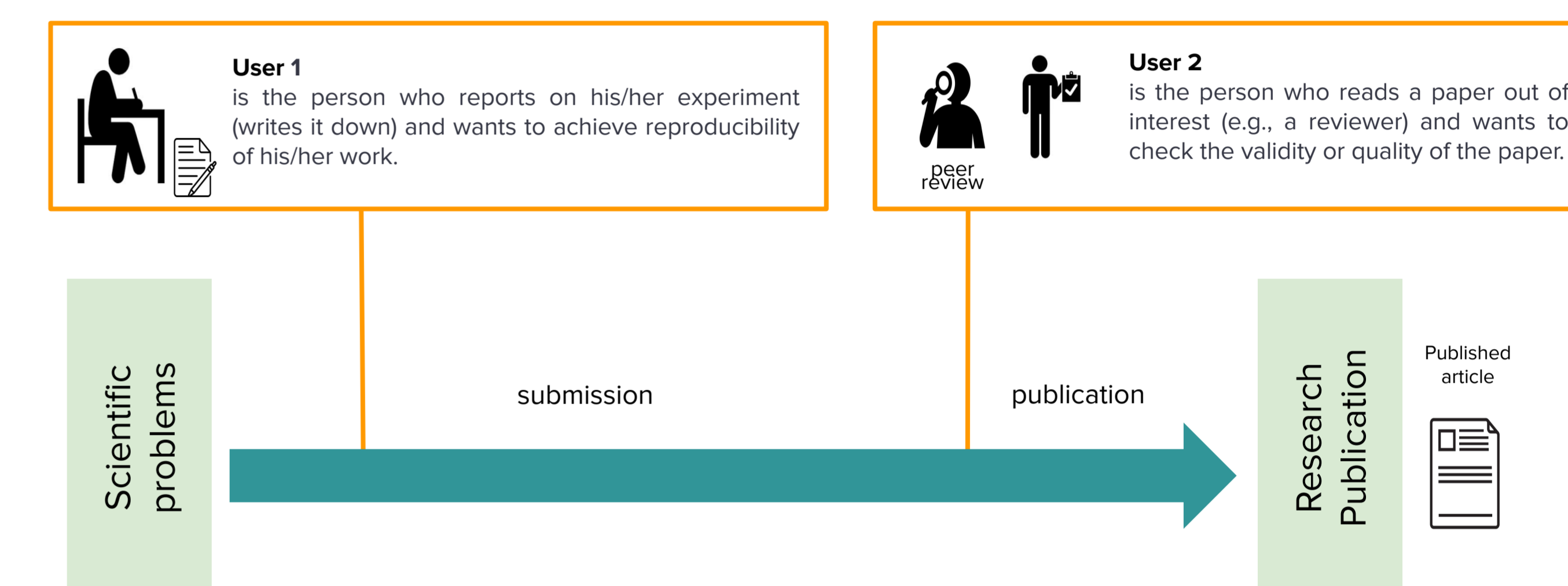
Authors: A. Ben Abbes, J. Machicao, L. Meneguzzi, P.L.P. Corrêa, A. Specht, R. David, G. Subsol, D. Vellenich, R. Devillers, S. Stall, N. Mouquet, M. Chaumont, L. Berti-Equille, D. Mouillot
Contact: pedro.correa@usp.br **Website:** <https://parsecproject.org/> **twitter:** @PARSEC_News

Reproducibility and Replicability (R&R):

The challenges of Reproducibility and Replicability have become a focus of attention in order to promote **open and accessible research**. Therefore, efforts have been made to develop good practices for R&R in the area of computer science. Nevertheless, **Deep Learning (DL)** based experiments **remain difficult to reproduce** by others due to the complexity of these techniques. In addition, several challenges concern the use of massive and heterogeneous data that contribute to the complexity of this R&R.

Checklist:

We compiled a checklist (Table 1) with the most relevant items for Reproducibility to improve DL experiments. This checklist is useful for: an **author reporting on an experiment**, and/or a **reviewer seeking to qualify the scientific contributions** of the work. This table is based on state-of-the-art guidelines from Pineau's Machine Learning checklist [2], the recommendations from Renard's [3], and the FAIR sub-principles (Hartley & Olsson) [4]. Besides that, we organized these criterias according to a DL workflow.



How to use?

We report a review of the reproducibility of three publications for Poverty estimation using DL and Remote sensing imagery. For each experiment, we identified the methods and workflows used, if the experiments were not fully reproducible. Although the three use cases were proposed for a specific task (poverty estimation), we believe that the evaluation methods could be applied to more general Deep Learning tasks, where difficulties might include (a) a lack of dataset specificity (and the metadata related with it), (b) inadequate description of the DL methodology, (c) the implementation methodology, and the infrastructure used. We also feel that these recommendations can be extended to other domains such as medical, climatic, biodiversity, industrial, military, etc.

References:

- [1] Peng, R. D. (2011), Reproducible Research in Computing Science. Science 334, 1226–1227.
- [2] Pineau, J. (2020b). The Machine Learning reproducibility checklist (v2.0, Apr.7 2020). www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist-v2.0.pdf
- [3] Renard, F., Guedria, S., Palma, N.D., & Vuillerme, N. (2020). Variability and reproducibility in deep learning for medical image segmentation. Scientific Reports 10(1), 1–16.
- [4] Hartley, M., & Olsson, T. S. G. (2020), dtoolAI: Reproducibility for Deep Learning. Patterns 1(5), 100073.

Acknowledgements:

PARSEC is funded by the Belmont Forum through the National Science Foundation (NSF), The São Paulo Research Foundation (FAPESP), the French National Research Agency (ANR). J.M. is grateful for the support from FAPESP (grant 2020/03514–9).

Author affiliations : Ali Ben Abbes (FRB-CESAB, Montpellier, FR) <https://orcid.org/0000-0001-5714-7562>; Jeaneth Machicao (University of São Paulo, BR) <https://orcid.org/0000-0002-1202-0194>; Leonardo Meneguzzi (University of São Paulo, BR) <https://orcid.org/0000-0002-4845-6758>; Pedro Pizzigatti Corrêa (University of São Paulo, BR) <https://orcid.org/0000-0002-8743-4244>; Alison Specht (The University of Queensland, AU) <https://orcid.org/0000-0002-2623-0854>; Romain David (ERINHA (European Research Infrastructure on Highly Pathogenic Agents) AISBL, FR) <https://orcid.org/0000-0003-4073-7456>; Gérard Subsol (Research-Team ICAR, LIRMM, CNRS, Univ. Montpellier, FR) <https://orcid.org/0000-0002-7461-4932>; Danton Ferreira Vellenich (University of São Paulo, BR) <https://orcid.org/0000-0002-3223-6996>; Shelley Stall, American Geophysical Union, USA) <https://orcid.org/0000-0003-2926-8353>; Nicolas Mouquet (FRB-CESAB, Montpellier, FR) <https://orcid.org/0000-0003-1840-6984>; Marc Chaumont (LIRMM, CNRS) <https://orcid.org/0000-0002-4095-4410>; Laure Berti-Equille (Espace-Dev (IRD-UM-UG-UR-UA-UNC), Montpellier, FR) <https://orcid.org/0000-0002-8046-0570>; David Mouillot (MARBEC, University of Montpellier) <https://orcid.org/0000-0003-0402-2605>.

