Reproducibility of the Methods in Medical Imaging with Deep Learning.

Attila Simkó, Anders Garpebring, Joakim Jonsson, Tufve Nyholm and Tommy Löfstedt
Correspondance email: attila.simko@umu.se

Our goal is to collect common issues in publicly available materials and use them as guidelines for future submissions.

Background

Concerns about the reproducibility of deep learning research are more prominent than ever, with no clear solution in sight.

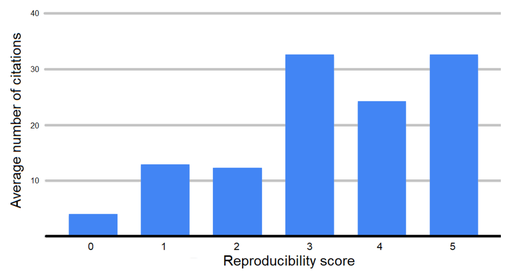
Providing public source code is a powerful solution for one aspect of reproducibility, however the quality of these repositories vary widely between projects.
A poll among the MIDL community ($N = 81$) showed that almost all of the participants have encountered issues with code repositories (96%) and they believe the quality of the code impacts the quality of the overall submission (98%). Although the poll shows that a large part of the community never receives any feedback regarding their code repositories (86%) we argue that by following a set of guidelines, the overall quality and popularity of the public code repositories could increase.

We aim to bring attention to the reproducibility concerns around MIDL submissions. And most importantly, we believe that starting a discussion about reproducibility would not only elevate the standard of individual MIDL submissions but also benefit the research community as a whole.

Approach

From the commonly encountered issues during our evaluations, we propose a set of guidelines to help the reproducibility of machine learning methods, adjusted to MIDL.
Each submission was evaluated whether each item in the guidelines were adequately addressed or not. For more details on the guidelines, take a look at the appendix of our paper.
Our evaluations (see figure on the right) show a clear increasing trend in the popularity of using code repositories and public data, and no signs of improvement for the repository related metrics.

Additionally, we have collected the number of citations for the papers submitted before 2022, and their average was plotted against our proposed reproducibility metric, see the figure below.

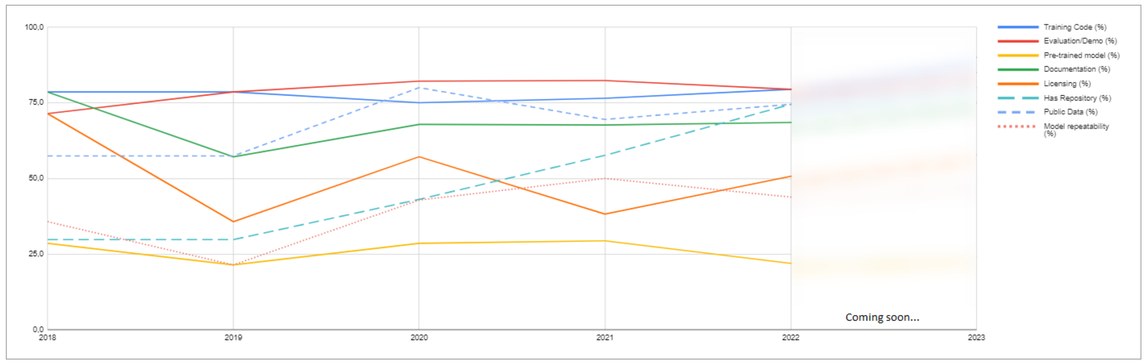


Additional motivation for thinking about reproducibility: The average number of citations for each paper collected from Google Scholar, plotted against our proposed reproducibility metric shows that with better reproducibility, the impact of the submissions increase.

Moving forward

We will continue having an open discussion about reproducibility, and adapt the guidelines as the field progresses.
Performing these evaluations yearly will show if there is a change in the attitude of the community towards reproducibility.
We would like to further explore how reproducibility metrics affect the impact of submissions.

Our evaluations are all publicly available, for more details follow the QR code.



The results for each year. The results "Has Repository", "Public Data" and "Model repeatability" evaluate all *submissions*, while the other metrics evaluate all *repositories*.

MIDL-adjusted reproducibility checklist

Manuscript specific:

- **Address reproducibility** ☐
- **Use public datasets** ☐
- **Make code public** ☐

Code specific:

- **List package dependencies** ☐ Yes/NA
  Store packages in .txt file
  List can be exported from pip/conda.
- **Trained model weights available** ☐ Yes/NA
  Standalone formats for later use (.pb, .onnx)
  or weights for retraining (.h5, .pt).
- **Code for model training** ☐ Yes/NA
  Guiding through building, compiling and training the proposed model.
- **Code for model evaluation** ☐ Yes/NA
  Guiding through how to use and assess the performance of the proposed model.
- **Code documentation** ☐ Yes/NA
  Describe everything that's available in the repository.
- **Repository licensing** ☐ Yes/NA
  Who can use your code and trained model, and in what settings.

← Find all paper evaluations online!

UMEÅ UNIVERSITET