Lessons from shortcomings in machine learning for medical imaging

Gaël Varoquaux*†‡, Veronika Cheplygina§
*INRIA, France
†McGill University, Montreal, Canada
‡Mila, Montreal, Canada
§IT University of Copenhagen, Denmark

Abstract

Machine learning for medical imaging data has many opportunities for improving patients' health, and has attracted a lot of attention in recent years. However, the progress of the field as a whole is being slowed down by the current incentives in (machine learning) research. In this report we summarize our findings based on literature and our own analysis, namely that larger datasets and more deep learning algorithms do not yet provide practical improvements in clinical problems. We provide recommendations for practices to adopt within research communities, as well as what we believe needs to change within research policy, to increase the impact of artificial intelligence in this field.

This report is a draft prepared for the OECD workshop on AI Productivity in Science. We welcome suggestions and comments.

I. Introduction

Machine learning has become inseparable from the field of medical imaging. One of the popular applications of machine learning in medical imaging is computer-aided diagnosis, where an algorithm is trained on existing images, for example brain scans of people with and without dementia, and later applied to previously unseen images to output a prediction of which category an image likely belongs to. Such algorithms have attracted a lot of attention and there are numerous reports about expert-level performance (for an overview see Liu et al. (2019)). Despite this popularity, the impact on the clinic has not been proportional to these claims: as an example, Roberts et al. (2021) found that out of 62 published studies on machine learning for COVID, none had potential for clinical use. Studies for other clinical applications of machine learning also failed to find reliable published prediction models: for pronosis after aneurysmal subarachnoid hemorrhage (Jaja et al., 2013), or stroke (Thompson et al., 2014).

In this report we build upon our recent publication (Varoquaux and Cheplygina, 2022) about evidence of the lack of progress of machine learning in medical imaging. For clarity, we include a box of concepts which might differ in their use in different communities in Table I. We summarize several examples of the lack of progress in Section II, and provide recommendations for researchers (Section III) and policy makers (Section IV) and IV on what we believe can be done to improve the situation.

Dataset	A collection of (image, label) pairs where the the label is either a category (such as disease or healthy), or another image (such as a segmentation map showing the locations of tumors).
Algorithm, classifier, model	Can refer to either a general concept (such as a neutral network), or a model trained on specific data.
Training	Fitting a model to a specific dataset, by learning parameters to transform the image into the label as well as possible.
Testing, predicting	Running a trained model on images, to output their predicted labels. Note that prediction does not imply forecasting, as the data is already available.
Overfitting	Fitting a model too closely to the data it is trained on, so that it does not generalize to previously unseen data.
Training, validation set	Parts of the dataset used for training. The validation set is used to mimic for previously unseen data, in order to avoid overfitting.
Test set	Part of the dataset reserved for evaluating the trained model. Should ideally be previously unseen data, but in practice is often available to the researcher, still leading to overfitting.

TABLE I
TERMS THAT MIGHT BE FREQUENTLY USED IN A MACHINE LEARNING IN MEDICAL IMAGING CONTEXT

II. IS AI RESEARCH MISSING ITS TARGET?

The increased popularity of machine learning in recent years is often explained by two factors: larger datasets becoming available, and deep learning which allows algorithm development without specialized domain knowledge, thus allowing more researchers to into the field. We believe the situation might not be as positive, as we illustrate below.

a) Large datasets are not a panacea: There is a tendency to expect that with a large enough dataset, a clinical task can be "solved". There are several problems here. First, not all clinical tasks can be neatly translated into machine learning tasks. Second, creating larger and larger datasets often relies on automatic methods that may introduce errors and bias into the data, see for example Oakden-Rayner (2020).

Finally, while improving algorithm training, large datasets also allow for more rigorous evaluation. Our analysis of prediction of Alzheimer's disease across six surveys and covering more than 500 publications (Fig. 1) shows that studies with larger sample sizes tend to report worse prediction accuracy. This is worrying since these studies are closer to real-life settings, though fortunately this effect is less visible in recent years.

b) Algorithm research may hit diminishing returns: A lot of research within medical imaging is focused on algorithm development, but the practical benefits of the reported accuracy gains are not always clear. We study four medical imaging competitions with significant incentives for the best algorithms. We compare the expected variability of an algorithm's performance to the gap between the performances of the top algorithms. We show that in three out of four cases, the performances of the top algorithms are within the expected variability, and are thus not practically better or worse than one another.

c) Who is included?: Deep learning studies are computational-resource intensive, and several studies in machine learning have noted how this affects who gets to do research. A method may win just because more computational resources were available (Hooker, 2020), and the representation of prestigious labs

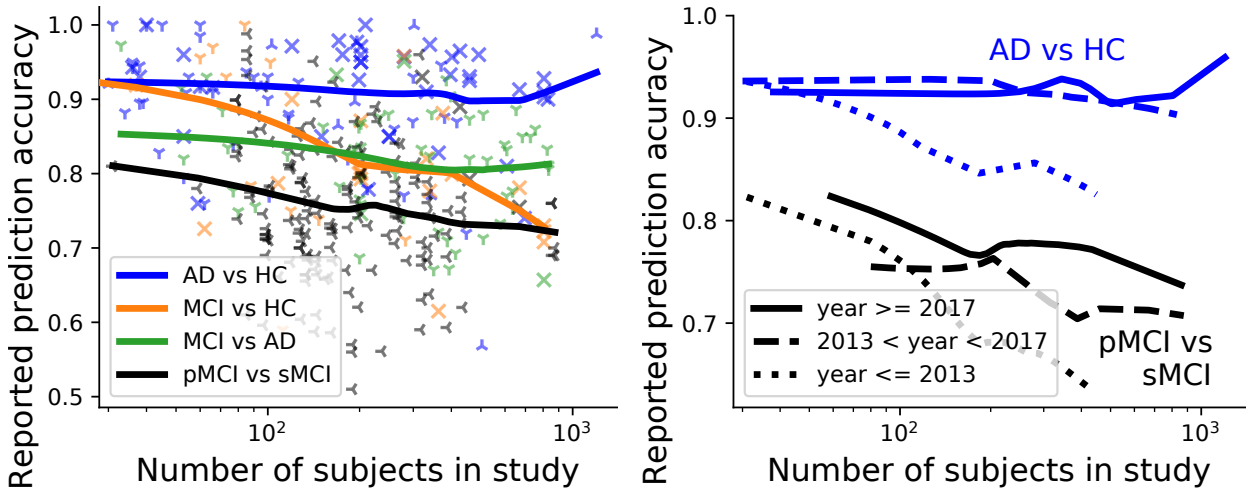

Fig. 1. **Larger brain-imaging datasets are not enough for better machine-learning diagnosis of Alzheimer's**. A meta-analysis across 6 review papers, covering more than 500 individual publications. The problem is typically formulated as distinguishing Alzheimer's disease (AD), healthy control (HC), and mild cognitive impairment (MCI), which can signal the onset of Alzheimer's. Distinguishing progressive mild cognitive impairment (pMCI) from stable mild cognitive impairment (sMCI) is the most relevant machine-learning task from the clinical standpoint. **Left:** Reported prediction accuracy as a function of the number of subjects in a study. **Right:** Same plot distinguishing studies published in different years.

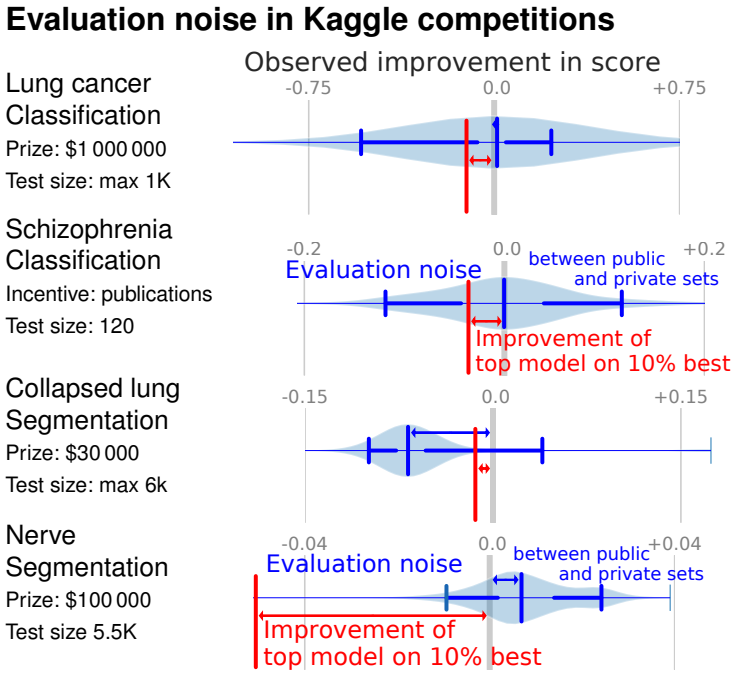

Fig. 2. **Kaggle competitions: improvements at the top of the leaderboard are often within the noise of the evaluation.** We investigated 4 medical-imaging competitions with significant incentives. The blue violin plot gives the distribution of differences between public and private leaderboards (positive means that private leaderboard is better than public leaderboard). A systematic shift between public and private set indicates overfittting or dataset bias. The width of this distribution gives the intrinsic evaluation noise of the challenge. The brown bar is the gap between the top-most model (the winner) and the 10% best model. If this gap is smaller than the width of the public-private differences, the 10% best models reached diminishing returns and did not lead to a actual improvement on new data.

and tech companies at conferences is increasing (Ahmed and Wahed, 2020). At a large medical imaging conference (MICCAI 2020), only 2% of accepted papers were from underrepresented regions (Latin America, South/South-East Asia, Africa, and Middle-East) [1], while the need for medical AI might be even greater in these regions.

III. Recommendations for research communities

There are a number of things we can already do as researchers within this community, especially those in positions of organizing conferences, and editing or reviewing papers.

a) Build awareness of data limitations: While it may not always be feasible to collect more data, it is important to understand the limitations of the data that we do have, such as the sample size or inclusion of different patient groups. On this note, we would recommend to report data characteristics and potential implications for the trained models, similar to Model Cards (Mitchell et al., 2019).

b) Reinvent benchmarking: If benchmarking algorithms is essential, a comparison needs to include both recent-and-competitive, and traditional-yet-effective methods. Furthermore, it needs to compare the range (rather than a single estimate) of each method's performance, ideally using multiple well-motivated metrics and statistical procedures (Bouthillier et al., 2021). But it might be worth considering more real-life effects of an algorithm, for example its carbon footprint, or how it affects the people it was designed to help (Thomas and Uminsky, 2020).

c) Improve publication norms: We need to let go of the idea that the only way to create impact is to publish a novel algorithm with state-of-the-art results - some of which may be overoptimistic due to the researchers' access to the data. Registered reports are a practice from psychology where the planned study reviewed and published *before* the experiments are done, and could therefore reduce publication bias. Another option is to focus on is to have different types of publications which focus on insight, such as replications or retrospective analyses of existing methods.

IV. Recommendations for research policy makers: setting incentives

As research positions and funding are often tied to publication, there are strong incentives for researchers to optimize for publication-related metrics. With the additional focus on novelty and state-of-the-art results, it is perhaps not surprising to see published methods which are over-engineered but under-validated. While some researchers might choose to opt out and/or try to change things as described in Section III, it is still a reality for many researchers in less secure positions. It is therefore important that there are also external incentives to speed up the change.

a) Quality rather than quantity: Several of the current problems stem from the way researchers are evaluated when applying for positions or for research funding. The focus on metrics like the h-index needs to be reduced in favor of, for example, evaluating five selected publications, in order to reduce the pressure to produce diminishing-returns research. This also holds for evaluating researchers based on previously acquired funding, where such biases would be propagated.

b) Funding for rigorous evaluation: Funding calls for evaluation of existing algorithms, replicating results of existing studies, prospective studies / registered reports could all be ways to support research methods which solve problems, and away from perceived novelty. Ideally these should be low-threshold schemes accessible to early career researchers.

c) Better recognition for open data and software: It should be more attractive to work on curated datasets and open-source software that everybody can use. Currently it is difficult to acquire funding, and often to publish, when working on such projects. Many team members are therefore volunteers, which creates biases against groups that are already under-represented, but which might have innovative ideas that would be vital for the field.

V. Conclusions

We present a number of problems which may be slowing down the progress of machine learning in medical imaging, based on both literature and our own analysis. In summary, not everything can be solved by larger datasets and developing more classifiers. The current focus on novelty and state-of-the-art results create methods which do not translate into practical improvements. We also provide a number of strategies to address this situation both from within the research community, and on the level of research policy. Given the huge efforts invested in AI research, failure to move the goal post may mean significant waste.

References

Ahmed, N. and Wahed, M. (2020). The de-democratization of AI: Deep learning and the compute divide in artificial intelligence research. *arXiv preprint arXiv:2010.15581*.

Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., Mohammadi Sepahvand, N., Raff, E., Madan, K., Voleti, V., Kahou, S. E., Michalski, V., Arbel, T., Pal, C., Varoquaux, G., and Vincent, P. (2021). Accounting for variance in machine learning benchmarks. In *Machine Learning and Systems*.

Hooker, S. (2020). The hardware lottery. *arXiv preprint arXiv:2009.06489*.

Jaja, B. N., Cusimano, M. D., Etminan, N., Hanggi, D., Hasan, D., Ilodigwe, D., Lantigua, H., Le Roux, P., Lo, B., Louffat-Olivares, A., et al. (2013). Clinical prediction models for aneurysmal subarachnoid hemorrhage: a systematic review. *Neurocritical care*, 18(1):143–153.

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., et al. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. In *Fairness, Accountability, and Transparency (FAccT)*, pages 220–229. ACM.

Oakden-Rayner, L. (2020). Exploring large-scale public medical image datasets. *Academic Radiology*, 27(1):106–112.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., et al. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3(3):199–217.

Thomas, R. and Uminsky, D. (2020). The problem with metrics is a fundamental problem for AI. *arXiv preprint arXiv:2002.08512*.

Thompson, D. D., Murray, G. D., Dennis, M., Sudlow, C. L., and Whiteley, W. N. (2014). Formal and informal prediction of recurrent stroke and myocardial infarction after stroke: a systematic review and evaluation of clinical prediction models in a new cohort. *BMC medicine*, 12(1):1–9.

Varoquaux, G. and Cheplygina, V. (2022). Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine*, 5(1):1–8.

NOTES

[1]MICCAI Society newsletter, August 18th, 2021