Automated machine learning as a partner in predictive modelling

Machine learning promises to underpin personalised medicine. However, the expertise required to develop and deploy state-of-the-art machine learning algorithms has contributed to the inconsistent quality of model development, the shallow range of methods considered, and the relatively poor penetrance of machine learning models in clinical use. In this Comment, we discuss the emerging field of automated machine learning and propose that it could have a central role in the future of clinical risk prediction. We argue that automated machine learning can empower both modelling experts and non-experts, democratise access to machine learning methods, and encode better standards in model development. Finally, we advocate that such frameworks be an initial step in model development to support practitioners to find the most suitable modelling approach for their question and to understand if machine learning shows benefit.

At present, the development of clinical risk prediction models is largely subject to the expertise of the modeller. The technical challenge of tuning[1] machine learning algorithms—a process of trial and error that requires an understanding of the function and range of suitable values for each algorithm-specific hyperparameter—is such that an estimated 95% of time in machine learning model development is spent programming, which requires substantial training.[2] Hyperparameters are a key component of controlling how machine learning algorithms work. However, less than a third of papers on clinical risk prediction using machine learning reported any relevant methods,[3] despite the importance of optimising machine learning models and the availability of increasingly sophisticated techniques to partly automate this process.[1,4] This paucity suggests that most developers use default settings, with the implication that resulting models will underperform.

Developing a clinical risk prediction model pipeline comprises multiple steps: imputation, predictor selection and pre-processing, model algorithm selection, training and optimisation, and fitting and calibration. Each stage has multiple possible methodological approaches, and so there might be hundreds—or even thousands—of potential pipeline combinations that could make up a complete risk prediction model. Manually searching for the most appropriate model pipeline from all the possible existing combinations is, therefore, impractical, meaning that relatively few approaches are trialled[3] and that each modelling stage is often considered independently of another. Questions could therefore arise as to why an approach was taken and whether alternatives were tested. Furthermore, over the past decade, there has been a rapid maturing of machine learning algorithms for risk prediction, increasing the practical challenge and expertise required to train the myriad statistical and machine learning options. However, as the "no free lunch" theorem[5] suggests, there exists no single method (or pipeline) that is ideal for all prediction problems. Substantial resources, data, and expertise are required to develop, evaluate, and deploy clinical risk prediction algorithms. Ethical development of these algorithms requires developers to harness the full range of modelling techniques at their disposal.

To solve these problems, software has been developed to support the application of a broad range of machine learning frameworks to any given prediction task, including the use of appropriate hyperparameter optimisation techniques.[6,7] This concept has recently been adapted for the specific challenges seen in health-care contexts and extended to optimise entire modelling pipelines.[8] Such an approach highly automates model development while keeping developers informed and, when necessary, in control of all key steps.

Within an open-source framework, automation of the technical processes of model development presents an opportunity to improve the quality and reproducibility of new models (panel). Software offering a high level of automation can efficiently select and train machine learning pipelines using any statistical or machine learning algorithm, performing a task that is currently impractical—if not impossible—even for individuals with substantial expertise. Such software can consider all interdependent stages of modelling pipelines holistically, encode state-of-the-art hyperparameter optimisation methods and model evaluation techniques, and be iteratively improved

Panel: Principles and recommendations for automated machine learning frameworks and their use in medicine

Open source
Automated machine learning software should be open source, with transparent code that is independently auditable

Clinical usefulness
Why the model will be used, by whom, in what circumstances, and with what software should be considered from the outset

Model performance
The relative performance of a wide range of different statistical and machine learning frameworks should be assessed for a given question

Transparent reporting
All stages of the model pipeline, including the management of missing data, variable pre-processing, and the statistical or machine learning framework (or frameworks) used should be clearly documented; why the final model was selected over other possible pipelines should also be logged

Deployment and independent validation
Automated machine learning software should support the deployment of resulting models, for example through an application programming interface or website, in such a way that that resulting models can also be independently validated without requiring specialist programming

by expert methodologists. Without these capabilities, model developers are unlikely to routinely apply a wide range of potential model frameworks to each clinical risk prediction task. As a result, the subsequent models might not be appropriate for the data and problem of interest. Trialling multiple statistical and machine learning approaches allows modellers to better understand where machine learning might provide advantages and where it is unnecessary. Furthermore, a high degree of automation democratises access to state-of-the-art machine learning algorithms that would otherwise require specialised knowledge that is not widely available, particularly within clinical domains. As precision medicine often requires bespoke solutions for different settings and health-care systems, such software can support the dissemination of relevant techniques, particularly in settings without access to enough biostatisticians and machine learning engineers.

Automation can also encode good practice. By focusing on improving the underlying software that is used for risk prediction problems, the research community can move on from a singular reliance on post-hoc review and the use of reporting guidelines at the stage of publication.[9] Like guidelines and checklists,[10] automated machine learning frameworks might not

cover all eventualities, but they can support best practice for both model development and evaluation. Further, applying an automated machine learning framework as an early step in model development can ensure there is a high-quality benchmark against which any alternative approach can be measured when using a given dataset for a specific prediction task.

Despite the promise of automation, the approach is relatively new and requires further development of existing tools. Importantly, automation does not remove key decisions from developers, which ultimately will underpin their clinical usefulness. In addition, no machine learning approach should be used in clinical practice without adequate model explanation; interpretability techniques should be included in automated machine learning software by default to support model debugging, development, and understanding.

As medicine becomes more personalised, the number and use of clinical risk prediction models will continue to grow. By democratising access to state-of-the-art techniques and encoding good practice to improve the quality of models, automated machine learning frameworks will probably have an increasingly important role in precision medicine. Automation should be considered as a way of augmenting practitioners, such that novel methods can become powerful tools in our arsenal, instead of languishing unused or even misused due to their complexity. Furthermore, open-source software should become an increasing focus of the research community in conjunction with modelling guidelines to enhance the clinical effect of work in this area and embed good practice.

*Thomas Callender, Mihaela van der Schaar
t.callender@ucl.ac.uk

Department of Respiratory Medicine, University College London, London WC1E 6JF, UK (TC); Department of Applied Mathematics and Theoretical Physics and Cambridge Centre for AI in Medicine, University of Cambridge, Cambridge, UK (MvdS)

1   Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. *Proc 25th ACM SIGKDD Int Conf Knowledge Discovery & Data Mining* 2019; published online July 25. https://doi.org/10.1145/3292500.3330701.

2   Sculley D, Holt G, Golovin D, et al. Hidden technical debt in machine learning systems. *Adv Neural Inf Process Syst* 2015; https://proceedings.neurips.cc/paper_files/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html.

3   Andaur Navarro CL, Damen JA, van Smeden M, et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. *J Clin Epidemiol* 2023; **154:** 8–22.

4   Bergstra J, Yamins D, Cox D. Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. *Proc 30th Int Conf Machine Learning* 2013; **28:** 115–23.

5   Wolpert DH, Macready WG. No free lunch theorems for optimization. *IEEE Trans Evol Comput* 1997; **1:** 67–82.

6   Kotthoff L, Thornton C, Hoos HH, Hutter F, Leyton-Brown K. Auto-WEKA 2.0: automatic model selection and hyperparameter optimization in WEKA. *J Mach Learn Res* 2017; **18:** 826–30.

7   Feurer M, Klein A, Eggensperger K, Springenberg J, Blum M, Hutter F. Efficient and robust automated machine learning. *Adv Neural Inf Process Syst* 2015; https://proceedings.neurips.cc/paper_files/paper/2015/hash/11d0e6287202fced83f79975ec59a3a6-Abstract.html.

8   Imrie F, Cebere B, McKinney EF, van der Schaar M. AutoPrognosis 2.0: democratizing diagnostic and prognostic modeling in healthcare with automated machine learning. *arXiv* 2022; published online Oct 21. https://doi.org/10.48550/arXiv.2210.12090 (preprint).

9   Zamanipoor Najafabadi AH, Ramspek CL, Dekker FW, et al. TRIPOD statement: a preliminary pre-post analysis of reporting and methods of prediction models. *BMJ Open* 2020; **10:** e041537.

10   Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med* 2020; **26:** 1320–24.