

Methodology and Statistics for the Behavioural, Biomedical and Social  
Sciences

## **PROPOSAL**

---

### **Tuning strategies for hyperparameters of random forests**

---

#### **Submitted by**

Judith Nève — 0070661

#### **Supervisors**

Dr. Maarten van Smeden  
Zoë Dunias

**Candidate journal:** *Statistics in Medicine*  
**FETC registration:** 22-1808  
**Word count:** 749/750

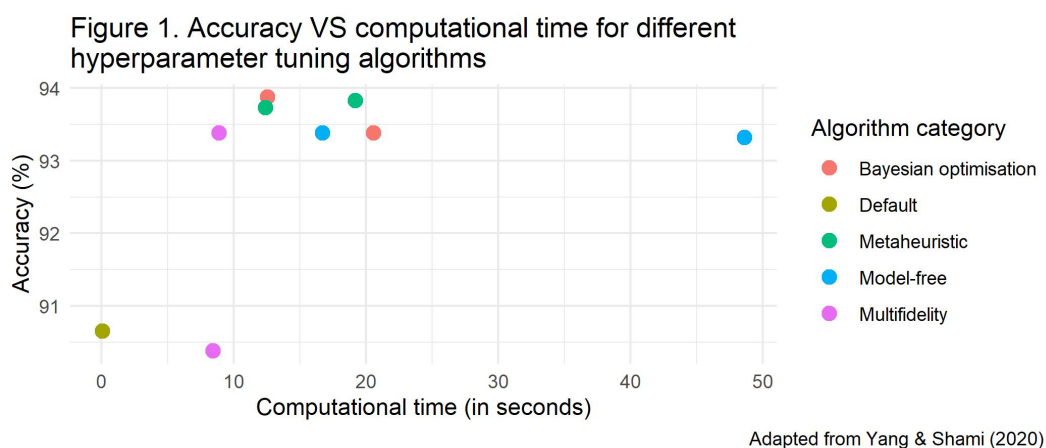
# 1 Introduction

Machine learning models are increasingly used in medicine for diagnostic and prognostic purposes [1] assisting patients' medical decision-making. For this purpose, high *classification accuracy* (i.e., the proportion of observations correctly classified) and *discrimination* (i.e., the ability to differentiate between classes) are insufficient: models also need to be well *calibrated* (i.e., risk predictions reflect patients' true risk) [2]. Miscalibrated models can lead to over- or undertreatment if over- or underestimating patients' risks.

A popular set of techniques is tree-based methods [3], among which random forests are particularly prevalent [4]. Random forests' results are influenced by hyperparameters (e.g. the minimum number of observations in a node) [5], which need to be set before fitting the model. While hyperparameter software defaults sometimes lead to appropriate discrimination [6], they can lead to severe miscalibration (work by Dunias, publication not yet available). It is then of utmost importance to identify hyperparameter tuning strategies improving calibration.

Tuning identifies hyperparameter values for which a metric (e.g. deviance) is optimal for a given model. Multiple elements need to be chosen before the tuning procedure can be carried out: the metric to optimise, the hyperparameters to tune, and the hyperparameter search algorithm. The latter two can greatly impact computation time: generally, the more hyperparameters are tuned, the larger computation time will be. Hyperparameter search algorithms may consider all possible values and combinations, or use information collected earlier in the tuning procedure to tune efficiently. Computation time is important to consider in conjunction to the improvement in model performance: greater computation times and longer training periods of models may lead to a strain on resources and increase carbon emissions. For complex models, this can be as much as five times the average emissions of a car over its lifetime [7].

Previous research suggests the most gain in random forest performance can be made by tuning the following hyperparameters: number of predictors sampled to make a split, proportion of the sample used to fit the tree, and minimum node size [8]. Most hyperparameter search algorithms improve accuracy compared to default hyperparameters, with various effects on computational time (Figure 1) [9].



While there has been some focus on comparing hyperparameter search algorithms, so far, no study has compared different optimisation metrics' effects on model performance or computational time. Additionally, hyperparameter tuning has only been studied with accuracy as the outcome, not calibration or discrimination, which are typically more important in clinical settings.

Larger computational time does not necessarily lead to an improvement in accuracy. It is unknown whether there is a point at which continuing a search algorithm no longer leads to improvements in model calibration and discrimination. The current project will thus address the following question: what are the optimal hyperparameter tuning strategies for balancing predictive performance and computational time of tree-based methods in low-dimensional settings?

## 2 Analytical strategy

There will be three simulation studies in this project to identify which hyperparameters to tune, which metric to optimise, and which search algorithm to use for optimal model performance. All studies will follow the ADEMP [10] approach to simulation studies. All studies will be conducted using 1000 simulated datasets per scenario, of which the following characteristics will vary: number of candidate predictors (8, 16, 32), event fraction (i.e., the proportion of individuals with the predicted outcome; 0.1, 0.3, 0.5), and sample size (required sample size calculated using [11], half the required sample size, twice the required sample size).

Study 1 will aim to identify hyperparameters with the most impact on model performance. We expect similar results to those of Probst and colleagues [8]. Study 2 will aim to identify an optimisation metric leading to better calibration performance. We expect all metrics to perform similarly. Study 3 will aim to identify the best hyperparameter search algorithm for model performance and computational time. We aim to identify a more efficient and better performing search algorithm than the standardly used grid search, extending prior findings [9] on model predictive performance.

Settings for the three simulation studies are summarised in Table 1. All analyses will be conducted in R on a high-performance computer, with the exception of search algorithms in study 3. These will be run in Python.

Table 1: settings for each simulation study

Study	Metric	Hyperparameters to tune	Search algorithm	Outcome measure
<b>Study 1</b>	Deviance	All hyperparameters	Grid search	Calibration slope AUC
<b>Study 2</b>	Deviance Brier score Accuracy Logarithmic loss	Selected from study 1	Grid search	Calibration slope AUC
<b>Study 3</b>	Selected from study 2	Selected from study 1	Grid search Random search Gradient-based optimisation BO-TPE Hyperband GA	Calibration slope AUC Computation time

Using results from each of the simulations, we will conclude the study with a recommendation regarding which metric to optimise, which hyperparameters to tune, and which search algorithm to use for common situation in medical prediction modeling. Information will however be available regarding the cost of other approaches such that future research will be able to make an informed decision best suited for the purpose of the studies.

## References

- [1] Benjamin S. Wessler et al. "Tufts PACE clinical predictive model registry: update 1990 through 2015". In: *Diagnostic and prognostic research* 1.1 (2017). Publisher: Springer, pp. 1–8.
- [2] Ben Van Calster et al. "Calibration: the Achilles heel of predictive analytics". en. In: *BMC Medicine* 17.1 (Dec. 2019), p. 230. ISSN: 1741-7015. DOI: 10.1186/s12916-019-1466-7. URL: <https://doi.org/10.1186/s12916-019-1466-7> (visited on 09/05/2022).
- [3] Tom M. Mitchell and Tom M. Mitchell. *Machine learning*. Vol. 1. 9. McGraw-hill New York, 1997.
- [4] Shahadat Uddin et al. "Comparing different supervised machine learning algorithms for disease prediction". en. In: *BMC Medical Informatics and Decision Making* 19.1 (Dec. 2019), p. 281. ISSN: 1472-6947. DOI: 10.1186/s12911-019-1004-8. URL: <https://doi.org/10.1186/s12911-019-1004-8> (visited on 10/11/2022).
- [5] Rafael Gomes Mantovani et al. *An empirical study on hyperparameter tuning of decision trees*. arXiv:1812.02207 [cs, stat]. Feb. 2019. DOI: 10.48550/arXiv.1812.02207. URL: <http://arxiv.org/abs/1812.02207> (visited on 09/08/2022).
- [6] Simon Bernard, Laurent Heutte, and Sébastien Adam. "Influence of Hyperparameters on Random Forest Accuracy". en. In: *Multiple Classifier Systems*. Ed. by Jón Atli Benediktsson, Josef Kittler, and Fabio Roli. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2009, pp. 171–180. ISBN: 978-3-642-02326-2. DOI: 10.1007/978-3-642-02326-2\_18.
- [7] Emma Strubell, Ananya Ganesh, and Andrew McCallum. *Energy and Policy Considerations for Deep Learning in NLP*. arXiv:1906.02243 [cs]. June 2019. DOI: 10.48550/arXiv.1906.02243. URL: <http://arxiv.org/abs/1906.02243> (visited on 10/07/2022).
- [8] Philipp Probst, Anne-Laure Boulesteix, and Bernd Bischl. "Tunability: Importance of hyperparameters of machine learning algorithms". In: *The Journal of Machine Learning Research* 20.1 (2019). Publisher: JMLR. org, pp. 1934–1965.
- [9] Li Yang and Abdallah Shami. "On hyperparameter optimization of machine learning algorithms: Theory and practice". en. In: *Neurocomputing* 415 (Nov. 2020), pp. 295–316. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2020.07.061. URL: <https://www.sciencedirect.com/science/article/pii/S0925231220311693> (visited on 09/26/2022).
- [10] Tim P. Morris, Ian R. White, and Michael J. Crowther. "Using simulation studies to evaluate statistical methods". en. In: *Statistics in Medicine* 38.11 (2019), pp. 2074–2102. ISSN: 1097-0258. DOI: 10.1002/sim.8086. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8086> (visited on 09/05/2022).
- [11] Richard D. Riley et al. "Calculating the sample size required for developing a clinical prediction model". en. In: *BMJ* 368 (Mar. 2020). Publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting, p. m441. ISSN: 1756-1833. DOI: 10.1136/bmj.m441. URL: <https://www.bmj.com/content/368/bmj.m441> (visited on 10/07/2022).

## Further Reading

- [12] Maarten van Smeden et al. "Sample size for binary logistic prediction models: Beyond events per variable criteria". en. In: *Statistical Methods in Medical Research* 28.8 (Aug. 2019). Publisher: SAGE Publications Ltd STM, pp. 2455–2474. ISSN: 0962-2802. DOI: 10.1177/0962280218784726. URL: <https://doi.org/10.1177/0962280218784726> (visited on 09/05/2022).
- [13] Ben Van Calster, Maarten van Smeden, and Ewout W. Steyerberg. *On the variability of regression shrinkage methods for clinical prediction models: simulation study on predictive performance*. arXiv:1907.11493 [stat]. July 2019. DOI: 10.48550/arXiv.1907.11493. URL: <http://arxiv.org/abs/1907.11493> (visited on 09/05/2022).
- [14] Ben Van Calster and Andrew J. Vickers. "Calibration of Risk Prediction Models: Impact on Decision-Analytic Performance". en. In: *Medical Decision Making* 35.2 (Feb. 2015). Publisher: SAGE Publi-

- cations Inc STM, pp. 162–169. ISSN: 0272-989X. DOI: 10.1177/0272989X14547233. URL: <https://doi.org/10.1177/0272989X14547233> (visited on 09/08/2022).
- [15] Gareth James et al. “Tree-Based Methods”. en. In: *An Introduction to Statistical Learning: with Applications in R*. Ed. by Gareth James et al. Springer Texts in Statistics. New York, NY: Springer US, 2021, pp. 327–365. ISBN: 978-1-07-161418-1. DOI: 10.1007/978-1-0716-1418-1\_8. URL: [https://doi.org/10.1007/978-1-0716-1418-1\\_8](https://doi.org/10.1007/978-1-0716-1418-1_8) (visited on 09/08/2022).
  - [16] Ben Van Calster et al. “A calibration hierarchy for risk models was defined: from utopia to empirical data”. en. In: *Journal of Clinical Epidemiology* 74 (June 2016), pp. 167–176. ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2015.12.005. URL: <https://www.sciencedirect.com/science/article/pii/S0895435615005818> (visited on 09/08/2022).
  - [17] Gary S. Collins and Douglas G. Altman. “Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2”. en. In: *BMJ* 344 (June 2012). Publisher: British Medical Journal Publishing Group Section: Research, e4181. ISSN: 1756-1833. DOI: 10.1136/bmj.e4181. URL: <https://www.bmj.com/content/344/bmj.e4181> (visited on 09/08/2022).
  - [18] Rafael García Leiva et al. “A Novel Hyperparameter-Free Approach to Decision Tree Construction That Avoids Overfitting by Design”. In: *IEEE Access* 7 (2019). Conference Name: IEEE Access, pp. 99978–99987. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2930235.
  - [19] Elad Hazan, Adam Klivans, and Yang Yuan. *Hyperparameter Optimization: A Spectral Approach*. arXiv:1706.00764 [cs, math, stat]. Jan. 2018. DOI: 10.48550/arXiv.1706.00764. URL: <http://arxiv.org/abs/1706.00764> (visited on 09/08/2022).
  - [20] “Developing a Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction | Journal of Applied Science & Process Engineering”. en-US. In: (Sept. 2021). URL: <https://publisher.unimas.my/ojs/index.php/JASPE/article/view/2639> (visited on 09/08/2022).
  - [21] Daud Muhajir et al. “Improving classification algorithm on education dataset using hyperparameter tuning”. en. In: *Procedia Computer Science*. Sixth Information Systems International Conference (ISICO 2021) 197 (Jan. 2022), pp. 538–544. ISSN: 1877-0509. DOI: 10.1016/j.procs.2021.12.171. URL: <https://www.sciencedirect.com/science/article/pii/S1877050921023954> (visited on 09/08/2022).
  - [22] M. T. Pratola and D. M. Higdon. “Bayesian Additive Regression Tree Calibration of Complex High-Dimensional Computer Models”. In: *Technometrics* 58.2 (Apr. 2016). Publisher: Taylor & Francis \_eprint: <https://doi.org/10.1080/00401706.2015.1049749>, pp. 166–179. ISSN: 0040-1706. DOI: 10.1080/00401706.2015.1049749. URL: <https://doi.org/10.1080/00401706.2015.1049749> (visited on 09/08/2022).
  - [23] Graeme L. Hickey et al. “Dynamic Prediction Modeling Approaches for Cardiac Surgery”. In: *Circulation: Cardiovascular Quality and Outcomes* 6.6 (Nov. 2013). Publisher: American Heart Association, pp. 649–658. DOI: 10.1161/CIRCOUTCOMES.111.000012. URL: <https://www.ahajournals.org/doi/full/10.1161/CIRCOUTCOMES.111.000012> (visited on 09/08/2022).
  - [24] B H Shekar and Guesh Dagnew. “Grid Search-Based Hyperparameter Tuning and Classification of Microarray Cancer Data”. en. In: (2019), p. 8.
  - [25] Wedad Alawad, Mohamed Zohdy, and Debatosh Debnath. “Tuning Hyperparameters of Decision Tree Classifiers Using Computationally Efficient Schemes”. en. In: (), p. 2.
  - [26] Rafael G Mantovani et al. “Hyper-Parameter Tuning of a Decision Tree Induction Algorithm”. en. In: (), p. 6.
  - [27] Ameema Zainab et al. “Performance Evaluation of Tree-based Models for Big Data Load Forecasting using Randomized Hyperparameter Tuning”. en. In: *2020 IEEE International Conference on Big Data (Big Data)*. Atlanta, GA, USA: IEEE, Dec. 2020, pp. 5332–5339. ISBN: 978-1-72816-251-5. DOI: 10.1109/BigData50022.2020.9378423. URL: <https://ieeexplore.ieee.org/document/9378423/> (visited on 09/08/2022).
  - [28] Nicole Ellenbach et al. “Improved Outcome Prediction Across Data Sources Through Robust Parameter Tuning”. en. In: *Journal of Classification* 38.2 (July 2021), pp. 212–231. ISSN: 1432-1343.

- DOI: 10.1007/s00357-020-09368-z. URL: <https://doi.org/10.1007/s00357-020-09368-z> (visited on 09/15/2022).
- [29] Leo Breiman. "Random Forests". en. In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: <https://doi.org/10.1023/A:1010933404324> (visited on 09/15/2022).
  - [30] Philipp Probst, Marvin N. Wright, and Anne-Laure Boulesteix. "Hyperparameters and tuning strategies for random forest". en. In: *WIREs Data Mining and Knowledge Discovery* 9.3 (2019). \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1301>, e1301. ISSN: 1942-4795. DOI: 10.1002/widm.1301. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1301> (visited on 09/15/2022).
  - [31] Philipp Probst and Anne-Laure Boulesteix. "To Tune or Not to Tune the Number of Trees in Random Forest". en. In: (), p. 18.
  - [32] Eric Brochu, Vlad M. Cora, and Nando de Freitas. *A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning*. arXiv:1012.2599 [cs]. Dec. 2010. DOI: 10.48550/arXiv.1012.2599. URL: <http://arxiv.org/abs/1012.2599> (visited on 09/23/2022).
  - [33] James Bergstra et al. "Algorithms for Hyper-Parameter Optimization". In: *Advances in Neural Information Processing Systems*. Vol. 24. Curran Associates, Inc., 2011. URL: <https://proceedings.neurips.cc/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html> (visited on 09/23/2022).
  - [34] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. "Practical Bayesian Optimization of Machine Learning Algorithms". In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc., 2012. URL: <https://www.proceedings.neurips.cc/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html> (visited on 09/23/2022).
  - [35] Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. "Sequential Model-Based Optimization for General Algorithm Configuration". en. In: *Learning and Intelligent Optimization*. Ed. by Carlos A. Coello Coello. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2011, pp. 507–523. ISBN: 978-3-642-25566-3. DOI: 10.1007/978-3-642-25566-3\_40.
  - [36] James Bergstra, Daniel Yamins, and David Cox. "Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures". en. In: *Proceedings of the 30th International Conference on Machine Learning*. ISSN: 1938-7228. PMLR, Feb. 2013, pp. 115–123. URL: <https://proceedings.mlr.press/v28/bergstra13.html> (visited on 09/24/2022).
  - [37] Mauro Birattari et al. "F-Race and Iterated F-Race: An Overview". en. In: *Experimental Methods for the Analysis of Optimization Algorithms*. Ed. by Thomas Bartz-Beielstein et al. Berlin, Heidelberg: Springer, 2010, pp. 311–336. ISBN: 978-3-642-02538-9. DOI: 10.1007/978-3-642-02538-9\_13. URL: [https://doi.org/10.1007/978-3-642-02538-9\\_13](https://doi.org/10.1007/978-3-642-02538-9_13) (visited on 09/24/2022).
  - [38] Xin Yao and Guojie Li. "General simulated annealing". en. In: *Journal of Computer Science and Technology* 6.4 (Oct. 1991), pp. 329–338. ISSN: 1860-4749. DOI: 10.1007/BF02948392. URL: <https://doi.org/10.1007/BF02948392> (visited on 09/25/2022).
  - [39] David Connolly. "General Purpose Simulated Annealing". In: *The Journal of the Operational Research Society* 43.5 (1992). Publisher: Palgrave Macmillan Journals, pp. 495–505. ISSN: 0160-5682. DOI: 10.2307/2583568. URL: <http://www.jstor.org/stable/2583568> (visited on 09/25/2022).
  - [40] Ihor O. Bohachevsky, Mark E. Johnson, and Myron L. Stein. "Generalized Simulated Annealing for Function Optimization". In: *Technometrics* 28.3 (1986). Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality], pp. 209–217. ISSN: 0040-1706. DOI: 10.2307/1269076. URL: <http://www.jstor.org/stable/1269076> (visited on 09/25/2022).
  - [41] Donald R. Jones, Matthias Schonlau, and William J. Welch. "Efficient Global Optimization of Expensive Black-Box Functions". en. In: *Journal of Global Optimization* 13.4 (Dec. 1998), pp. 455–492. ISSN: 1573-2916. DOI: 10.1023/A:1008306431147. URL: <https://doi.org/10.1023/A:1008306431147> (visited on 09/25/2022).