Evaluating the Performances of Tree-Based Machine Learning Methods for Detecting Interaction Effects in Clinical Research

Xinhui Ran[1], Andrea Bellavia[1]

[1]TIMI Study Group, Division of Cardiovascular Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA (US)

INTRODUCTION

- With the rise of precision medicine, it is important to account for interactive effects that might describe biological mechanisms of clinical relevance, when assessing the joint effect of several predictors on a given health endpoint.
- Assumptions-free machine learning (ML) methods offer a suitable framework for the assessment of complex interactions in clinical research. Nevertheless, if overtrained and poorly controlled they might increase the risk of data overfitting and the identification of spurious interactions with limited clinical relevance.
- We conducted a simulation study, generating several datasets with varying levels of interaction complexity and compared the performances of selected Tree-based ML methods (Random Forest[1], GBM[2], and XGBoost[3]), to detect these interaction effects and distinguish clinical vs spurious interaction.

METHOD

- Four data scenarios with different levels of interactive effects are generated.
- • Each simulated dataset with 14 predictors, both continuous and binary, with heterogenous ranges and distributions. Outcome variable Y is continuous.
- For each scenario of increased complexity, all evaluated under two sample sizes: n=1000, n=10000.
- Data is randomly split to training set and testing set (7:3). 5-fold cross-validation was used for hyper-parameter tuning. Final models were fit on the training data, and predictions were made on testing dataset.
- Results presentation:
  - the plot of overall interaction strength (H-statistics[4])
  - the plot of pairwise interaction strength (H-statistics[4])

References:

1 Wright, M. N. & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw 77:1-17.

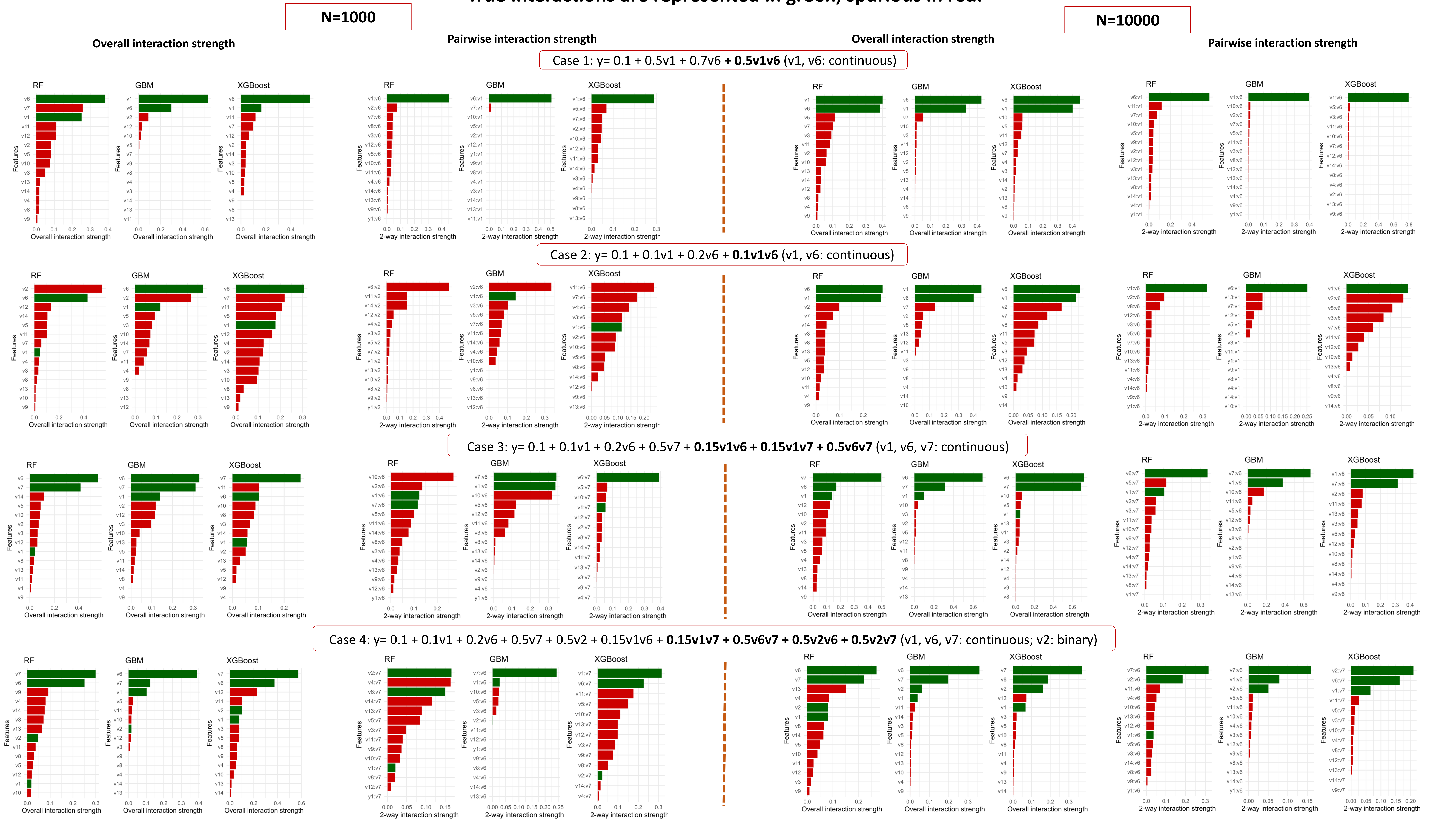
2 J.H. Friedman (2001). "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics 29(5):1189-1232.

3 Tianqi Chen and Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System", 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, 2016.

4 Molnar, C., Casalicchio, G., & Bischl, B. (2018). "iml: An R package for Interpretable Machine Learning." Journal of Open Source Software, 3(26), 7861.

Contact: xran@bwh.harvard.edu timi.org/biostatistics

Figure. H-statistics representing overall interactions and pairwise interactions as estimated by tree-based approaches in several simulated scenarios. True interactions are represented in green, spurious in red.



CONCLUSIONS

- With larger sample size and/or larger interaction effect sizes, all methods can identify true/most important interaction effects. With a larger effect size (i.e., case 1), even with N=100, both GBM and XGBoost still correctly identify the most important interaction effect (result not shown). Large sample sizes also reduce the risk of spurious interactions in the case of high correlations (result not shown).
- With smaller effect sizes and sample sizes there is a higher chances of false positives (i.e., spurious interactions). In such cases, gradient boosting outperforms random forests.
- Well-trained and tuned ML approaches can distinguish true from spurious interactions in most settings. Next steps will focus on binary and time-to-event outcomes.