Deriving Interpretable Thresholds For Variable Importance In Random Forests By Permutation

Blanco M[1], Müller T[1], Schlieker L[1], Ott A[2], Hornung R[3, 4] and Buchner H[1]

[1]Staburo GmbH, Munich, Germany
[2]Roche Diagnostics GmbH, Penzberg , Germany
[3]Institute for Medical Information Processing, Biometry and Epidemiology, University of Munich
[4] Munich Center for Machine Learning (MCML)

MOTIVATION

- Variable Importance (VIMP) in Random Forest (RF) is relevant for variable selection, interpretability, domain knowledge and decision making.
- **However**, there is no theoretical null-distribution or thresholds for significance testing.

Different objectives!

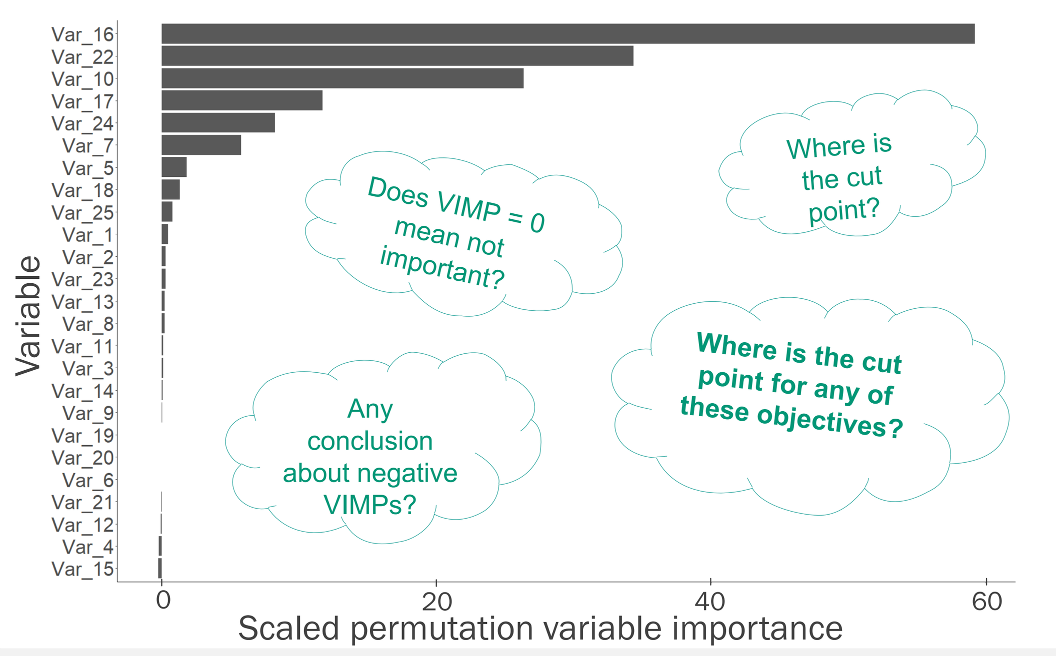

Figure 1: *VIMP output of a RF.*

Boruta (Kursa et al. (2010)) is a current method that:

- Is well performing.
- Permutes original variables independently (shadows).
- Considers a variable as informative if its VIMP is consistently higher than the maximum of the shadows.

GOALS

- **Interpretable Testing** → clear null-hypothesis.
- **Flexible Thresholds** → support for multiple testing scenarios.
- **Bias-Adjusted Comparisons** → each variable is compared to its own shadow, addressing variable-specific biases (Strobl et al. (2007) & Nicodemus et al. (2010)).
- **Visual Insights** → easily interpretable outputs.

PROPOSED APPROACH

H_0: The VIMP of X ≤ the VIMP of its own shadow.

	Original			Shadow		
y	x_1	x_2	x_3	$x_1^{(s)}$	$x_2^{(s)}$	$x_3^{(s)}$
0	1	4	2	3	2	6
1	2	3	4	1	4	2
1	3	2	6	4	1	8
0	4	1	8	2	3	4

1. Copy the set of predictors and randomly permute its rows.
2. Paste the resulting dataset to the original one.
3. Run RF and calculate VIMP (Scaled Mean Decrease in Accuracy) on the new merged dataset.
4. Repeat the process to obtain *n* VIMP scores for each variable.

DECISION CRITERION:

$$p_j = 1 - \hat{F}_{VI_j^{(s)}}\left(median(VI_j)\right) \le \alpha \rightarrow x_j \text{ is informative}$$

POOLING:

Non-parametric estimates of small p-values.

$$p_j^{(pooled)} = 1 - \hat{F}_{\{\widetilde{VI}_1^{(s)},\ldots,\widetilde{VI}_j^{(s)},\ldots,\widetilde{VI}_m^{(s)}\}}\left(median(\widetilde{VI}_j)\right) \le \alpha$$

Where

$$\widetilde{VI}_j^{(s)} = \frac{VI_j^{(s)} - mean(VI_j^{(s)})}{sd(VI_j^{(s)})} \quad \text{and} \quad \widetilde{VI}_j = \frac{VI_j - mean(VI_j^{(s)})}{sd(VI_j^{(s)})}$$

PRE-SELECTION OF VARIABLES:

Increases sensitivity and **reduces runtime.**

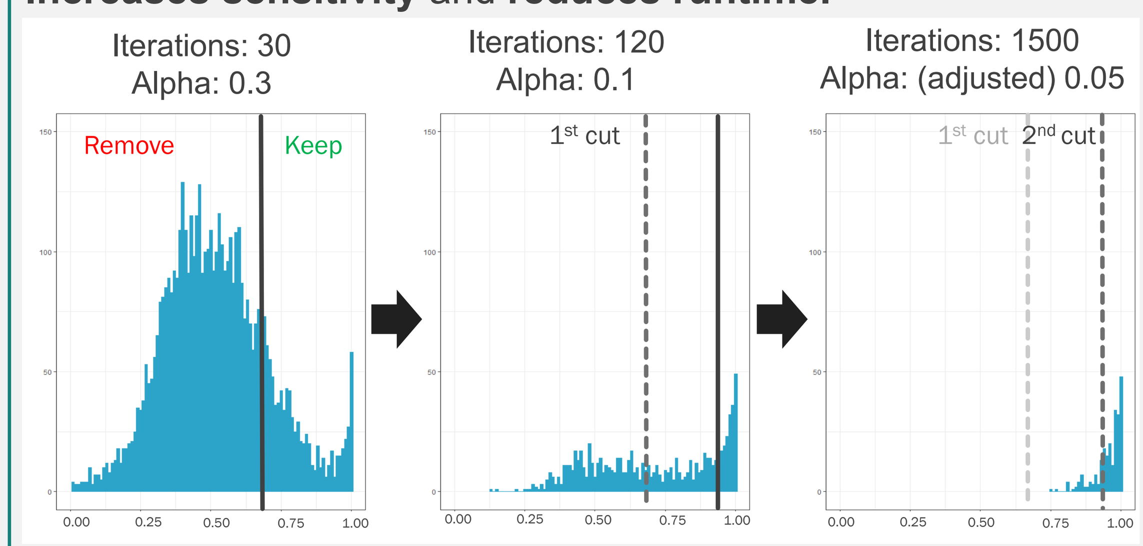

Figure 2: *P-values density and thresholds through the pre-selection process.*

SIMULATION

DESIGN:

It was used by Degenhardt et al. (2019).

$$y = 0.25(4x_1) + \frac{4}{1 + exp(-20(x_2 - 0.5))} + 3x_3 + \varepsilon$$

Where $\varepsilon \sim N(0, 0.2)$, $x_1, \ldots, x_6 \; i.i.d \sim U(0,1)$ and used to generate the correlated predictor variables according to:

$$v_i^{(j)} = x_i + 0.001 + \left(\frac{0.5(j-1)}{p-1}\right) \cdot N(0, 0.3)$$

$v_i^{(j)}$ is the *j-th* variable in group *i*, for $j = 1, \ldots, p$ and $i = 1, \ldots, 6$.

- Variables in the same group are noisy measurements of latent effect (x_i).
- **Informative variables:** 3 groups of $p = 10$ variables with correlation within group.
- **Uninformative variables**: 30 correlated + 4940 uncorrelated.
- Total of 5000 variables.
- 50 replicates.

RESULTS WITH PRE-SELECTION AND FALSE DISCOVERY RATE (FDR) ADJUSTMENT:

Benjamini-Hochberg (BH) adjustment method.

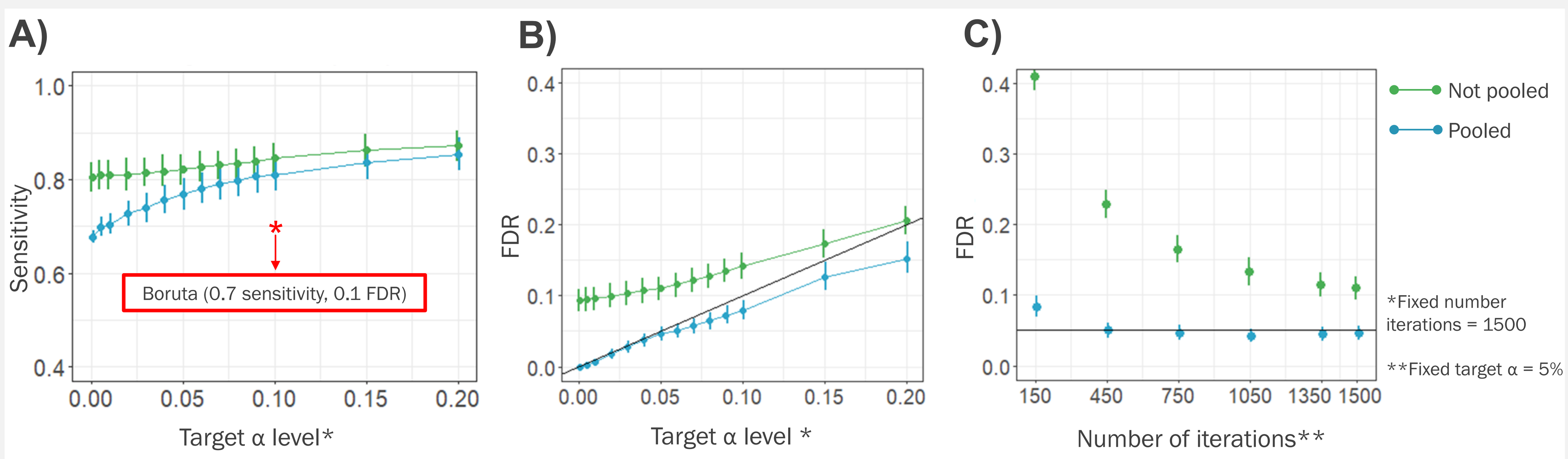

Figure 3: *A) Sensitivity vs. target α level. B) Observed FDR vs. target α level. C) Observed FDR vs. number of iterations.*

ILLUSTRATION: ALZHEIMER DISEASE STUDY

- Craig-Shapiro et al. (2011).
- **Goal**: **Identify** potential cerebrospinal fluid (CSF) **biomarkers** to improve the **early detection of Alzheimer.**
- 190 analytes in 333 CSF samples from **cognitively normal** and **mildly demented** patients.

Confirmatory analysis:

- **Four different machine learning algorithms** used: Boosted Trees, Nearest Shrunken Centroids, Random Forest and Partial Least Squares.
- Assessment of the **top 15 predictors** based on each algorithm's **built-in important measure**.

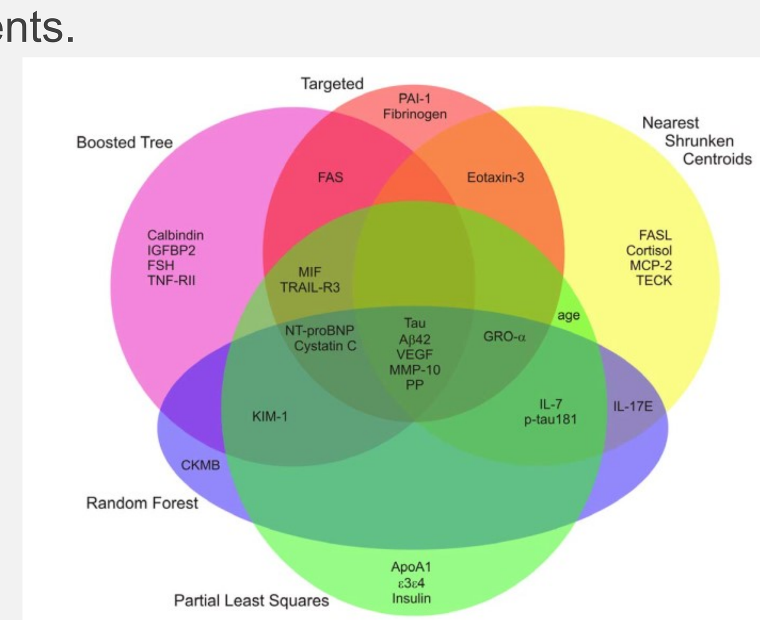

Figure 4: *Result of confirmatory analysis (from Craig-Shapiro et al. (2011))*

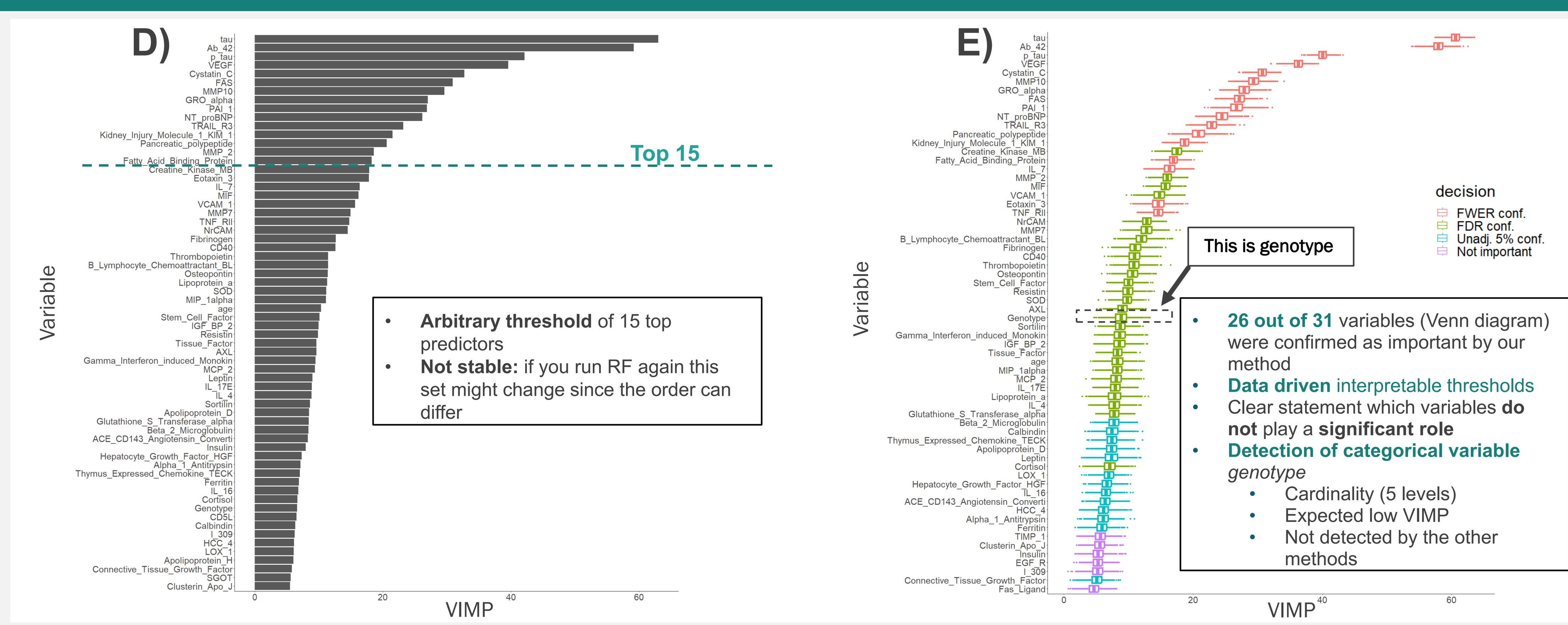

Figure 5: *D) RF VIMP. E) Proposed approach (with pre-selection of variables) VIMP results.*

CONCLUSION

- **Main method:** "p-values" that can be **adjusted** and are **interpretable**.
- **Extensions:** Pre-selection and pooling **improve performance** on high-dimensional data.
- **Bias correction:** direct comparison criterion improves handling **bias in VIMPs** found in literature (Strobl et al. (2007), Nicodemus et al. (2010) and illustration).
- **Thresholds: flexible** to the **user's need**.
- **Visual guidance: extends** the usual RF VIMP plot to **3 levels of significance.**

REFERENCES

- Kursa, M. B., Jankowski, A., & Rudnicki, W. R. (2010). Boruta–a system for feature selection. *Fundamenta Informaticae*, *101*(4), 271-285.
- Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, *8*(1), 1-21
- Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC bioinformatics*, *11*, 1-13.
- Degenhardt, F., Seifert, S., & Szymczak, S. (2019). Evaluation of variable selection methods for random forests and omics data sets. *Briefings in bioinformatics*, *20*(2), 492-503.
- Craig-Schapiro, R., Kuhn, M., Xiong, C., Pickering, E. H., & Liu, J. (2011). Multiplexed Immunoassay Panel Identifies Novel CSF Biomarkers for Alzheimer's.