Allergy

WILEY

Allergy EUROPEAN JOURNAL OF ALLERGY AND CLINICAL IMMUNOLOGY EAACI

CORRESPONDENCE

Correspondence: Accuracy Is Not Enough: Stability-Aware Feature Selection for Reproducible Biomarker Discovery

Yoshiyasu Takefuji

Faculty of Data Science, Musashino University, Tokyo, Japan

Correspondence: Yoshiyasu Takefuji (takefuji@keio.jp)

Received: 19 August 2025 | **Revised:** 2 September 2025 | **Accepted:** 17 September 2025

Funding: The author received no specific funding for this work.

Keywords: biomarker discovery | feature selection | random forest | reproducibility | stability

ABSTRACT

Random forest (RF) models can achieve high predictive accuracy, yet their model-specific feature importances may be unstable and misleading. Using an allergy benchmark dataset (10,000 instances, 11 features), we compared five selection strategies—RF, logistic regression, feature agglomeration (FA), highly variable gene selection (HVGS), and Spearman correlation—evaluating cross-validated accuracy with the top five features and after removing the top two (reselecting the top three). RF attained 0.9999 accuracy with the top five but fell to 0.8836 and showed unstable rankings; logistic regression maintained 0.9116 but was also unstable. FA, HVGS, and Spearman achieved near-perfect accuracy (0.9999) with the top five and modest declines (0.9076–0.9116) with stable rankings. Results underscore that accuracy does not imply reliable importance; stability-aware, model-agnostic, or unsupervised methods better support reproducible biomarker discovery.

Shahbazi Khamas et al. investigated complementary predictors of asthma attacks in children, including the salivary microbiome, serum inflammatory mediators, and prior attack history [1]. They trained a random forest (RF) model on a training set and optimized feature selection using fivefold cross-validation with 10 repetitions. This strategy aims to balance bias and variance in performance estimation and helps ensure that each fold contains a representative distribution of outcome classes. Classification accuracy was used as the primary performance metric [1].

However, using RF for feature selection raises important theoretical and empirical concerns due to its model-specific nature, which can lead to misleading interpretations. While supervised learning models like RF use ground-truth labels to validate target prediction accuracy, the feature importance values derived from these models do not have an analogous ground truth for validation, leading to erroneous interpretations [2–5]. Variable importance measures in RF show a bias towards correlated predictor variables [3]. It is critical to distinguish between two types of accuracy in supervised machine learning models: target prediction accuracy and the reliability of feature importance estimates. Feature importance reflects a variable's contribution to a given model's predictions, not necessarily a true causal or associative relationship. As a result, high predictive accuracy does not guarantee that the inferred feature importances are reliable or stable across resamples, models, or slight perturbations of the data [6–9].

To probe these issues, this study analyzes an allergy benchmark dataset comprising 10,000 instances and 11 features to assess the effectiveness of feature selection and the stability of feature rankings [10]. Allergy prediction refers to using the model to distinguish allergic from non-allergic outcomes based on measured independent variables, while feature importance quantifies the associations between each predictor and the target outcome, with the rankings indicating the relative influence of those predictors within the model. The dataset has the following

https://doi.org/10.1111/all.70075

variables: Age, Gender, Family_History, Previous_Reaction, Symptoms, Food_Type, Food_Frequency, Medical_Conditions, IgE_Levels, Severity_Score, and Allergic (target).

Five approaches are compared: supervised models (RF and logistic regression), unsupervised methods such as feature agglomeration (FA) and highly variable gene selection (HVGS), and a non-target, supervised statistical procedure such as Spearman's correlation with *p*-values. Top five features are selected by each method, reduced datasets are constructed, and cross-validation performance is evaluated on these reduced sets. The premise is that superior feature selection should yield higher cross-validation accuracy. In a complementary stability test, the top two features are removed from the original dataset, the top three features are then reselected, and rankings are compared against the initial top five to evaluate consistency and robustness of feature selection.

Table 1 shows that several methods achieved near-perfect cross-validation performance with the top five features but differed in stability and robustness when the top two features were removed. RF reached 0.9999 accuracy with the top five features but dropped to 0.8836 with the top three and exhibited unstable feature rankings, indicating sensitivity to feature set changes. Logistic regression achieved 0.9116 accuracy with both the top five and top three features, yet its rankings were unstable, suggesting multiple near-equivalent feature subsets. FA and HVGS each attained 0.9999 accuracy with the top five features and 0.9076 with the top three while maintaining stable rankings, reflecting consistent selection under perturbation. Spearman correlation likewise delivered 0.9999 accuracy with the top five and 0.9116 with the top three, with stable rankings, highlighting reliable, reproducible feature prioritization. Overall, stability favored the unsupervised and correlation-based approaches, whereas RF and logistic regression produced less consistent rankings despite competitive accuracy. For purposes of reproducibility and transparency, Python code, allergy.py, is publicly available at GitHub [11].

In summary, high predictive accuracy alone does not ensure reliable feature importance or stable feature rankings. On the benchmark dataset, RF achieved excellent accuracy with the top five features but showed instability and performance degradation when top features were removed, underscoring its sensitivity to feature set changes. Logistic regression maintained accuracy but also yielded unstable rankings, suggesting nonunique solutions. In contrast, unsupervised methods (FA, HVGS) and Spearman correlation combined near-perfect accuracy with stable rankings, indicating more reproducible feature prioritization. These findings caution against relying on model-specific importances for inference and support using stability-aware, model-agnostic, or unsupervised selection for transparent, reproducible biomarker discovery. Reproducible code (allergy.py) is available on GitHub [11].

TABLE 1 | Summarizes results of cross-validation and feature ranking stability.

Model	Top 5 cross-validation accuracy	Top 3 cross-validation accuracy	Feature ranking stability
Random forest	0.9999	0.8836	Unstable
Logistic regression	0.9116	0.9116	Unstable
Feature agglomeration	0.9999	0.9076	Stable
HVGS	0.9999	0.9076	Stable
Spearman	0.9999	0.9116	Stable

Author Contributions

Yoshiyasu Takefuji completed this research and wrote this article.

Ethics Statement

The author has nothing to report.

Consent

The author has nothing to report.

Conflicts of Interest

The author declares no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are openly available in food-allergy-dataset at https://github.com/RuthvikUppala30/food-allergy-dataset.

Linked Articles

This article is linked to Shahbazi Khamas et al. papers. To view these articles, visit https://doi.org/10.1111/all.70004.

References

1. S. Shahbazi Khamas, P. Brinkman, A. H. Neerincx, et al., "Complementary Predictors for Asthma Attack Prediction in Children: Salivary Microbiome, Serum Inflammatory Mediators, and Past Attack History," *Allergy* (2025): 1–14, https://doi.org/10.1111/all.70004.

2. M. L. Wallace, L. Mentch, B. J. Wheeler, et al., "Use and Misuse of Random Forest Variable Importance Metrics in Medicine: Demonstrations Through Incident Stroke Prediction," *BMC Medical Research Methodology* 23, no. 1 (2023): 144, https://doi.org/10.1186/s12874-023-01965-x.

3. C. Strobl, A. L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional Variable Importance for Random Forests," *BMC Bioinformatics* 9 (2008): 307, https://doi.org/10.1186/1471-2105-9-307.

4. R. Dunne, R. Reguant, P. Ramarao-Milne, et al., "Thresholding Gini Variable Importance With a Single-Trained Random Forest: An Empirical Bayes Approach," *Computational and Structural Biotechnology Journal* 21 (2023): 4354–4360, https://doi.org/10.1016/j.csbj.2023.08.033.

5. M. Huti, T. Lee, E. Sawyer, and A. P. King, "An Investigation Into Race Bias in Random Forest Models Based on Breast DCE-MRI Derived Radiomics Features," *Clinical Image-Based Procedures, Fairness of AI in Medical Imaging, and Ethical and Philosophical Issues in Medical Imaging* 14242 (2023): 225–234, https://doi.org/10.1007/978-3-031-45249-9_22.

6. T. Parr, J. Hamrick, and J. D. Wilson, "Nonparametric Feature Impact and Importance," *Information Sciences* 653 (2024): 119563, https://doi.org/10.1016/j.ins.2023.119563.

7. D. S. Watson and M. N. Wright, “Testing Conditional Independence in Supervised Learning Algorithms,” *Machine Learning* 110, no. 8 (2021): 2107–2129, https://doi.org/10.1007/s10994-021-06030-6.

8. Z. C. Lipton, “The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability Is Both Important and Slippery,” *Queue* 16, no. 3 (2018): 31–57, https://doi.org/10.1145/3236386.3241340.

9. A. Fisher, C. Rudin, and F. Dominici, “All Models Are Wrong, but Many Are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously,” *Journal of the American Statistical Association* 20 (2019): 177.

10. “GitHub,” food_allergy_dataset.csv, https://github.com/RuthvikUppala30/food-allergy-dataset.

11. “GitHub,” allergy.py, https://github.com/y-takefuji/6_skills_for_data_scientists/.

Biography

Yoshiyasu Takefuji holds notable global rankings in several fields. He ranks 54th out of 395,884 scholars in neural networks (AI), 23rd out of 47,799 in parallel computing, and 14th out of 7222 in parallel algorithms. Furthermore, he ranks the highest in AI tools and human-induced error analysis, underscoring his significant contributions to these domains.