

Decision Trees and Random Forests: A Comparative Overview (2021–2025)

Machine learning practitioners frequently use **Decision Trees** and **Random Forests** for classification and regression tasks. A Decision Tree is simple to interpret: it splits data by feature thresholds to route instances to leaf nodes with predictions. However, deep trees can overfit training data, capturing noise instead of general patterns. Researchers mitigate this via pruning or restricting depth. In contrast, Random Forests are ensembles of many trees built from bootstrap samples and feature subsets (*bagging* + *random subspace*). This ensemble averaging reduces variance and often yields better generalization.

Decision Trees:

Decision Trees offer interpretability and simplicity, but deep trees risk **overfitting**. Recent research emphasizes pruning strategies such as ACV-CCP (Halabaku & Bytyçi, 2024) to reduce this issue and improve generalization.

Random Forests:

Random Forests, introduced by Breiman, build on earlier ideas such as Ho's random subspace method. By averaging predictions from multiple randomized trees, they reduce variance and improve accuracy. However, the ensemble behaves like a "black box," making interpretability a challenge. Recent surveys (Haddouchi & Berrado, 2024) provide taxonomies of interpretability techniques, while approaches like *majoritary reasons* (Audemard et al., 2021) aim to balance explanation complexity and accuracy.

Applications and Challenges:

Random Forests have been widely applied across fields. For instance, in ecology, they uncover nonlinear interactions in ecosystems (Simon et al., 2023). In medicine, however, fully grown RF trees may yield overly sharp probability peaks, sometimes harming calibration (Barreñada et al., 2024). An improved RF model accounting for correlation among trees enhances accuracy (Sun et al., 2024). Other works compare RF with XGBoost under different sampling techniques (Imani et al., 2025).

Summary:

Decision Trees provide interpretability but are prone to overfitting. Random Forests generally outperform them in predictive accuracy due to reduced variance, but at the cost of interpretability. Ongoing research (2021–2025) focuses on bridging this gap with improved pruning methods, interpretability frameworks, and hybrid ensemble strategies, making Random Forests more transparent and robust.

References (2021–2025)

1. Halabaku, E., & Bytyçi, E. (2024). Overfitting in Machine Learning: A Comparative Analysis of Decision Trees and Random Forests. *Intelligent Automation & Soft Computing*, 39(6).
2. Haddouchi, M., & Berrado, A. (2024). A survey and taxonomy of methods interpreting random forest models. *arXiv preprint*.
3. Audemard, G., Bellart, S., Bounia, L., Koriche, F., Lagniez, J.-M., et al. (2021). Trading Complexity for Sparsity in Random Forest Explanations. *arXiv preprint*.
4. Simon, S. M. et al. (2023). Interpreting random forest analysis of ecological models to identify mechanisms. *Scientific Reports*.
5. Barreñada, L. et al. (2024). Understanding overfitting in random forest for probability estimation. *Diagnostic Prognostic Research*.
6. Sun, Z. et al. (2024). An improved random forest based on the classification accuracy and correlation measurement of decision trees. *Expert Systems with Applications*.
7. Imani, M., Beikmohammadi, A., & Arabnia, H. R. (2025). Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS. *Technologies*, 13(3), 88.
8. Kaveh, N. et al. (2023). Comparative analysis of random forest, exploratory methods, and predictor selection. *Ecological Modelling*.