Random Forest Classifier Performance in Low False Alarm and Low Missed Detection Regimes

Kush R. Varshney[1,2], Ryan J. Prenger[2], Barry Y. Chen[2], Tracy D. Lemmond[2], and William G. Hanley[2]

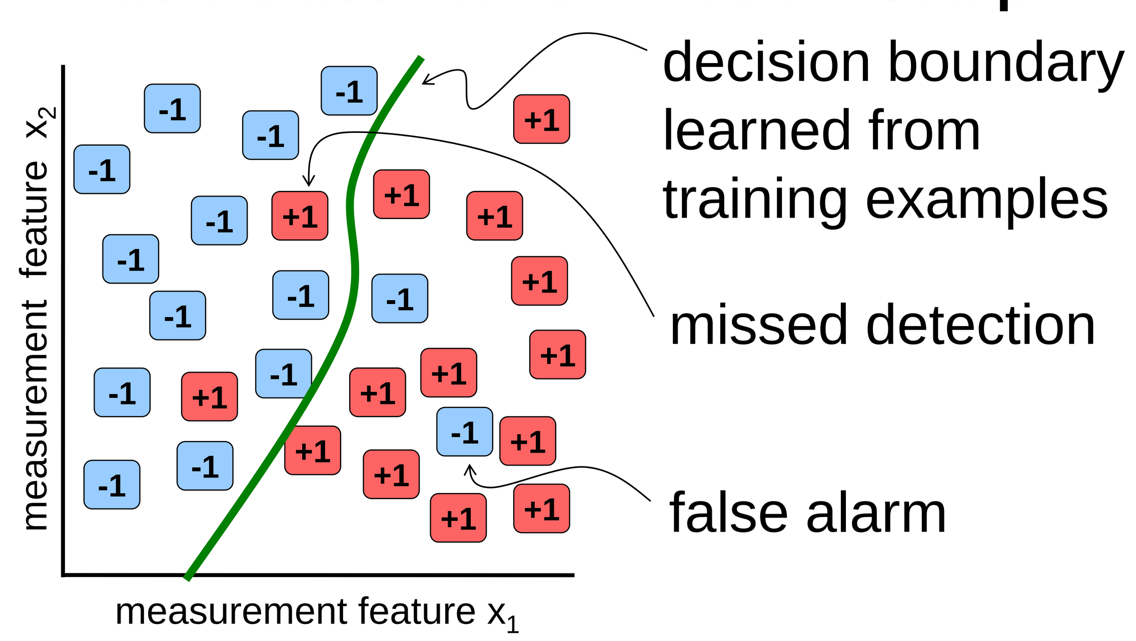
(1) Massachusetts Institute of Technology; (2) National Security Engineering Division, Lawrence Livermore National Laboratory

Overview

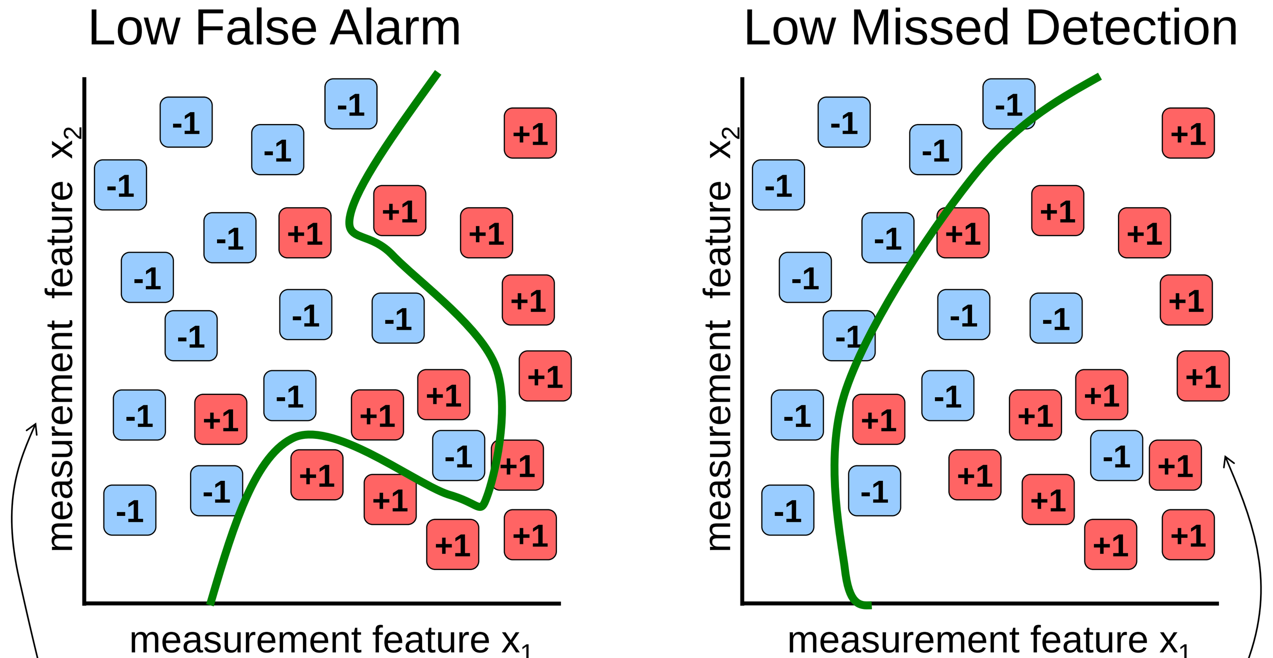
- Different types of errors have different costs in most decision-making problems
- Binary classification problems have two types of errors: false alarms and missed detections
- We present new analysis of the Random Forest, a state-of-the-art ensemble classifier, that takes the two types of errors into account
- The theoretical analysis is supported by comparisons to empirical classification performance

Detection and Classification

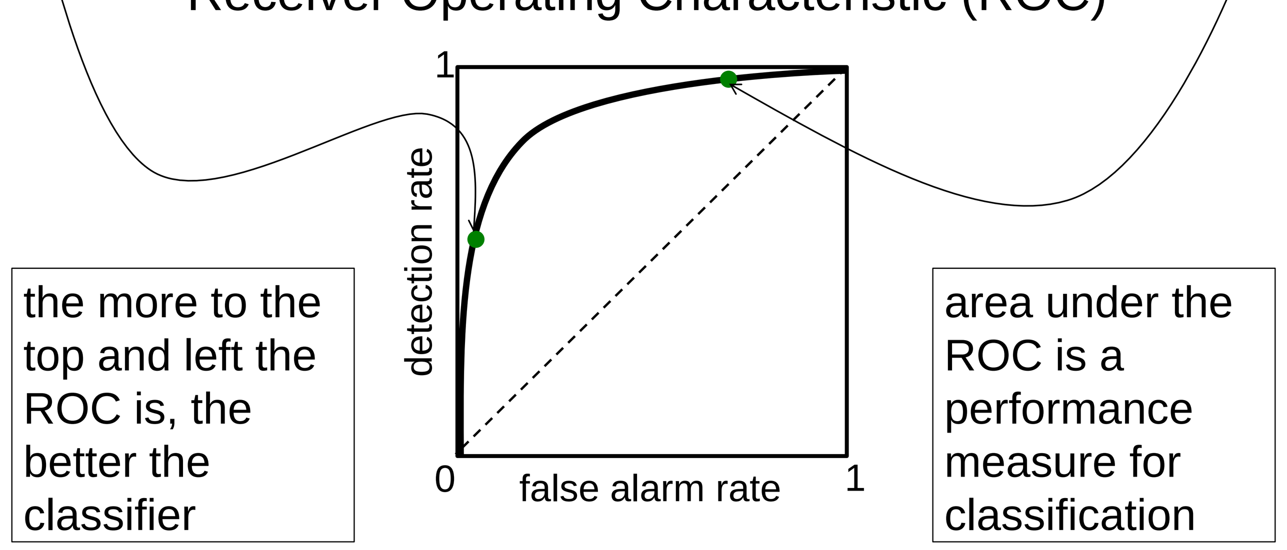
Basic Classification Problem Setup



Different Operating Points

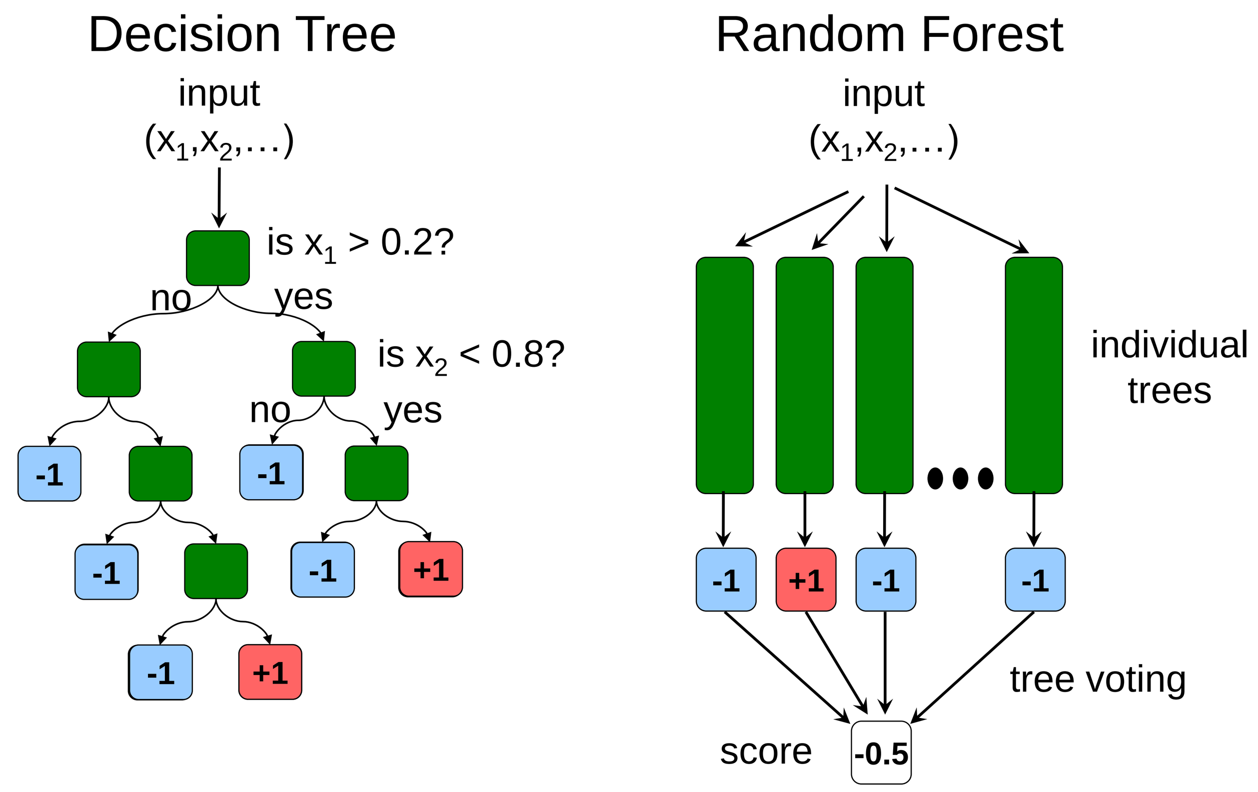


Receiver Operating Characteristic (ROC)



Random Forest Classifier

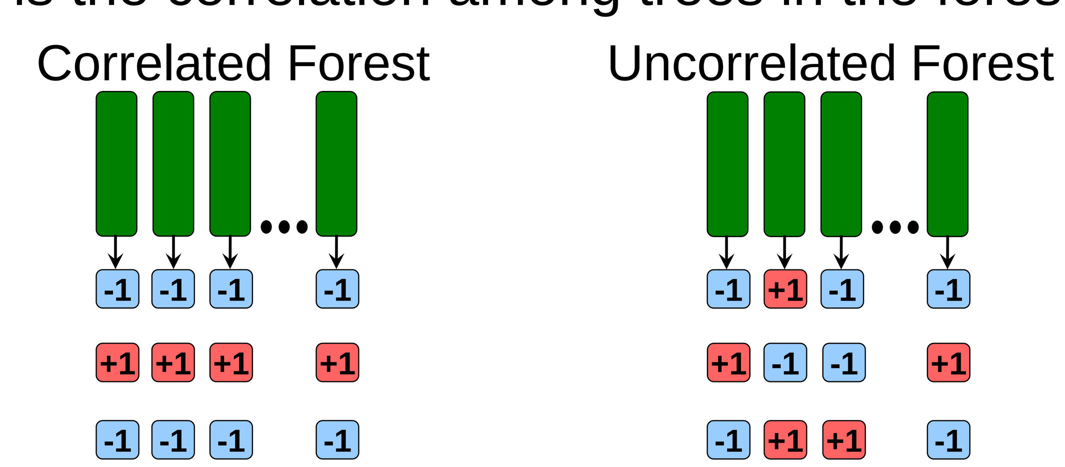
- A Random Forest is an ensemble of decision trees trained using bootstrapped samples



- At each node in the decision tree, m features are randomly selected and used to determine a linear decision boundary
  - m is known as the split dimension
  - linear splits can be done by various methods: Gini impurity, Fisher's linear discriminant analysis, and Anderson-Bahadur linear discriminant analysis
- Average score from tree votes is compared to a threshold to determine classification
  - different thresholds are different operating points

Generalization Error Bound

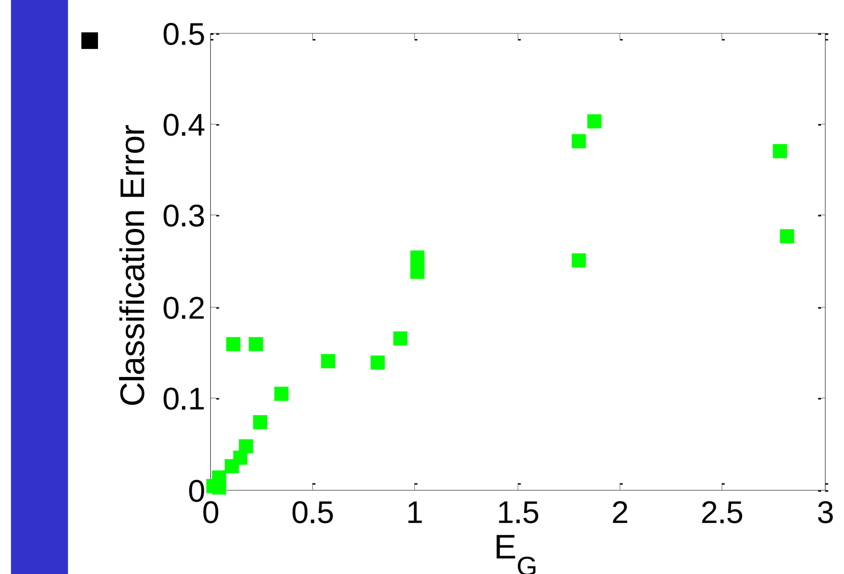
- total error rate ≤ E_G
- $E_G = \rho(1-s^2)/s^2$
  - based on Chebyshev inequality [Breiman, 2001]
- ρ is the correlation among trees in the forest



- s is the strength of the forest
- For small overall classification error, want high strength and low correlation
- Result is only for overall error
  - false alarms and missed detections are not considered separately

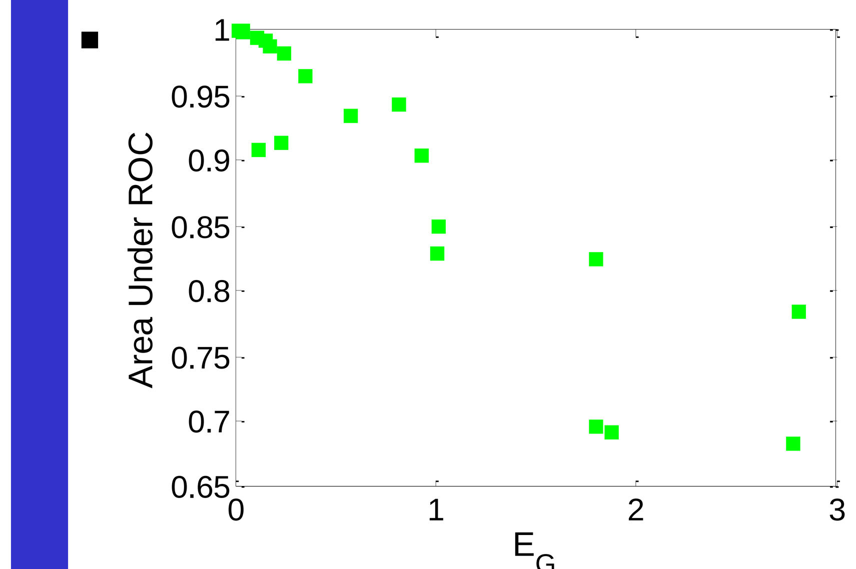
Empirical Comparison To Bound

- Comparison of classification error and E_G for 20 different datasets with Gini Random Forest



Breiman's bound is loose, but has a strong correlation with the classification error over many datasets. Thus the bound is predictive of performance.

- Comparison of area under ROC and E_G for 20 different datasets with Gini Random Forest



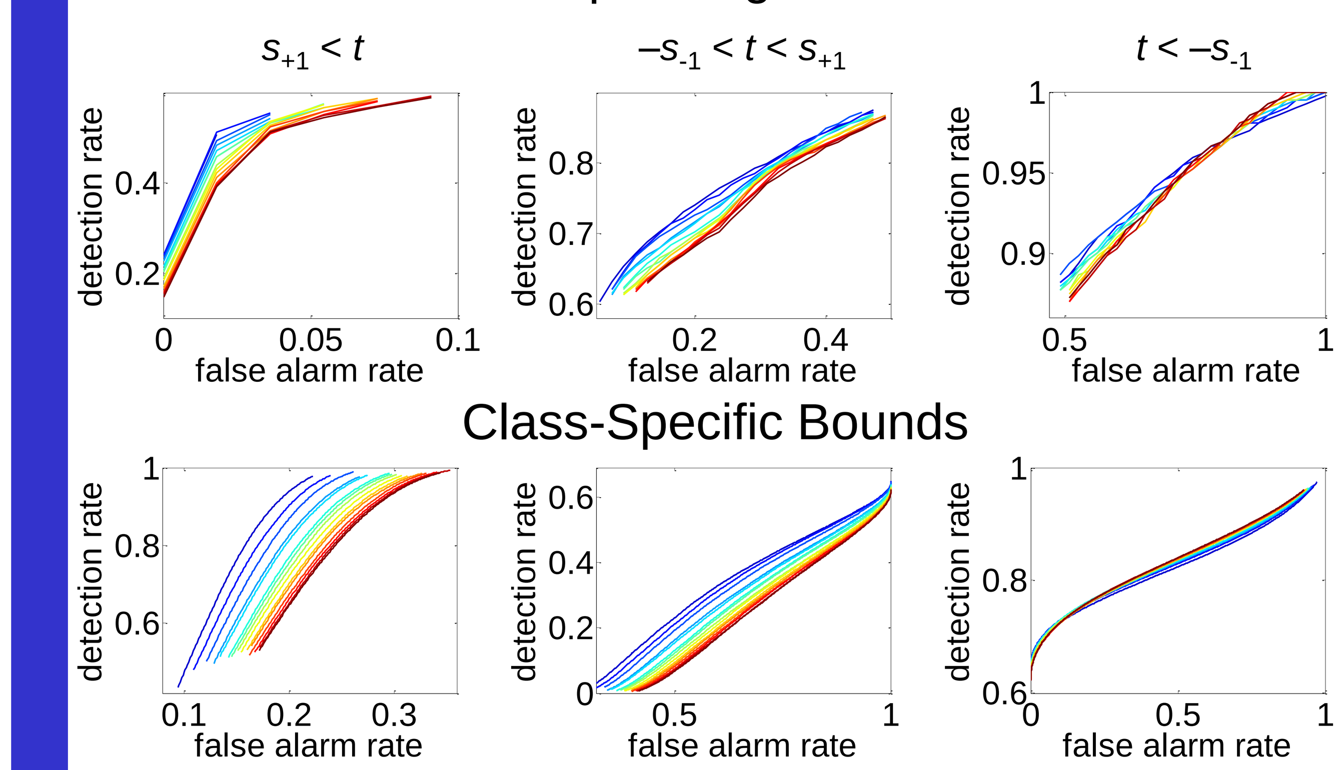
Breiman's bound also has a strong correlation with the area under the ROC over many datasets. (Larger area implies better classifier.)

Class-Specific Error Bounds

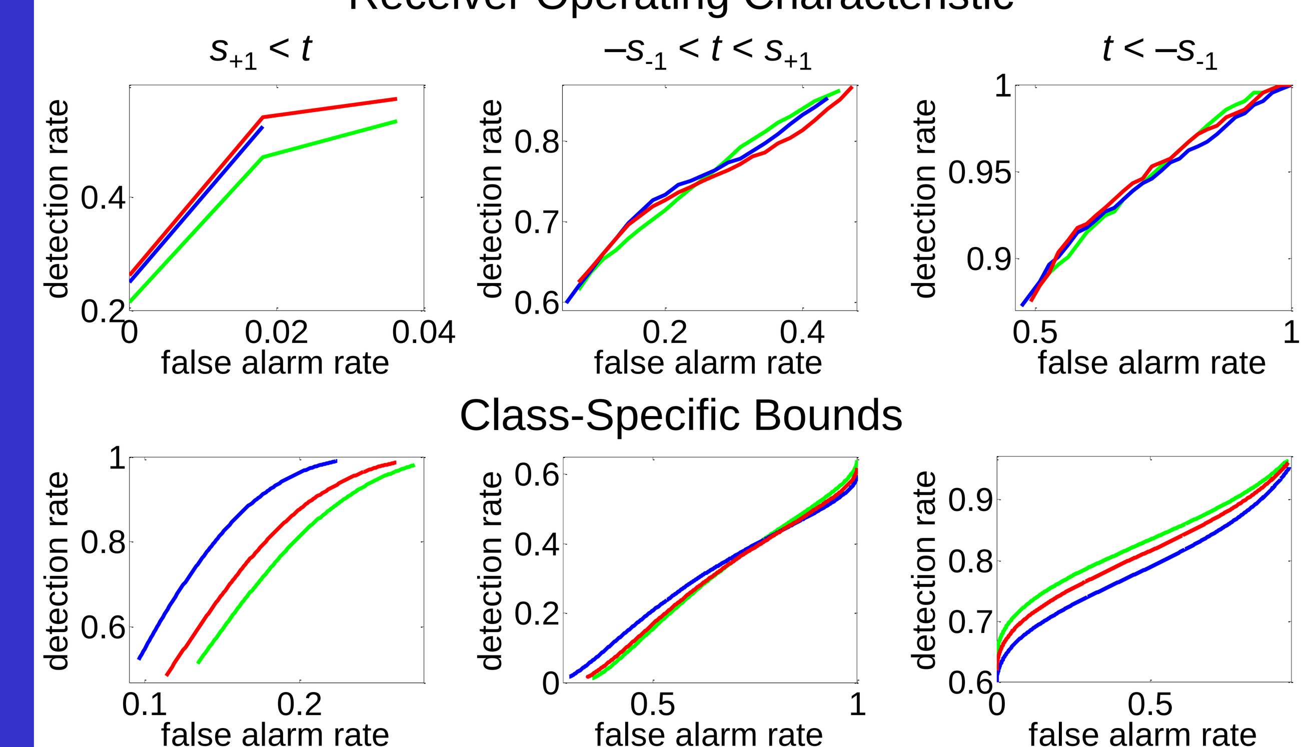
- Would like error bounds for false alarm rate and missed detection rate separately
- Based on class-specific strengths s_{-1}, s_{+1}, and correlations ρ_{-1}, ρ_{+1}
  - calculated for subset of samples of prescribed class
- Denote score threshold as t, where $-1 \le t \le +1$
- false alarm rate ≤ $\rho_{-1}(1-s_{-1}^2)/(\rho_{-1}(1-s_{-1}^2)+(t+s_{-1})^2)$
  - valid when $t > -s_{-1}$
- false alarm rate ≥ $(t+s_{-1})^2/(\rho_{-1}(1-s_{-1}^2)+(t+s_{-1})^2)$
  - valid when $t < -s_{-1}$
- missed det. rate ≥ $(t-s_{+1})^2/(\rho_{+1}(1-s_{+1}^2)+(t-s_{+1})^2)$
  - valid when $t > s_{+1}$
- missed det. rate ≤ $\rho_{+1}(1-s_{+1}^2)/(\rho_{+1}(1-s_{+1}^2)+(t-s_{+1})^2)$
  - valid when $t < s_{+1}$
- Can state lower bound on entire ROC using the parameters E_F and E_M
  - $E_F = \rho_{-1}(1-s_{-1}^2)/(s_{-1}+s_{+1})^2$
  - $E_M = \rho_{+1}(1-s_{+1}^2)/(s_{-1}+s_{+1})^2$

Class-Specific Comparison

- Comparison of ROC and class-specific bounds for SPECTF dataset with Gini Random Forest and 15 different split dimensions (blue = 1, red = 15)



- Comparison of ROC and class-specific bounds for SPECTF dataset with 3 different Random Forest methods and m = 5 (green = Gini, blue = Fisher, red = Anderson-Bahadur)



- Comparisons of ROCs and bounds illustrate that the bounds are predictive of ROC behavior
- Correlation between class-specific areas of ROC (A_F,A_M) and (E_F,E_G,E_M) shows that class-specific bound parameters are predictive of classification performance in low false alarm and low missed detection regimes (correlation over three Random Forest methods):

	E_F	E_G	E_M
A_F	0.6976 largest in the row	0.3878	–0.6401 smallest in the row
A_M	–0.6905 smallest in the row	0.2994	0.8392 largest in the row

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

LLNL-POST-415164