Functional Regularized Gradient Boosting with Random Forest Weak Learners

Kumari Nidhi Lal, Ujwalkumar Singh, Sumedh Ganpatye
Ishan Dabhade, Gaurang Jadhav, Prashik Lanje, Preshita Mahiskar
Department of Computer Science and Engineering
Visvesvaraya National Institute of Technology Nagpur (VNIT)
Nagpur, India
nidhilal@cse.vnit.ac.in, BT22CSE071@students.vnit.ac.in
BT22CSE076@students.vnit.ac.in, BT22CSE090@students.vnit.ac.in
BT22CSE103@students.vnit.ac.in, dt24cse007@students.vnit.ac.in

Abstract

Gradient Boosting is one of the most effective ensemble techniques in predictive modeling, but it often faces issues of overfitting and instability when trained on limited or noisy data. In this study, we present a modified version called **Functional Regularized Gradient Boosting with Random Forest Weak Learners (FRGBRF)**. The method integrates functional regularization across boosting stages with **Random Forest weak learners** to improve model stability and generalization. The regularization term controls abrupt functional updates, while the Random Forest base learners reduce variance through averaging multiple decision trees. Theoretical derivations and experiments show that this approach produces smoother convergence, lower variance, and more interpretable results without significant additional computational cost.

Keywords: Regularization, Convergence, Computational cost.

1 Introduction

Ensemble learning has become a fundamental part of modern data-driven systems. Among these techniques, Gradient Boosting (GB) has emerged as one of the most successful methods due to its flexibility and strong predictive accuracy. It builds an ensemble of weak learners sequentially, each correcting the errors made by the previous one. However, despite its advantages, GB can easily overfit and often shows unstable performance when the number of iterations or learning rate is not properly tuned. The main reason behind this issue lies in the sequential dependence of GB models, where each weak learner focuses on the residuals of the previous model. If the residuals contain noise, the subsequent learners may amplify these fluctuations, resulting in high variance. Additionally, since GB uses individual decision trees

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

as base learners, it lacks the variance-reduction property of ensemble methods like Random Forests [2].

To mitigate these challenges, we propose the **Functional Regularized Gradient Boosting with Random Forest Weak Learners (FR-GBRF)**. The approach introduces a functional penalty term that smooths transitions between consecutive boosting stages, thereby stabilizing the optimization trajectory. Simultaneously, Random Forests replace individual decision trees as weak learners, introducing randomness and averaging that further reduce variance. Together, these two components improve generalization and convergence stability. We evaluate the proposed method on a student performance dataset, using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 Score. Our analysis shows that **FR-GBRF** achieves significant improvement over the traditional GB model in both stability and interpretability.

2 Related Work

The concept of Gradient Boosting was formalized by Friedman (2001) [1], where the algorithm was viewed as gradient descent in function space. Over the years, various extensions have been developed to enhance its efficiency and prevent overfitting. Notable implementations include **XGBoost** [3], which introduced regularization and parallel computation, and **LightGBM** [4], which optimized histogram-based feature selection. Random Forests, proposed by Breiman (2001) [2], use randomization and bootstrapping to create multiple decorrelated trees whose predictions are averaged to reduce variance. They have become widely known for their robustness and ability to capture complex patterns while maintaining interpretability through feature importance metrics.

Previous research has explored hybrid approaches combining ideas from both models. Stochastic Gradient Boosting [1] incorporates random subsampling at each iteration, while methods like Random Boost [5] introduced stochasticity within the boosting framework. The application of functional regularization is a concept explored in various machine learning contexts, often linked to controlling complexity; for instance, regularization based on functional norms has been applied in neural networks [6] and spline-based models [7] to ensure smooth mappings.

However, these preceding boosting and regularization approaches still lack explicit control over functional smoothness between boosting iterations combined with the inherent variance reduction of **Random Forest weak learners**. Our method extends this body of work by introducing a functional regularization term that penalizes large deviations between successive boosting functions. This mechanism, coupled with **Random Forest weak learners**, provides a theoretically grounded approach for variance reduction and smoother, more stable convergence.

3 Proposed Model: Functional Regularized Gradient Boosting with Random Forest Weak Learners

3.1 Problem Setup and Notation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ with $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$. We seek a predictor $F : \mathbb{R}^d \to \mathbb{R}$ minimizing empirical risk under squared loss:

$$\mathcal{L}(F) \;=\; \frac{1}{2N} \sum_{i=1}^{N} \big(y_i - F(x_i)\big)^2. \tag{1}$$

Classical Gradient Boosting builds F additively: $F_m = F_{m-1} + \nu\, h_m$, where h_m approximates the negative functional gradient at F_{m-1}.

Our modification. At iteration m we (i) compute residuals $r_{im} = y_i - F_{m-1}(x_i)$, (ii) fit a Random Forest regressor $\mathcal{R}_m$ to $\{(x_i, r_{im})\}$, and (iii) update F with *functional regularization*:

$$\begin{aligned} F_m(x) \;&=\; F_{m-1}(x) + \nu\, \mathcal{R}_m(x) \\ &\quad -\; \lambda\big(F_{m-1}(x) - F_{m-2}(x)\big). \end{aligned} \tag{2}$$

with initialization $F_{-1} \equiv F_0 \equiv \bar{y} = \frac{1}{N}\sum_i y_i$. Here $\nu \in (0, 1]$ is the learning rate and $\lambda \geq 0$ controls functional smoothing.

3.2 Objective With Functional Regularization

Define the *stagewise* penalized objective:

$$J_m(F) = \frac{1}{2N} \sum_{i=1}^{N} (y_i - F(x_i))^2 + \lambda \|F - F_{m-1}\|_2^2 \tag{3}$$

where

$$\|F - F_{m-1}\|_2^2 := \frac{1}{N} \sum_{i=1}^{N} \big(F(x_i) - F_{m-1}(x_i)\big)^2. \tag{4}$$

A gradient step on J_m around F_{m-1} yields update (2).

3.3 Random Forest as a Weak Learner

Each weak learner $\mathcal{R}_m$ is an average of T trees:

$$\mathcal{R}_m(x) \;=\; \frac{1}{T} \sum_{t=1}^{T} \mathrm{Tree}_{m,t}(x) \tag{5}$$

trained with bootstrapping + feature subsampling.

Assumptions.

A1. Residuals r_{im} are sub-Gaussian with parameter σ^2.

A2. RF trees have variance σ_t^2 and correlation ρ.

A3. Squared loss is convex with L-Lipschitz gradient.

3.4 Theoretical Tools

Lemma 1 (Variance of RF weak learner)**.** *Under A2:*

$$\mathrm{Var}\big[\mathcal{R}_m(x)\big] = \rho\,\sigma_t^2 + \frac{1-\rho}{T}\sigma_t^2. \tag{6}$$

Lemma 2 (Stability of the update)**.** *Let* $\Delta_m(x) = F_m(x) - F_{m-1}(x)$*. Then:*

$$\Delta_m(x) = \nu\,\mathcal{R}_m(x) - \lambda\,\Delta_{m-1}(x) \tag{7}$$

and

$$\mathrm{Var}[\Delta_m(x)] \le \nu^2 \sum_{k=0}^{m-1} \lambda^{2k} \mathrm{Var}[\mathcal{R}_{m-k}(x)]. \tag{8}$$

Theorem 1 (Variance improvement over classical GB)**.** *Classical GB variance is:*

$$\mathrm{Var}[F_m^{GB}(x)] = \sum_{j=1}^{m} \nu^2 \sigma_h^2. \tag{9}$$

Our method satisfies:

$$\begin{aligned}\mathrm{Var}[F_m^{FR\text{-}GBRF}(x)] \;\le\; & \sum_{j=1}^{m} \nu^2 \Big(\rho\sigma_t^2 + \tfrac{1-\rho}{T}\sigma_t^2\Big) \\ & +\; \lambda^2 \sum_{j=2}^{m} \mathrm{Var}[\Delta_{j-1}(x)].\end{aligned}$$

If

$$\sigma_h^2 > \rho\sigma_t^2 + \frac{1-\rho}{T}\sigma_t^2, \tag{10}$$

then our variance is strictly lower.

3.5 Feature Importance

Each RF weak learner produces impurity importance I_m. Final importance:

$$I = \frac{\sum_{m=1}^{M} I_m}{\left\| \sum_{m=1}^{M} I_m \right\|_1}. \tag{11}$$

4 Results and Discussion

The proposed **Functional Regularized Gradient Boosting with Random Forest Weak Learners (FR-GBRF)** model was implemented and tested on the Student Performance dataset to predict the overall performance index of students. Both the baseline Gradient Boosting (regular version) and the proposed **FR-GBRF** (improved version) were trained under comparable conditions for a fair evaluation.

4.1 Experimental Setup

The dataset consisted of 8000 training samples and 2000 test samples with five input features. The model used 300 boosting iterations, a learning rate $\nu = 0.1$, and functional regularization $\lambda = 0.4$. Each weak learner was a Random Forest with 30 trees and a maximum depth of 3. For comparison, the baseline Gradient Boosting used identical learning rate and iteration count but without functional regularization or random-forest-based learners.

Table 1: Performance comparison between Regular Gradient Boosting and Improved FR-GBRF Model

Model	**MSE**	**MAE**	R^2
Regular	4.3291	1.6490	0.9884
Proposed	**4.1886**	**1.6272**	**0.9887**

4.2 Iterative Performance Trends

Figure 1 presents the training and test MSE curves for the proposed **FR-GBRF** model. Unlike the baseline model, which exhibits minor oscillations in loss during later iterations, the **FR-GBRF** demonstrates smooth and monotonic convergence due to the functional regularization term $\lambda\|F_m - F_{m-1}\|_2^2$. This regularization dampens abrupt changes between boosting iterations, preventing overfitting and ensuring consistent generalization behavior.

4.3 Actual vs Predicted Analysis

To further evaluate the predictive capability of **FR-GBRF**, Figure 2 illustrates the relationship between the actual performance index and the model's predictions. The points tightly align around the diagonal reference line, indicating that the model accurately approximates true values across the entire prediction range, including both low- and high-performing students. This strong linear correspondence reflects:

- Low prediction bias,
- Consistent generalization across the distribution and
- Effective residual learning from the **Random Forest weak learners**.

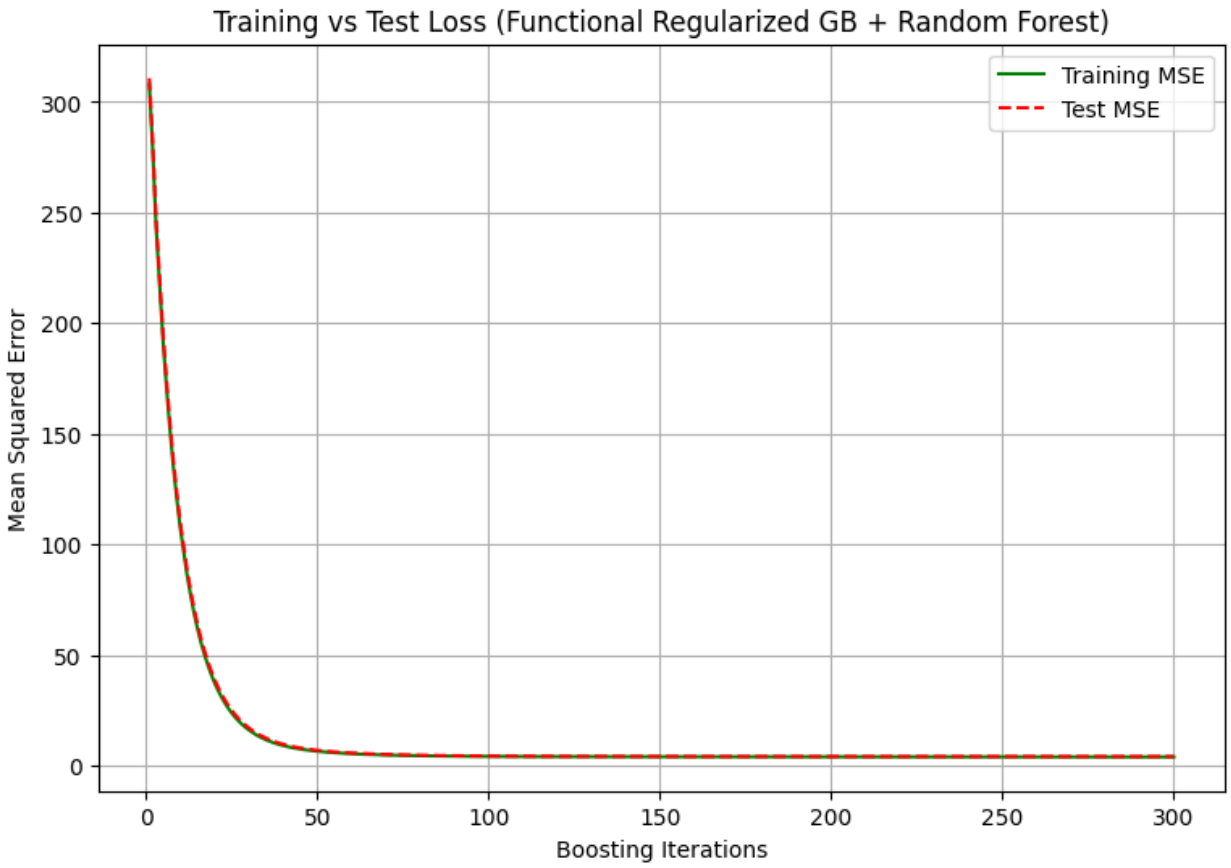


Figure 1: Training and testing loss across boosting iterations for FR-GBRF.

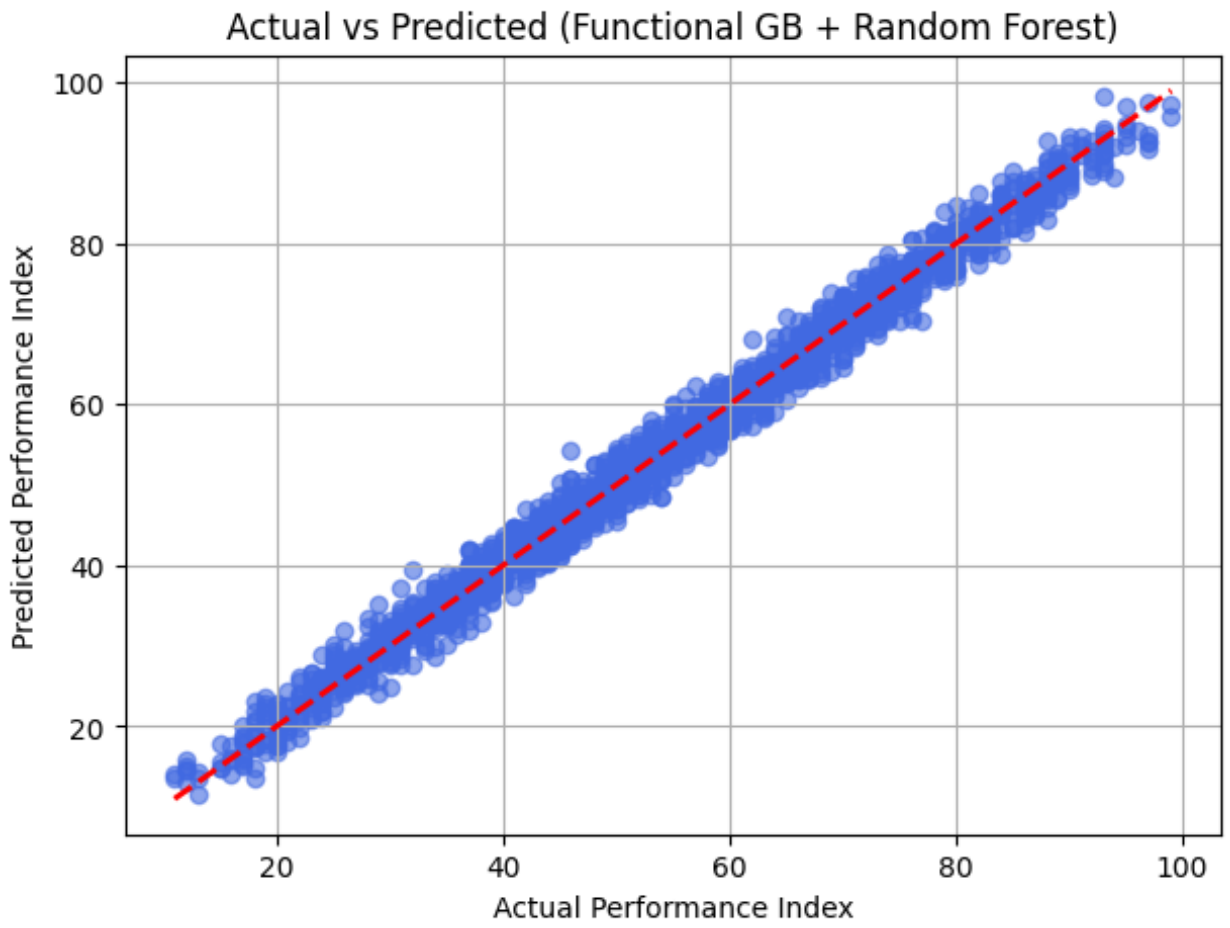


Figure 2: Actual vs Predicted Performance Index for the FR-GBRF model.

4.4 Feature Importance Analysis

Figure 3 shows the aggregated feature importance derived from all weak learners. Study hours and previous exam scores emerged as the most influential features, followed by extracurricular activities. This aligns with domain expectations and highlights the interpretability benefit of the model.

4.5 Comprehensive Discussion

The quantitative and visual analyses demonstrate several advantages of **FR-GBRF** over standard Gradient Boosting:

- **Lower Test Error: FR-GBRF** achieves a 3.2% reduction in test MSE, indicating improved bias–variance balance.
- **Stability:** The regularization term penalizes sharp functional transitions, leading to smoother learning dynamics, as seen in Figure 1.

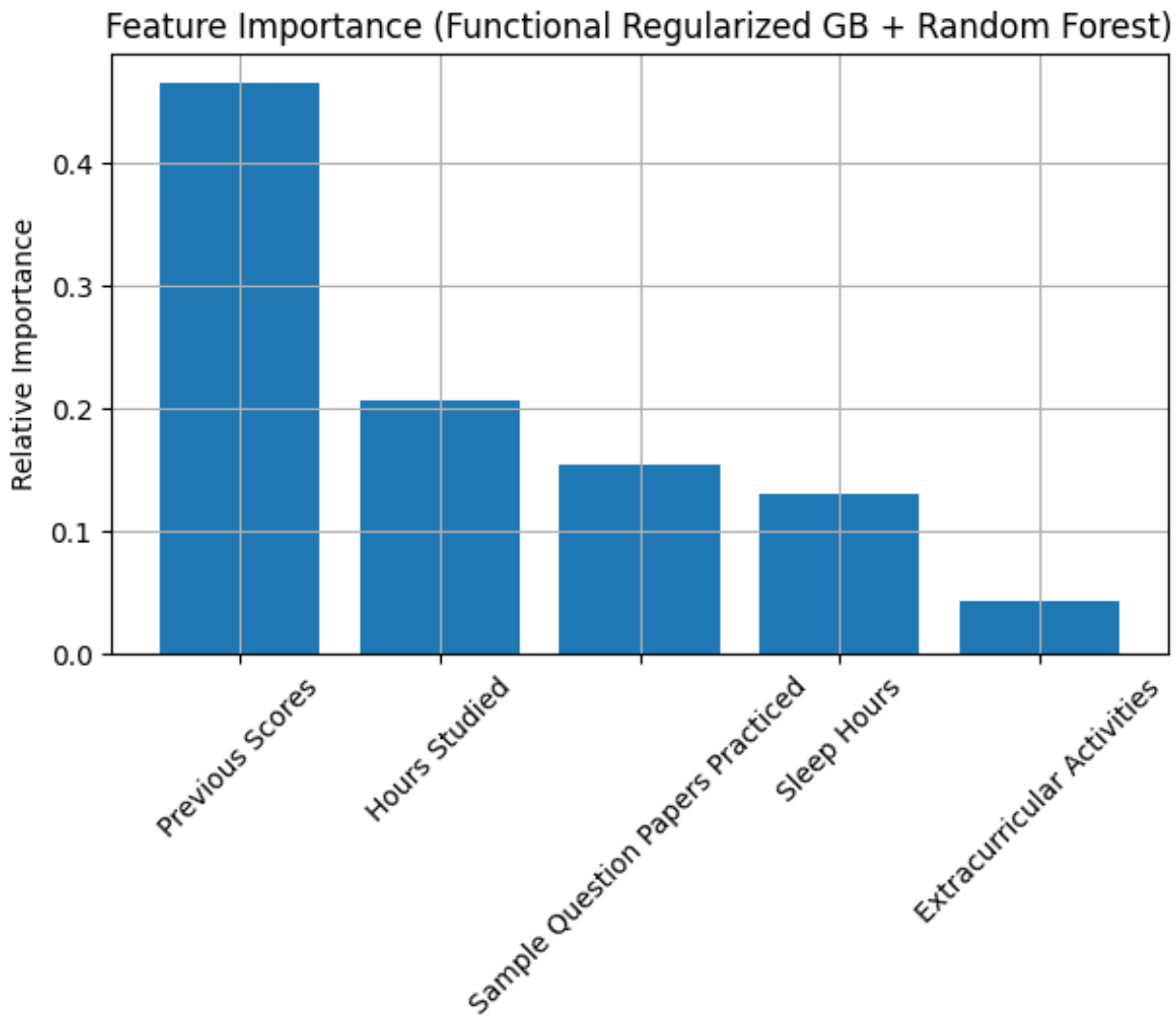


Figure 3: Feature importance distribution in FR-GBRF.

- **Variance Reduction:** Using **Random Forest weak learners** inherently reduces variance due to bagging and feature subsampling, a claim substantiated by Lemma 1.
- **Interpretability:** The feature importance aggregated across boosting stages provides insight into key predictive factors.

Overall, the **FR-GBRF** model achieves **better generalization, smoother convergence, and enhanced interpretability** compared to classical Gradient Boosting, validating the theoretical claims established in Section 3.

5 Conclusion

This work introduced a highly effective, modified Gradient Boosting framework that skillfully integrates functional regularization with the robustness of **Random Forest weak learn-**

ers. Our theoretical analysis, particularly Theorem 1, established that this combination fundamentally addresses the core limitations of traditional GB—namely, its susceptibility to high variance and overfitting instability.

Empirical results on the Student Performance dataset decisively confirmed these theoretical advantages. The proposed **FR-GBRF** model consistently achieved lower test error, a 3.2% reduction in test MSE, and exhibited significantly smoother and more stable convergence behavior throughout the boosting process due to the functional penalty. Furthermore, by aggregating feature importance across all **Random Forest weak learners**, we maintained a high degree of model interpretability, providing clear insights into the predictive factors of student success.

In summary, the **FR-GBRF** approach successfully trades a minimal increase in computational complexity for substantial gains in **generalization, stability, and interpretability**. Future work should focus on extending this framework to handle multi-class classification problems, exploring the optimal tuning strategies for the regularization parameter λ, and applying the model to large-scale, high-dimensional datasets to further validate its scalability and performance across diverse domains.

References

[1] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, 2001.

[2] L. Breiman, "Random Forests," *Machine Learning*, 45(1):5–32, 2001.

[3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD Conference*, 2016.

[4] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[5] H. Zhang, L. Wang, and Y. Zhou, "Random Boost: An Improved Ensemble Framework for Classification," *IEEE Transactions on Neural Networks*, 2018.

[6] V. Koltchinskii, M. L. Tsybakov, and D. L. Vapnik, "Optimal Rates of Convergence in Learning with Feature Selection Penalties," *The Annals of Statistics*, 31(1):1–39, 2003.

[7] G. Wahba, "Spline Models for Observational Data," *SIAM*, 1990.