

Signal vs Background Events in High-Energy Physics

Tolga Kuntman and Michel Moussally

Introduction

- Modern scientific experiments generate large, high-dimensional datasets
- In high-energy physics, rare signal events must be identified among overwhelming background noise
- The HIGGS dataset captures this challenge using simulated particle collision events
- Complex, nonlinear feature interactions limit classical analysis methods
- Machine learning models are therefore widely used in this domain
- In this project, we study a binary classification task to distinguish signal from background events in the HIGGS dataset

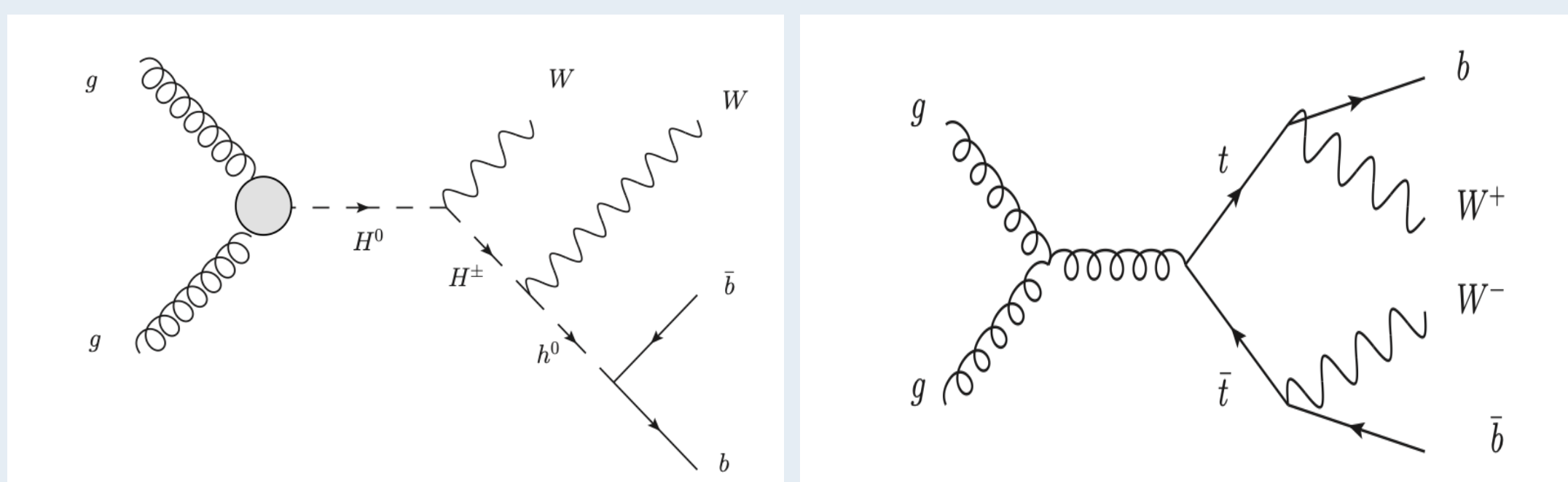


Figure 1: Diagrams for Higgs Benchmark that show the Higgs Particle (left figure) and Background Noise (right figure)

Dataset and Features

- Dataset: HIGGS dataset (UCI Machine Learning Repository)
- Origin: Simulated high-energy physics experiments for Higgs boson detection
- Task: Binary classification (signal vs. background)
- Sample size: 500,000 randomly sampled instances
- Data split: 80% training, 10% validation, 10% test
- Features: 28 continuous variables, including **raw kinematic measurements** (e.g. particle momentum and angular variables) and **engineered physics features** (e.g. invariant masses, angular separations, and event-level quantities)
- Labels: Binary indicator of whether an event corresponds to a Higgs signal

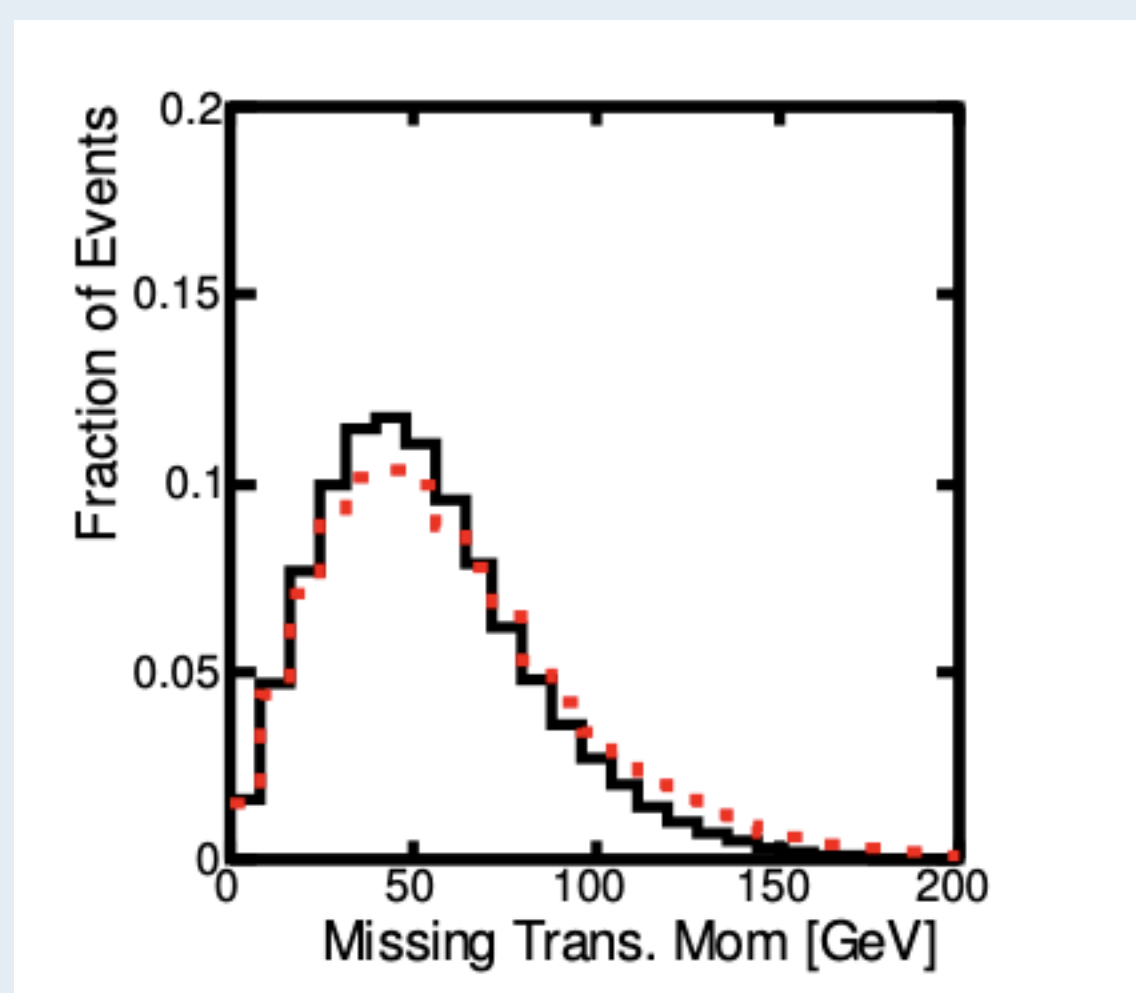


Figure 2: Distribution of Higgs Particle (black) and Background Noise (red) for one of the High-Level Features

Methods

Models:

- Logistic Regression:** linear baseline with L2 regularization, trained using gradient-based optimization.
- Gradient Boosted Trees:** ensemble model capturing nonlinear feature interactions in tabular data
- Feed-Forward Neural Network (MLP):** fully connected neural network with 5 hidden layers and 300 units per layer.
- All models follow the same training, validation, and test workflow, illustrated on the right.

Evaluation metrics:

- ROC-AUC

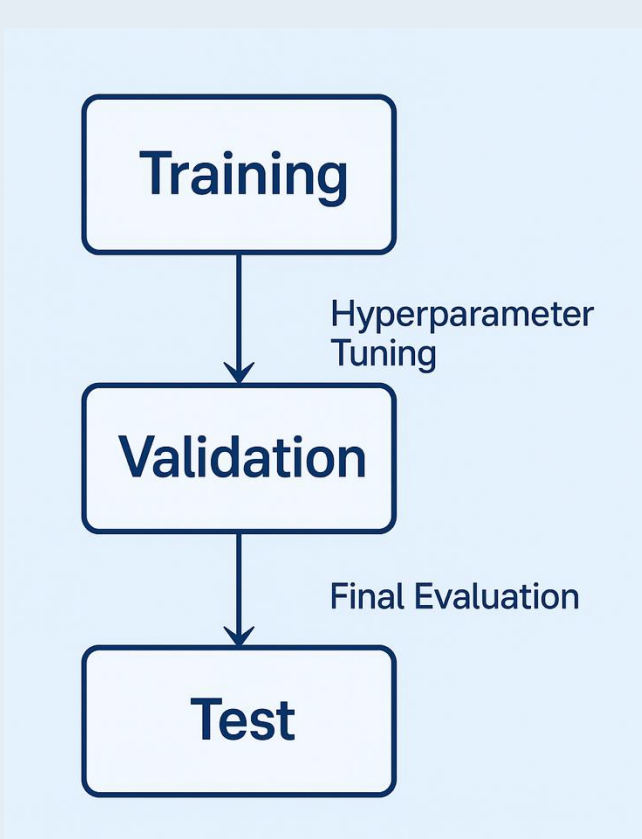


Figure 3: Experimental Workflow

Results

We evaluate model performance using the Area Under the ROC Curve (AUC), which measures a classifier's ability to distinguish signal from background across all decision thresholds and is well suited for rare-event classification problems.

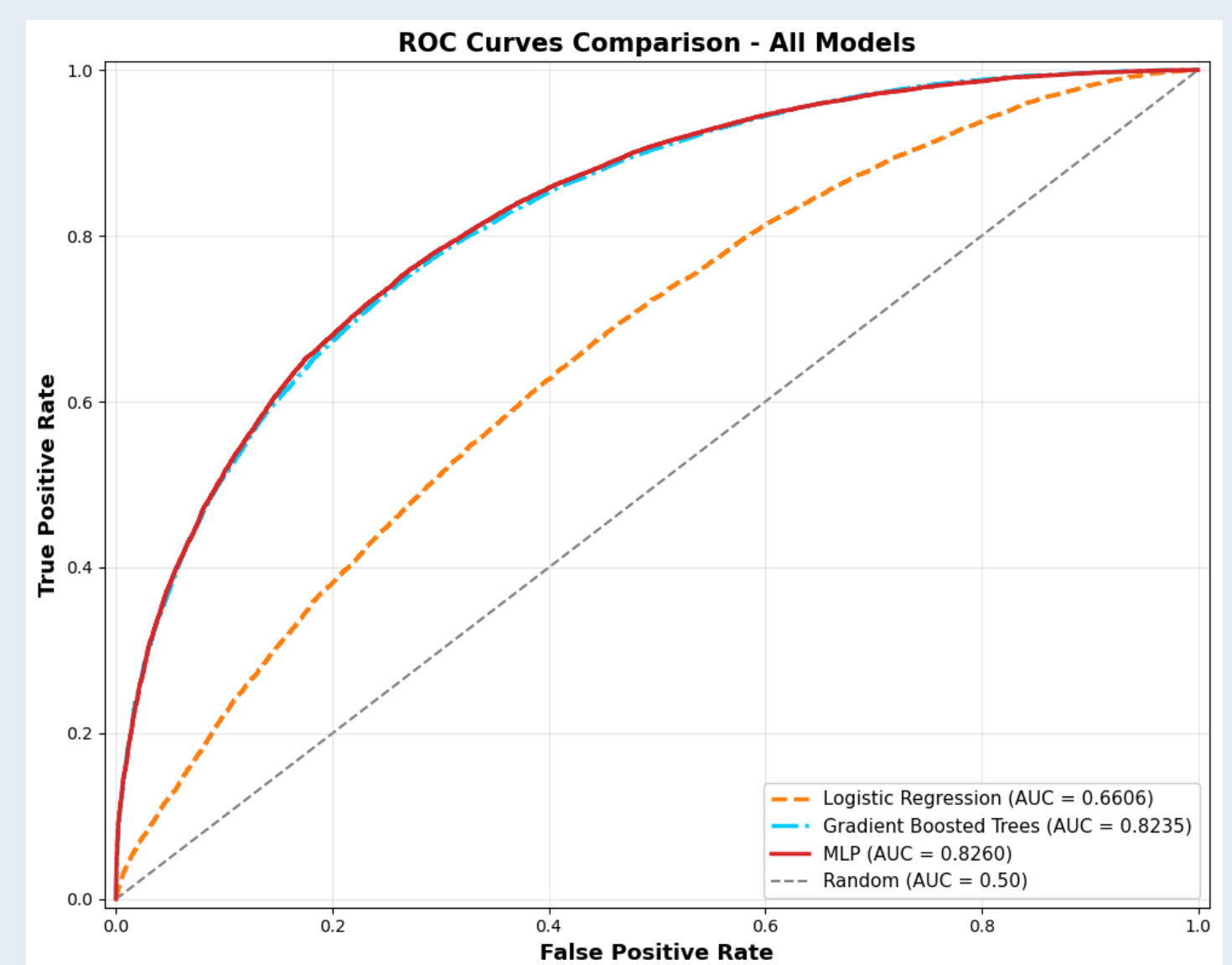


Figure 4: ROC Curves Comparison for All the Trained Models

Model	AUC
Logistic Regression	0.661
Gradient Boosted Trees	0.824
Feed-Forward MLP	0.828

Figure 5: AUC Comparison for All the Trained Models

- Gradient Boosted Trees and MLPs significantly outperform the linear baseline
- The MLP achieves the highest AUC, with a marginal improvement over trees

Discussion

- Logistic Regression is limited by its linear decision boundary and cannot fully exploit nonlinear feature interactions
- Gradient Boosted Trees perform strongly on tabular data by modeling complex and localized patterns
- The small performance gap between MLP and Gradient Boosted Trees suggests diminishing returns from increased model capacity under limited training data
- Lower absolute AUC values compared to the reference study are primarily due to reduced dataset size and computational constraints

Conclusion

- Nonlinear models significantly outperform linear baselines on the HIGGS classification task
- Gradient Boosted Trees and Feed-Forward Neural Networks achieve comparable performance on the reduced dataset
- The small performance gap highlights the impact of data scale and model capacity in deep learning
- ROC-AUC proves to be an effective metric for comparing models in rare-event classification problems

Future Work

- Threshold optimization based on physics-driven cost functions
- Deeper neural architectures or alternative activations
- Feature importance and interpretability analysis