Predicting Physics Observables by Analysing LHC Data

Nissim Sahoo
Roll No.: 24B1818

Mentor: Deependra Sharma

WiDS, Analytics Club, IIT Bombay

February 1, 2026

Abstract

This report presents a physics-informed machine learning workflow for analysing LHC data and separating prompt, non-prompt, and background D^* meson candidates. A Boosted Decision Tree is used as a baseline classifier, followed by a neural network incorporating particle identification variables and a physics-aware loss to enforce physical consistency.

1 Introduction

High-energy physics experiments at the Large Hadron Collider (LHC) produce large datasets containing reconstructed particle candidates. These data are stored in the ROOT framework and analysed to extract physics observables such as heavy- flavor production mechanisms.

The objective of this work is to distinguish prompt D^* mesons, non-prompt D^* mesons originating from B-hadron decays, and combinatorial background using machine learning techniques guided by physical principles.

2 ROOT Data Structure

Candidate-level information is stored in ROOT files containing a single `TTree` named `treeMLDstar`. Each entry corresponds to one reconstructed D^* candidate and includes kinematic, topological, and particle identification variables suitable for supervised learning.

3 Physics Background

3.1 Heavy-Flavor Production

- Prompt D^*: directly produced in the primary pp collision.
- Non-prompt D^*: produced via $B \to D^* + X$ decays.
- Background: random track combinations without physical decay origin.

Prompt and non-prompt candidates differ mainly in decay topology and displacement from the primary vertex.

4 Feature Construction

The feature vector excludes invariant mass and transverse momentum to avoid sculpting physics distributions. It consists of

$$x = (\text{topology},\ \text{geometry},\ \text{PID}),$$

where topology includes decay length and impact parameters, geometry includes pointing angles, and PID variables are derived from TPC and TOF detector responses.

5 Machine Learning Strategy

Boosted Decision Tree Baseline

A Boosted Decision Tree (BDT) classifier is trained first using only topological and geometric variables. Gradient boosting is chosen due to its robustness and strong performance in heavy-flavor analyses. The BDT provides a transparent baseline and is evaluated using class-wise efficiencies obtained by applying a threshold on the prompt-class score.

Neural Network Extension

To incorporate particle identification information and capture higher-order correlations between variables, a neural network classifier is trained using the full feature set. The network outputs class probabilities via a softmax layer,

$$p_i(x) = \frac{e^{z_i}}{\sum_j e^{z_j}},$$

and is trained using cross-entropy loss.

6 Physics-Aware Loss

Standard cross-entropy loss treats all misclassifications equally. However, combinatorial background candidates cannot originate from the primary vertex and therefore should not be classified as prompt.

To encode this physical constraint, a physics-aware penalty term is introduced:

$$\mathcal{L}_{\text{phys}} = \frac{1}{N_{\text{bkg}}} \sum_{n:\, y_n = 2} p_0(x_n),$$

where $p_0(x_n)$ is the predicted prompt probability for background samples. This term penalizes prompt-like behavior in background candidates.

The total loss is defined as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{phys}} \mathcal{L}_{\text{phys}},$$

where λ_{phys} controls the strength of the physics constraint. This provides a soft, data-driven enforcement of known physical behavior without hard selection cuts.

7 Model Evaluation

Classifier performance is evaluated using ROC curves and the area under the curve (AUC). The true positive rate (TPR) and false positive rate (FPR) are defined as

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}.$$

Efficiencies for prompt, non-prompt, and background classes are reported as a function of classifier threshold.

7.1 Comparison of Boosted Decision Tree and Neural Network Performance

A Boosted Decision Tree (BDT) was trained as a baseline classifier using topological and geometric variables. A working point corresponding to a prompt efficiency of 75% was selected, giving a score threshold of 0.2466 and a background efficiency of 4.1%. When applied to the full test sample, the BDT achieved a prompt efficiency of 83.4%, non-prompt efficiency of 21.4%, and background efficiency of 6.9%.

A neural network trained with an extended feature set including particle identification (PID) variables achieved a test accuracy of 85.1%. The network shows strong background rejection and improved separation between prompt and non-prompt candidates.

ROC-based comparisons show similar performance for background rejection, with AUC values of 0.957 (prompt vs background) and 0.975 (non-prompt vs background). For prompt versus non-prompt separation, the neural network achieves an AUC of 0.864, outperforming the BDT. This improvement is attributed to the inclusion of PID information and non-linear feature correlations.

The BDT provides a robust baseline, while the neural network offers superior performance for prompt versus non-prompt discrimination.

7.2 Model Performance Comparison

The classification performance of the Boosted Decision Tree (BDT) and the neural network (NN) is evaluated using Receiver Operating Characteristic (ROC) curves. Figure 1 shows the ROC curve obtained with the BDT classifier, demonstrating strong separation power between signal and background.

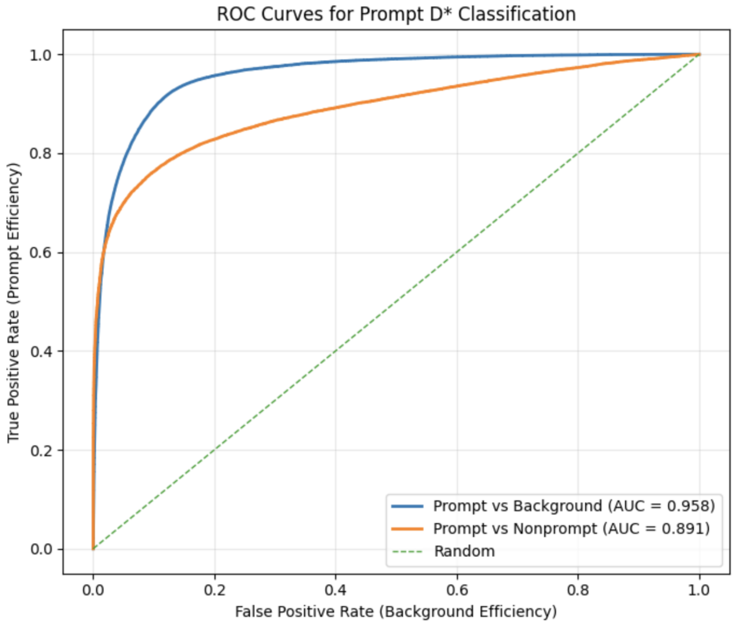


Figure 1: ROC curve for the Boosted Decision Tree classifier.

Figure 2 presents the ROC curve for the neural network. The NN achieves a higher area under the curve (AUC) compared to the BDT, indicating an overall improvement in classification performance, driven by its ability to model non-linear correlations and incorporate particle identification information.

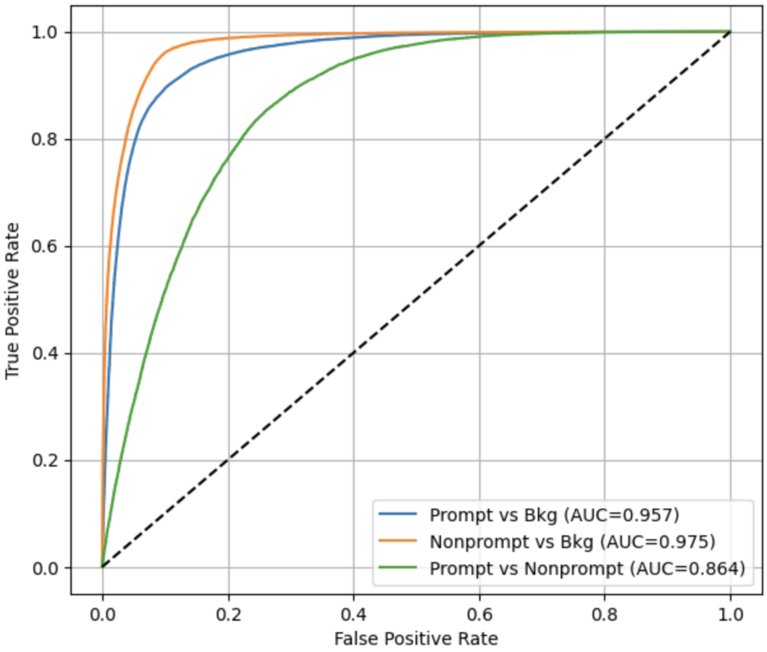


Figure 2: ROC curve for the neural network classifier.

The training behaviour of the neural network is illustrated in Figure 3. The loss decreases smoothly for both training and validation samples, indicating stable convergence. The total loss includes the standard cross-entropy term as well as the physics-aware penalty, ensuring that background candidates are discouraged from being classified as prompt.

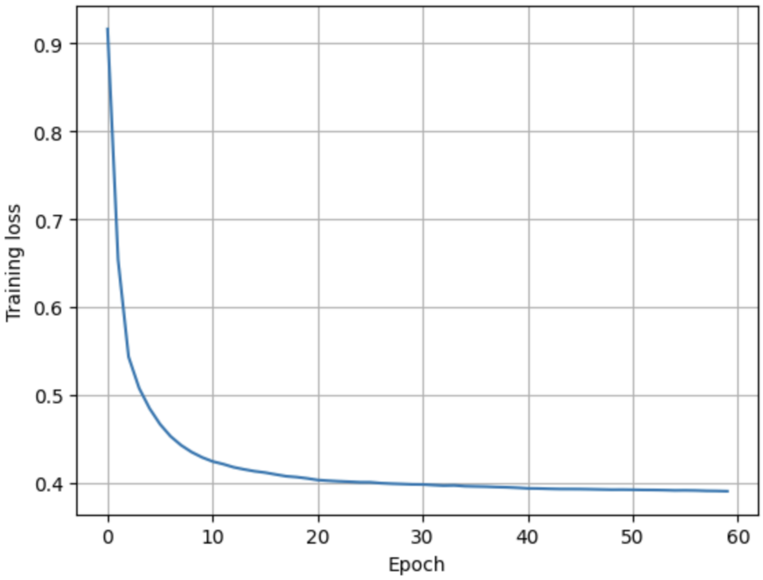
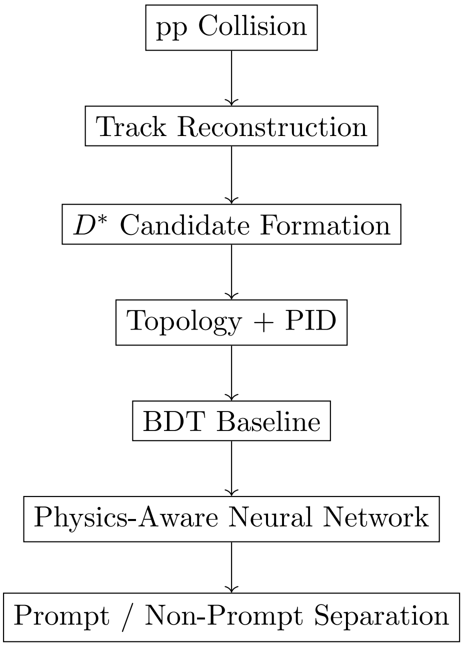


Figure 3: Neural network training and validation loss as a function of epoch.

8 Analysis Workflow


pp Collision
Track Reconstruction
D^* Candidate Formation
Topology + PID
BDT Baseline
Physics-Aware Neural Network
Prompt / Non-Prompt Separation

9 Conclusion

A two-stage machine learning strategy was implemented for D^* meson analysis. A BDT provides a strong and interpretable baseline, while a physics-aware neural network improves discrimination by incorporating PID information and enforcing physical constraints. This approach achieves improved performance while preserving the integrity of physics observables.