A novel machine learning method to identify heavy-flavor decay leptons

Raghunath Sahoo,* Kangkan Goswami, and Suraj Prasad
Department of Physics, Indian Institute of Technology Indore, Simrol, Indore 453552, India

Introduction

Heavy-flavor hadrons, containing charm and beauty quarks, are produced in the early stages of proton-proton (pp) and heavy-ion collisions via hard partonic scatterings. They traverse the medium before hadronizing, making them valuable probes of the Quark-Gluon Plasma (QGP). Experimentally, heavy-flavor production is studied via hadronic decays, di-leptons, or semi-leptonic decays ($B, D \rightarrow l^{\pm} + \nu_l + X$). However, separating heavy-flavor decay leptons (HFLs) from inclusive samples is challenging and typically requires computationally costly cocktail fits.

We present a novel machine learning (ML) approach for direct, track-level identification of heavy-flavor decay electrons and muons. Using the eXtreme Gradient Boosting (XGBoost) algorithm, trained on PYTHIA8-generated pp collisions at $\sqrt{s} = 13.6$ TeV, we show that topological features such as the distance of closest approach (DCA), transverse momentum (p_T), and pseudorapidity (η) are sufficient to distinguish leptons from heavy- and light-flavor decays with high accuracy.

Methodology

We simulate approximately 10^9 minimum bias pp events with PYTHIA8, including hard-QCD processes and vertex smearing. Electrons are selected in the midrapidity region ($|\eta| < 0.8$), while muons are considered at forward rapidity ($2.5 < \eta < 4.0$). Heavy-flavor leptons are tagged by their parent hadrons, while the background contains contributions from photon conversions, Dalitz decays, light-flavor hadron decays, and quarkonia. To address strong class imbalance between heavy and light-flavor leptons, we use SMOTE oversampling, Bayesian hyperparameter optimization, and train models separately in p_T bins. Model performance is evaluated using precision, recall, F1 score, and confusion matrices.

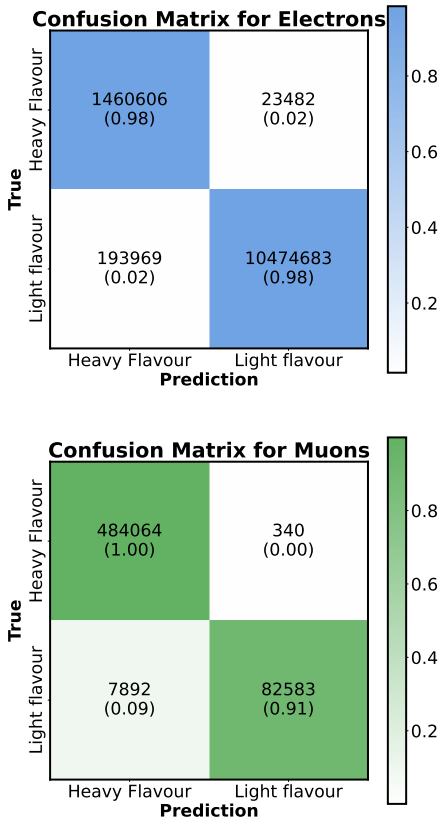

FIG. 1: Confusion matrix for classification of electrons (top) and muons (bottom) into heavy and light flavor decay categories [1].

*Electronic address: raghunath.sahoo@cern.ch

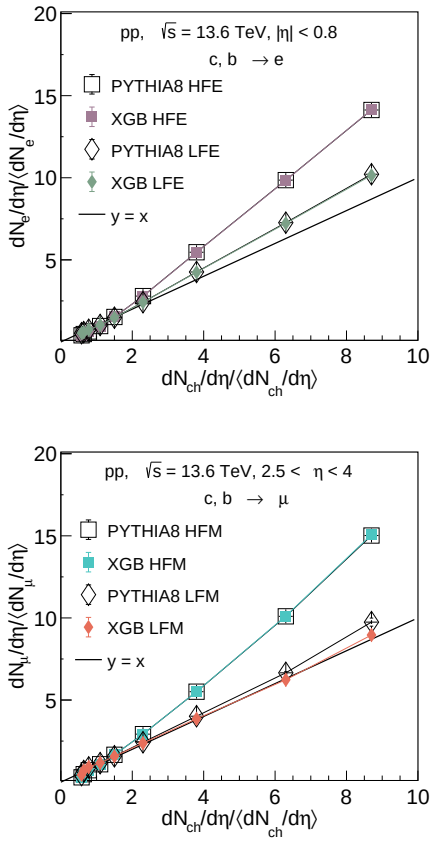

FIG. 2: Self-normalized yield of electrons as a function of normalized charged-particle multiplicity in *pp* collision at $\sqrt{s} = 13.6$ TeV [1].

Results

The XGBoost classifier achieves $\sim$ 98% accuracy for heavy-flavor decay electrons and nearly 100% for heavy-flavor decay muons. Figure 1 shows the confusion matrices for electrons and muons, where the vertical axis corresponds to the true labels from PYTHIA8 and the horizontal axis represents the predictions from the XGBoost classifier. The matrices demonstrate that the majority of leptons are correctly classified, with only a very small fraction of heavy-flavor decay electrons misidentified as light-flavor.

Figure 2 shows the self-normalized yields of heavy and light flavor decay electrons and muons as a function of charged-particle multiplicity. Light-flavor leptons rise linearly, while heavy-flavor leptons display a stronger-than-linear increase, especially in high-multiplicity events, indicating their origin in hard partonic scatterings. The XGBoost model preserves these trends and matches PYTHIA8 predictions closely, validating its robustness for electrons and muons, as well as for different multiplicity classes. This strong non-linear rise of the heavy flavor decay leptons is consistent with the non-linear behavior of prompt and non-prompt charmonium and open charm states [2, 3].

Summary

We demonstrate, for the first time, a machine learning approach for track-level identification of heavy-flavor decay leptons in *pp* collisions. Using PYTHIA8 simulations for training, XGBoost achieves high accuracy for the classification of heavy-flavor electrons and muons. The model performance is robust across transverse momentum intervals and rapidity ranges, yielding $\sim$ 98% accuracy for electrons and nearly 100% for muons. Importantly, the model is able to reproduce transverse momentum spectra, multiplicity-dependent self-normalized yields, and azimuthal correlations of heavy-flavor decay leptons, in close agreement with PYTHIA8 predictions and consistent with existing ALICE measurements [1]. This ML-based method provides an independent alternative to cocktail subtraction and offers a powerful tool for future experimental studies at the LHC and upcoming facilities.

References

[1] R. Sahoo, K. Goswami and S. Prasad, arXiv:2509.00712.
[2] S. Prasad, N. Mallick and R. Sahoo, Phys. Rev. D **109**, 014005 (2024).
[3] K. Goswami, S. Prasad, N. Mallick, R. Sahoo and G. B. Mohanty, Phys. Rev. D **110**, 034017 (2024).