Boosted Decision Tree Analysis for WW Signal vs Top Background

Manan Makhija

January 31, 2026

Abstract

A multivariate analysis using Boosted Decision Trees (BDTs) is performed to separate the WW signal from dominant top-quark backgrounds using the TMVA framework. Multiple kinematic and event-level observables are combined into a single classifier. The performance of the BDT is evaluated using input variable distributions, correlation matrices, ROC curves, and overtraining checks. Signal and background efficiencies are reported for an optimized BDT selection.

1 Introduction

The production of WW events at hadron colliders is an important electroweak process and constitutes a background to many new physics searches. A major challenge in isolating the WW signal arises from large backgrounds originating from top-quark processes, particularly $t\bar{t}$, tW, and $\bar{t}W$ production.

Traditional cut-based analyses often fail to exploit correlations among kinematic variables. Multivariate techniques such as Boosted Decision Trees (BDTs) combine multiple observables into a single discriminant and offer superior separation power. In this work, a BDT-based analysis using the ROOT TMVA framework is presented.

2 Datasets and Input Variables

The analysis uses the following Monte Carlo datasets:

- WW signal
- $t\bar{t}$ background
- tW background
- $\bar{t}W$ background

The BDT is trained using kinematic and event-level variables that are well motivated by the physics of WW and top-quark production:

- Leading and subleading lepton transverse momenta (p_{T1}, p_{T2})
- Lepton pseudorapidities (η_1, η_2)
- Dilepton invariant mass ($m_{\ell\ell}$)
- Dilepton transverse momentum ($p_T^{\ell\ell}$)
- Azimuthal separation between leptons ($\Delta\phi_{\ell\ell}$)

- Missing transverse momentum (p_T^{miss})
- Transverse masses (m_{T1}, m_{T2})
- Jet multiplicity (n_{Jet})
- b-jet multiplicity (n_{BJet})

3 Input Variable Distributions

The distributions of the input variables for the training samples are shown in Figures 1 and 2. Due to the large number of variables, the plots are split into two parts.

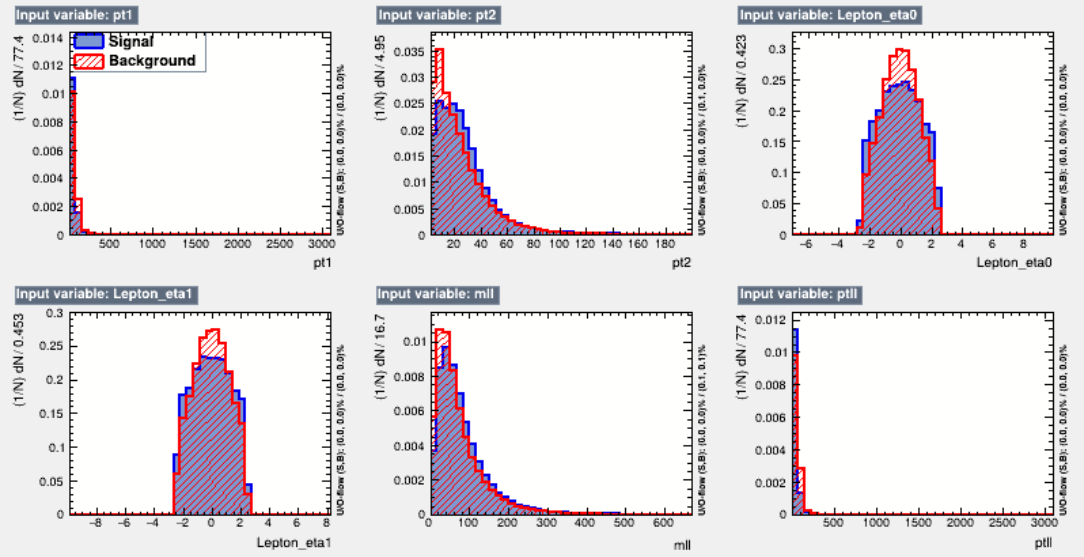

Figure 1: Input variable distributions used for BDT training (Part 1).

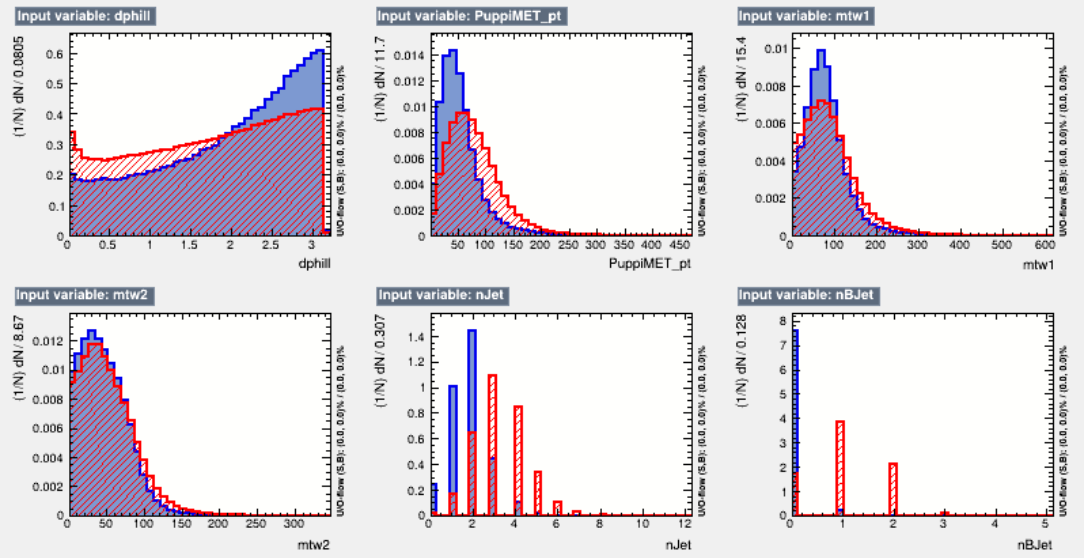

Figure 2: Input variable distributions used for BDT training (Part 2).

Variables related to jet and b-jet multiplicities show strong discrimination between signal and background, while kinematic variables provide additional separation through correlations

exploited by the BDT.

4 Correlation Matrix

Correlations among the input variables are studied using the linear correlation matrix provided by TMVA. This allows identification of redundant variables and ensures that the BDT can effectively exploit non-linear correlations.

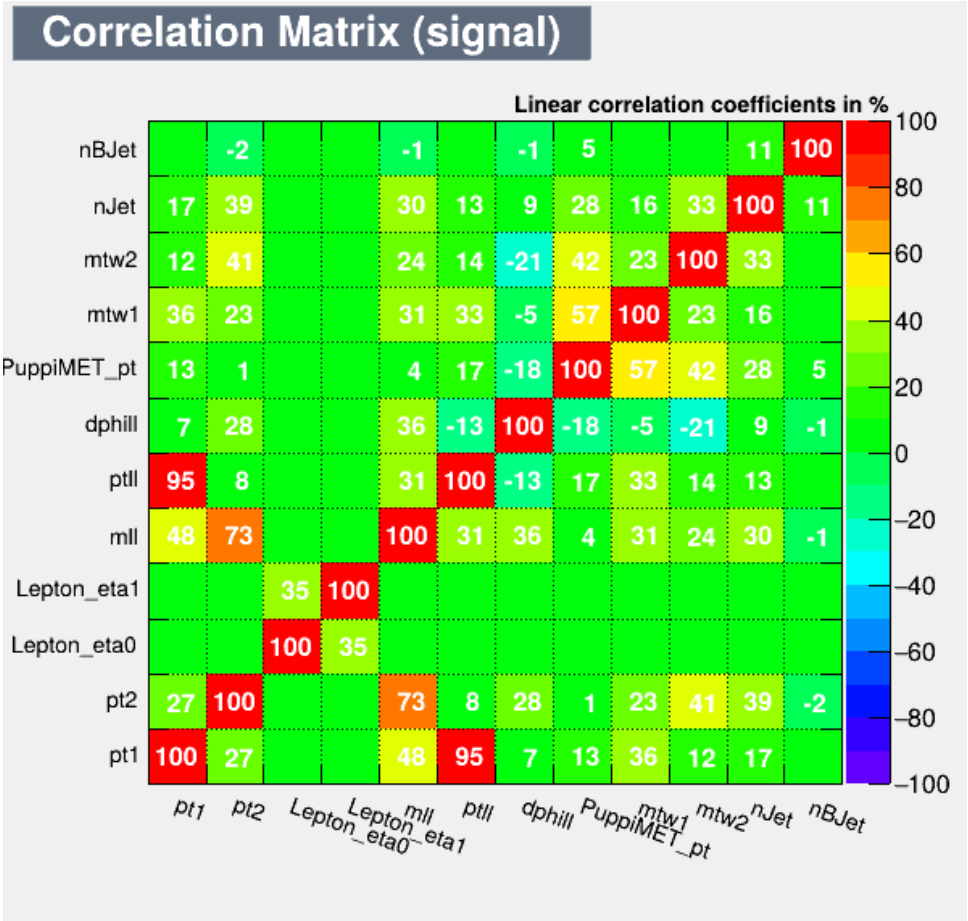

Figure 3: Linear correlation matrix of input variables for the training sample (signal).

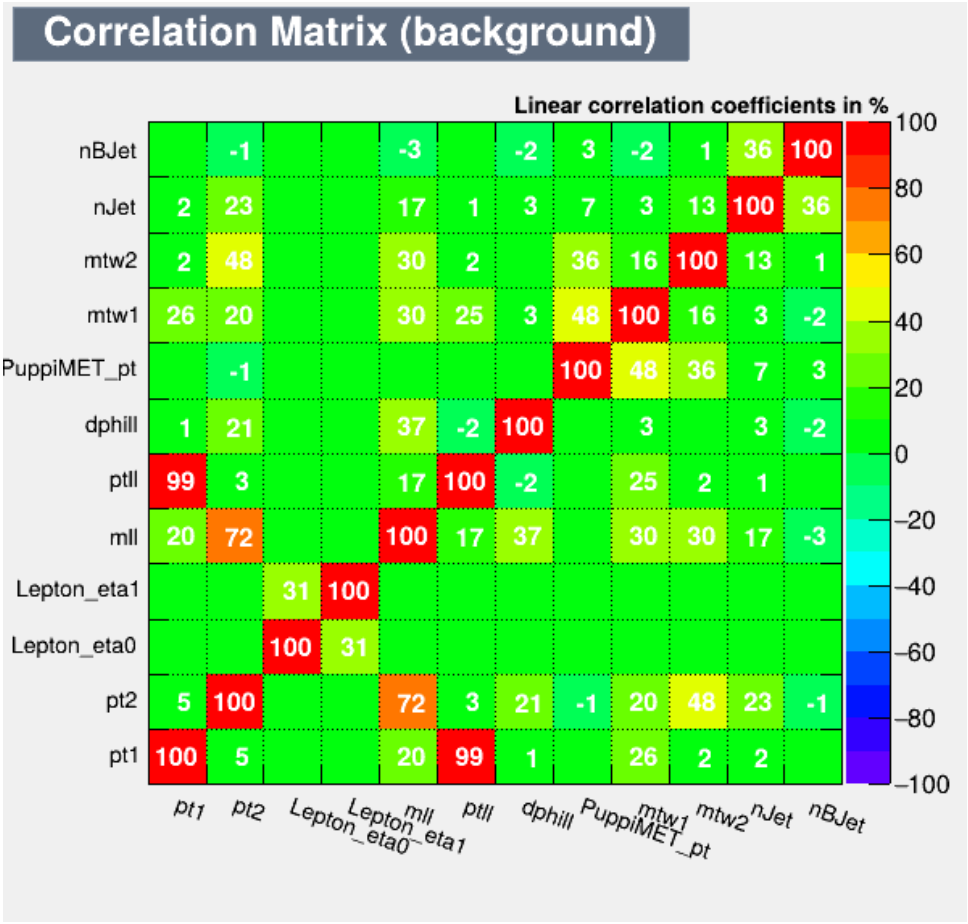

Figure 4: Linear correlation matrix of input variables for the training sample (background).

Moderate correlations are observed among kinematic observables, while jet and b-jet multiplicities remain largely uncorrelated with most leptonic variables.

5 BDT Output Distributions

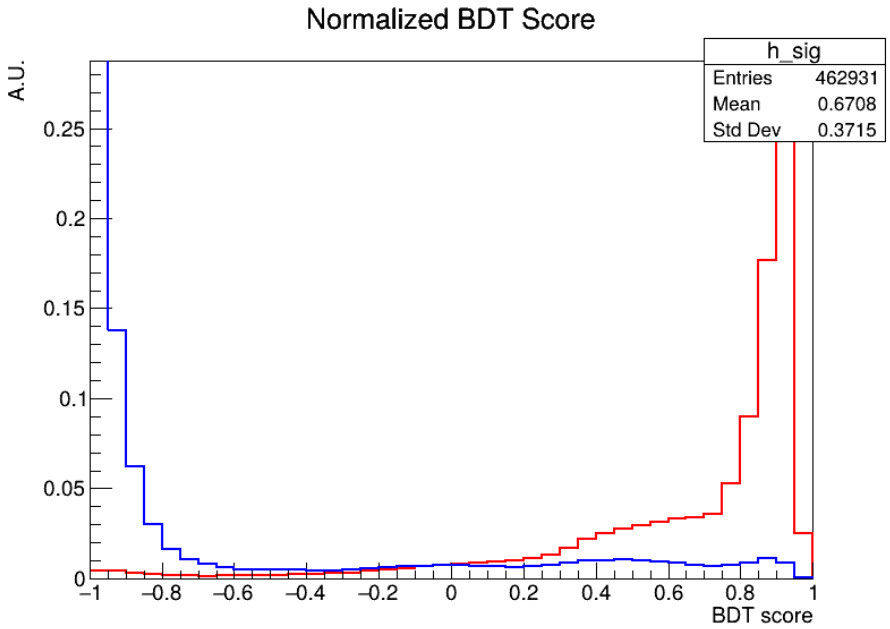

Figure 5: Normalized BDT output distribution for WW signal and combined top backgrounds.

Figure 5 shows the normalized BDT score distribution. The signal peaks at higher BDT values, while the background is concentrated at lower values, indicating good separation.

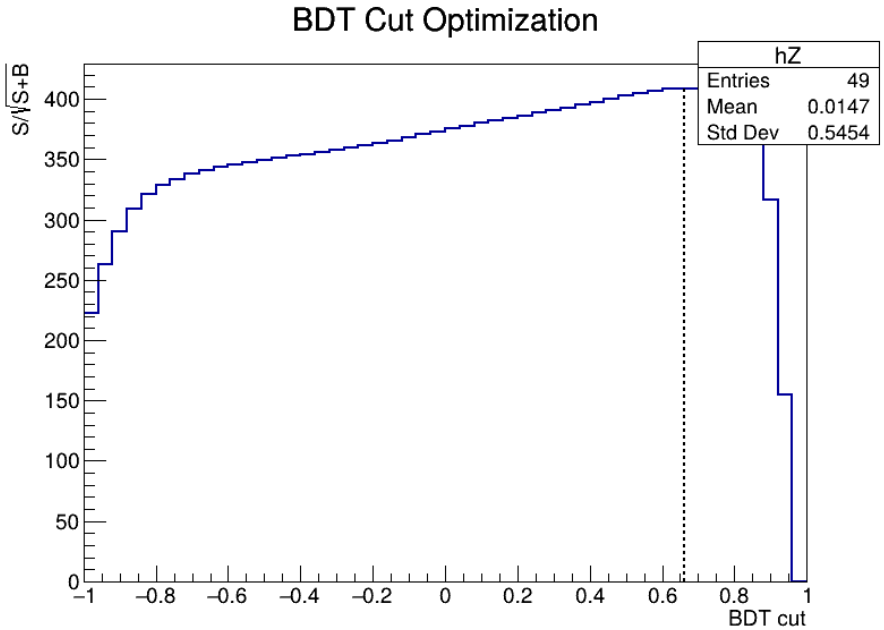

Figure 6: Stacked BDT output distribution for individual background processes compared with the WW signal.

The $t\bar{t}$ process is the dominant background, with smaller contributions from tW and $\bar{t}W$.

6 BDT Performance

6.1 ROC Curve

The performance of the classifier is quantified using the Receiver Operating Characteristic (ROC) curve, shown in Figure 7. The ROC curve illustrates the trade-off between signal efficiency and background rejection.

The area under the ROC curve (AUC) indicates strong overall classifier performance.

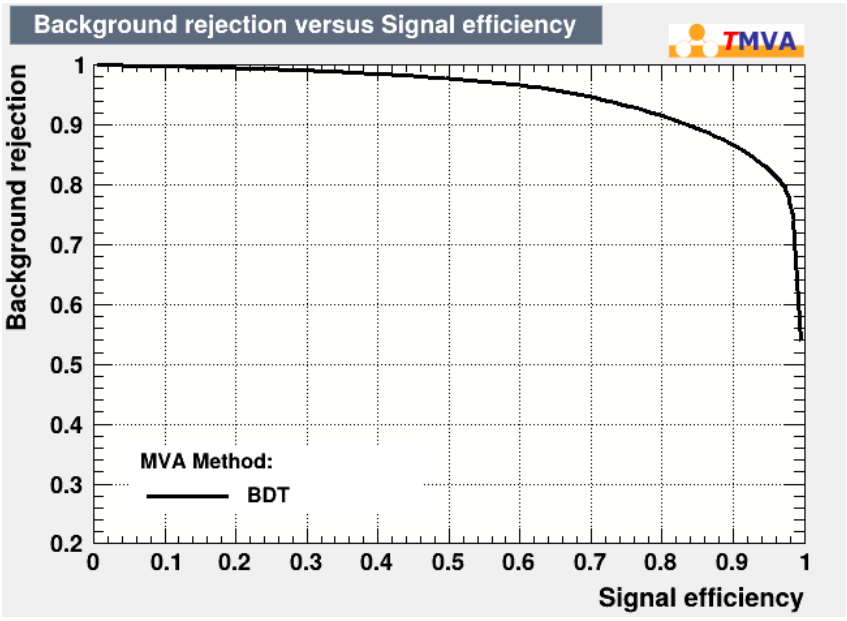

Figure 7: ROC curve for the BDT classifier.

6.2 Overtraining Check

An overtraining check is performed by comparing the BDT response for training and testing samples. Good agreement between the two distributions indicates that the classifier generalizes well.

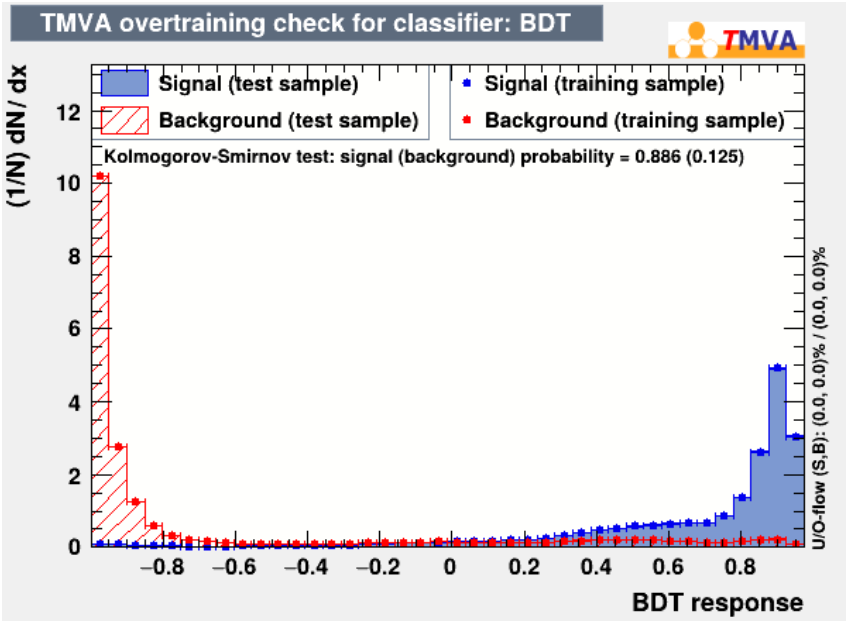

Figure 8: Overtraining check for the BDT classifier.

No significant overtraining is observed.

7 Variable Importance

TMVA provides a ranking of input variables based on their separation power. The most important variables are found to be the b-jet multiplicity and jet multiplicity, followed by missing transverse momentum and dilepton kinematics. This ranking is consistent with the physical expectation that top-quark backgrounds are enriched in b-jets relative to the WW signal.

8 BDT Cut Optimization

The optimal BDT cut is determined by studying the signal and background efficiencies as a function of the BDT score. The selected working point is:

$$\text{BDT cut} = 0.66 \tag{1}$$

$$\epsilon_S = 0.68 \tag{2}$$

$$\epsilon_B = 0.049 \tag{3}$$

Since the samples are not normalized to physical cross sections, the analysis focuses on efficiencies rather than absolute signal significance.

9 Conclusion

A Boosted Decision Tree analysis using TMVA has been successfully applied to separate WW signal events from dominant top-quark backgrounds. The BDT exploits both kinematic and event-level variables and demonstrates strong discrimination power. The classifier performance has been validated using input variable studies, correlation matrices, ROC curves, and overtraining checks. This work highlights the effectiveness of multivariate techniques in modern high-energy physics analyses.