Boosted Decision Tree Analysis for WW Signal vs Top Background

Manan Makhija

January 20, 2026

Abstract

A multivariate analysis based on Boosted Decision Trees (BDT) is performed to separate the WW signal from dominant top-quark backgrounds using the TMVA framework. Signal and background discrimination is achieved using kinematic and event-level observables. The optimal BDT cut is determined by maximizing the statistical significance, and the final signal efficiency and background rejection are reported.

1 Introduction

The production of WW events at hadron colliders constitutes an important electroweak process and serves as both a signal and a background in many physics analyses. 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), provide superior discrimination power by combining multiple variables into a single classifier. In this work, a BDT-based analysis using the TMVA framework is presented.

2 Datasets and Variables

The analysis is performed using the following datasets:

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

Each dataset is processed to include the BDT output score. The BDT is trained using the following input variables:

- 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 BDT Output Distributions

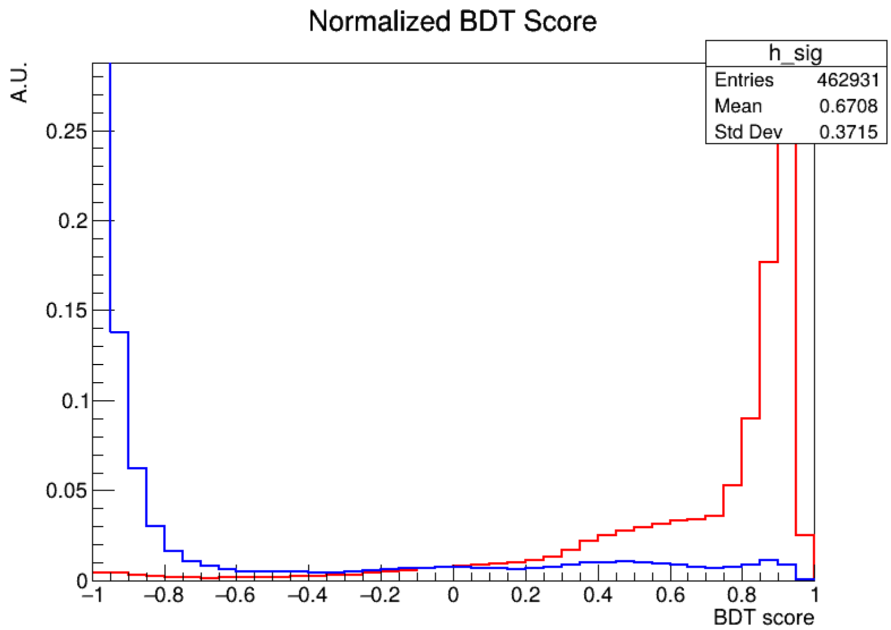

Figure 1: Normalized BDT output distribution for WW signal and top backgrounds.

Figure 1 shows the normalized BDT output distributions for the WW signal and the combined top backgrounds. The signal distribution peaks at higher BDT values, while the background is concentrated near lower values, indicating effective separation.

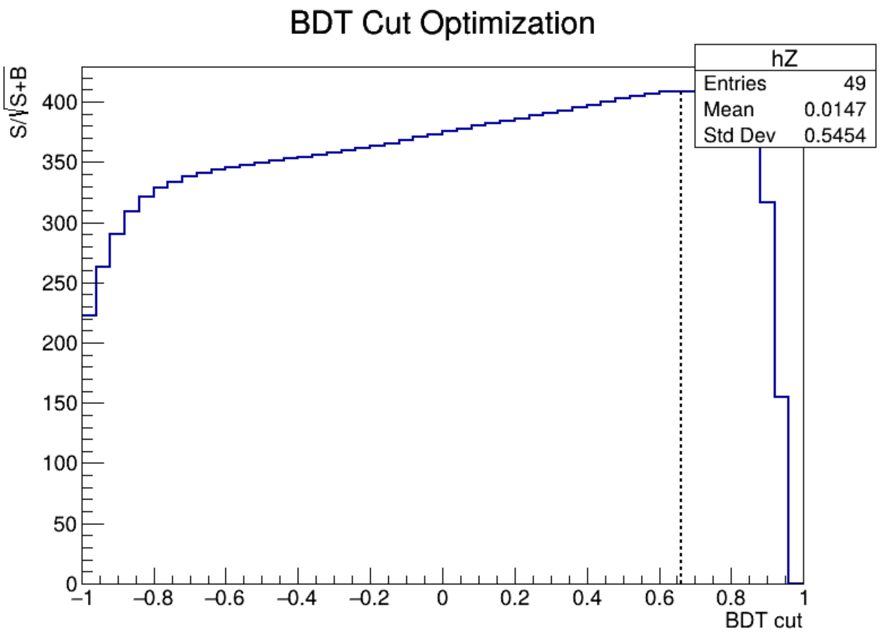

Figure 2: Stacked BDT output distribution for individual background processes compared with the WW signal.

Figure 2 shows the stacked background composition compared with the signal. The $t\bar{t}$ process is the dominant background, with smaller contributions from tW and $\bar{t}W$.

4 Cut Optimization and Significance

To determine the optimal BDT selection, the statistical significance $Z = \frac{S}{\sqrt{S+B}}$ is evaluated as a function of the BDT cut value, where S and B denote the number of signal and background events, respectively.

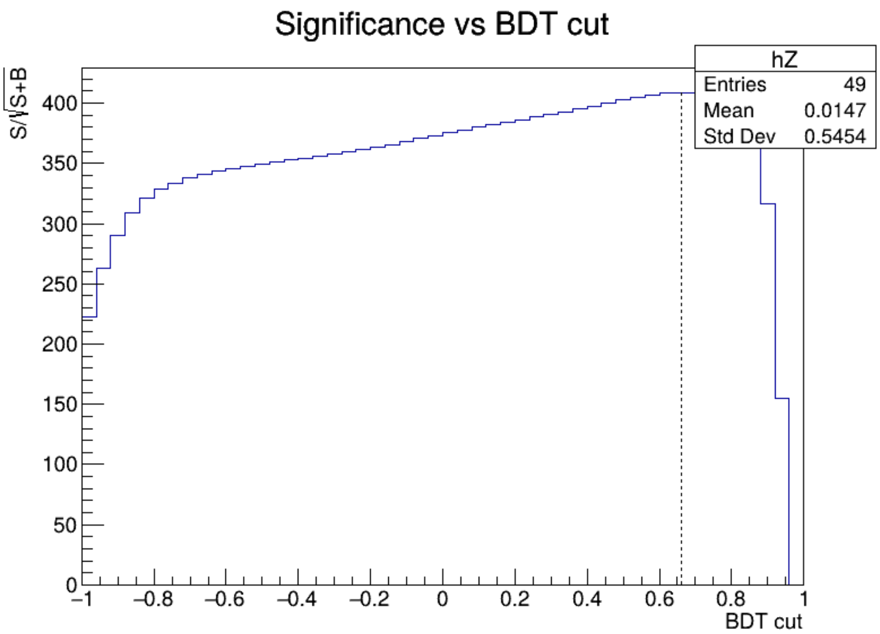

Figure 3: Statistical significance as a function of the BDT cut value.

Figure 3 shows the significance as a function of the BDT cut. The optimal cut is found at:

$$\text{Optimal BDT cut} = 0.66 \tag{1}$$
$$\text{Maximum significance} = 408.7 \tag{2}$$

At the optimal cut, the signal and background efficiencies are:

$$\epsilon_S = 0.681 \tag{3}$$
$$\epsilon_B = 0.049 \tag{4}$$

The final event yields after the BDT selection are:

$$S = 315{,}144 \tag{5}$$
$$B = 279{,}471 \tag{6}$$

The corresponding Asimov significance is:

$$Z_A = 517.3 \tag{7}$$

5 Conclusion

A Boosted Decision Tree analysis using TMVA has been successfully implemented to separate the WW signal from dominant top-quark backgrounds. The BDT classifier provides strong discrimination power, achieving high signal efficiency while suppressing background contributions. The optimized BDT cut significantly enhances the statistical significance compared to a cut-based approach. This study demonstrates the effectiveness of multivariate techniques in modern high-energy physics analyses.