Boosted particle tagging at the LHC based on large R jet substructure using machine learning techniques

Author: Carlos Buitrago Cárdenas. cabuitrago@unal.edul.co
Director: Carlos Sandoval Usme. cesandovalu@unal.edu.co
FENYX-UN
Physics Bachelor Thesis

Abstract

As the searches for new physics at the LHC continue to look into previously unexplored kinematic regimes, the importance of boosted particles constantly increases. These particles, whose transverse momentum is far greater than their rest mass, need to be efficiently reconstructed and identified (”tagged”). When decaying hadronically, these kind of objects have a unique detector signature. As their decay products are highly collimated, they are reconstructed as a single large R jet. In this work, a tagging mechanism for boosted particles based on the substructure information of these large R jets is explored. Specifically, the fat jet *N-subjettiness* shape and the fat jet mass are used to train different multivariate classifiers. The performance of the resulting algorithm is examined for the tagging of boosted top quarks and W bosons. The algorithm’s response to various inputs is also analysed.

Introduction

Boosted particles decaying hadronically result in multiple collimated (and unresolved) jets which will merge into a single large R jet, usually called fat jet [1]. The key to identifying and measuring boosted particles lies in the internal structure of the reconstructed fat jets, whose discriminating power is the object of this study.

The energy pattern of boosted hadronically decaying particles is fundamentally different from that of QCD jets of a similar invariant mass. *N-subjettiness* [2] is a jet shape which exploits said difference. It is denoted by $\tau_N$and it is defined using N candidate subjets identified during the clustering process as follows:

$$\tau_N = \frac{1}{d_0}\sum_k p_{T,k} \times \min(\Delta R_{1,k}, \Delta R_{2,k}, \ldots, \Delta R_{N,k})$$

Where k runs over the constituent particles and d_0 is a normalization factor which depends on the radius parameter used in the jet algorithm. The τ_N variables give us information about how well the jet substructure is defined by N subjets, effectively working as discriminant variables which can be employed to develop a tagging mechanism by making use of machine learning algorithms.

Methodology

The generation of the Monte Carlo samples for signal and background events consists of the same simulation chain. Initially, matrix elements for the signal processes are produced at LO using *MadGraph 5*, after which datasets of 500000 tree-level events are generated via *MadEvent* at $\sqrt{s}$ = 14 TeV with a generator-level cut of p_T> 350 GeV on all partons. Afterwards, the events are showered using *Pythia 8* and the response of the ATLAS detector is simulated via *Delphes 3*.

The signal processes being considered correspond to the production of top quarks and *W* bosons, both decaying in the fully hadronic channel (see figure 1). On the other hand, for the background sample the process consists of QCD dijet production (top jets excluded).

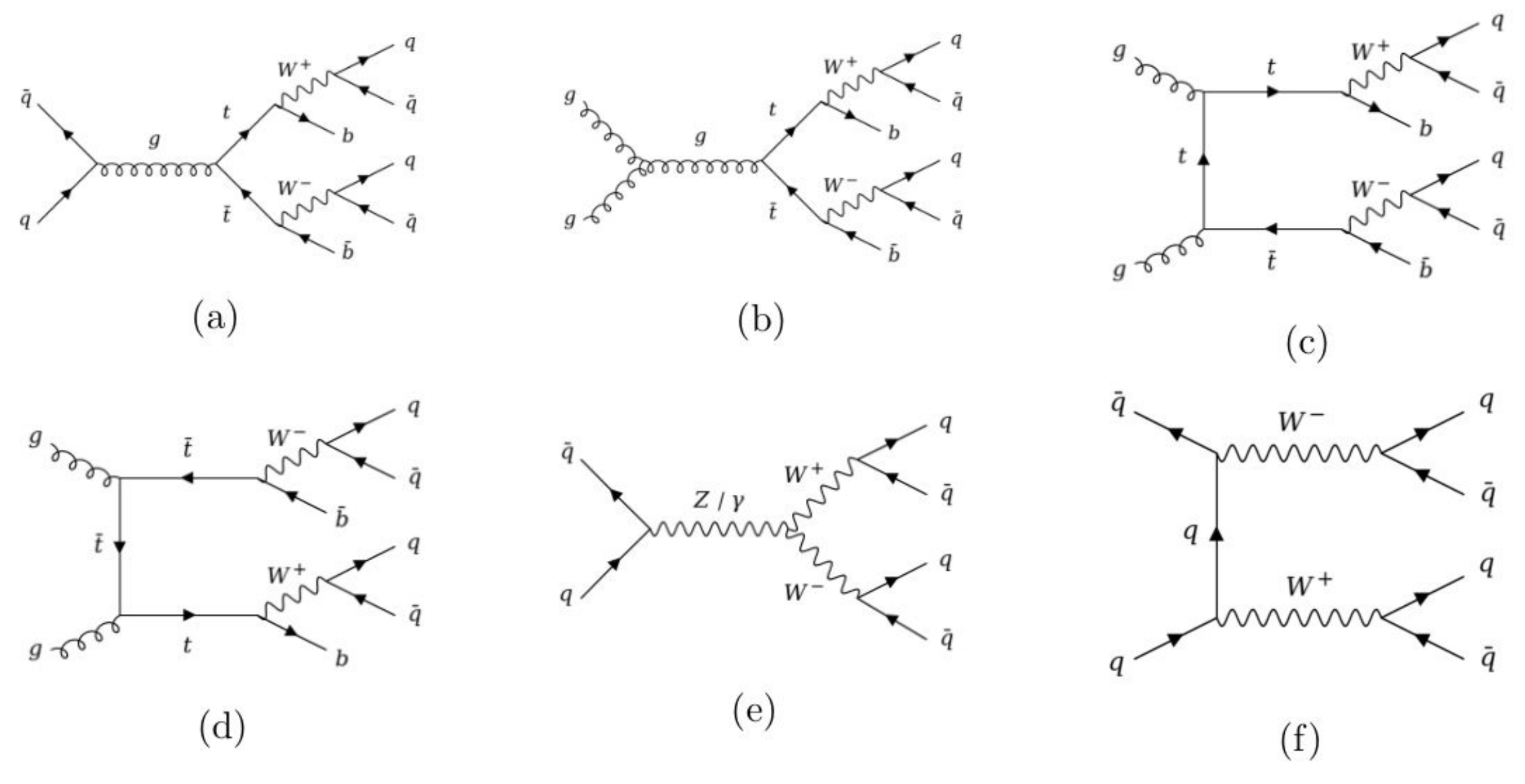

Figure 1: Signal processes used for the simulation of (a) - (d) boosted top production and (e) - (f) boosted *W* production.

The jets are reconstructed using the anti- k_T algorithm with *R* = 1. After their τ_N values and mass are computed using *FastJet*, they are used to train multiple multivariate algorithms via *TMVA 4*. The performance of the resulting tagging mechanism is evaluated when different inputs are used during the training. For both the signal and background datasets, half of the events (randomly selected) are used as the training sample and the other half as the testing sample.

Results

Initial assessment of the discriminant capabilities of the substructure variables:

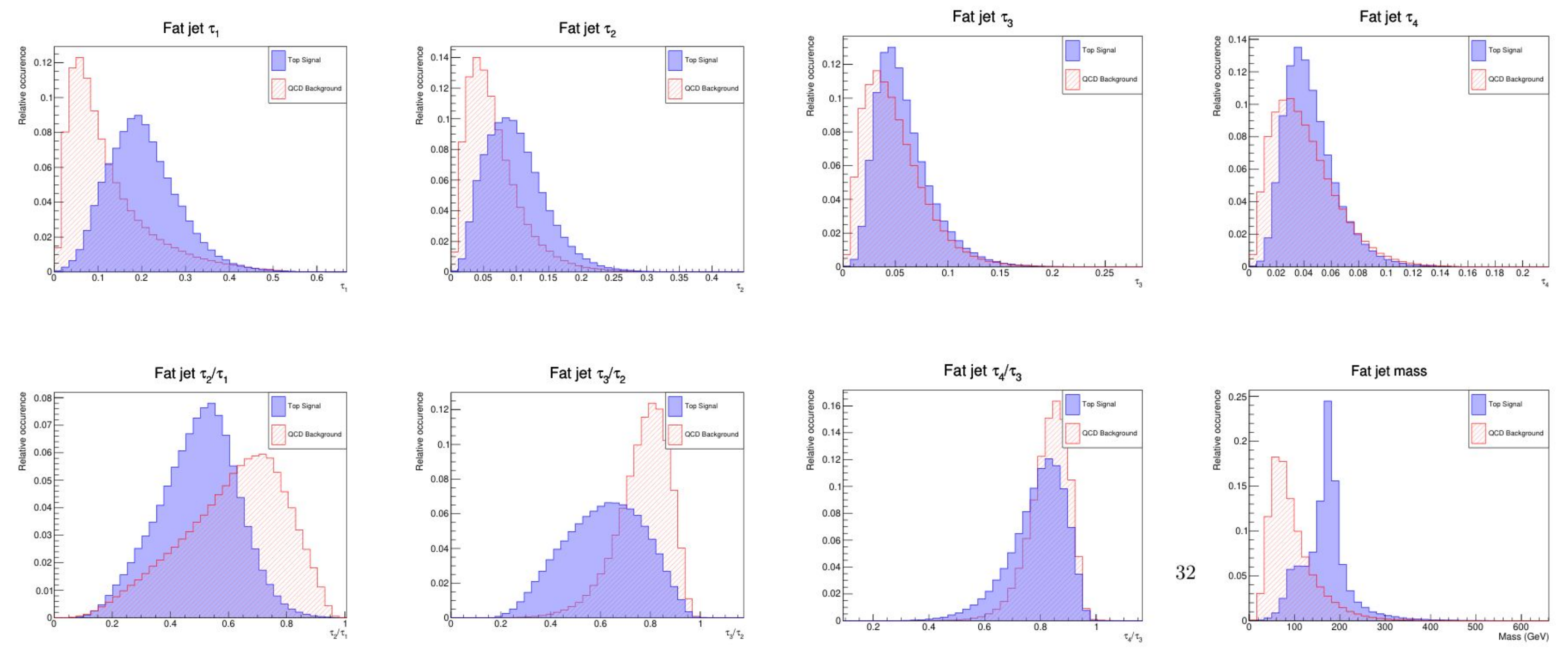

Figure 2: Distribution of τ_N variables, their ratios and the mass for the fat jets in the signal and background training samples (only shown for the top signal).

Classifier training results, performance measured by the area under the ROC curve (background rejection vs. signal efficiency):

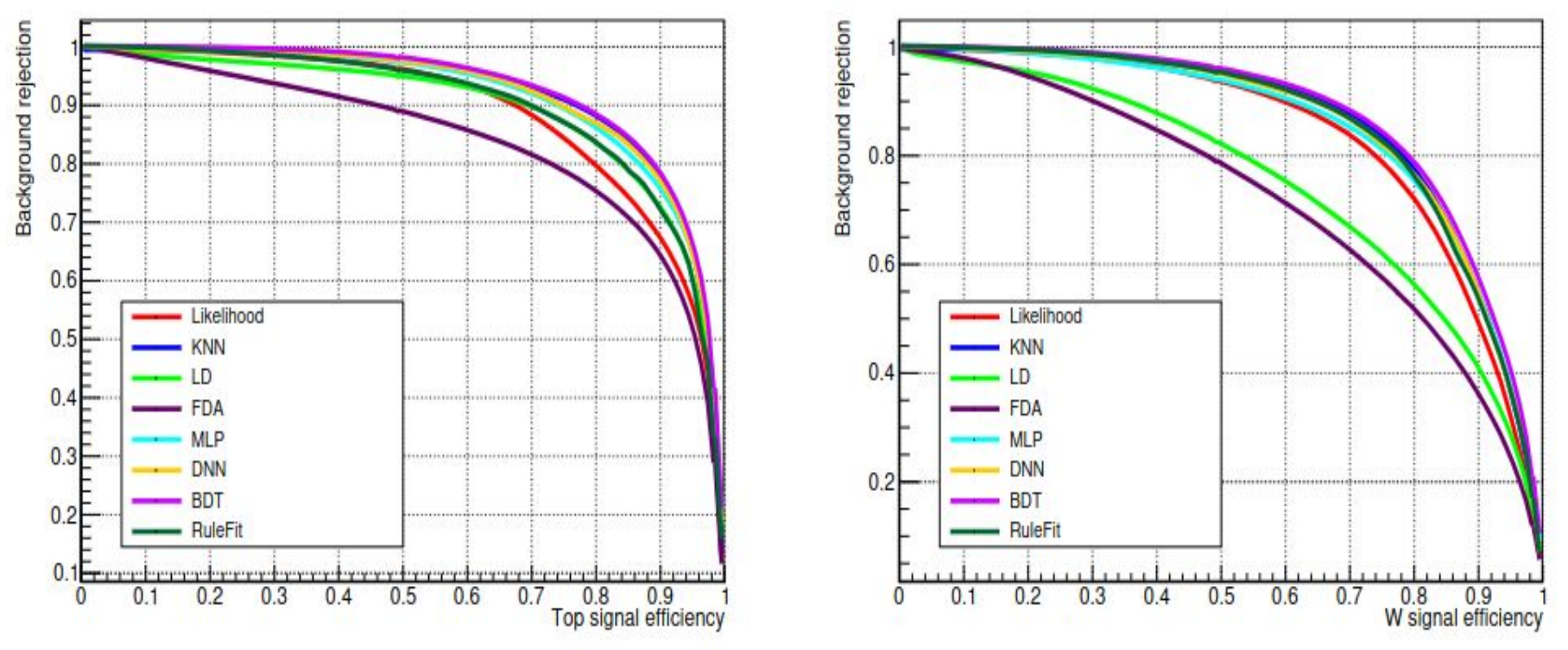

Figure 3: Performance of the different multivariate algorithms trained.

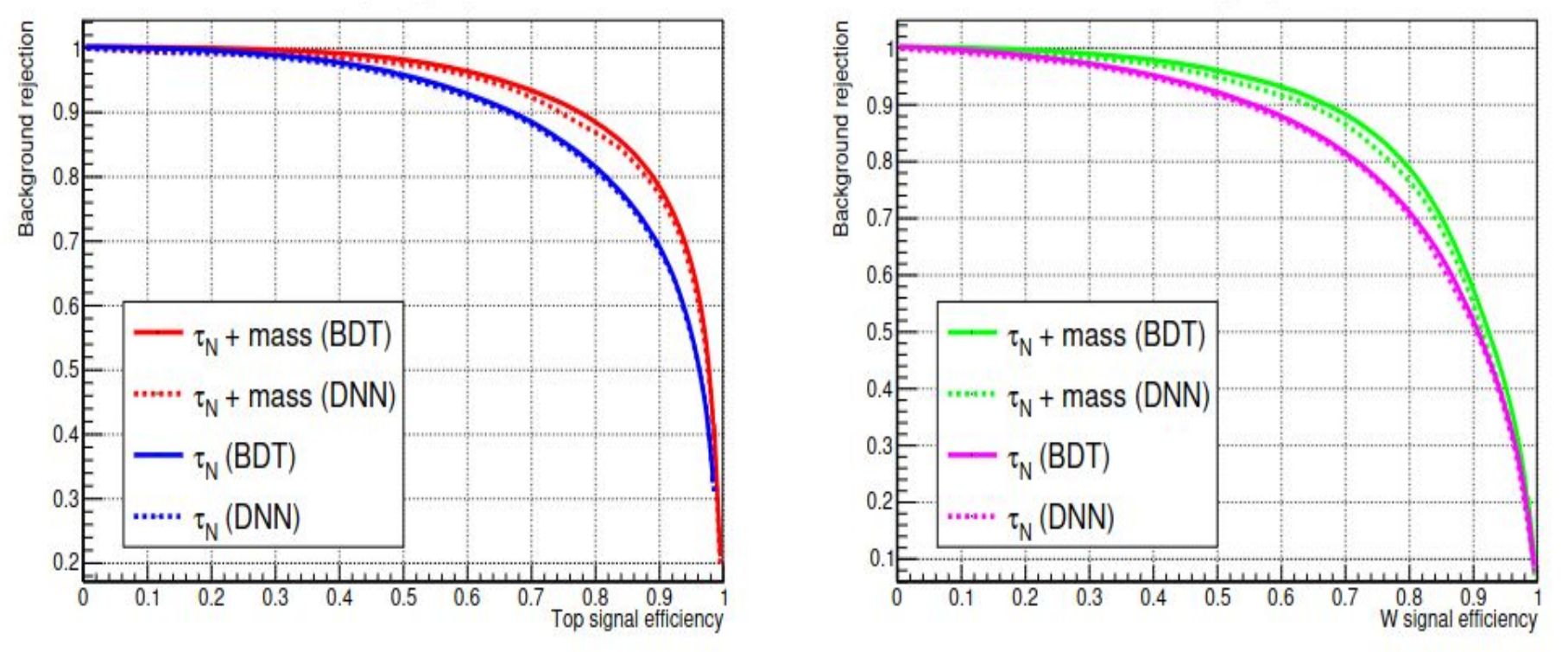

Figure 4: Effect of not including the mass as an input during the training.

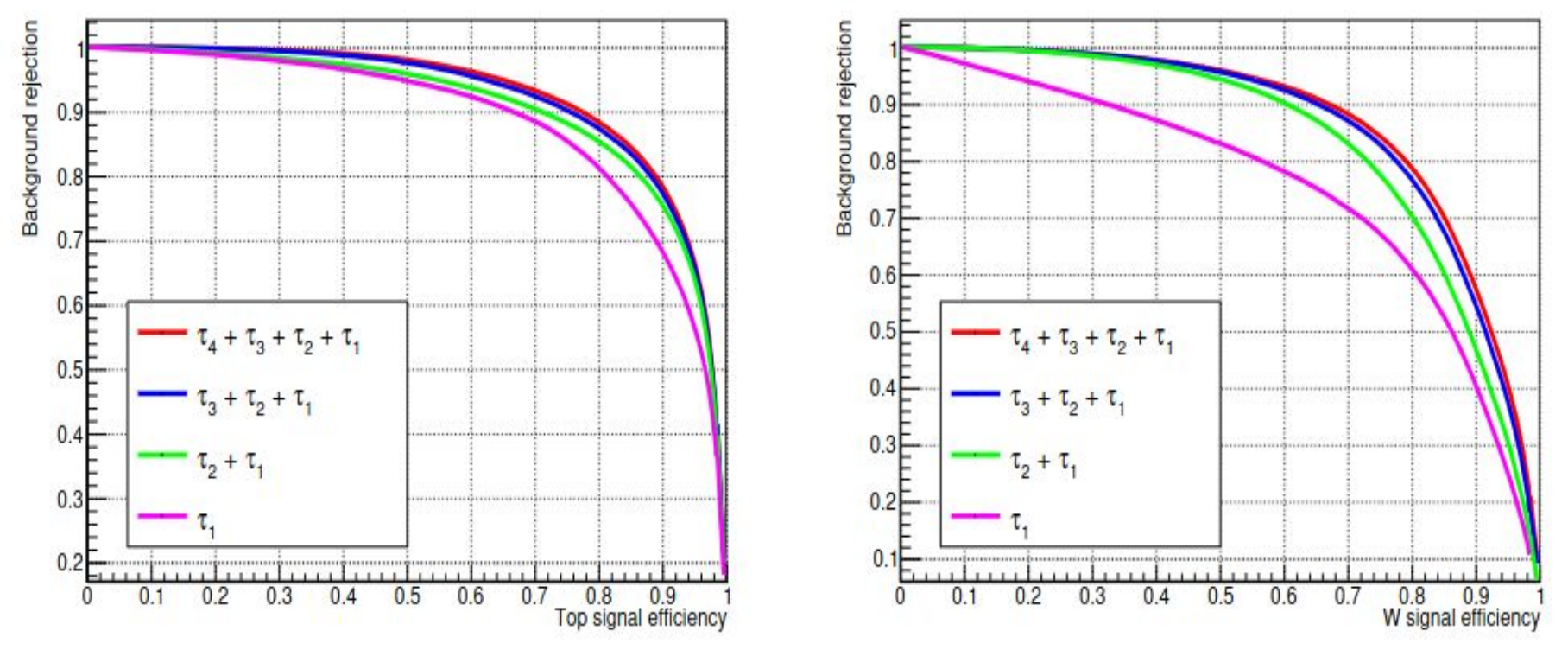

Figure 5: Effect of using different number of τ_N variables as input for the algorithm training (BDT).

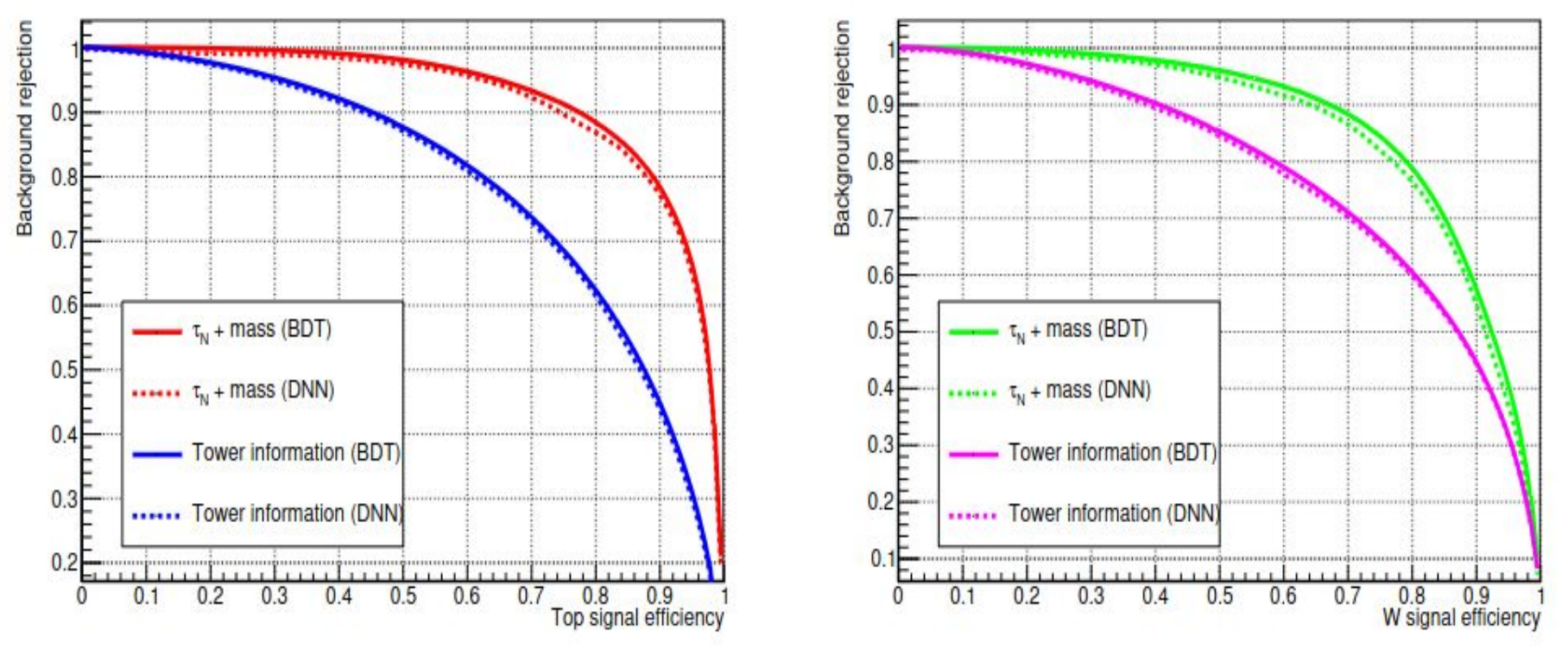

Figure 6: Performance obtained using substructure variables vs. using kinematic variables.

Conclusions

The discriminant capabilities of the τ_{N+1}/τ_N ratios were noticeable from the signal and background distributions of the training set. The boosted top and *W* tagging method proposed was carried out using the first four τ_N variables and the mass of the resulting fat jets. The best performance was achieved using the BDT classifier. The relevance of the mass in the tagging method was addressed and the potential of the τ_N variables by themselves was notorious. The gain in performance decreased noticeably with the increase of τ_N inputs being used, indicating a saturation of the model being trained. Finally, a clear difference in the performance achieved using substructure and kinematic variables was observed.

References

1 Altheimer, Andrew, et al. (2014). Boosted objects and jet substructure at the LHC. Report of BOOST2012, held at IFIC Valencia, 23rd–27th of July 2012. *The European Physical Journal C*, *74*, 1-24.

2 Thaler, J., & Van Tilburg, K. (2011). Identifying boosted objects with N-subjettiness. Journal of High Energy Physics, 2011(3), 1-28.