PROCEEDINGS OF SCIENCE

b

Gradient Boosting MUST taggers for highly-boosted jets

Rosa María Sandá Seoane[a,*] **and J. F. Seabra**[a,b]

[a] *Instituto de Física Teórica UAM-CSIC,*
C/ Nicolás Cabrera 13-15, Campus de Cantoblanco, 28049, Madrid, Spain

[b] *Departamento de Física and CFTP, Instituto Superior Técnico, Universidade de Lisboa,*
Av. Rovisco Pais,1, 1049-001 Lisboa, Portugal

E-mail: r.sanda@csic.es, joao.f.seabra@tecnico.ulisboa.pt

Identification of multi-pronged jets plays a fundamental role at the LHC and other colliders, especially in the forthcoming years when the energy frontier will reach unexplored regions. Within multivariate tools, the strategy of MUST (Mass Unspecific Supervised Tagging) has proven to be successful in implementing generic jet taggers capable of discriminating signals over a wide range of jet masses. We extend the MUST concept, by using eXtreme Gradient Boosting (XGBoost) classifiers instead of neural networks (NN's). We build both fully generic and specific multi-pronged taggers, to identify 2, 3, and/or 4-pronged signals from SM QCD background. We show that XGBoost-based taggers are not only easier to optimize and faster to train than those based in NN's, but also show quite similar performance, even when testing with signals not used in training, providing an efficient alternative machine-learning implementation for generic jet taggers.

The Eleventh Annual Conference on Large Hadron Collider Physics (LHCP2023)
22-26 May 2023
Belgrade, Serbia

[*] Speaker

1. Introduction

In the years to come, the energy frontier will reach unexplored regions at the LHC, and among other crucial tasks, jet identification plays a fundamental role when searching BSM physics at the LHC. Most jets are produced from QCD processes, but when sufficiently boosted, both SM particles (W, Z, t, h) as well as possible new particles can produce jets too.

Instead of crafting tools based on machine learning to tag a specific signal, an alternative approach involves building more generalized taggers that can differentiate a broader spectrum of multi-pronged signals from the QCD background. Introduced in [1] and further explored in [2–4], the Mass Unspecific Supervised Tagging (MUST) method is designed for training generic jet-tagging tools. In this approach, the jet mass and its transverse momentum, allowed to vary over wide ranges, serve as input variables for a machine-learning (ML) algorithm. As demonstrated in [1], the results indicate that neural networks (NNs) similar to those employed in [5] exhibit strong discriminative capabilities for any kind of multi-pronged signal throughout the entire training ranges of jet mass and transverse momentum.

Our aim within the work presented in this proceeding, part of the work submitted in [7], is to construct MUST taggers employing eXtreme Gradient Boosting (XGBoost) [6]. XGBoost is an open-sourced tool that has been selected for its scalable and exceptionally accurate implementation of the gradient boosting technique, pushing the frontiers of computing performance tailored for boosted tree algorithms. In comparison to previously constructed MUST taggers based on neural networks (NNs), the taggers presented here exhibit similar performance in signal-to-background discrimination. However, they offer advantages in terms of ease of optimization and significantly faster evaluation on test sets. This enhanced speed becomes particularly crucial for using the taggers in anomaly detection analyses over extensive data samples, as anticipated in the high luminosity phase of the LHC. In this context, the marginally superior performance of NN-based taggers is not decisive for detecting a potential new signal. A plausible approach involves employing a faster XGBoost-based tagger in the initial stages of analysis, with the incorporation of an NN only if a significant excess is detected. Lastly, XGBoost-based taggers provide an alternative machine learning implementation that may be essential to assess the analyses' robustness, especially in the presence of a potential new physics signal.

2. Setup

We use XGBoost to build a fully generic tagger, dubbed XGenT, designed to discriminate QCD jets (one-pronged, labeled as background) from multi-pronged ones (2, 3 and 4-pronged jets, labeled as signal). We also build for comparison a NN-based tagger, dubbed GenT, with the hyperparameters set as in [1]. Additionally, we build three multi-pronged taggers named $\texttt{XGenT}_{nP}$ ($n = 2, 3, 4$), trained specifically for nP signal discrimination.

All the data used to build the training and testing sets for the taggers was obtained from Monte Carlo simulations (see [7]). To generate background QCD jets, we considered the inclusive process $pp \to jj$. The distribution of jet mass m_J is continuous, with $m_J \in [10, 500]$ GeV, divided into 10 bins of 50 GeV (except the first one, within $[10, 50]$ GeV). The generation is made in 100 GeV bins of transverse momentum p_T, with $p_T \in [200, 2200]$ GeV. Even though events populate mostly the lower end of each bin, bins are narrow enough to provide a smooth p_T dependence. We should emphasize at this point that the m_J and p_T ranges are chosen arbitrarily and can be modified. Signal events were generated with the $pp \to ZS$ process, where S is a scalar and $Z \to \nu\nu$. Event samples are collected for each of the decay modes of S: for 4-pronged jets, $S \to u\bar{u}u\bar{u}$, $S \to b\bar{b}b\bar{b}$; for 3-pronged jets, $S \to F\,\nu$; $F \to udd$, $F \to udb$; and for 2-pronged jets, $\to u\bar{u}$, $S \to b\bar{b}$, where F is a color-singlet fermion. The decays of both S and F were implemented with a flat matrix element to make the jet substructure as model-independent as possible. We set $M_{S,F} \in [10, 800]$ GeV (and $M_{S,F} < p_T R/2$, with $R = 0.8$).

For each sample event, we obtain the mass m_J and transverse momentum p_T of jets, as well as 17 N-subjettiness observables, $\left\{\tau_1^{(1/2)}, \tau_1^{(1)}, \tau_1^{(2)}, \ldots, \tau_5^{(1/2)}, \tau_5^{(1)}, \tau_5^{(2)}, \tau_6^{(1)}, \tau_6^{(2)}\right\}$ (computed following Ref. [8]), which all together characterize jet substructure. These 19 variables are the input features of our taggers. We balance the number of background and signal events in the two-dimensional space defined by the m_J and the p_T bins. For further details on training and taggers architecture, see [7].

3. Results

We test our taggers with jet signals produced from boosted SM particles and new scalars with different masses, both included and not included in the training, to test whether the taggers are capable of classifying as signal novel classes of complex jets. In the left panels of Fig. 1, we show the results for a 2-pronged signal included in the training sample of a W boson decaying into two quarks ($W \to q\overline{q}$), where the W boson is pair-produced by a heavy Z' resonance with mass $M_{Z'} = 1.1$, 2.2 or 3.3 TeV. In the right panels of Fig. 1, we show the results for a 3-pronged signal not included in the training sample, of heavy neutrino N with a mass of 80 GeV decaying into two quarks and a hard electron ($N \to eq\overline{q}$), also initiated by a Z'.

In the upper row of Fig. 1 we compare the performance of the fully-generic tagger `XGenT` (dashed lines) with the performance of NN-based tagger `GenT` (solid lines). In all cases, the `XGenT` and the `GenT` taggers have a remarkably similar performance, which is also achieved in other benchmark points considered in [7], related to several multi-pronged signals produced both from SM and different BSM particles with different masses. The results for the heavy neutrino decay in the right-upper panel of Fig. 1, as well as the results for other examples exhibited in [7], show that `XGenT` taggers are also capable of identifying new signals not included in the training. As expected, from the lower row of Fig. 1 we see that for all the considered examples the multi-pronged taggers provide higher efficiencies than the generic taggers, but at the cost of losing generality. For the full set of results (including a prongness selection tagger), see [7].

4. Conclusions

The results presented in this proceeding, part of a work submitted in [7], show that the performance of the fully-generic tagger `XGenT` closely rivals that of its NN counterpart, `GenT`, introduced in Refs. [1, 3], not only on multi-pronged jets used in the training but also on other types of complex jets not included in the training set. This outcome is noteworthy in itself, given the limited availability of generic jet taggers in existing literature. Employing alternative methods

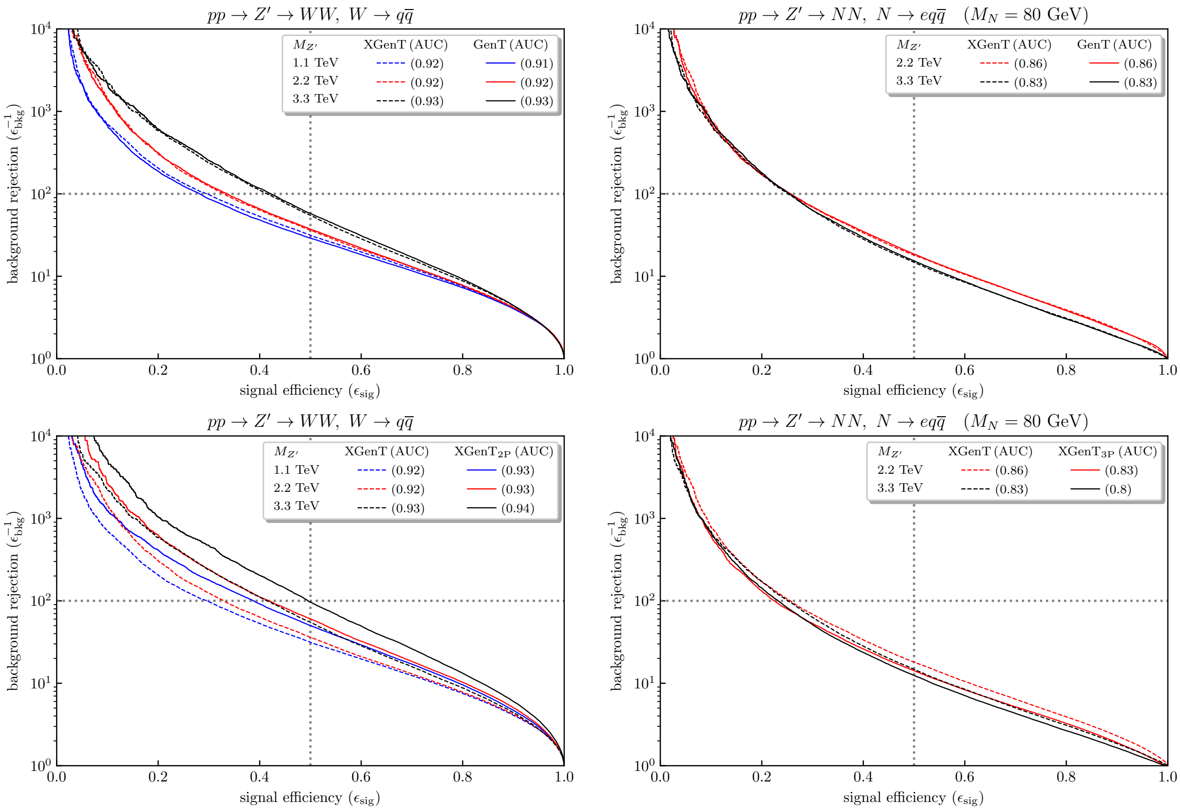

Figure 1: `XGenT` vs. `GenT` (upper row) and `XGenT` vs. `XGenT`$_{2P/3P}$ ROC curves (lower row) for 2/3-pronged jet signals used/not used in training.

for exploring new physics in data is essential to assess the robustness of searches, and becomes imperative if any deviation from the Standard Model expectation is identified. Furthermore, it has been determined that `XGenT` significantly outpaces its NN counterpart in terms of processing speed. In scenarios where jet tagging is computationally intensive and speed is critical, `XGenT` offers a substantial advantage.

Acknowledgments

RMSS would like to thank Diego Mouriño. This work is partially supported by the "Atracción de Talento" program (Modalidad 1) of the Comunidad de Madrid (Spain) under the grant number 2019-T1/TIC-14019 (RMSS), by the Spanish Research Agency (Agencia Estatal de Investigación) through the Grant IFT Centro de Excelencia Severo Ochoa No CEX2020-001007-S (RMSS) and by the grant PID2021-124704NB-I00 (RMSS) funded by MCIN/AEI/10.13039/501100011033. The work of JFS is supported by the FCT grant SFRH/BD/143891/ 2019.

)

References

[1] J. A. Aguilar-Saavedra, F. R. Joaquim and J. F. Seabra, JHEP **03** (2021), 012 [erratum: JHEP **04** (2021), 133] doi:10.1007/JHEP03(2021)012 [arXiv:2008.12792 [hep-ph]].

[2] J. A. Aguilar-Saavedra, Eur. Phys. J. C **81** (2021) no.8, 734 doi:10.1140/epjc/s10052-021-09530-w [arXiv:2102.01667 [hep-ph]].

[3] J. A. Aguilar-Saavedra, Eur. Phys. J. C **82** (2022) no.2, 130 doi:10.1140/epjc/s10052-022-10058-w [arXiv:2111.02647 [hep-ph]].

[4] J. A. Aguilar-Saavedra, Eur. Phys. J. C **82** (2022) no.3, 270 doi:10.1140/epjc/s10052-022-10221-3 [arXiv:2201.11143 [hep-ph]].

[5] J. A. Aguilar-Saavedra, J. H. Collins and R. K. Mishra, JHEP **11** (2017), 163 doi:10.1007/JHEP11(2017)163 [arXiv:1709.01087 [hep-ph]].

[6] T. Chen and C. Guestrin, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16), 785 (2016) [arXiv:1603.02754 [hep-ph]].

[7] J. A. Aguilar-Saavedra, E. Arganda, F. R. Joaquim, R. M. Sandá Seoane and J. F. Seabra, [arXiv:2305.04957 [hep-ph]].

[8] K. Datta and A. Larkoski, JHEP **06** (2017), 073 doi:10.1007/JHEP06(2017)073 [arXiv:1704.08249 [hep-ph]].