Study of novel pileup mitigation techniques in quark/gluon and tau jets using machine learning

Summer student report

Nafija Ibrišimović, University of Sarajevo

August 23, 2024

Abstract

The goal of this project is testing new Pileup Per Particle Identification (PUPPI) algorithm for identification and reconstruction of hadronic decays of tau leptons, alongside so far used algorithms with the same purpose, hadron-plus-strips (HPS) and Charged Hadron Subtraction (CHS) algorithm. Tests are done on the set of data for the next run, simulated by Monte Carlo (MC). This report presents preliminary results of these tests, showing efficiency and purity of algorithms for for several requirements on the tau identification.

1 Introduction

With the increasing luminosity of the LHC, the number of pileup interactions per bunch crossing is increasing as well, and will reach up to 200 pileup interactions during high-luminosity LHC operation. This results in more noise around interesting physics objects, such as jets initiated by quark/gluon or tau leptons. Removing the pileup from an event is therefore essential, since pileup does not only affect the jet energy but also other event observables. In the CMS Collaboration a new machine learning based pileup mitigation technique, Pileup Per Particle Identification (PUPPI) algorithm [1], has been introduced and extensively tested for jets. Testing these types of algorithms also for hadronic decays of tau leptons is crucial in CMS physics analyses that test the Standard Model or search for signals of new physics. The PUPPI algorithm was recently adjusted to improve efficiencies for hadronic tau decays as described in Ref. [2].

For tau identification, CMS uses Charged Hadron Subtraction (CHS) algorithm [3]. As the products of hadronic tau decay (denoted τ_h) leave particular signatures in the CMS detector, they can be used to reconstruct these decays as jets in data. In CMS, this is done by the hadron-plus-strips (HPS) algorithm [4], which uses AK4 jets as seeds. Apart from these, CMS is currently using DeepTau [4] identification algorithm, a multiclassifier algorithm based on a convolutional deep neural network, to reject jets that are falsely reconstructed as τ_h decays.

Recently, CMS developed new jet tagging algorithms which include τ_h classifiers. This study uses PNet [5] and ParT [6], which are trained on and applied to CHS and PUPPI jets, respectively.

In this project, efficiency and purity of HPS, CHS and PUPPI algorithms are tested for several requirements on the tau identification performed on data for two regions of pseudorapidity (η), $0.0 \leq |\eta| < 1.3$ and $1.3 \leq |\eta| < 2.4$, which correspond to the central and forward detector regions, respectively. Tests are performed on a sample of simulated $\mu\tau_\mathrm{h}$ events that are generated by the Drell–Yan process $\mathrm{Z}/\gamma^* \to \tau\tau$, simulated for CMS detector conditions of year 2024.

2 Efficiency

In the following, we denote "HPS", "CHS", and "PUPPI" as three overlapping collections of jets that were reconstructed by the different algorithms, as well as identified as τ_h jets. For HPS, the identification algorithm used is DeepTau, for CHS that is PNet, and for PUPPI ParT. Table 1 shows the requirements that were used in the code [7] to identify the τ_h jets belonging to each collection.

Algorithm	Identifier [8]	VVVLoose	Medium
HPS	idDecayModeNewDMs	True	True
	&& idDeepTau2018v2p5VSjet	>= 1	>= 5
CHS	rawPNetVSjet	> 0.05	> 0.5
PUPPI	rawPNetVSjet	> 0.05	> 0.5

Table 1: Identifiers for HPS, DeepTau, CHS and PUPPI algorithms in the NANOAOD data format of CMS [7].

The same identifiers are used in calculating purity as well.

Absolute efficiency is defined as the number of τ_h jets reconstructed by each algorithm with respect to all generated τ_h jets in the MC simulation, while relative efficiencies are calculated for each combination of algorithms with respect to the number of τ_h jets reconstructed by HPS, CHS or PUPPI that are paired with a generated τ_h jet. Absolute efficiencies for all three algorithms and both identification requirements are shown in Figure 1. Relative efficiencies for $0.0 \leq |\eta| < 1.3$ are shown in diagrams in Figure 2 and for $1.3 \leq |\eta| < 2.4$ in Table 2.

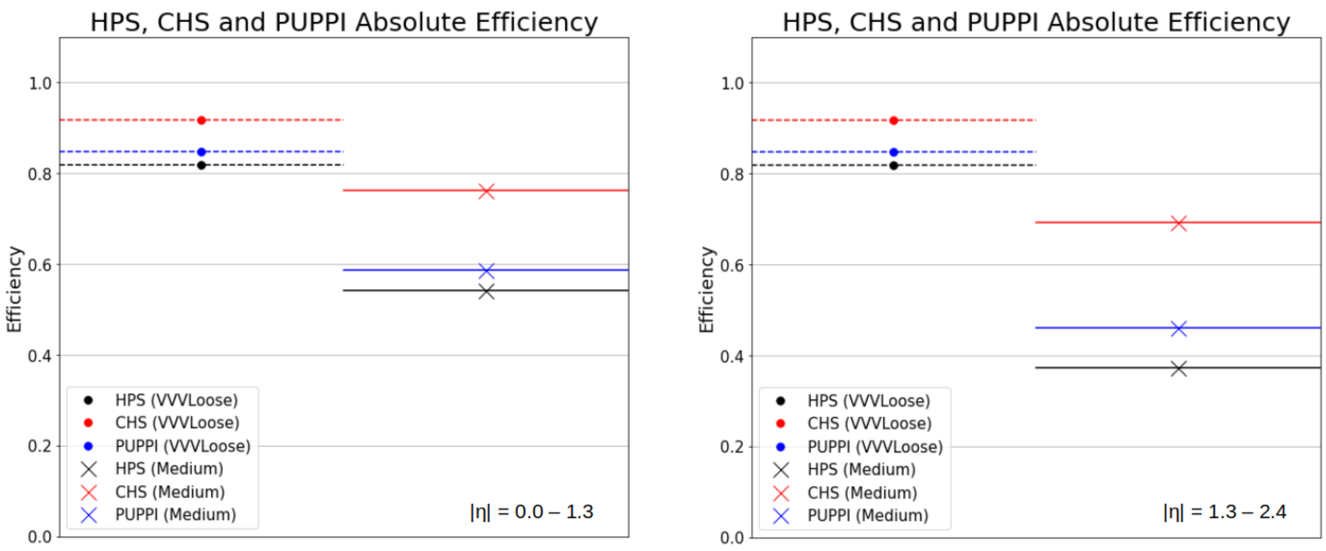

Figure 1: Absolute efficiencies of HPS, CHS and PUPPI for VVVLoose and Medium identification requirement in η regions $0.0 \leq |\eta| < 1.3$ (left) and $1.3 \leq |\eta| < 2.4$ (right).

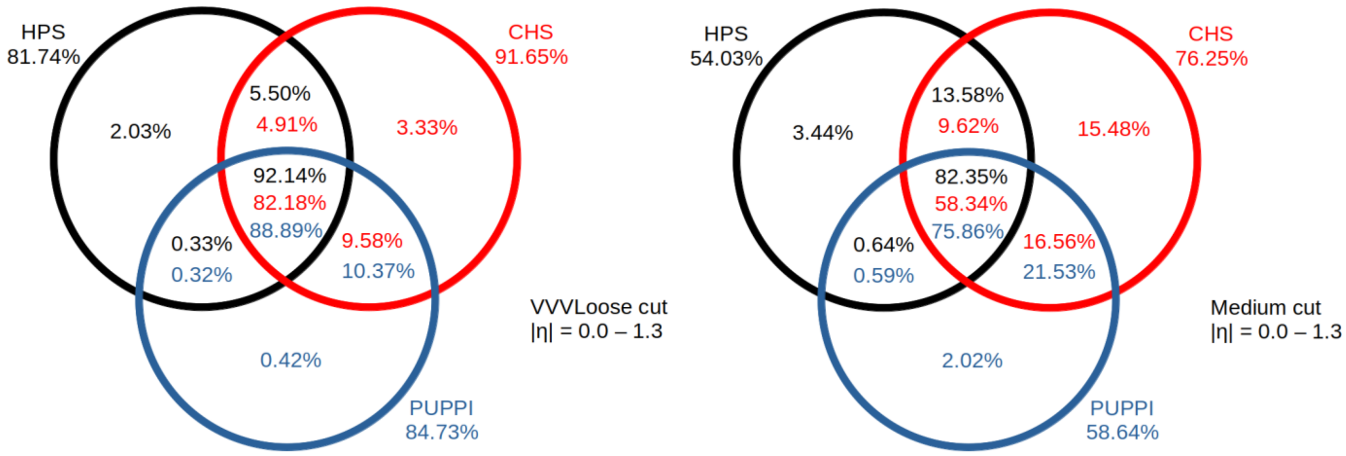

Figure 2: Relative efficiencies for $0.0 \leq |\eta| < 1.3$ with respect to HPS (black), CHS (red) and PUPPI (blue) for VVVLoose (left) and Medium identification requirement (right). Absolute efficiencies for HPS, CHS and PUPPI are shown below the respective labels for each identification requirement.

Collection	w.r.t. HPS		w.r.t. CHS		w.r.t. PUPPI	
	VVVL (71.43%)	M (37.37%)	VVVL (87.90%)	M (69.13%)	VVVL (78.36%)	M (46.12%)
HPS	1.97%	3.92%	6.08%	9.22%	0.44%	0.61%
CHS	7.49%	17.05%	5.84%	26.43%	16.70%	33.04%
PUPPI	0.48%	0.75%	14.89%	22.04%	0.76%	2.93%
All three	90.06%	78.27%	73.18%	42.31%	82.10%	63.43%

Table 2: Relative efficiencies for $1.3 \leq |\eta| < 2.4$ with respect to HPS, CHS and PUPPI for VVVLoose (VVVL) and Medium identification requirement (M). Absolute efficiencies for each identification requirement are shown in parentheses.

From these results, it is notable that there is a certain number of hadronic taus that are visible only to one of the algorithms, while the other two do not reconstruct them, meaning that they are inevitably lost when excluding the respective algorithm that reconstructs them from this stage of data analysis. Another thing worth noticing is the high dependence of both absolute and relative efficiencies on the chosen identification requirement, where absolute efficiencies of HPS and PUPPI drop and relative efficiencies with respect to CHS rise significantly when applying tighter identification requirement.

3 Purity

Purity is defined as the ratio of the number of generated τ_h jets paired with a reconstructed and identified τ_h jet and the total number of τ_h jets the respective algorithm reconstructed. The purity of HPS, CHS and PUPPI is shown in Figure 3.

These plots show the expected jump in purity with the use of tighter identification requirement. The increase in purity is significantly high for PUPPI, which has the highest purity out of three for the Medium identification requirement in the $1.3 \leq |\eta| < 2.4$ region.

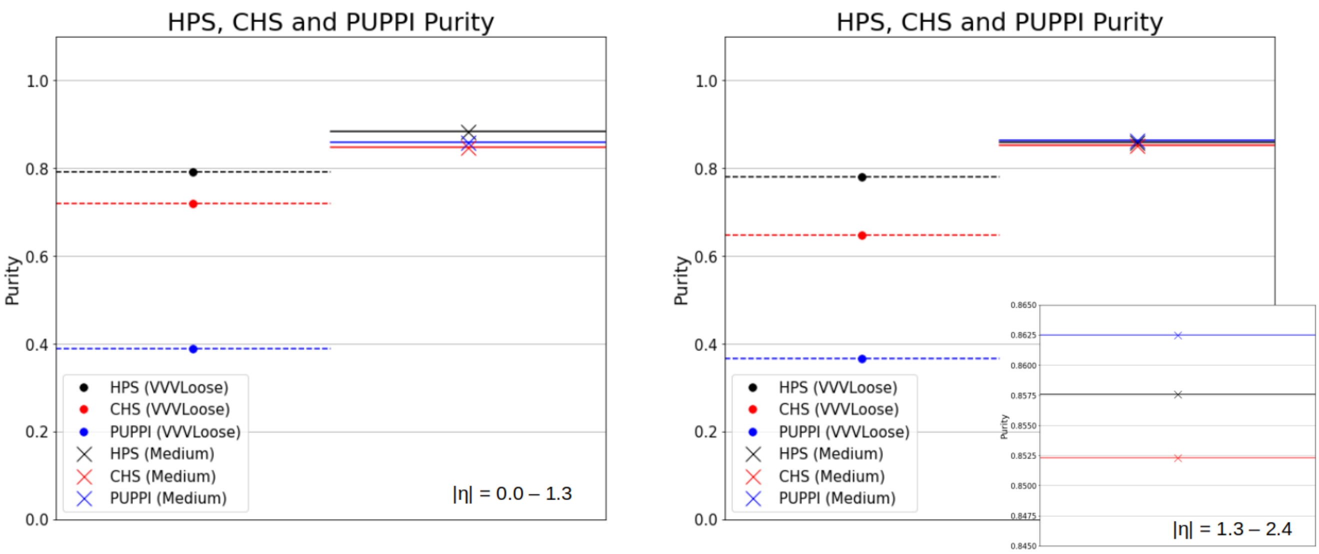

Figure 3: Purity of HPS, CHS and PUPPI for VVVLoose and Medium identification requirements for η regions $0.0 \leq |\eta| < 1.3$ (left) and $1.3 \leq |\eta| < 2.4$ (right). The purity in the second region for the Medium identification requirement is shown zoomed in (bottom right corner).

4 Conclusion and Outlook

Looking at the relative efficiencies and taking into account the relatively large number of taus reconstructed only by CHS, it is clear that PUPPI still needs more testing and adjusting before it can completely take over the reconstruction of hadronic taus from CHS. However, the performance of the jet reconstruction algorithms still have to be studied further with tau identification requirements that are more comparative before one can draw out a concrete conclusion. Another thing worth looking into is the reason behind varying in dependence of the applied identification requirement among the algorithms.

The code used for getting the results in this report is preserved on GitLab in Ref. [7].

Acknowledgments

First of all, I would like to thank CERN for giving me the opportunity to work here as a summer student. This project would also be unimaginable without everyone in CMS Collaboration who worked on developing, testing and improving algorithms for tau decay identification. In the end, most of all I would like to thank my supervisors, Dr. Izaak Neutelings and Dr. Anna Benecke, who came up with this project and whose help and guidance led this project to where it is now.

References

[1] **CMS** Collaboration, "Pileup mitigation at CMS in 13 TeV data", *JINST* **15** (2020), no. 09, P09018, `doi:10.1088/1748-0221/15/09/P09018`, `arXiv:2003.00503`.

[2] **CMS** Collaboration, "Optimizing the pileup per particle identification algorithm in the context of τ_h lepton identification in Run3",.

[3] **CMS** Collaboration, "Particle-flow reconstruction and global event description with the CMS detector", *JINST* **12** (2017), no. 10, P10003, `doi:10.1088/1748-0221/12/10/P10003`, `arXiv:1706.04965`.

[4] **CMS** Collaboration, "Identification of hadronic tau lepton decays using a deep neural network", *JINST* **17** (2022) P07023, `doi:10.1088/1748-0221/17/07/P07023`, `arXiv:2201.08458`.

[5] H. Qu and L. Gouskos, "ParticleNet: Jet Tagging via Particle Clouds", *Phys. Rev. D* **101** (2020), no. 5, 056019, doi:10.1103/PhysRevD.101.056019, arXiv:1902.08570.

[6] H. Qu, C. Li, and S. Qian, "Particle Transformer for Jet Tagging", arXiv:2202.03772.

[7] "GitLab – Nafija Ibrisimovic/jme-validation, branch confusion_matrix". https://gitlab.cern.ch/nibrisim/jme-validation/-/tree/confusion_matrix?ref_type=heads. (Retrieved Aug 21, 2024).

[8] "Documentation for EGamma1_Run2024D-PromptReco-v1.root". https://cms-nanoaod-integration.web.cern.ch/autoDoc/NanoAODv14/2024Prompt/doc_EGamma1_Run2024D-PromptReco-v1.html#Tau. (Retrieved Aug 23, 2024).