Tau identification in CMS during LHC Run 2

Mohammad Hassan Hassanshahi[1*] on behalf of the CMS Collaboration

1 Imperial College London
* mhh18@ic.ac.uk

November 20, 2021

16th International Workshop on Tau Lepton Physics (TAU2021),
September 27 – October 1, 2021
doi:10.21468/SciPostPhysProc.?

Abstract

The LHC Run 2 data-taking period was characterized by an increase in instantaneous luminosity and center-of-mass energy. Several techniques have been deployed in the CMS experiment to reconstruct and identify tau leptons in this environment. The DeepTau identification algorithm is used to identify hadronically decaying tau leptons from quark and gluon induced jets, electrons, and muons. Compared to previously used MVA identification algorithms, the use of deep-learning techniques brought a noticeable improvement in the tau identification and rejection of contaminating sources. Low transverse momentum topologies were addressed separately with a dedicated identification algorithm, while machine learning techniques were implemented to improve the identification of the tau hadronic decay channels. These algorithms have been already used for several published physics analyses in CMS. The algorithms are presented together with their measured performances.

Contents

---

1 Introduction

During the Run-2 of the Large Hadron Collider (LHC), which took place in 2015-18, the center-of-mass energy and instantaneous luminosity of proton-proton (pp) collisions increased. These increases provided the opportunity for probing new physics through precise measurement of the Standard Model (SM) parameters or by directly searching for physics beyond the Standard Model (BSM). Several BSM theories can be probed at the LHC through processes with tau leptons in the final state [1–3]. Besides, various properties of the Higgs boson can be measured through its decay to tau letpons [4–6], thanks to the large branching fraction of the decay and the relatively clear signal taus provide in the detector.

Tau leptons decay hadronically (to hadrons and neutrinos) more than half of the times. Identifying such decays is challenging since they can be faked by other objects such as quark or gluon jets, especially with the higher luminosity and hence more soft pp interactions in the Run-2 of the LHC. Needless to say, in order to exploit searching for new physics in tau final state, we need to reduce such contamination to the greatest extent possible. In this note, we report three techniques developed in CMS to improve tau identification: A deep convolutional neural network (CNN) for identifying taus which decay hadronically (τ_h) from quark or gluon jets, electrons and muons, a boosted decision tree to identify the decay modes of τ_h, and an attention-based graph neural network to reconstruct 3-charged-prong decays of τ_h in low-p_T regime.

In section 2, we illustrate the algorithm used in CMS to reconstruct τ_h. In section 3, the techniques developed for improving τ_h identification are presented, and finally, the last section contains conclusion.

2 τ_h reconstruction in CMS

CMS employs particle-flow (PF) [7] algorithm in order to reconstruct individual physics objects: neutral and charged hadrons, muons, electrons, and photons. To this end, the PF algorithm uses an optimized combination of the information from all CMS subdetectors (see [8] for more information on the CMS detector). The objects which are reconstructed by the PF algorithm are called PF candidates. More complex objects such as τ_h, which usually consist of multiple PF candidates, are reconstructed by means of dedicated algorithms.

In CMS, τ_h candidates are reconstructed with the Hadron-Plus-Strip (HPS) algorithm [9–11]. As the first step, the algorithm uses a hadronic jet as a seed. These jets are reconstructed by clustering PF candidates using the anti-k_T algorithm [12,13] with distance parameter 0.4. All PF candidates within a cone size of $\Delta R = 0.5$ (where $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ and η is pseudorapidity) around the jet axis are considered for the next steps. For each jet, at most one τ_h is eventually reconstructed; only the highest-p_T τ_h is kept if more than one passes all of the HPS requirements.

Next, neutral pions (π^0) and charged hadrons ($h^\pm$) are reconstructed from the jet constituents. To reconstruct neutral pions, the 4-momenta of all photons and electrons within a "strip" in $\eta - \phi$ space are added. PF charged hadrons are selected as charged hadron candidates. Based on the number of π^0 and $h^\pm$ candidates, a decay mode is assigned to the τ_h.

Further conditions are applied to the reconstructed τ_h. The candidate is rejected if it contains reconstructed π^0 or $h^\pm$ outside the signal cone defined by $\Delta R = 3/p_T$ (GeV) with respect to the τ_h axis. The signal cone is bounded between 0.05 and 0.1. In addition, charge and mass conditions are imposed on the τ_h and intermediate resonances, respectively, to ensure the compatibility of τ_h with a genuine hadronic tau decay.

3 τ_h identification techniques

3.1 Deep CNN for τ_h identification (DeepTau)

An object whose kinematic quantities are similar to those of a τ_h candidate can pass the HPS algorithm requirements and be misidentified as a τ_h. For example, quark and gluon jets can fake any decay mode of τ_h while electrons and muons can be misidentified mainly as 1-charged-prong decays of τ_h . In order to improve the identification of genuine τ_h , CMS has developed a deep convolutional neural network, named *DeepTau*, in which low-level and high level features of τ_h are combined to achieve the optimal performance.

There are a total of 47 high-level features in the network. These features include τ_h candidate properties, such as its four-momentum, its compatibility with primary vertex, the number of charged and neutral particles used to reconstruct τ_h , as well as general event properties such as the estimated pileup density.

To define low-level features, two overlapping grids centered on τ_h axis are defined in $\eta-\phi$ space. The inner (outer) grid contains 11 × 11 (21 × 21) cells with cell size of 0.02 × 0.02 (0.05 × 0.05) which covers the signal (isolation cone), see Fig. 1. For each cell, seven object types are considered: PF candidates including muons, electrons, photons, charged and neutral hadrons along with muons and electrons from standalone reconstruction algorithms, which provide more information about the objects. In each cell and for a given object type, only the highest-p_T object is retained. In total, 188 features are extracted from each cell.

A summary of the network architecture is shown in Fig. 2 . For the low-level features, in each cell from inner or outer grid, the associated features are pre-processed using four neural networks: Three for incorporating the features of electrons and photons combined, muons, and hadrons, independently, and one for concatenating and combining their outputs. The pre-processing step reduces the number of features per cell from 188 to 64. At this step, there are 64 grids of size 11 × 11 and 64 of size 21 × 21. Each of these grids are fed into a convolutional neural network (CNN) which eventually reduces the grid to a single value. In parallel to low-level features, the 47 high-level features are pre-processed in a neural network with 57 outputs. Finally, a neural network with 5 dense layers is used to combine high- and low-level features. This network receives 64 × 2 low-level and 57 high-level features as input and four scores corresponding to τ_h , muons, electrons and hadronic jets as output.

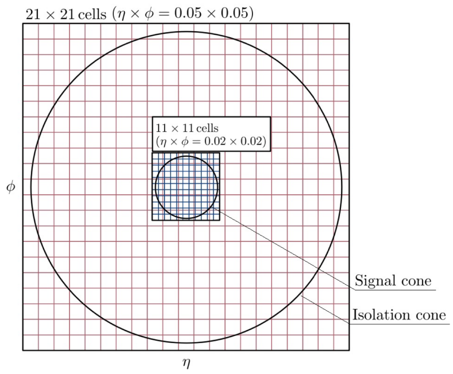

Figure 1: The grids used in DeepTau for low-level feature extraction.

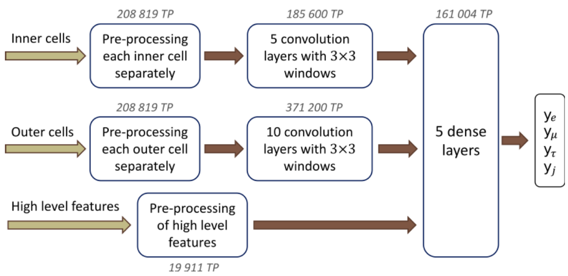

Figure 2: The architecture of DeepTau [14].

The final discriminator is defined as

$$D_\alpha(y) = \frac{y_\tau}{y_\tau + y_\alpha} \tag{1}$$

The loss function includes a regular cross-entropy term in addition to two binary focal-loss terms. The training was performed with NAdam algorithm.

The performance of the DeepTau discriminator for identifying τ_h against jets is shown in Fig. 3. The performance is studied using two processes dominated with quark and gluon jets in the final state, namely $t\bar{t}$ and W+jet. The DeepTau discriminator significantly outperforms the previous ones used in CMS in all working points. We observed similar enhancement when using DeepTau for discriminating τ_h against muons and electrons. More information about DeepTau and its performance can be found in [15].

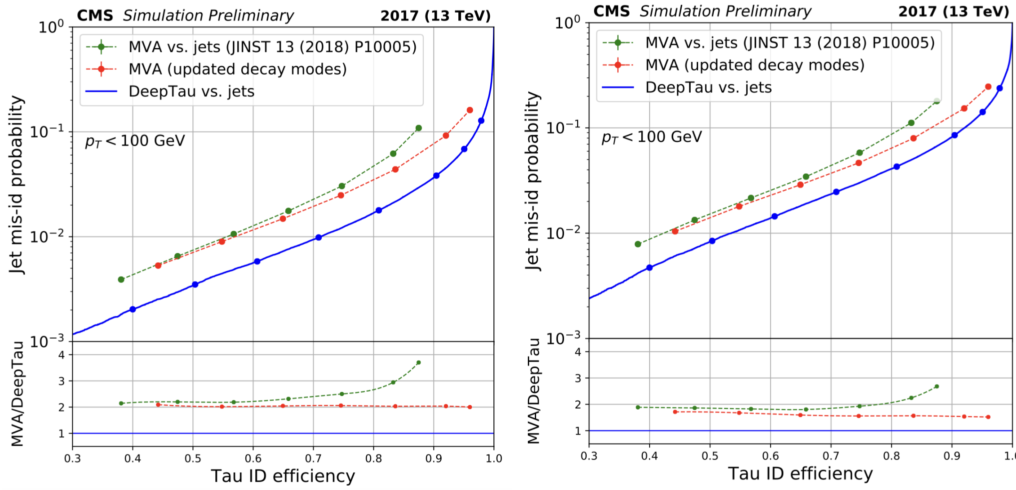

Figure 3: The receiver operating characteristic (ROC) curves comparing the performance of the DeepTau and previous MVA discriminants in identifying τ_h against jets. In the left (right) plot, $t\bar{t}$ (W+jet) sample is used for hadronic jet production [15].

3.2 τ_h decay mode identification (MVA decay mode)

The unprecedented amount of pp collision data that the experiments at the LHC have recorded during Run-2 data-taking period enables precise measurement of fundamental Standard Model parameters in the processes containing tau leptons in the final state. Some of these measurements require a strong identification power of τ_h decay modes because their observables are decay mode sensitive, for example the measurement of the CP structure of Higgs-tau Yukawa coupling [5].

Although the HPS algorithm, combined with the DeepTau discriminator, can provide pure samples of τ_h, the HPS is not particularly optimized for decay mode identification. As an example, the HPS algorithm consolidates $\tau \to \pi\pi^0$ and $\tau \to \pi 2\pi^0$ [1] into a single reconstructed decay mode, meaning that only one strip is reconstructed. To improve decay mode identification, we developed two classifiers targeting 1- and 3-charged-prong decays, independently, using boosted decision tree (BDT) algorithm from XGBoost library. The decay modes reconstructed with these classifiers are called *MVA decay mode*, in which MVA stands for multivariate analysis. The HPS algorithm is already very effective in identifying the number of charged prongs in a τ_h decay, which means that our classifier primarily targets finding the number of π^0. Therefore, it is sensible to have separate classifiers based on the number of charged prongs in the decay.

The main 1-charged-prong decays of τ_h are $\tau \to a_1 \to \rho\pi^0 \to \pi 2\pi^0$, $\tau \to \rho \to \pi\pi^0$, and $\tau \to \pi$. They differ by the number of π^0 in the final state and the number and types of

[1] In this section, neutrinos are not shown for simplicity as they either do not interact with or are not practically detectable in our detector.

intermediate resonances. We exploited these differences as features in the classifier to improve decay mode identification. The main features include the invariant masses of the reconstructed strip as well as the reconstructed ρ meson, the kinematic and angular quantities associated to the τ_h decay products, and the decay mode reconstructed by the HPS algorithm (*HPS decay mode*).

A similar approach is taken for the 3-charged-prong decay classifier, in which the dominant decays are $\tau \to a_1 \to \rho^0\pi \to 3\pi$ and $\tau \to 3\pi\pi^0$. Likewise, the invariant mass, kinematic and angular quantities along with the HPS decay mode are incorporated into the features of the classifier. Thanks to the presence of more pions in the final state compared to the 1-charged-prong case, several features associated to different combinations of pions are also added to the classifier. Both classifiers have multiple outputs each representing the score for one of the decay modes. An "other" output category is also added to each classifier to collect a small fraction of objects not similar to the other categories.

Fig. 4 compares the performance of MVA and HPS decay modes. The purity of all decay modes has improved by 10 to 55%-points and the efficiency of the decay modes containing at least one π^0 in the final state has enhanced by 5 to 40%-points. The efficiency of 3π decay mode is retained, while it is reduced by 7%-points in the π decay mode. For the first time in CMS, a collection of $\tau \to \pi 2\pi^0$ decay with decent efficiency and purity is provided. More information on the classifier can be found [16].

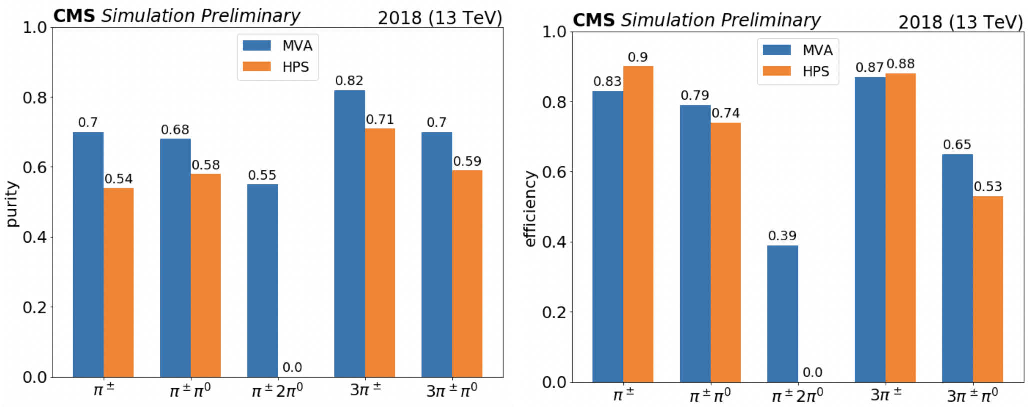

Figure 4: The purity (left) and efficiency (right) of τ_h decay modes reconstructed as MVA (blue) and HPS (orange) decay modes. The τ_h candidates used for producing the figures are collected from $H \to \tau\tau$ decay with one tau decaying to a muon and neutrino and the other to a τ_h [16].

3.3 Low-pt tau reconstruction

There has been recently a growing interest for lepton universality tests through measuring the decay rate of B mesons. τ_h from such decays are mainly produced in low transverse momentum regime ($p_T < 10$ GeV), as shown for the $B_c \to J/\psi\tau\nu$ decay in Fig. 5 (left). The HPS algorithm is not efficient in this p_T regime as the strong magnetic field of the CMS detector largely spread τ_h decay products in the $\eta - \phi$ plane and hence they are not contained within the cone of a seeding jet (see section 2). Therefore, we used a machine learning algorithm to identify low-p_T 3-charged-prong decays without using jets for seeding. This algorithm is optimized for B meson decay studies but could potentially be extended to other analyses with low-p_T τ_h in the final state.

In this reconstruction algorithm, firstly all PF charged pions are collected. After that, the tracks not originating from the vicinity of primary vertex (PV) are removed from the collection.

The PV is defined as the closest proton-proton (pp) collision point to the extrapolation of J/ψ direction in the $B_c \to J/\psi \tau \nu$ decay. This choice is analysis-specific but it provides optimal efficiency for selecting the pp collision from which B_c is produced.

Even after vertex requirement, a large number of charged pions are left, which are mainly from soft interactions in the pp collisions. In order to reduce this contamination, an attention-based graph neural network (ABCNet) [17] is employed. The benefit of graph neural nets for such analyses is that the data is treated the same way as they are recorded by the detector. Moreover, this network takes advantage of attention-based mechanism to improve local feature extraction, leading to a more efficient architecture. The input variables to the network are the 4-momentum of charged pions, their distance from PV and their charge. ABCNet assigns a probability to each of the charged pions for originating from a real τ_h decay. Pions are required to have an ABCNet score (probability) of more than 0.1443, which corresponds to 80% efficiency.

And finally, among the charged pions which survive the previous conditions, there are different ways to choose three to be a candidate for τ_h. In order to find the right combination, the highest-p_T τ_h candidate which satisfies the following conditions is chosen:

- τ_h vertex compatibility of more than 10%
- More than 3σ significance for τ_h vertex flight length with respect to the PV
- The sum of the ABCNet scores of the three pions be above 2.3

The first condition is a score showing the goodness of fit to the pion tracks and the second one is the distance between the vertex of the reconstructed three charged pions and the PV, divided by their vertex reconstruction uncertainty.

The efficiency of identifying charged pions originating from a real τ_h decay as a function of generator-level visible τ_h p_T is shown in Fig. 5 (right). The efficiency is defined as [18]:

$$\epsilon = \frac{\text{Tau is reconstructed and three charged pions are the right combination}}{\text{All events with 3-charged-prong tau at the generator-level}} \tag{2}$$

The new algorithm significantly outperforms the HPS algorithm in low-p_T regime. This promising result opens a window to the analyses of B meson decays with tau final state. More information on this algorithm can be found in [18].

4 Conclusion

The large amount of data taken during Run-2 of the LHC and recorded by the CMS experiment provides a great opportunity for measuring the SM parameters and probing BSM physics, in particular in processes with tau leptons in the final state. In order to achieve optimal sensitivity, one needs to maximize the power of identification and reconstruction of taus. In this note, we summarized the techniques developed in CMS for identifying and reconstructing hadronic decays of taus (τ_h). A deep convolutional neural network was designed for identifying τ_h, which showed a significant improvement in discriminating τ_h against hadronic jets, electrons and muons. In addition, a boosted decision tree algorithm was developed to identify decay modes of τ_h. This algorithm enhanced the purity in all decay modes and increased the efficiency in decay modes with at least one π^0 in the final state. Besides, in order to identify 3-charged-prong decays of τ_h in the low-p_T regime, we used an attention-based graph neural network which remarkably enhanced the identification efficiency compared to the existing method.

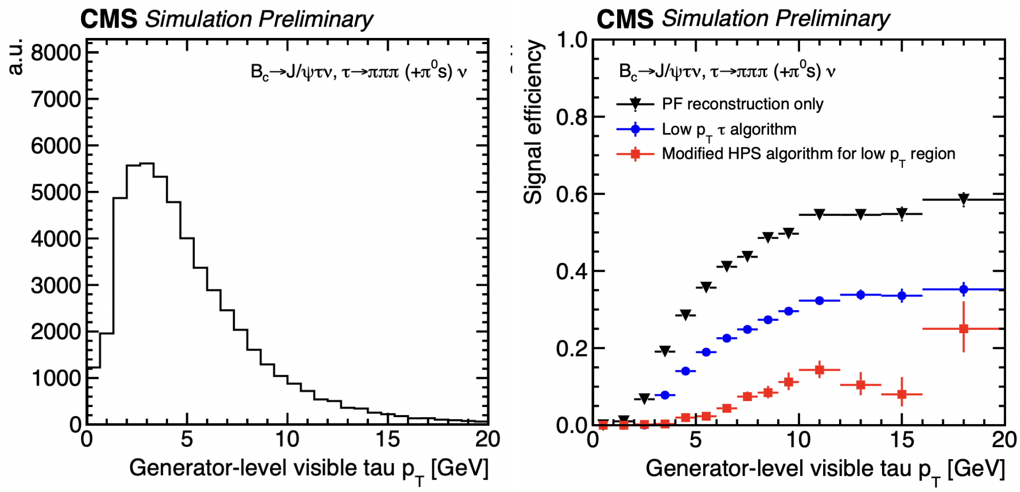

Figure 5: Left: The generator-level distribution of τ_h p_T in the $B_c \to J/\psi\tau\nu$ with $\tau \to \pi\pi\pi$ $(+\pi^0)$. Right: The efficiency, as defined in Eq. 2, for identifying τ_h as a function of τ_h p_T when using modified HPS algorithm – with distance parameter increased from 0.4 to 0.8 – (in red), and when using dedicated low-p_T τ_h reconstruction algorithm (in blue). The black points show the efficiency for reconstructing all three PF charged pions [18].

Author contributions The classifiers for τ_h decay mode finding, as described in section 3.2, are initiated, trained, and developed to the final stage by the author.

References

[1] CMS Collaboration, *Search for additional neutral MSSM Higgs bosons in the $\tau\tau$ final state in proton-proton collisions at $\sqrt{s}$ = 13 TeV*, JHEP **09**, 007 (2018), doi:10.1007/JHEP09(2018)007.

[2] ATLAS Collaboration, *Searches for third-generation scalar leptoquarks in $\sqrt{s}$ = 13 TeV pp collisions with the ATLAS detector*, JHEP **06**, 144 (2019), doi:10.1007/JHEP06(2019)144.

[3] CMS Collaboration, *Search for a low-mass $\tau^+\tau^-$ resonance in association with a bottom quark in proton-proton collisions at $\sqrt{s}$ = 13 TeV*, JHEP **05**, 210 (2019), doi:10.1007/JHEP05(2019)210.

[4] CMS Collaboration, *Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector*, Phys. Lett. B **779**, 283 (2018), doi:10.1016/j.physletb.2018.02.004.

[5] CMS Collaboration, *Analysis of the CP structure of the Yukawa coupling between the Higgs boson and τ leptons in proton-proton collisions at $\sqrt{s}$ = 13 TeV*, Tech. rep., CMS-HIG-20-006, CERN, Geneva (2021).

[6] ATLAS Collaboration, *Cross-section measurements of the Higgs boson decaying into a pair of τ-leptons in proton-proton collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, Phys. Rev. D **99**, 072001 (2019), doi:10.1103/PhysRevD.99.072001.

[7] CMS Collaboration, *Particle-flow reconstruction and global event description with the CMS detector*, JINST **12**(10), P10003 (2017), doi:10.1088/1748-0221/12/10/P10003.

[8] CMS Collaboration, *The CMS Experiment at the CERN LHC*, JINST **3**, S08004 (2008), doi:10.1088/1748-0221/3/08/S08004.

[9] CMS Collaboration, *Performance of tau-lepton reconstruction and identification in CMS*, JINST **7**, P01001 (2012), doi:10.1088/1748-0221/7/01/P01001.

[10] CMS Collaboration, *Reconstruction and identification of τ lepton decays to hadrons and ν_τ at CMS*, JINST **11**(01), P01019 (2016), doi:10.1088/1748-0221/11/01/P01019.

[11] CMS Collaboration, *Performance of reconstruction and identification of τ leptons decaying to hadrons and ν_τ in pp collisions at $\sqrt{s}$ = 13 TeV*, JINST **13**(10), P10005 (2018), doi:10.1088/1748-0221/13/10/P10005.

[12] M. Cacciari, G. P. Salam and G. Soyez, *The anti-k_t jet clustering algorithm*, JHEP **04**, 063 (2008), doi:10.1088/1126-6708/2008/04/063.

[13] M. Cacciari, G. P. Salam and G. Soyez, *FastJet User Manual*, Eur. Phys. J. C **72**, 1896 (2012), doi:10.1140/epjc/s10052-012-1896-2.

[14] A. Cardini, *Tau identification exploiting deep learning techniques*, PoS **ICHEP2020**, 723 (2021), doi:10.22323/1.390.0723.

[15] CMS Collaboration, *Performance of the DeepTau algorithm for the discrimination of taus against jets, electron, and muons* (2019), CMS-DP-2019-033.

[16] CMS Collaboration, *Identification of hadronic tau decay channels using multivariate analysis (MVA decay mode)* (2020), CMS-DP-2020-041.

[17] V. Mikuni and F. Canelli, *ABCNet: An attention-based method for particle tagging*, Eur. Phys. J. Plus **135**(6), 463 (2020), doi:10.1140/epjp/s13360-020-00497-3.

[18] CMS Collaboration, *Performance of the low-pT tau identification algorithm* (2020), CMS-DP-2020-039.