IDENTIFICATION OF TAU LEPTONS USING DEEP LEARNING TECHNIQUES AT CMS

K. Androsov[1,a] for the CMS Collaboration

[1] *Istituto Nazionale di Fisica Nucleare Sezione di Pisa, Pisa, Italy*

E-mail: [a] konstantin.androsov@cern.ch

The reconstruction and identification of tau leptons decaying into hadrons are crucial for analyses with tau leptons in the final state. To discriminate hadronic τ decays from the three main backgrounds (quark or gluon induced jets, electrons, and muons), with a low rate of misidentification and with high efficiency on the signal at the same time, the information of multiple CMS sub-detectors is combined. The application of deep machine learning techniques allows to exploit the available information in a very efficient way. The introduction of a new multi-class DNN-based discriminator at CMS provides a considerable improvement of the tau identification performance with respect to the previously used BDT and cut-based discriminators.

Keywords: LHC, CMS, machine learning, tau lepton

Konstantin Androsov

1. Introduction

The tau is the heaviest Standard Model (SM) lepton with a mass of 1776.86 ± 0.12 MeV [1]. It decays into hadrons + neutrino in about 64.8% of all cases. Taus play an important role for Higgs physics, where the scalar couplings to fermions are proportional to the mass of the fermions. Other physics analyses, such as measurements of the properties of SM particles or searches for new BSM particles (W', Z', leptoquarks, …), also involve tau leptons. A good performance in reconstruction and identification of the hadronic tau decays (τ_h) is a crucial ingredient for achieving optimal results in such analyses. For proton-proton collisions at the Large Hadron Collider (LHC), the main backgrounds that can be misidentified as τ_h are quark or gluon induced jets, electrons and muons that can be produced by Drell-Yan, leptonic W decays, and other SM processes. In this article, we introduce a new machine learning (ML) based algorithm, DeepTau, to identify τ_h decays in the CMS experiment [2].

2. Tau reconstruction and identification in CMS

The distinct feature of the CMS detector [2] is a superconducting solenoid of 6 m internal diameter, providing a magnetic field of 3.8 T. A silicon pixel and strip tracker, a lead tungstate crystal electromagnetic calorimeter (ECAL), and a brass and scintillator hadron calorimeter are located within the solenoid volume. Muons are detected in gas-ionization detectors embedded in the steel flux-return yoke outside the solenoid.

Individual particles (electrons, muons, photons, and neutral and charged hadrons) in the event are reconstructed by the particle-flow (PF) algorithm [3], which combines the information from all CMS subdetectors. Jets are reconstructed based on an anti-k_T algorithm [4, 5], clustering neutral and charged PF candidates with a distance parameter of 0.4. Hadronically decaying taus are reconstructed with the hadron-plus-strip (HPS) algorithm [6, 7], seeded by anti-k_T jets. This algorithm uses information from PF candidates belonging to the jet and reconstructs τ_h candidates based on the number of charged hadrons and the number of ECAL strips in the $\eta - \varphi$ plane. Tau candidates are rejected if their absolute charge is other than 1 or if they have charged particles or strips outside the signal cone. The signal cone is defined in the $\eta - \varphi$ plane by $R_{sig} = 3.0\ \text{GeV} / p_T^{\tau}$, and is limited to the range $0.05 - 0.10$. In [7] four modes to reconstruct τ_h decays are defined: 1 charged prong + 0, 1, 2 π^0 and three charged prongs + 0 π^0 with tight matching conditions. Recently more inclusive decay mode definitions (further referred to as the "updated decay modes") have been introduced, adding three charged prongs + 0 or 1 π^0 with relaxed matching conditions to the previously available reconstruction modes.

Before the introduction of DeepTau, to discriminate τ_h decays against each type of background three dedicated algorithms were used within CMS [7]. The rate of quark or gluon induced jets that were reconstructed as tau candidates (τ_j) was reduced by a multivariate (MVA) discriminator based on boosted decision trees (BDT) trained on 22 high-level input variables like the sums of energy depositions in the tau isolation cone ($R_{iso} = 0.5$) and the tau lifetime. An ensemble of 8 BDT discriminators, each trained on $\mathcal{O}(30)$ variables that characterize ECAL clusters and track quality, was used to reject electrons reconstructed as tau candidates (τ_e). A cut-based selection, using summary information about hits in muon chambers and energy deposited in the calorimeters, was applied to discriminate true taus against muons reconstructed as tau candidates (τ_μ).

3. DeepTau: a new ML-based tau identification algorithm

To further improve the identification of τ_h decays, low-level information from multiple CMS sub-detectors is combined. The application of ML techniques has been proven to provide superior results for such multi-dimensional problems. DeepTau is a new multiclass tau identification algorithm based on a convolutional deep neural network (DNN) that combines information from the high-level variables attributed to the reconstructed hadronic tau candidate with low-level information from the inner tracker, calorimeters and muon sub-detectors using particle candidates reconstructed within the τ_h signal and isolation cones. DeepTau also takes advantage from using the updated decay mode definitions.

The training is performed on a balanced mix of $\mathcal{O}(1.4 \cdot 10^8)$ τ_e, τ_μ, τ_h and τ_j candidates coming from Drell-Yan, $t\bar{t}$, W+jets and Z' Monte Carlo (MC) simulation. Training, validation and testing sets are composed of reconstructed tau candidates with a minimal preselection: $p_T \in [20,\ 1000]$ GeV, $|\eta| < 2.3$, and $|dz| < 0.2$ (the longitudinal impact parameter of the tau with respect to the primary vertex), which makes it suitable for a wide range of CMS analyses with hadronic taus in the final state. The ground truth is based on MC truth matching.

The inputs are separated into sets of high-level and low-level features. As high-level inputs, the algorithm takes 42 variables that are used during tau reconstruction or proven to provide discriminating power by previous tau discriminators, and one global event variable – the average energy deposition density (ρ). For each candidate reconstructed within the tau signal or isolation cones, information of 4-momentum, track quality, relation with the primary vertex, calorimeter clusters, and muon stations is used, if available. The tau signal and isolation cones define two regions of interest in vicinity of the tau candidate. Based on the angular distance between the reconstructed tau 4-momentum, all available candidates are split into two $\eta \times \varphi$ grids of 11×11 (21×21) cells with a cell size of 0.02×0.02 (0.05×0.05) for the signal (isolation) cone. In cases where there is more than one object of the given type that belong to the same cell, only the object with the highest p_T is considered as input. Within each cell, the input variables are split into 3 blocks: e-gamma, muon, hadrons. One input cell is represented by 188 inputs: 34 variables in the hadrons block, 60 variables in the muon block, and 82 variables in the e-gamma block, plus four high-level features, which are added for each block.

As a result, the total number of inputs is 105 699: 43 high-level features and 105 656 from the two grids. The high dimensionality of the inputs is compensated by a low occupancy: the average number of non-empty cells in the training set is around 1.7% (7.1%) for the signal (isolation) grid.

The organization of the low-level inputs into two 2D grids allows to first process the local patterns originating from the tau or jet structure, and then iteratively to combine the obtained information covering bigger $\eta \times \varphi$ regions up to the point where the whole tau signal or isolation cones are covered. This approach is inspired by similar techniques that are widely used in the modern ML-based image recognition with convolutional DNNs. Considering the high dimensionality of the input space (188 inputs per cell), a pre-processing step with several fully connected dense layers allows us to reduce the dimensionality before processing the signal (isolation) grid with 5 (10) convolutional layers with 3×3 windows each, on each step extracting 64 features from nine alongside cells until the entire grid is convoluted into an array of 64 features. Also, the information from the high-level features is pre-processed by three fully connected dense layers. It is then combined with the convoluted representations of the signal and isolation cones and passed through 5 dense layers. The four outputs, p_i, of the network represent estimates of the probabilities of the reconstructed tau candidate to be τ_e, τ_μ, τ_j, or a genuine τ_h. The overall number of trainable parameters is 1 555 352.

In order to ensure the best performance for a wide tau identification efficiency range, we define a custom loss function based on the focal loss [8] for the training. The loss function is

minimized using the Adam algorithm with the Nesterov momentum [9]. The DNN structure is implemented using the Tensorflow package [10] and the training is run for 10 epochs. The best performance on the validation set is achieved after 7 epochs and the corresponding DNN is chosen as the final discriminator. The discriminator score against each background source is chosen to be of the form $D_\tau^\alpha = p_\tau/(p_\tau + p_\alpha)$, where $\alpha \in \{e, \mu, j\}$.

4. Results

The performance of the algorithm is evaluated using MC simulation and, applying the following preselection on the reconstructed tau candidates: $p_T \in (20, 1000)$ GeV, $|\eta| < 2.3$, $|dz| < 0.2$ cm. The tau ID efficiency is estimated from $H \to \tau\tau$ MC using reconstructed tau candidates that match hadronically decaying taus at the generator level (the simulation step just before modelling of interactions of the particles with the detector). The results in Figure 1 show the DeepTau performance in form of the receiver operating characteristic (ROC) curve on 2017 MC. The jet misidentification probability is estimated from $t\bar{t}$ MC using reconstructed tau candidates that match quarks or gluons at the generator level and do not overlap with generated prompt electrons, muons or products of hadronic tau decays. The probability for an electron (muon) to be misidentification as τ_h is estimated from Drell-Yan MC using reconstructed tau candidates that match to electrons (muons) at the generator level. DeepTau shows consistent improvement at both low and high p_T ranges for all sources of backgrounds.

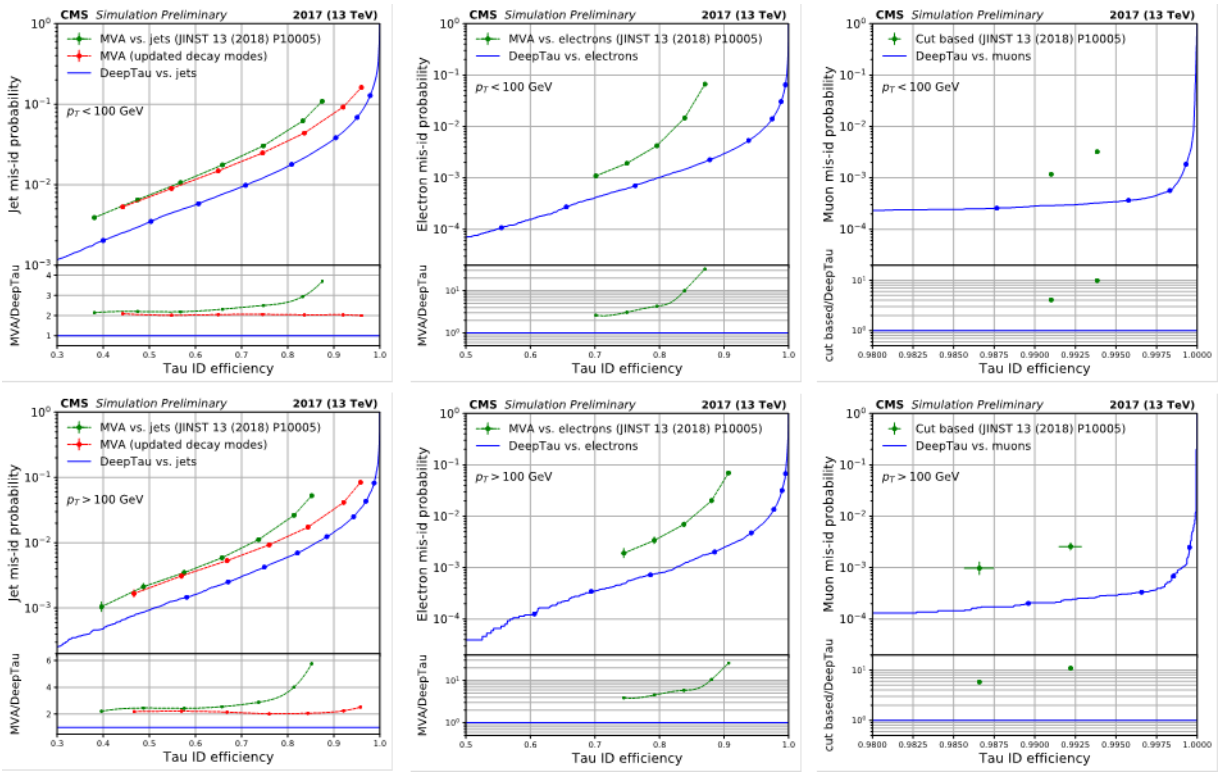

Figure 1. Performance of tau discrimination against quark and gluon induced jets (left), electrons (middle), and muons (right) for DeepTau and the previously available discriminators from [7]. Working points of the discriminators are indicated by the dots. These plots are split by τ p_T ranges

To evaluate the DeepTau performance on data, events with well reconstructed muon and tau candidates are selected. The visible $\mu\tau$ mass is reconstructed as the sum of 4-momenta of the muon

and visible tau decay products. Figure 2 shows a comparison of the distributions of the visible $\mu\tau$ mass for 2018 data between the selection using the previously available discriminators from [7] and the selection using DeepTau. With the DeepTau selection, the yield from genuine τ_h increases by 20%, while the yield from fakes decreases by 23%.

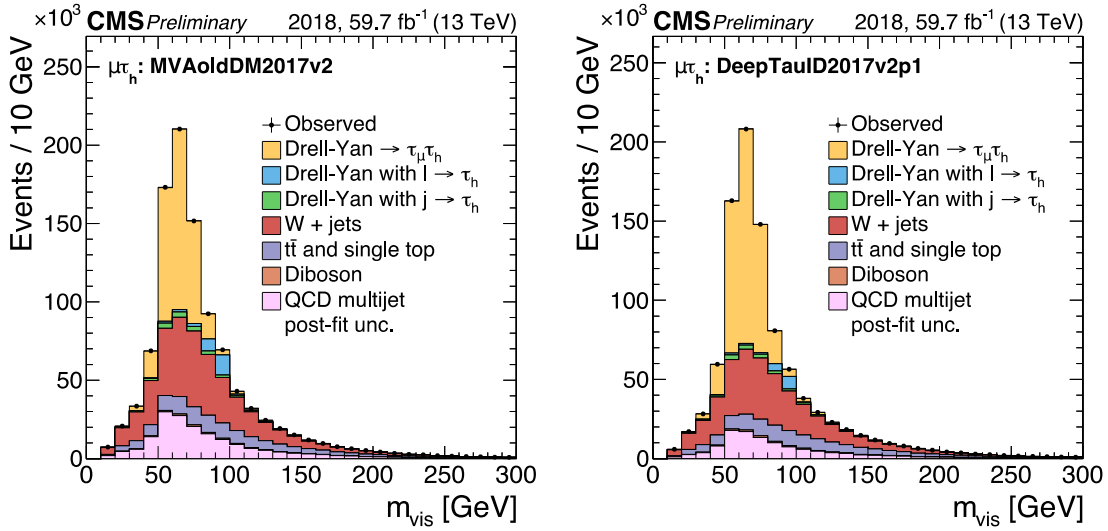

Figure 2. Distributions of the visible $\mu\tau$ mass for 2018 data with the selection using previously available discriminators from [7] (left) and the selection using DeepTau (right)

5. Conclusion

A new ML-based algorithm to discriminate hadronic tau decays against all main sources of backgrounds has been developed. The introduction of DeepTau provides a considerable improvement of the tau identification performance. Compared to the previously used discriminators, for the same efficiency to reconstruct hadronic tau decays, the jet misidentification probability is reduced by more than 50%, and the probability to misidentify an electron (muon) as a τ_h is reduced by up to 95% (90%).

References

[1] Particle Data Group. 2019 Review of Particle Physics // Phys. Rev. D 98 (2018) 030001
[2] CMS Collaboration. The CMS Experiment at the CERN LHC // 2008 JINST 3 S08004
[3] CMS Collaboration. Particle-flow reconstruction and global event description with the CMS detector // 2017 JINST 12 P10003
[4] M. Cacciari, G.P. Salam and G. Soyez. The anti-k_t jet clustering algorithm // JHEP 04 (2008) 063
[5] M. Cacciari, G.P. Salam and G. Soyez. FastJet User Manual // Eur. Phys. J. C 72 (2012) 1896
[6] CMS Collaboration. Reconstruction and identification of τ lepton decays to hadrons and ν_τ at CMS // 2016 JINST 11 P01019
[7] CMS Collaboration. Performance of reconstruction and identification of τ leptons decaying to hadrons and ν_τ in pp collisions at $\sqrt{s} = 13$ TeV // 2018 JINST 13 P10005
[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár. Focal Loss for Dense Object Detection // arXiv:1708.02002
[9] T. Dozat, Incorporating Nesterov Momentum into Adam // CS 229 Machine Learning (2015) 054
[10] M. Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.