PROCEEDINGS OF SCIENCE

# Machine learning techniques for heavy flavour identification

**Barbara Chazin Quero**[1], on behalf of the CMS Collaboration
*Instituto de Física de Cantabria (CSIC-Universidad de Cantabria), Spain*
*E-mail:*barbara.chazin.quero@cern.ch

Reliable and performant heavy flavour identification is of prime importance for the physics program of the CMS experiment. During the last years the CMS collaboration has dedicated a considerable effort to improve and expand its capabilities in this sector by applying several machine learning techniques well established in industry, but still experimental in HEP. These proceedings describe a poster focused on a selection of these techniques and the description of the implementation details as well as the resulting gains.

*Sixth Annual Conference on Large Hadron Collider Physics (LHCP2018)*
*4-9 June 2018*
*Bologna, Italy*

[1] Speaker



PoS(LHCP2018)066

## 1. Introduction

The efficient identification of heavy-flavour jets (b or c tagged jets) is essential both for the study of the standard model processes and searches for new physics. The heavy-flavour jet identification techniques developed by CMS exploit the properties of the hadrons in the jet in order to discriminate between jet originating from b or c quarks (heavy-flavour jets) and those originating from light-flavour quarks or gluons (light-flavour jets). In particular, the heavy hadrons from the hadronization of a b or c quark carry special properties that increase the efficiency of its identification.

## 2. Heavy-flavour jet identification algorithms developed by CMS

The most performant algorithms for heavy-flavour jet identification are based on multivariate combination of the b and c hadrons properties in the jet. Due to the long lifetime of these particles, the decays result in displaced tracks with large impact parameter and secondary vertex; their large mass and hard fragmentation allow for decay products with larger transverse momentum relative to the jet axis than the other jet constituents; and the possible semi-leptonic decays enable the presence of soft muons or electrons in the jet. The new algorithms developed during the Run2 combine a large number of variables related to these features using different machine learning techniques.

### 2.1. Algorithm developments in Run 2 for b-jet identification

Based on the Combined Secondary Vertex (CSV) algorithm widely used during Run 1, a new release has been developed during Run 2. The new version of this algorithm, the CSVv2 [1], also involves the use of secondary vertex and track-based lifetime information but combining all the variables with a group of artificial neural networks (ANNs) instead of using a likelihood ratio as discriminator value. Each ANN is a Feed-fordward multilayer perceptron (n:1:1) trained against c and light jets giving one discriminant value per training. The input variables are classified in three independent categories as a function of the secondary vertex information. The outputs values are the discriminator values of each vertex category which, combined with a likelihood ratio, end up in only one discriminator per training. The final discriminator value is a linear (1:3) combination of the two training discriminators.

A little more complicated is the DeepCSV tagger that inherits the same observables used by CSVv2 algorithm but includes more charged tracks (up to six are used, instead of just two) in the track-based variables. It is based on a deep-feed neural network (DNN) with 4 hidden layers of 100 nodes each and an output layer of 5 nodes corresponding to 5 jets categories used in the training. The nodes of the hidden layers use a rectified linear unit as activation function while the nodes of the final layer use a normalized exponential function to be able to interpret the output value as a probability of a given jet flavour category.

A different approach than the machine learning techniques for the heavy flavour identification is adopted in the combined Multivariate Algorithm version 2, cMVAv2, based on the same cMVA algorithm used in Run 1, which uses as input 6 b-jet discriminators outputs of CSVv2 and other simple taggers in a gradient boosting classifier (GBC) as a Boosted Tree Decision (BDT).

The performance of these b-jet tagging algorithms are shown in Figure 1 [1]. The DeepCSV discrimination against c and light jets outperforms all other algorithms for b-tagging efficiencies below 70%, while the cMVAv2 tagger performs better against light jets for b-tagging efficiencies above 70%. Both taggers improve the CSVv2 performance by ~4% for a mistag rate for light jets of 1% .

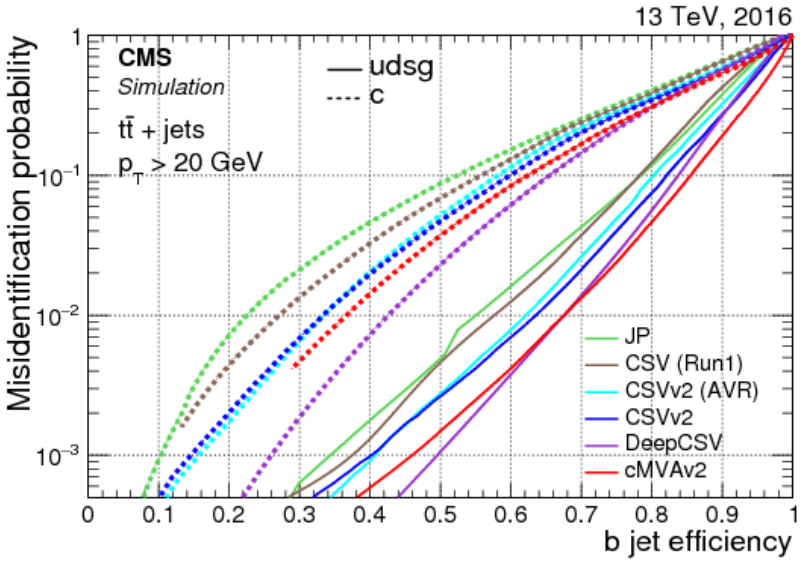


*Figure 1:* Performance of b-jet tagging algorithms in 2016 demonstrating the probability for non-b jets to be misidentified as b jet, as a function of the efficiency to correctly identify b jets.

After the CMS Phase 1 upgrade the new pixel detector includes now an additional layer closer to the beam spot [2]. A comparison of the DeepCSV performance with the 2016 detector, Phase 1 detector and 2016 training, and with Phase 1 detector and new dedicated trainning is done in Figure 2 [3].

Finally, the DeepFlavour tagger with a more complex architecture has been recently included in the list of b-jet identification algorithms of Run 2. This algorithm is based on a deep neural network using 16 properties of up to 25 charged and 8 properties of up to 4 neutral particle-flow jet constituents, as well as 17 properties of up to 4 secondary vertices associated with the jet. For each category of particles and vertices separate hidden convolutional layers are trained. Each layer has several filters that act on each particle or vertex individually. All the outputs and a set of global jet properties work as input of a dense layer of 350 nodes using a rectified linear unit as its activation function. Then 7 hidden layers with 100 nodes each using the former activation function and the last layer using a normalized exponential function produce the input of the last layer with 4 nodes corresponding to the 4 flavour jet categories

used in the training. The output value is a probability of given jet flavour category p(f). Figure 3 show a schematic illustration of this DNN. The DeepFlavour tagger gives a 4% absolute improvement in b-tag efficiency for a mistag rate for light jets of 0.1% against DeepCSV [3].

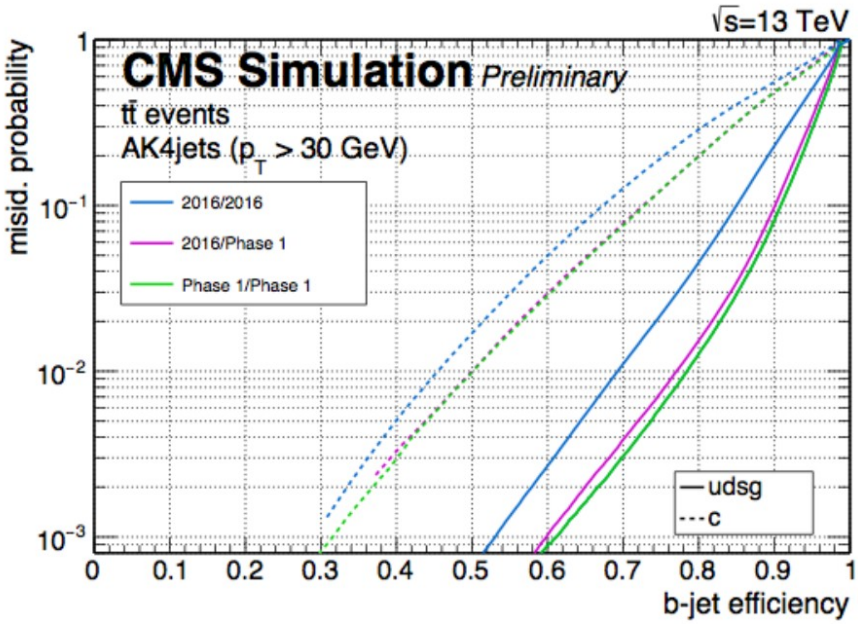


*Figure 2:* DeepCSV b-tagging algorithm performances in 2017 demonstrating the probability for non-b jets to be misidentified as b jet, as a function of the efficiency to correctly identify b jets. The performance of the algorithm is shown for the following scenarios: the training before the Phase 1 upgrade with a simulation of the detector before the upgrade (2016/2016), the training from before the upgrade with a simulation of the upgraded detector (2016/Phase 1), and a re-trained algorithm with the a simulation of upgraded detector (Phase 1/Phase 1)

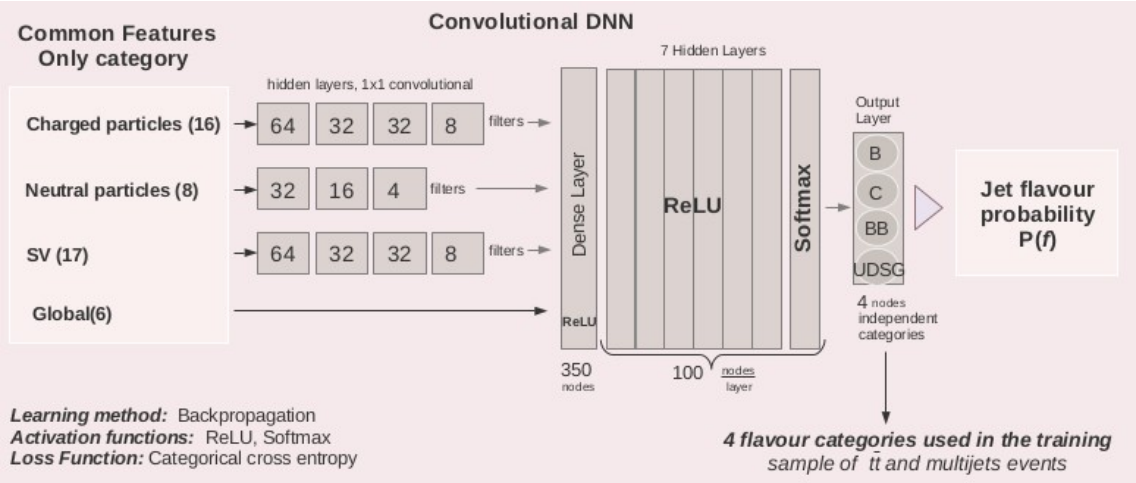


*Figure 3:* A deep-neural-network algorithm is based on different properties of charged (≤ 25) and neutral (≤ 4) particle jet constituents, SV (≤ 4) and global jet variables. Squematic illustration.

### 2.2 Algorithm developments in Run 2 for c-jet identification

The algorithm for c jet identification is based on similar input variabales and categories that entered in CSVv2, with the addition of some soft lepton information (up to two tracks or

leptons). It is based on a GBC used for two trainings to dicriminate c-jets against light (CvsL) and b (CvsB) jets. The c jet tagging can be also done with other algorithms. Figure 4 [1] shows the performance of different c-tagging algoritms. It can be seen that DeepCSV is already outperforming the dedicated c-tagger.

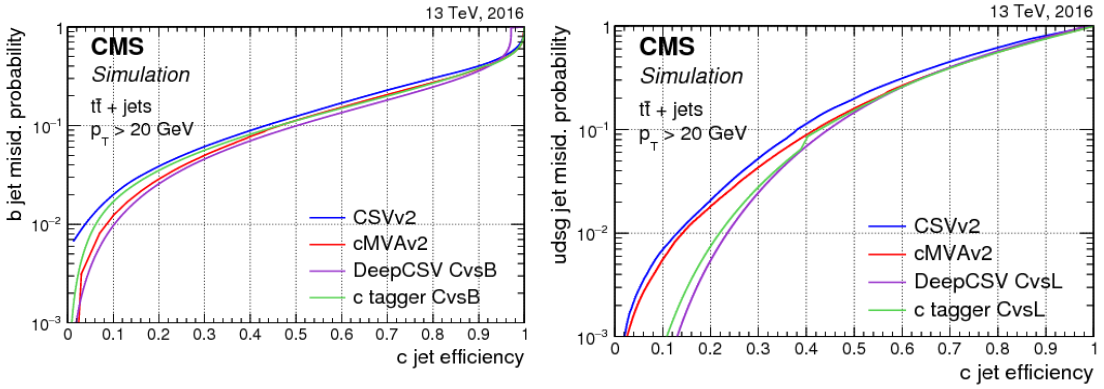


*Figure 4:* Performance of c-jet tagging algorithms in 2016 demonstrating the probability for b (light) jets to be misidentified as c jet on the left(right), as a function of the efficiency to correctly identify b jets.

## 3. Summary

During the last years the CMS Collaboration has developed a variety of algorithms based on machine learning techniques for heavy-flavour (bottom or charm) jet identification. A detailed description of the implementation of these techniques and the studies of their performance on simulations of different final states with heavy and light-flavour quarks has been presented in this note.

The most efficient algorithms to correctly identify b jets in tt and multijets events in CMS are DeepCSV and cMVAv2, which improve the b-tag efficiency by about the 4% with respect to the CSVv2 tagger for a misidentification probability for light-flavour jets of 1%. The DeepCSV algorithm is also outperforming the dedicated c-jet algorithm identification. The DeepFlavour tagger, with an absolute improvement of about 4% in the b-tag efficiency for a light jets mistag rate of 0.1% compared to DeepCSV, is the next step of deep learning based heavy flavour identification algorithms.

## References

[1] CMS Collaboration, *Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV*, JINST 13 (2018) P05011 [arXiv:1712.07158], doi: 10.1088/1748-0221/13/05/P05011.

[2] CMS Collaboration, *CMS Technical Design Report for the Pixel Detector Upgrade*, CERN-LHCC-2012-016, CMS-TDR-11, doi: 10.2172/1151650.

[3] CMS Collaboration, *CMS Phase 1 heavy flavour identification performance and developments*, CERN-CMS-DP-2017-013, https://cds.cern.ch/record/2263802.