Low Momentum Taus.

BDT Analysis for reconstuction of low transverse momentum tau leptons

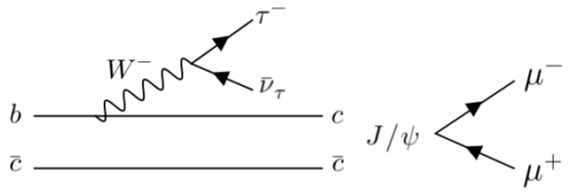
Agni Keyoor Purani[1], Elisabetta Gallo[2,3], Andrea Cardini[3]

Low transverse momentum taus?

Low pT taus can act as a probe for Lepton Flavour Violation (LFV). The only way for the b quarks to directly couple with tau leptons is through the exchange of a W boson. So, according to the Standard Model (SM), the process must be lepton flavour independent.

The Hadron Plus Strip (HPS) algorithm for tau reconstruction in the CMS experiment targets high pT taus and is inefficient at low pT. Hence, there is a need for a dedicated tau reconstruction algorithm.

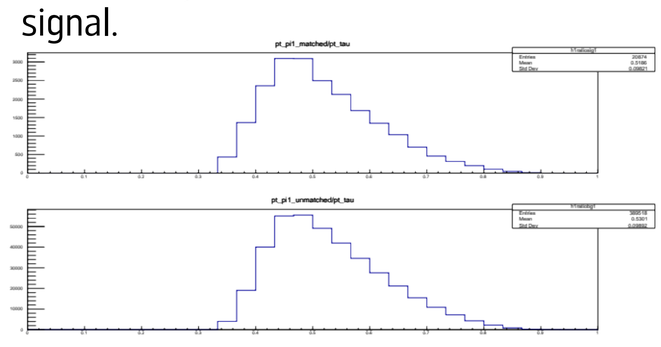
$$\frac{\mathrm{BR}(B_C \to \tau\nu J/\psi)}{\mathrm{BR}(B_C \to \mu\nu J/\psi)} \approx 1$$

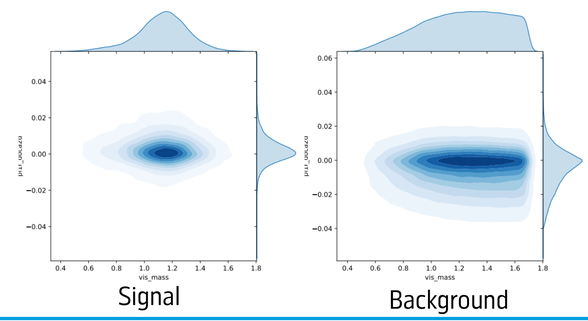


Why need machine learning?

Reconstruction of taus involves the identification of the signal pions (3 prongs) from the background. The background is combinatorial by construction and is similar to the signal. No single variable provides sufficient discrimination power.

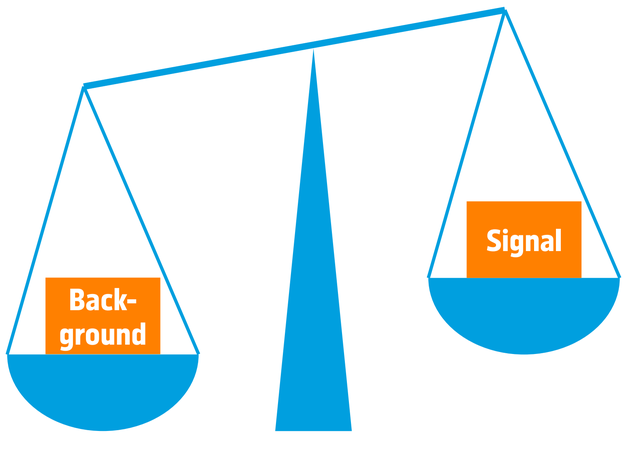
This is why multivariate methods like BDT are implemented to better identify the signal.





Signal Background

Data Imbalance!



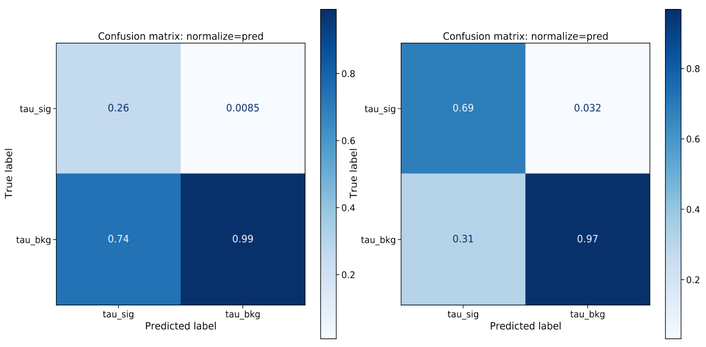
Since the data is a collection of charged pion triplets and only one of the combinations is the correct one, the background will naturally be larger than the signal (20 times in the training dataset)!

The background consists of triplets where up to 2 pions can be originated from a tau decay. Artificially increasing or decreasing the number of samples for a given class cannot solve this imbalance. Therefore, class weights are used to focus the training on a particular class to improve its identification efficiency.

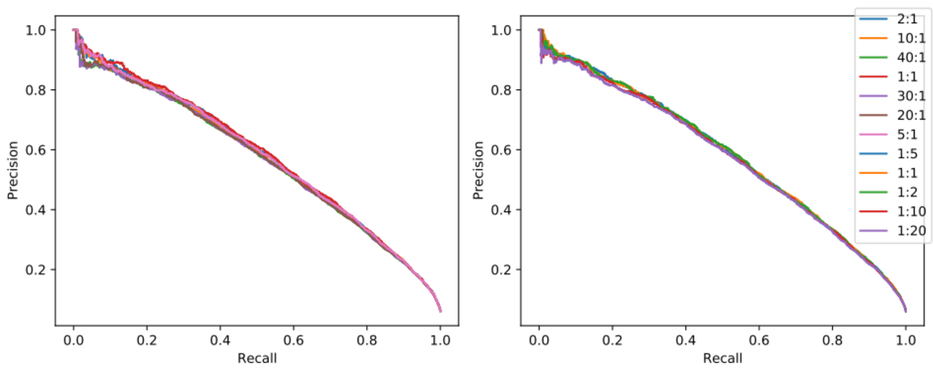
The metrics to evaluate the BDT performance can be affected by the class imbalance. The Precision Recall (PR) curve is one of the metrics that is more sensitive to the data imbalance.

Class Imbalance Weights

Class weights created some difference in the results for the same threshold (below). As we moved to a more balanced case, we find that the efficiency of the model increases as expected but at the price of purity.



With Weights Without Weights

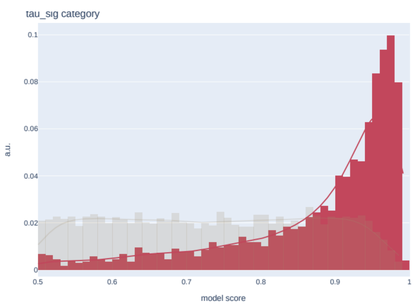


PR curves with signal (background) havig higher weight on the left(right). Both the plots contain the model with no weights.

The background triplets dominate the signal region for most models. This leads to a lower purity which is not desirable.

Decision Threshold

The PR curves show it is possible to achieve similar results with certain combinations of class weights and thresholds but not both efficiency and purity as the curves are different. Choosing the right threshold is important to get the best performance.



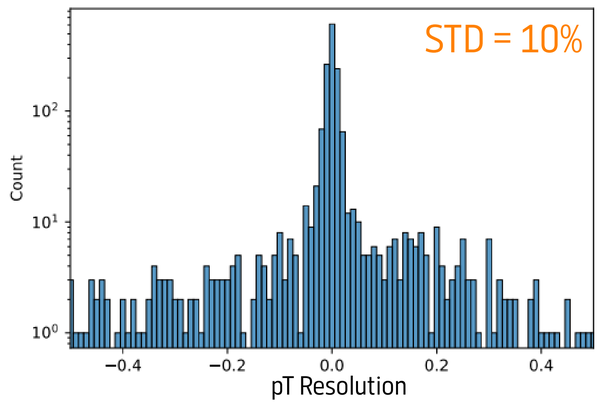
By setting a higher threshold here we can get more purity at the price of a lower efficiency.

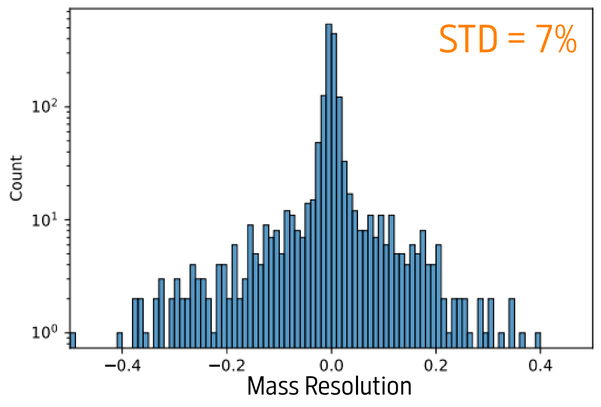
Conclusion

Based on the purity and efficiency of the signal compared across thresholds and class balance weights and the AUC score of PR curves, a 1:2 signal to background ratio with 0.45 as the signal threshold seems to perform the best.

The purity is 78±5% and the efficiency is 31±2%.

The results from the test data were compared with generated pions in the plots on the right and show how well tau properties can be reconstructed.





DAAD
Deutscher Akademischer Austauschdienst
German Academic Exchange Service

[1]Indian Institute of Technology Kharagpur
[2]University of Hamburg
[3]Deutsches Elektronen-Synchrotron DESY