Investigation of machine learning techniques to uncover the Higgs CP nature in CMS

DESY Summer Student Programme, 2021

Sahar Abdelhay Ali FARRAG

Assiut University

Supervisors:

Andrea Cardini
Oleg Filatov

September 9, 2021

Abstract

CP-violation in the Higgs sector remains a possible source of the baryon-asymmetry in the universe. The aim of the project is to uncover the Higgs CP nature in CMS using machine learning techniques and compare our boosted decision tree (BDT) model with the previously used neural network (NN) one. I used data and simulation of proton-proton collision at the LHC during 2018 (HTT data). After data pre-processing and training the model with the features used in the published analysis, I tuned some hyperparameters to get the best model, and then I added extra input features to improve the model. The expected significance for the $H \rightarrow \tau\tau$ process shows that the BDT performs noticeably better than NN even when using the same input features.

Contents

1. Introduction

The properties under CP symmetry of the Higgs boson are an important test for the Standard Model (SM) of particle physics. The SM predicts the existence of one Higgs boson, and its coupling to fermions and vector bosons are expected to be invariant under CP symmetry. The presence of CP violation in the Higgs sector, if found, would provide a strong indication for physical phenomena not predicted by the SM and be used to constrain theories beyond the SM. The main goal of the project is to uncover the Higgs CP nature in CMS using machine learning (ML) techniques and compare our BDT model with the previously used NN one.

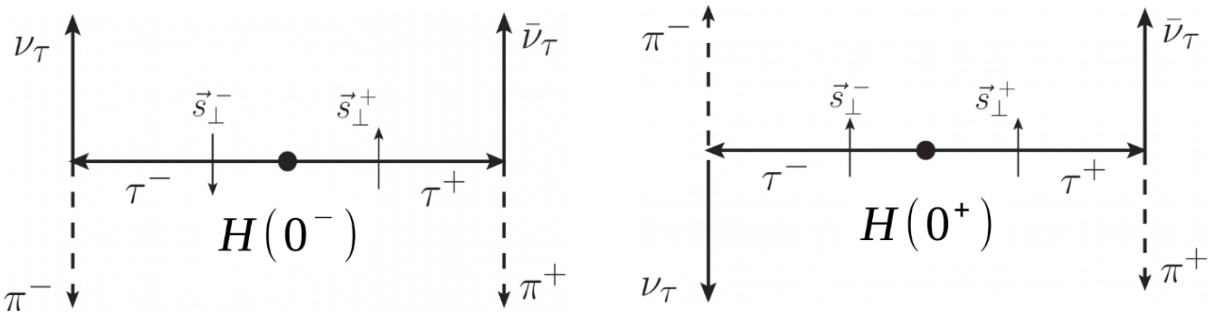

Fig. 1: Left (Right): schematic depiction of a CP-odd (-even) H$\rightarrow\tau\tau$ decay

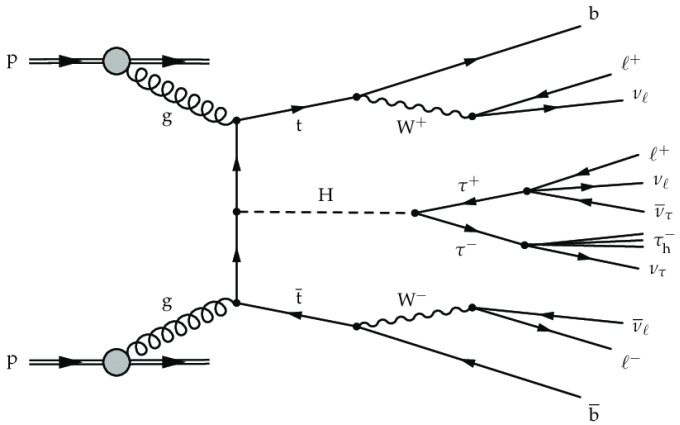

Fig. 2: An example of a Feynman diagram for ttH production, with subsequent decay of the Higgs boson to a pair of τ leptons

2. Methodology

2.1. Input data

This project used hard scattering events recorded during proton-proton collisions at the LHC by the CMS experiment. I used recorded and simulated events relative to 2018 detector conditions (HTT data). The targeted decay of Higgs is the decay of Higgs into two tau leptons, one tau decaying into a muon and neutrinos and the other decaying into hadrons. I defined three categories, each one collecting physical processes with common features.

Initially, input features were investigated using three notable samples, each corresponding to a different category of physical processes:

1. ***VBF*** is a Higgs production mechanism
2. ***DYJets*** and ***WJets*** are the two major backgrounds

2.2. Input features

I initially studied the same input features used in the published analysis. Then, I look at distributions of physical quantities to investigate their potential to separate signal and backgrounds. I looked at quantities such as kinematics of reconstructed leptons and jets like transverse momentum (p_T) and pseudorapidity (η), and more complex variables such as the invariant mass of lepton and jet systems. Fig. 3 shows clearly that di-tau mass is very interesting because DYJets and VBF peaks are at different values while Wjets has a smoother distribution.

Fig. 4: shows other variables which were tested for their potential of differentiating between signal and backgrounds. Physical processes can also differ based on how their respective features are correlated with each other. This can be investigated by looking at 2-D distributions as shown in Fig. (5 and 6). Based on the investigated plots, I decided which features I will add into the BDT model afterwards.

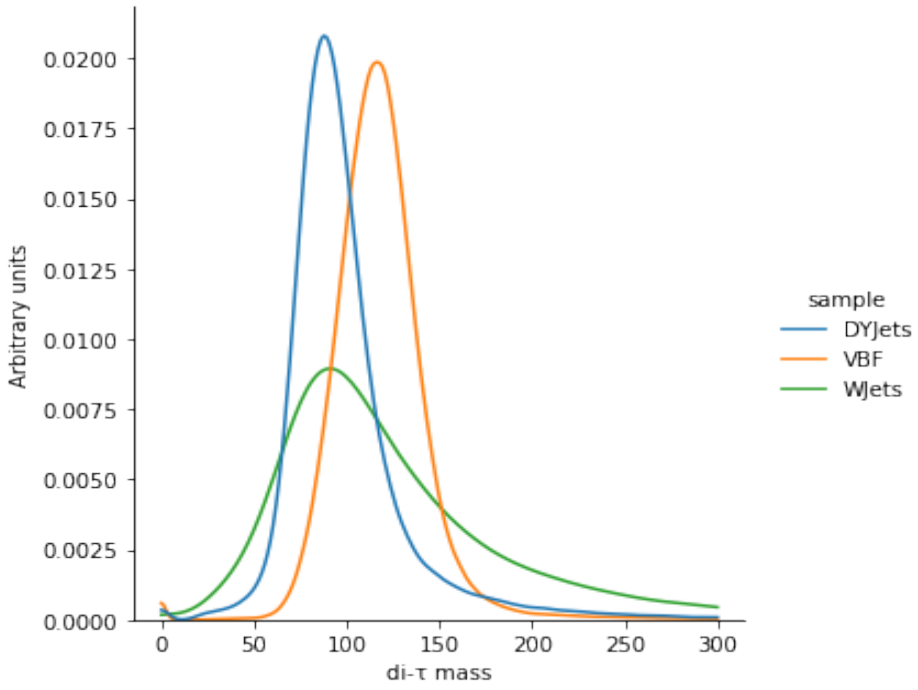

Fig. 3: Overlayed 1-D distribution for di-tau mass for the three different categories

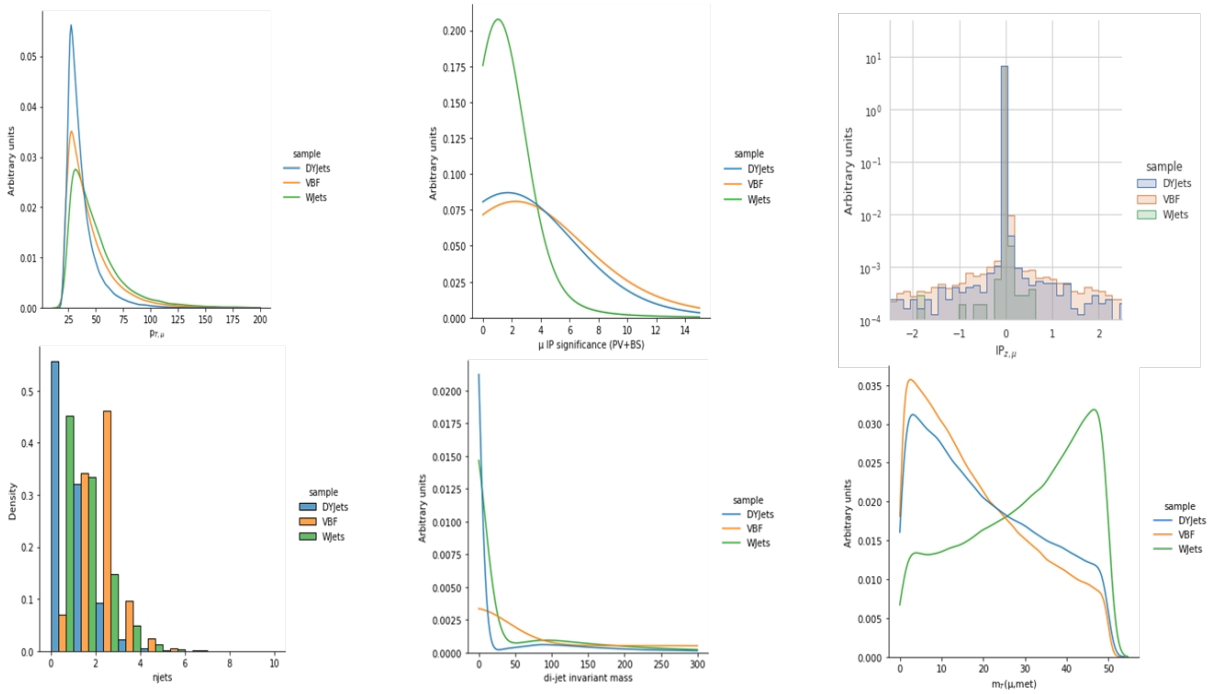

Fig. 4: Top row: transverse momenta (left), impact parameter (IP) significance (center) and longitudinal IP (right) of the muon. Bottom row: number of jets (left), invariant mass of the two leading jets (center), and transverse mass of the muon and MET (right)

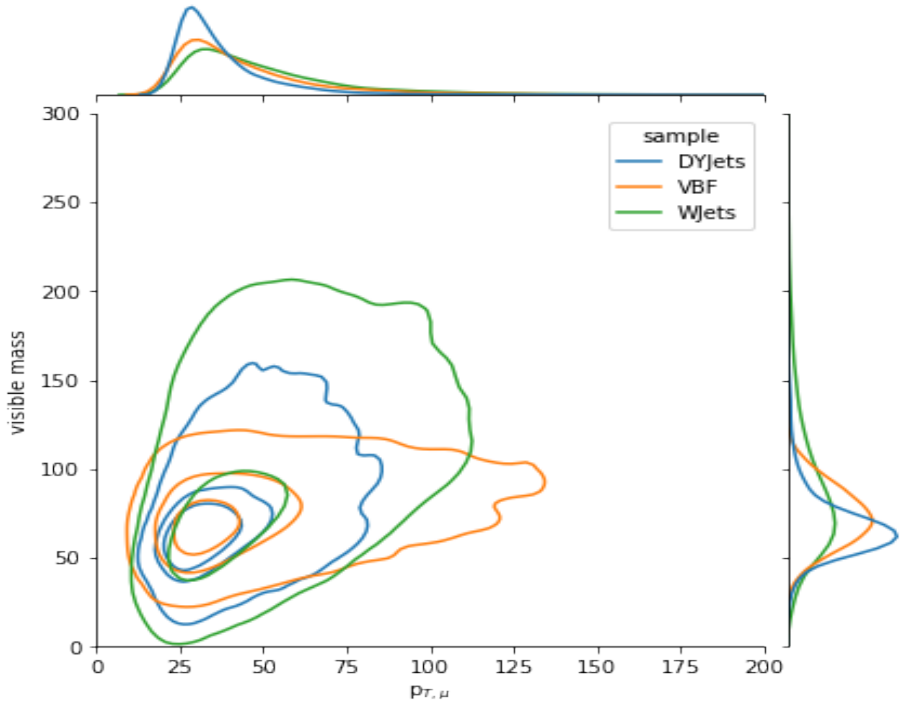

Fig. 5: overlayed 2-D distribution for transverse momentum of the muon and invariant mass of the visible decay products of the two tau leptons for the three different categories

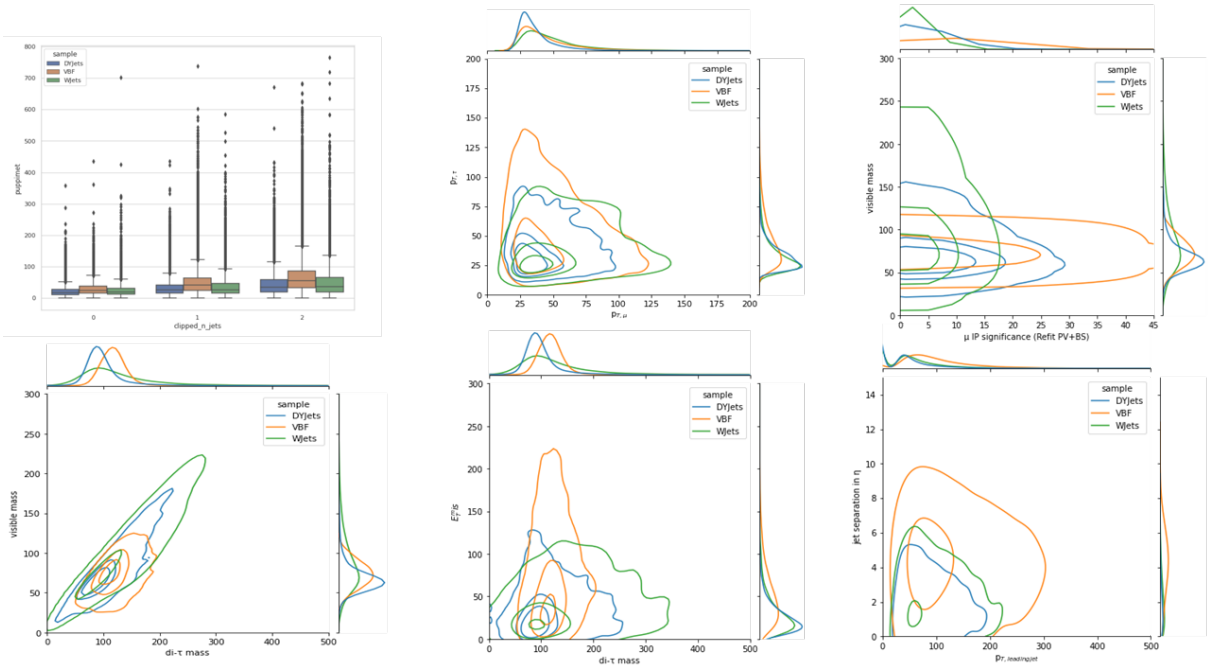

Fig. 6: more overlayed 2-D distributions for various features

2.3. Architecture

2.3.1. Boosted decision trees (LightGBM)

In our model, I used a Light Gradient Boosting Machine (lightGBM) which is a gradient boosting framework that uses tree based learning algorithm. Below a diagram explains the implementation of LightGBM and how it works.

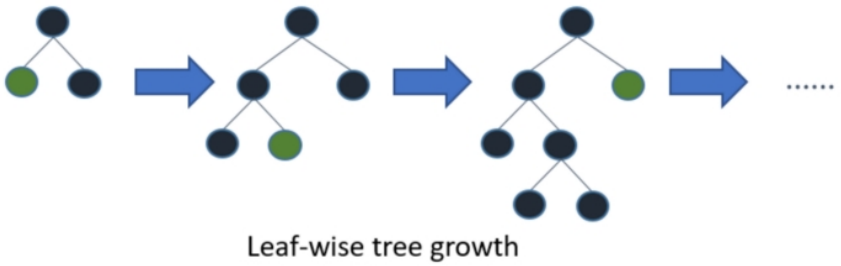

Fig. 7: At each step of the tree iteration, two new leaves are associated to a random existing one which is at the end of an open branch. The process is repeated till a maximum lenght of the branch, i.e. the max_depth parameter of the architecture, is reached.

2.3.2. Hyperparameter optimization

Hyperparameter optimization is consists in choosing a set of parameters for the BDT architecture in order to maximize its performance. In our model, I tuned several

hybeparameters as shown in the following diagram to get the best BDT model. The performance of the BDT was evaluated by looking at the loss function for the validation dataset. Several values were tested for each hyperparameter, the one yielding the best results was kept when optimizing other parameters in the sequence shown in Fig.8.

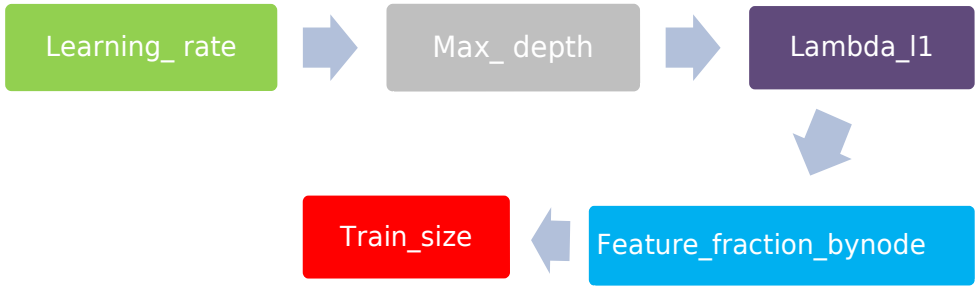

Fig. 8: Hyperparameter optimization's sequence

3. Results

3.1. Feature importance

The two lists below show the input features used initially in our model as published analysis and the feature importance based on the BDT model after adding extra input features.

Observable
p_T of leading τ_h
p_T of τ_μ
p_T of visible di-τ
Visible di-τ mass
di-τ mass
Leading jet p_T
Trailing jet p_T
Jet multiplicity
Dijet invariant mass
Dijet p_T
Dijet $\lvert\Delta\eta\rvert$
p_T^{mis}
$m_T(\tau_\mu, p_T^{mis})$

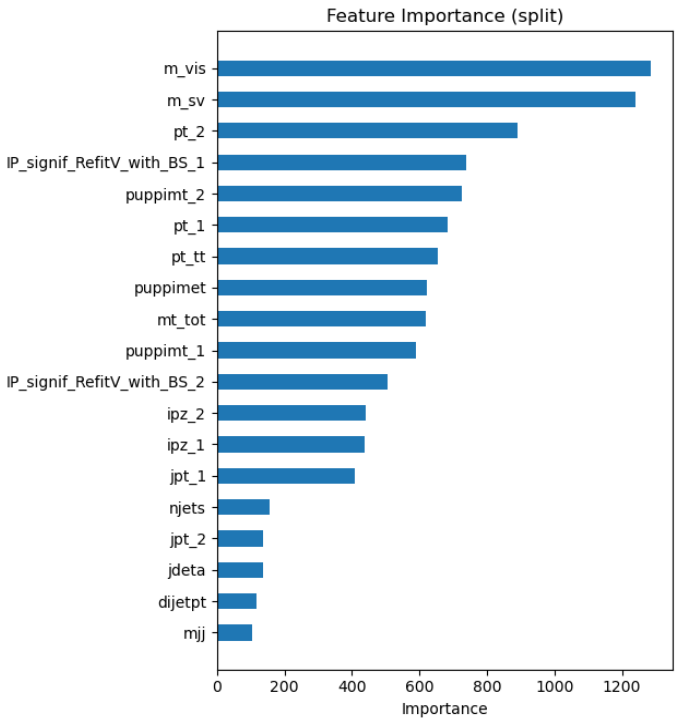

Fig. 9: Left: observable used for the training of the NN used in [Ref. 1] and for the hyperparameter optimization. Right: feature ranking after adding additional features to the BDT training

The IP significance of the muon and the transverse mass of the hadronic tau and MET are ranked 4th and 5th respectively, meaning that their addition to the BDT was considered useful.

3.2. Confusion matrices for the best model

Fig.10: shows the confusion matrices for the best model normalized by rows (efficiency matrix) or by columns (purity matrix). The diagonal elements are all above 0.6 showing that the BDT achieves a relatively good efficiency and purity for all categories.

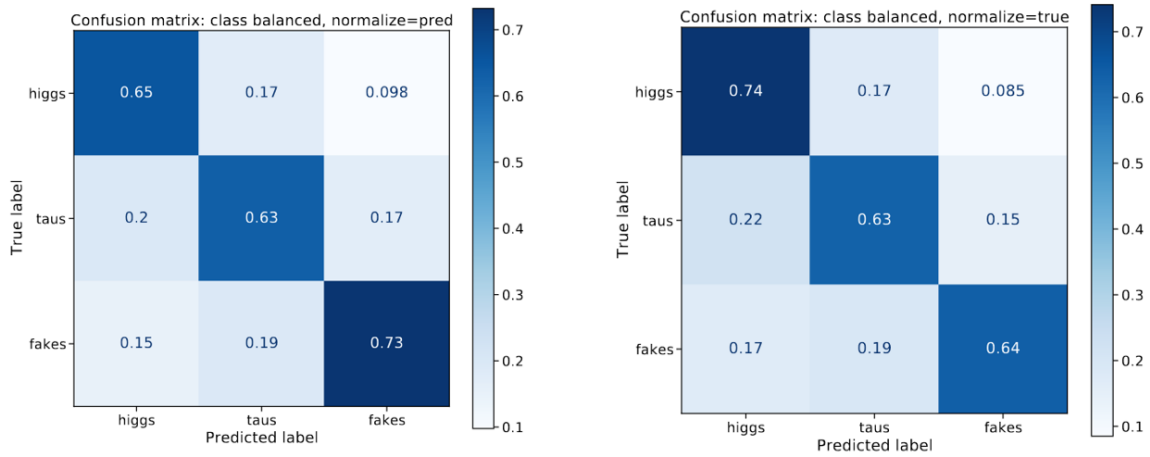

Fig. 10: The confusion matrices for the best model

3.3. Comparison with NN: $H \rightarrow \tau\tau \rightarrow \tau_\mu \tau_h$ signal strength

Results relative to the measurement of the Higgs CP properties have been obtained as a parametric fit of simualted signal to an Asimov dataset. Figure 11 shows how the expected limits for the inclusive Higgs production signal strength have improved by ~40% with the new ML algorithm.

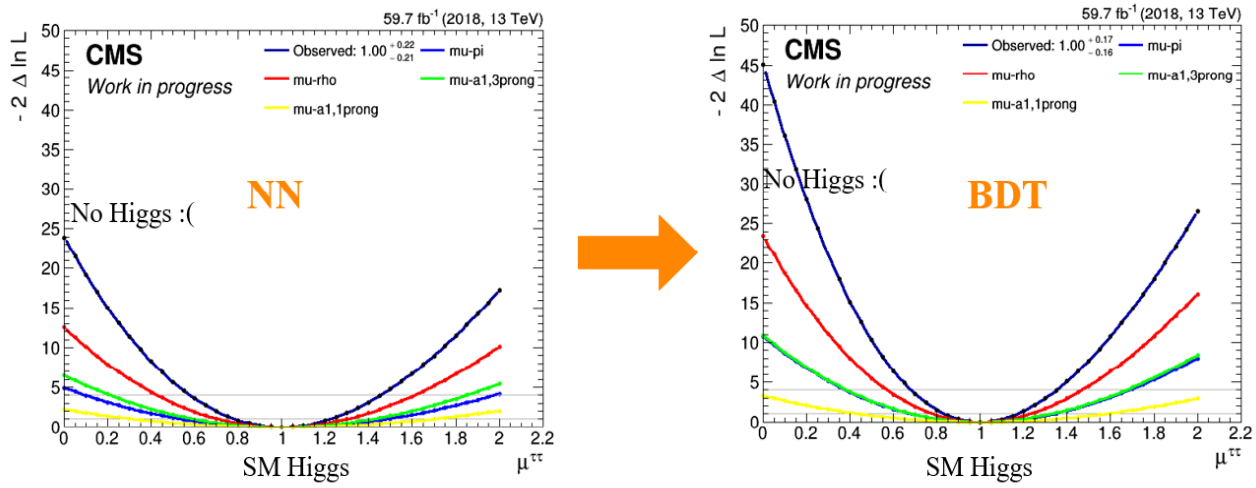

Fig. 11: Comparison with NN: $H \rightarrow \tau\tau \rightarrow \tau_\mu \tau_h$ signal strength

3.4. CP-even vs CP-odd Higgs

As shown in Fig. 12, the significance for the exclusion of a pure CP-odd Higgs increases by ~37% with the new BDT model.

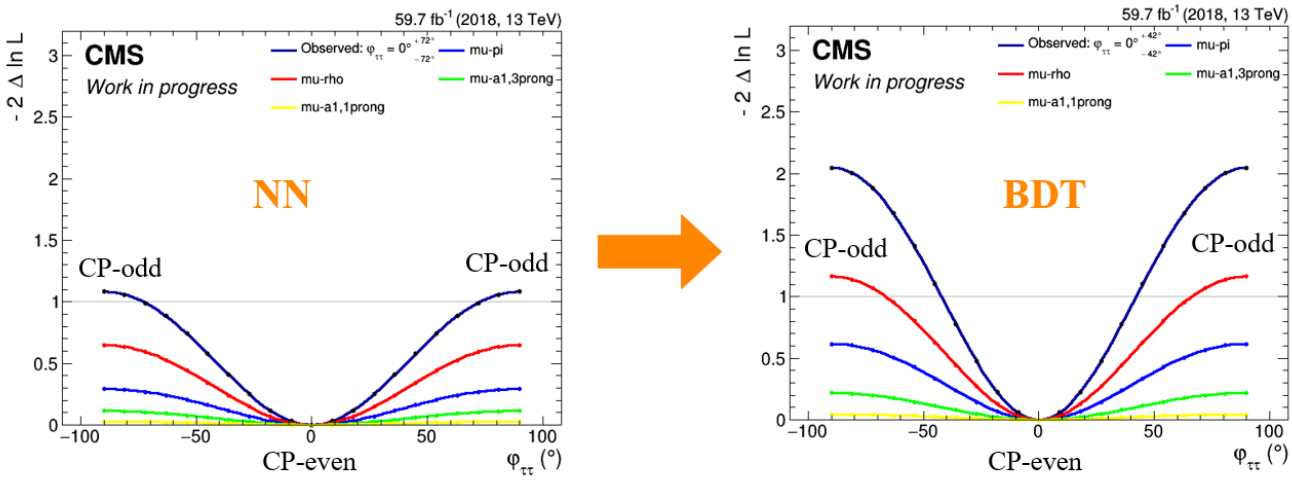

Fig. 12: CP-even vs CP-odd Higgs

Conclusions

Table 1: shows a comparison between previously used NN (Old NN), the BDT obtained by simply optimizing the hyperparameters (BDT-best architecture) and the one which inclused extra input features (BTD-best architecture + extra features). The comparison is made in terms of the significance for the identification of the $H \rightarrow \tau\tau \rightarrow \tau_\mu \tau_h$ process and for the exclusion of a pure CP-odd Higgs hypothesis.

Model	Significance H→ττ	CP-odd exclusion
Old NN	4.9σ	1.04σ
BDT- best architecture	6.0σ	1.21σ
BDT- best architecture + Extra input features	6.9σ	1.43σ

Through the project, it's been proven that $H \rightarrow \tau\tau$ identification has improved a lot and this means that BDT performs noticeably better than NN even when using the same features. The performance was evaluated using the expected significance for the $H \rightarrow \tau\tau$ process and

for the exclusion of a pure CP-odd hypothesis estimated on an Asimov dataset. The noticeable increase in significance brought by adding additional input feature and by optimizing the hyperparameters indicates that there is still room for improvement. This will allow a more precise measurement of the Higgs CP nature and constrain additional sources of baryon-asymmetry in the Universe.

References

1. *Cardini, A. (2021). Measurement of the CP properties of the Higgs boson in its decays to τ leptons with the CMS experiment [Doctoral Dissertation, Universität Hamburg]. Deutsches Elektronen-Synchrotron DESY. https://bib-pubdb1.desy.de/record/462769*