Vector Boson Fusion Tagging with Graph Neural Networks

CERN Summer Student Project Report

Jetë Selmani

Supervisors : Andrea Malara, Santeri Laurila

July - August 2024

Abstract

Despite its relatively low production cross-section, the vector boson fusion (VBF) production mode stands out as an exceptionally promising channel for investigating the properties of the Higgs boson. Using data from CMS, the aim of the project was to study the performances of the traditional discrimination method and a neural network designed to distinguish between gluon-gluon fusion (ggF) and VBF production modes. The training of ParticleNet graph neural network algorithm using Monte Carlo simulated events produced promising results in distinguishing VBF from ggF production modes, showing that this new method is superior to the old one.

1 Introduction

Since the Higgs boson was discovered, it has become a powerful tool for exploring the manifestations of the SM and probing physics landscapes beyond it. The dominant production mechanisms for Higgs boson at the LHC are gluon-gluon fusion (ggF), vector boson fusion (VBF), production in association with a massive vector boson (VH), and associated production with a pair of $t\bar{t}$ quarks ($t\bar{t}H$) or with a single top quark (tHq) [1]. The Feynman diagrams of the first two Higgs boson production processes are shown on Figure 1.1.

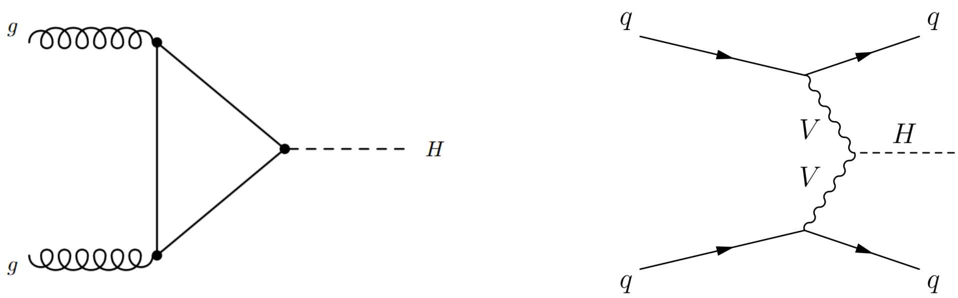


Figure 1.1. Feynman diagrams for ggF (left) and VBF (right) [2, 3].

The diagram for ggF shows two gluons that, through a loop of quarks, produce a Higgs boson. While all quarks theoretically contribute to the loop, in practice, including only the top quark is sufficient. This is because the Higgs boson couples approximately 35 times more strongly to the top quark than to the next heaviest fermion, the bottom quark, resulting in the bottom quark's contribution being suppressed by a factor of 35^2 [4].

In the VBF diagram, each initial-state quark radiates off a heavy vector boson (Z or W), which fuse to create a Higgs boson. The quarks that emit the vector boson travel roughly along their initial directions, resulting in two forward jets. These two jets are widely separated from each other, resulting in a large angle η. The properties of the reconstructed jet can reveal the type of particle that initiated it, which is crucial for distinguishing between different physics processes and improving the sensitivity of both searches for new particles and measurements of Standard Model processes. The ggF is the most dominant process at the LHC, i.e. it has the largest cross section for Higgs boson production. Figure 1.2 (left) presents the cross section for various Higgs boson production mechanisms as a function of the center-of-mass energy, while

Figure 1.2 (right) illustrates the branching ratios for the main decays of the Higgs boson.

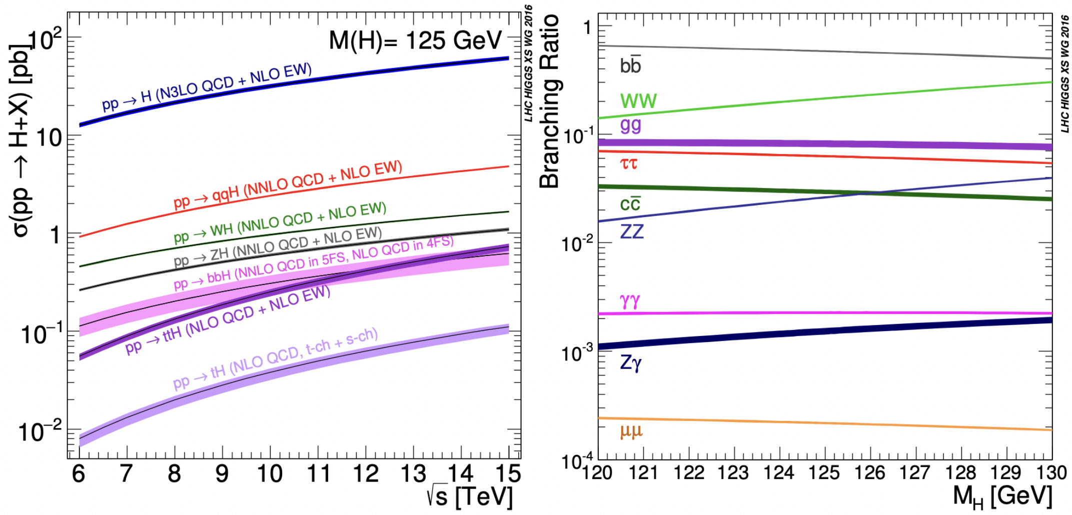


Figure 1.2. (Left) The cross section as a function of the center-of-mass energy for *pp* collisions. The ggF process is indicated as *H*, whereas the VBF process as *qqH*. (Right) The branching ratios of the Higgs boson decay [1].

The Higgs boson decays most frequently into a bottom quark-antiquark ($b\bar{b}$) pair with a branching ratio of about 58%. Despite that, this decay mode is challenging to explore experimentally. In the dominant gluon-gluon fusion production mode, the $H \to b\bar{b}$ signal is overwhelmed by backgrounds produced through quantum chromodynamics multijet events [5].

So far, at the LHC all main production modes and most of the decay channels (bb, WW, $\tau\tau$, ZZ, $\gamma\gamma$) are discovered [6]. One of the more rare channels is the decay of the Higgs boson into invisible particles. Within the SM, the only invisible decay of the Higgs boson is $H \to ZZ \to 4\nu$ with branching ratio around 0.1%. In several theoretical models, the 125 GeV Higgs boson is proposed to serve as a portal between a dark sector and the Standard Model. As a result, the Higgs boson could potentially decay into a pair of dark matter (DM) particles, which would not interact with the detector material. Consequently, this would contribute to the Higgs boson's invisible decay. Therefore, directly searching for these invisible decays of the Higgs boson is a key method for probing dark matter production.

Observed and expected upper limits on $(\sigma/\sigma_{\text{SM}}) \times \mathcal{B}(H \to \text{inv})$ at 95% C.L. for 2012-2018 data are presented in Figure 1.3. The combination of data collected in 2012, 2015, 2016, 2017 and 2018 sets an observed (expected) upper limit on the invisible decay branching ratio of the Higgs boson, $\mathcal{B}(H \to \text{inv})$ at less than 18% (10%) with 95% confidence. This is the most precise constraint on $\mathcal{B}(H \to \text{inv})$ achieved so far [7].

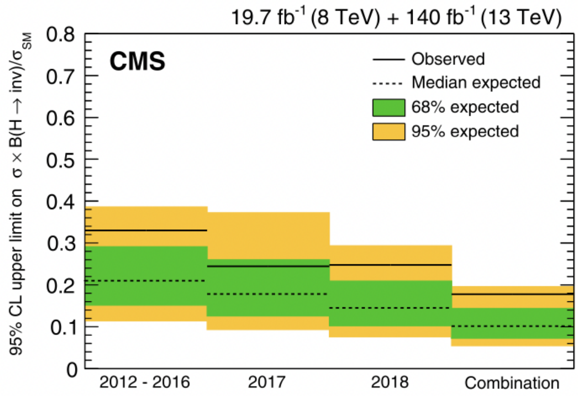


Figure 1.3. Observed and expected 95% C.L upper limits on $(\sigma/\sigma_{\text{SM}}) \times \mathscr{B}(H \to \text{inv})$ across all data-taking years, including their combination, assuming a 125.38 GeV SM Higgs boson [7].

This summer project is part of the effort of developing a neural network capable of distinguishing between ggF from VBF production modes. The traditional discrimination method that uses high-level kinematic variables, and a new approach using a graph neural network trained with low-level variables, are described, and the discrimination power of these different methods between ggF and VBF production modes is compared using simulated event samples.

2 Analysis methods

The simulated samples used in this analysis include the gluon-gluon fusion and vector boson fusion production modes. The analysis is done on the Higgs to invisible decay mode. The simulated samples used for performance comparisons are based on LHC Run 3 data, at a center-of-mass energy of $\sqrt{s} = 13.6$ TeV, while the training of the ParticleNet [8] was done using LHC Run 2 simulated samples at a center-of-mass energy of $\sqrt{s} = 13$ TeV. This approach is reasonable because the detectors are quite similar in both runs.

Two classification methods for discriminating ggF from VBF are compared. The first method is with high-level variables and the second one with graph neural networks.

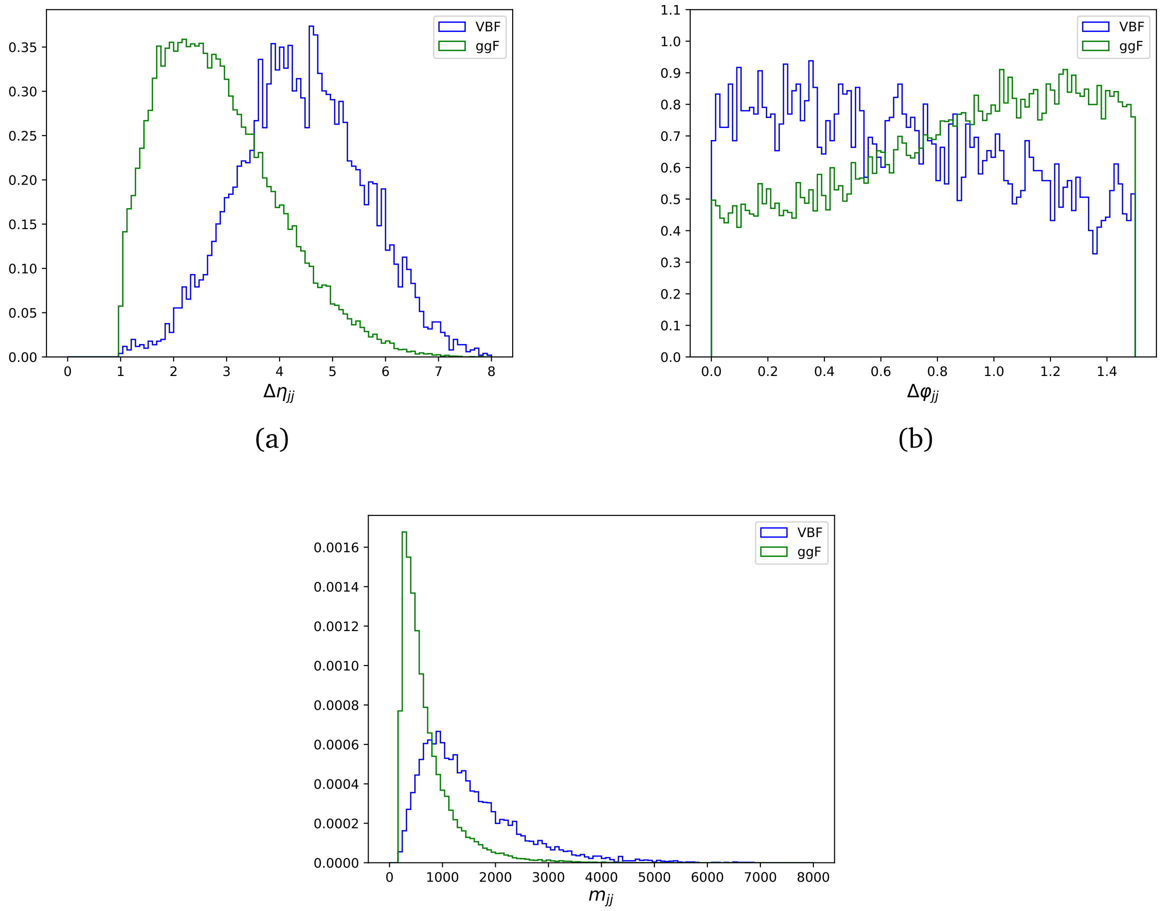
The performance of the classifiers is illustrated using ROC curves. Key components of these curves are the True Positive Rate TPR, and the False Positive Rate FPR. The TPR is defined as the ratio of events that were classified as signal to all events that should have been classified as signal. The FPR is the ratio of background events incorrectly classified as signal. When choosing a threshold in the plot, everything on the right of the threshold is classified as a signal event, and everything to the left as a background event. Altering the threshold, the FPR and TPR values change. Moving the threshold in very small steps, an array of values for FPR and TPR will be received. Plotting the FPR versus TPR on a graph, generates the ROC curve.

The performance of the classification model is evaluated by the area under the curve (AUC-Score) of the ROC curve. It indicates how well the model distinguishes between classes. A higher AUC-Score denotes better model performance. The closer the ROC curve is to the lower right corner of the graph, i.e where the TPR is high and the FPR is low, the higher the model's accuracy and ability to distinguish between classes.

Classification with high-level variables

To distinguish between the ggF and VBF processes, histograms of various variables are plotted. From these histograms, three variables were found to effectively differentiate between the two processes: the difference in pseudorapidity $\Delta\eta_{jj}$, the difference in azimuthal angle $\Delta\phi_{jj}$, and the invariant mass m_{jj} of the the two jets identified as the VBF jet candidates. The distributions of these variables are shown for both ggF and VBF production modes in a single plot in Figure 2.1.

To improve the selection of jets with VBF properties and minimize the contamination from jet pairs in QCD multijet events, various preselections are applied to the jets. The transverse momentum of the leading jet is required to be $p_T > 80$ GeV, whereas for the subleading jet $p_T > 40$ GeV. For both the leading and subleading jet, the mass should be $m \geq 0$. The two jets are required to be in opposite hemispheres of the detector, $\eta_{\mathrm{j}1}\eta_{\mathrm{j}2} < 0$. The difference in pseudorapidity and azimuthal angle between the two jets must satisfy the condition $\Delta\eta > 1$ and $\Delta\varphi < 1.5$, and the invariant mass of the jets must be $m_{jj} > 200$ GeV. Further details on the applied preselections can be found in Reference [7].



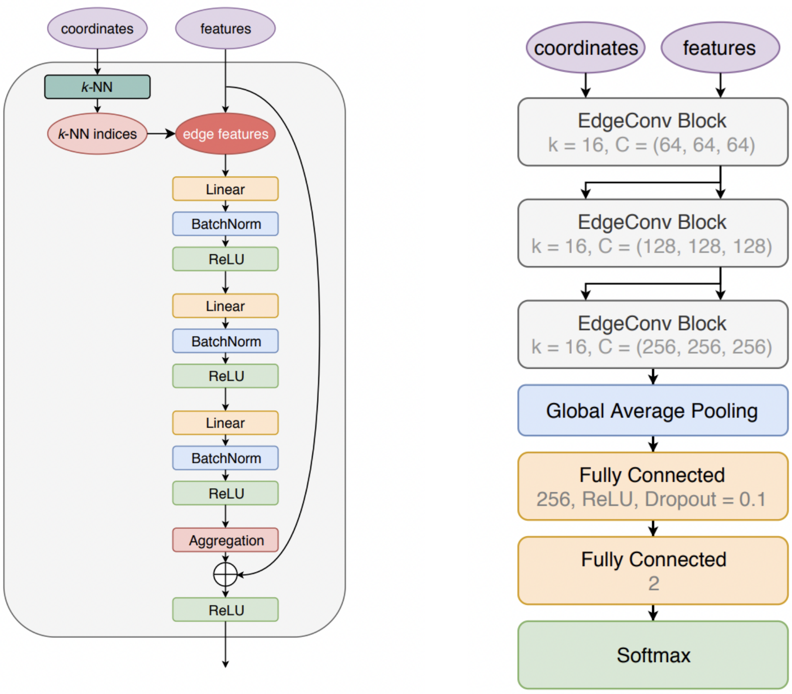
(a) (b) (c)

Figure 2.1. The histograms of (a) $\Delta\eta_{jj}$, (b) $\Delta\phi_{jj}$, and (c) m_{jj}, for ggF and VBF processes, with each histogram normalized to unit area.

Classification with graph neural networks

In the study of jet tagging, where the goal is to identify the particles that initiate a jet, recent advancements in machine learning have introduced a new approach called ParticleNet. ParticleNet is a type of graph neural network (GNN) architecture, based on the Dynamic Graph Convolutional Neural Network (DGCNN). This model treats a jet as an unordered permutation-invariant set of particles, each carrying a feature vector, much like a 'point cloud'. Using this technique, ParticleNet has shown significant improvements in jet tagging accuracy compared to previous methods. To train the neural network, weaver [9] - a machine learning R&D framework is used.

A key component of ParticleNet is the edge convolution (EdgeConv) operation, which starts by representing a point cloud as a graph. Each vertex corresponds to a point, and edges connect each point to its k nearest neighbors. In this way, a local patch needed for convolution is defined for each point as the k nearest neighboring points connected to it.



EdgeConv ParticleNet

Figure 2.2. The structure and architecture of EdgeConv and ParticleNet [8].

Figure 2.2 shows that EdgeConv needs two classes of features as input: the "coordinates", which include the relative values of η and φ for each particle, and "features" (η, φ, p_T, charge...). The features are The EdgeConv block begins by identifying the k nearest neighbors for each particle, using the input coordinates to calculate distances. The "edge features" for the EdgeConv operation are then created from the "features" input, based on the indices of these nearest neighbors. The EdgeConv operation uses a three-layer multilayer perceptron (MLP), where each layer consists of a linear transformation, batch normalization, and ReLU activation. It includes a shortcut connection inspired by ResNet, allowing input features to bypass the transformation layers.

On the other hand, ParticleNet architecture is composed of three EdgeConv blocks, where the first block uses particle coordinates in $\eta-\varphi$ space to compute distances, and the subsequent blocks use learned feature vectors. Each block considers 16 nearest neighbors, with increasing channel sizes across the blocks. After the EdgeConv blocks, global average pooling aggregates the features, followed by a fully connected layer with ReLU activation and dropout to prevent overfitting. The final output for binary classification is produced by another fully connected layer, followed by a softmax function.

Regarding the classification with graph neural network, the performance of the ParticleNet architecture is evaluated on all events. The training of the ParticleNet model was done by the collaborators in the analysis team, while its ouptut, the DNN score was subsequently used to generate the ROC curve presented in Figure 3.1b. The DNN score is a number between 0 and 1, and represents the probability that the input belongs to a particular class.

Training with b-hive

The b-hive framework combines state-of-the-art tools for HEP data processing and and training in a Python-based ecosystem [10]. It uses Python packages like law, coffea, and PyTorch bundled in conda-environment. The tasks that can be performed with this framework are the DatasetConstructor Task, Training Task, Inference Task, ROCCurve Task and WorkingPoint Task. The DatasetConstructor Task takes a ROOT file as input and converts it into numpy arrays, which serve as the training format and can then be utilized by training libraries. Then, the training of the network is done with the Training Task. The Inference Task performs the prediction, while the ROCCurve Task computes and plots the ROC curves, and the WorkingPoint Task calculates the working points.

As part of the analysis process, the training for a smaller sample was carried out using the b-hive framework. The training is evaluated on top tagging and quark-gluon tagging tasks. The corresponding ROC curves are shown in Figure 2.3. These ROC curves are used to identify whether a single jet originates from a b-quark or a c-quark. The TPR measures how effectively the algorithm identifies b-jets or c-jets as part of the top quark tagging process, while the FPR the probability of light-flavour quark and gluon jets being misidentified as b-jets or c-jets.

The obtained performance is quite far from previous trainings performed in CMS, most likely due to the limited size of the training samples used. Nonetheless, the primary goal of this training was to gain hands-on experience in training an advanced deep learning algorithm.

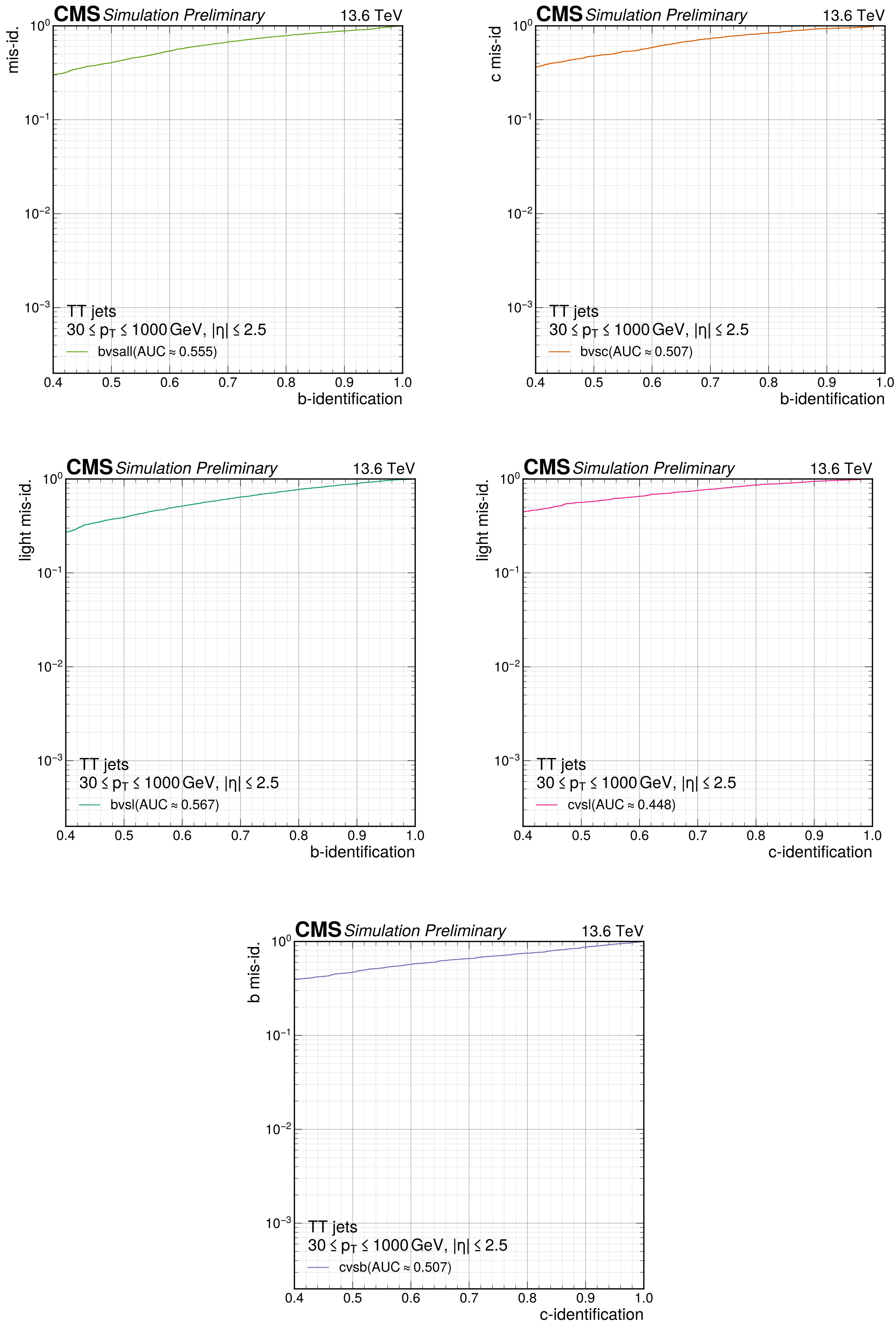
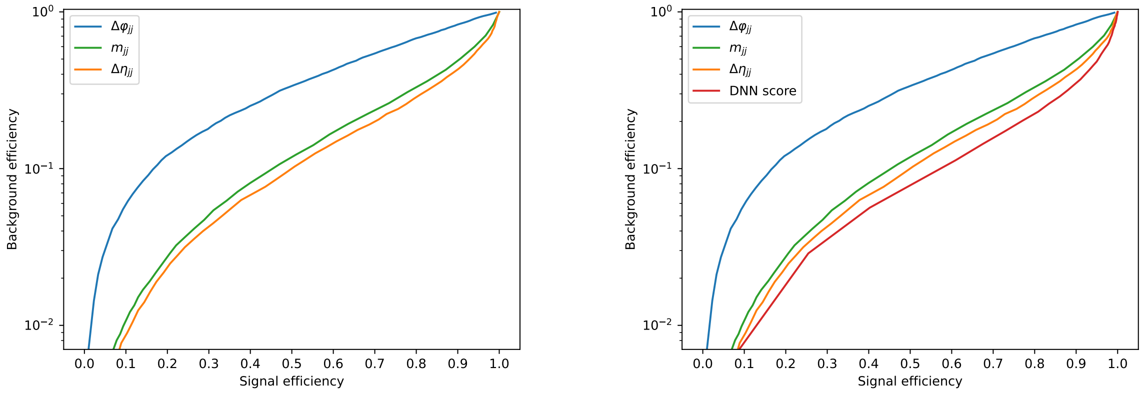


Figure 2.3. The ROC curves for the training done with b-hive.

3 Results

In the histograms in Figure 2.1, the ggF is considered as the background, while the VBF is treated as the signal. By making cuts at different bin intervals, the amount of signal or background included can be observed as one moves through the histogram. The ROC curves in Figure 3.1 are obtained by plotting the background efficiency against the signal efficiency.



(a) (b)

Figure 3.1. The ROC curves using (a) high-level variables, with (b) the ROC curve for the DNN score from the ParticleNet model, trained for VBF tagging, added for performance comparison.

From Figure 3.1a it can be noticed that among the other variables, the $\Delta\eta_{jj}$ variable has the best performance for distinguishing the signal from the background, while in the Figure 3.1b one can observe that the ROC curve for the DNN score has slightly a better performance than the m_{jj}. The values of AUC-Score for the corresponding ROC curves are summarized in Table 3.1.

Table 3.1. AUC-Scores for each ROC curve of Figure 3.1b.

Variable	AUC-Score
$\Delta\phi_{jj}$	0.3921
m_{jj}	0.8089
$\Delta\eta_{jj}$	0.8320
DNN score	0.8637

4 Conclusions

In this project the discrimination between ggF from VBF production modes is investigated. This discrimination is done with both high-level variables and graph neural networks classification. Numerous variables are inspected to find the best ones for distinguishing these two processes. From the classification with high-level variables the best variable identified was the pseudorapidity $\Delta\eta_{jj}$ of the jets with the value for AUC-Score 0.832. The AUC score (0.8637) from ParticleNet demonstrates its effectiveness in distinguishing between the two production modes and indicates better performance compared to the use of high-level variables. However, the project aims to improve it further. This improvement can be done by increasing the number of input variables for the training, increasing the number of particles, and training an ensemble of models with different initial parametrization.

Given the limited size of the training samples and the time limitation of the summer project, the performance achieved is not yet at its final level.

Acknowledgements

I would like to thank my supervisors Andrea Malara and Santeri Laurila for their guidance and support throughout this project. Their expertise and patience have greatly contributed to my learning experience. I am also deeply thankful to those who made it possible for me to participate in this summer program, including the funding bodies and coordinators. This has been an invaluable experience, and I am grateful for the opportunity to have been part of it.

Bibliography

[1] P A Zyla *et al.* “Review of Particle Physics, 2020-2021. RPP”. In: *PTEP* 2020.8 (2020), p. 083C01. DOI: 10.1093/ptep/ptaa104. URL: https://cds.cern.ch/record/2729066.

[2] ATLAS Collaboration. *Gluon Fusion*.

[3] ATLAS Collaboration. *Vector Boson Fusion*.

[4] Stan Bentvelsen, Eric Laenen, and Patrick Motylinski. “Higgs production through gluon fusion at leading order”. In: *NIKHEF* 7 (2005), p. 2005.

[5] CMS Collaboration *et al.* “Measurement of the Higgs boson production via vector boson fusion and its decay into bottom quarks in proton-proton collisions at $\sqrt{s} = 13$ TeV”. In: *arXiv preprint arXiv:2308.01253* (2023).

[6] Changqiao Li. *Combined measurements of Higgs couplings, cross-sections and interpretation at the ATLAS experiment*. Tech. rep. ATL-COM-PHYS-2024-072, 2024.

[7] CMS collaboration *et al.* “Search for invisible decays of the Higgs boson produced via vector boson fusion in proton-proton collisions at $\sqrt{s} = 13$ TeV”. In: *arXiv preprint arXiv:2201.11585* (2022).

[8] Huilin Qu and Loukas Gouskos. “Jet tagging via particle clouds”. In: *Physical Review D* 101.5 (2020), p. 056019.

[9] Huilin Qu. *Weaver*. https://github.com/hqucms/weaver.

[10] Niclas Eich. *b-hive*. GitLab, 2024. URL: https://gitlab.cern.ch/cms-btv/b-hive.