Classification of Higgs Boson Production Modes using Machine Learning

Imperial College London

Ekaterini Veroe Protopapa
Philipp Sonnenschein

Supervisor: Dr. Jonathon Langford
High Energy Physics Group, Imperial College London

1. Motivation

- The Standard Model (SM) of Particle Physics is the current best description of nature but is considered to be incomplete (e.g. lack of dark matter candidates and prediction of massless neutrinos) [1]
- This inspires the search for physics Beyond the Standard Model (BSM), where the **Higgs boson** is of significant importance
- The Higgs boson diphoton decay mode is interesting because it is sensitive to the 5 major production modes due to its clean final state signature in the detector [2]
- The Higgs boson couples differently to other fundamental particles, therefore measuring production modes targets the different parameters of the SM
- By categorising events according to the different production modes we can measure them independently [2]
- The large amounts of data collected at the LHC, in addition to the interesting kinematic features of each mode motivate machine learning (ML) techniques

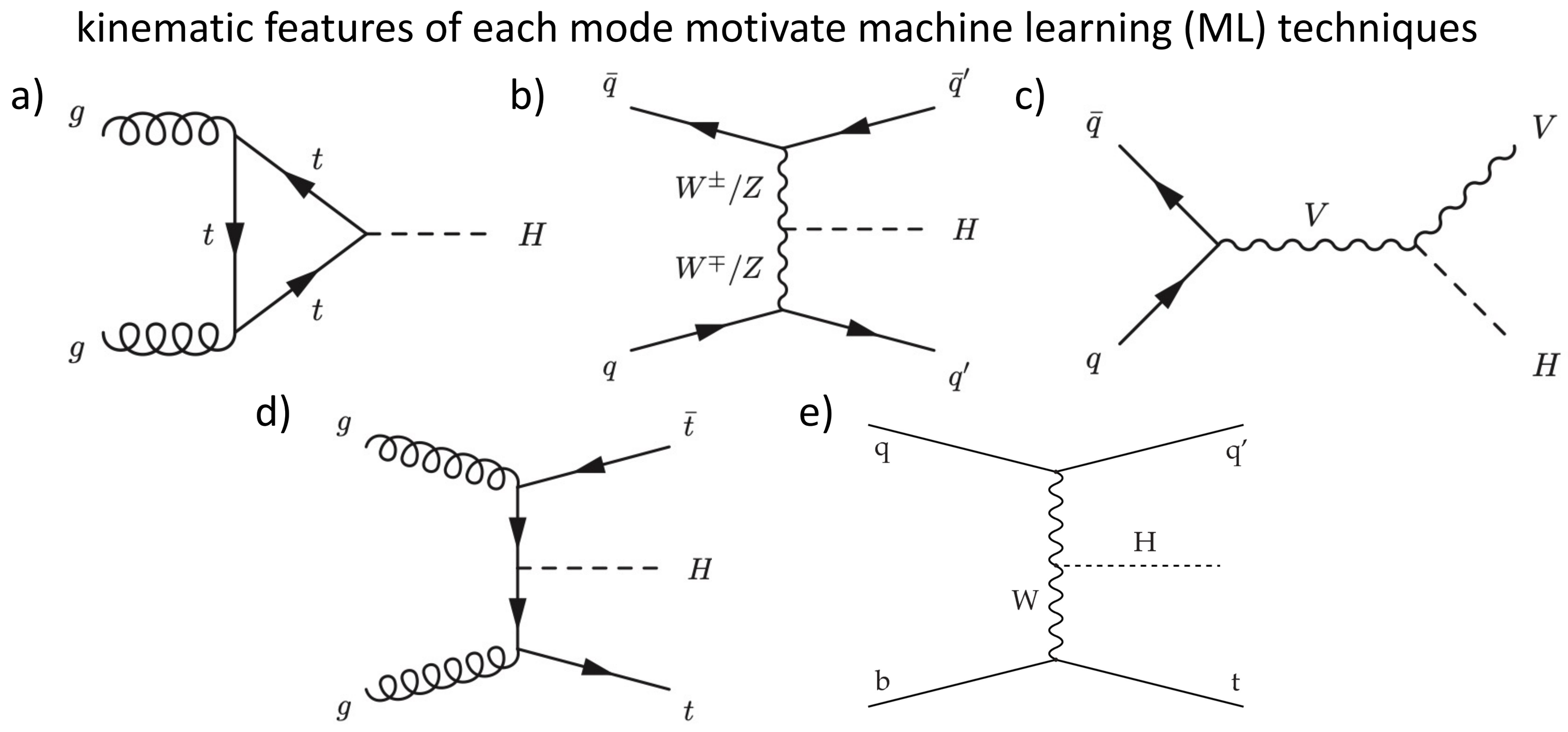

Fig. 1: Feynman diagrams of the main Higgs boson production modes according to decreasing cross section a) ggH b) qqH c) VH d) ttH e) tH [3]

Aim: Improve the current classification of the Higgs boson production modes in the diphoton decay channel by using different machine learning techniques in a unified approach.

2. Approach

- Apply ML techniques to predict the probabilities of each production mode for every event
- Train and test ML models on Monte Carlo (MC) simulated datasets with 50 input features covering jet, photon and lepton information
- Different behaviour and correlations between the input variables aid the classification process of the ML algorithms
- Optimize hyperparameters and node architecture
- Compare ML models to traditional cut-based approach

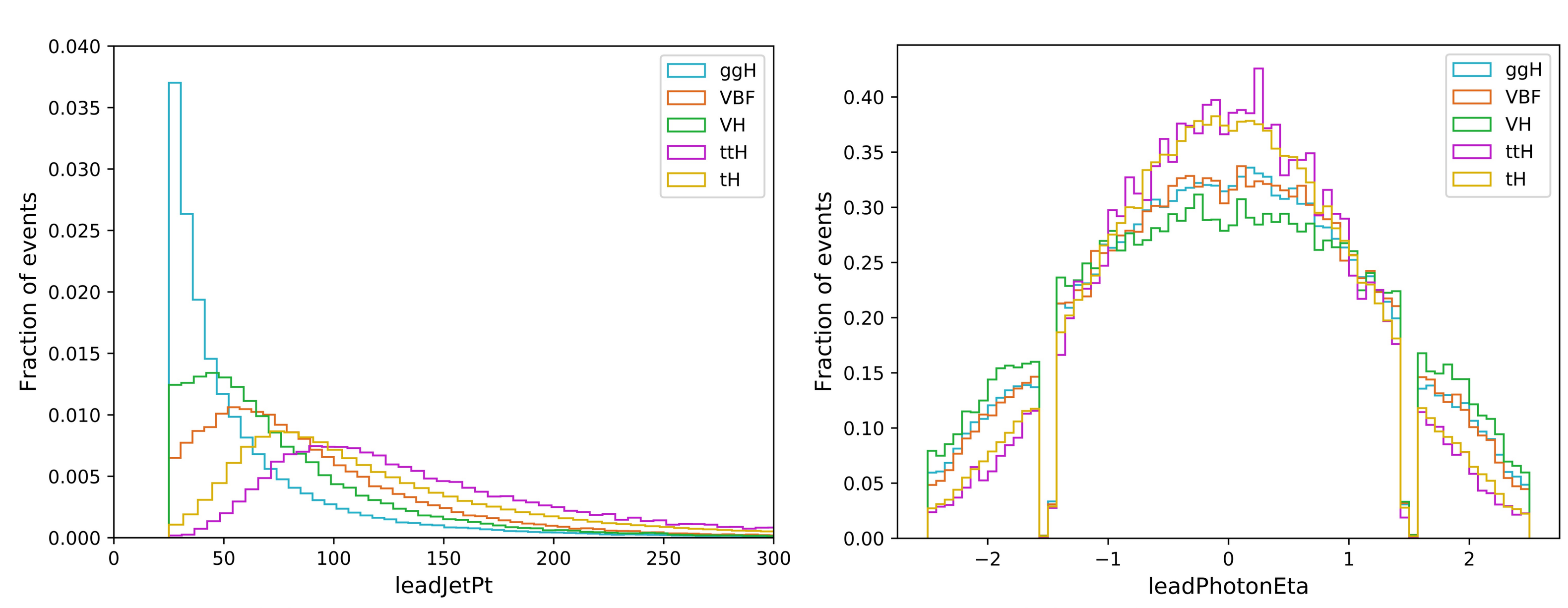

Fig. 2: (a) Lead jet transverse momentum distribution (b) lead photon eta distribution

Two different ML methods are used for the classification:

a. Boosted Decision Trees (BDTs)

- Sequential combination of decision trees with a set of input features and a recursive split of the input data [4]
- Aims to optimize predictions made by the model
- Commonly used in the high energy physics community for classification tasks

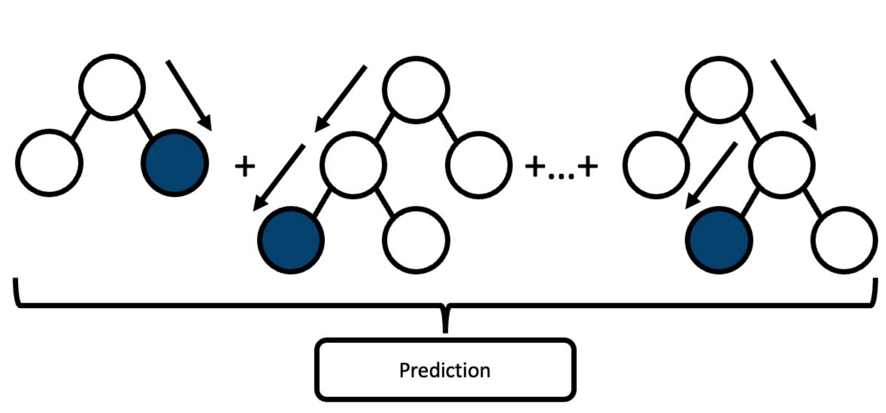

Fig. 3: Schematic diagram representing a Boosted Decision Tree

b. Deep Neural Networks (DNNs)

- Consists of an input layer, a series of hidden intermediate layers and an output layer
- Nodes are described by weight and bias that encode the characteristics of the data [5]
- Have not been applied to the classification of the production modes in the diphoton decay channel

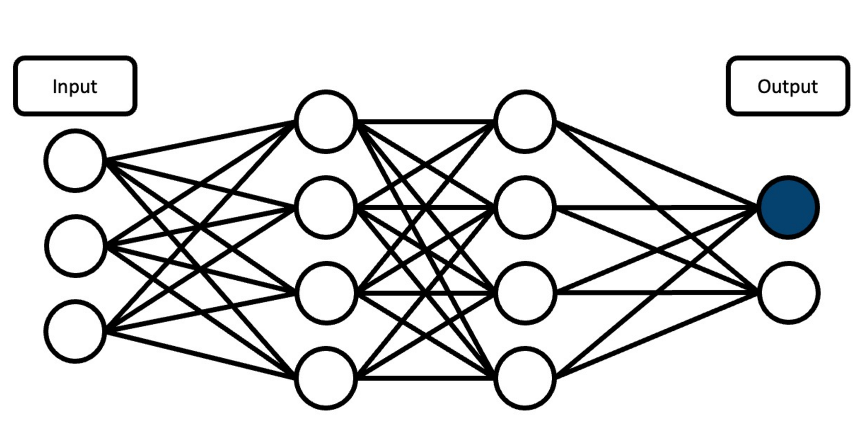

Fig. 4: Schematic diagram representing a Deep Neural Network

3. Results

Confusion matrices capture the performance of classifiers, illustrating the fraction of events of a given (true) production mode resulting in each predicted production mode (modes with the highest probability).

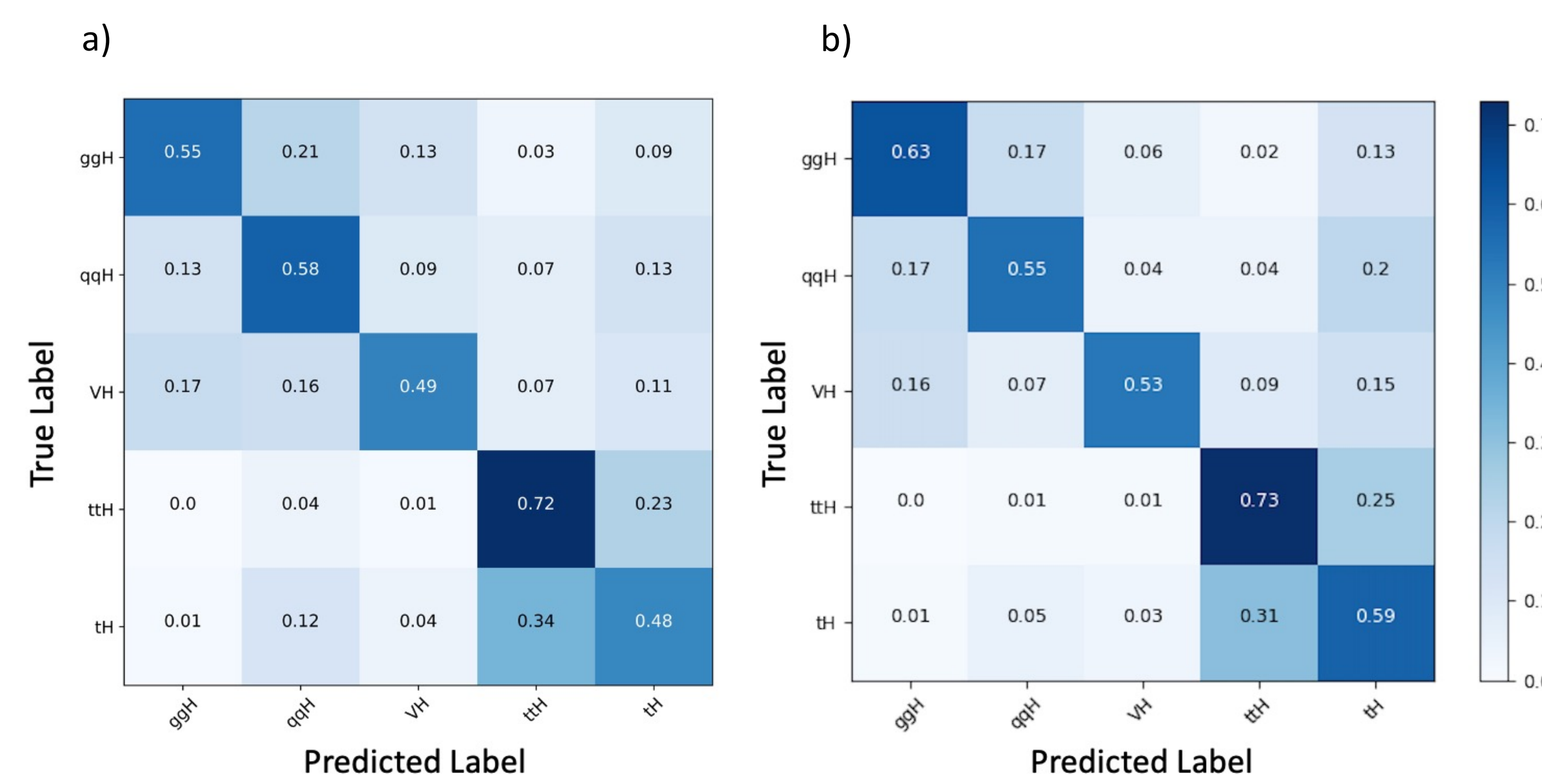

Fig. 5: Confusion matrices generated by the a) BDT and b) DNN. Matrices are normalized according to their true labels

- Both ML methods present similar performances, but have varying strengths for the classification of different production modes
- The BDT has an accuracy score of 0.564 and NN an accuracy score of 0.591

The results can also be visualized by looking at the relative fraction of events assigned to each label:

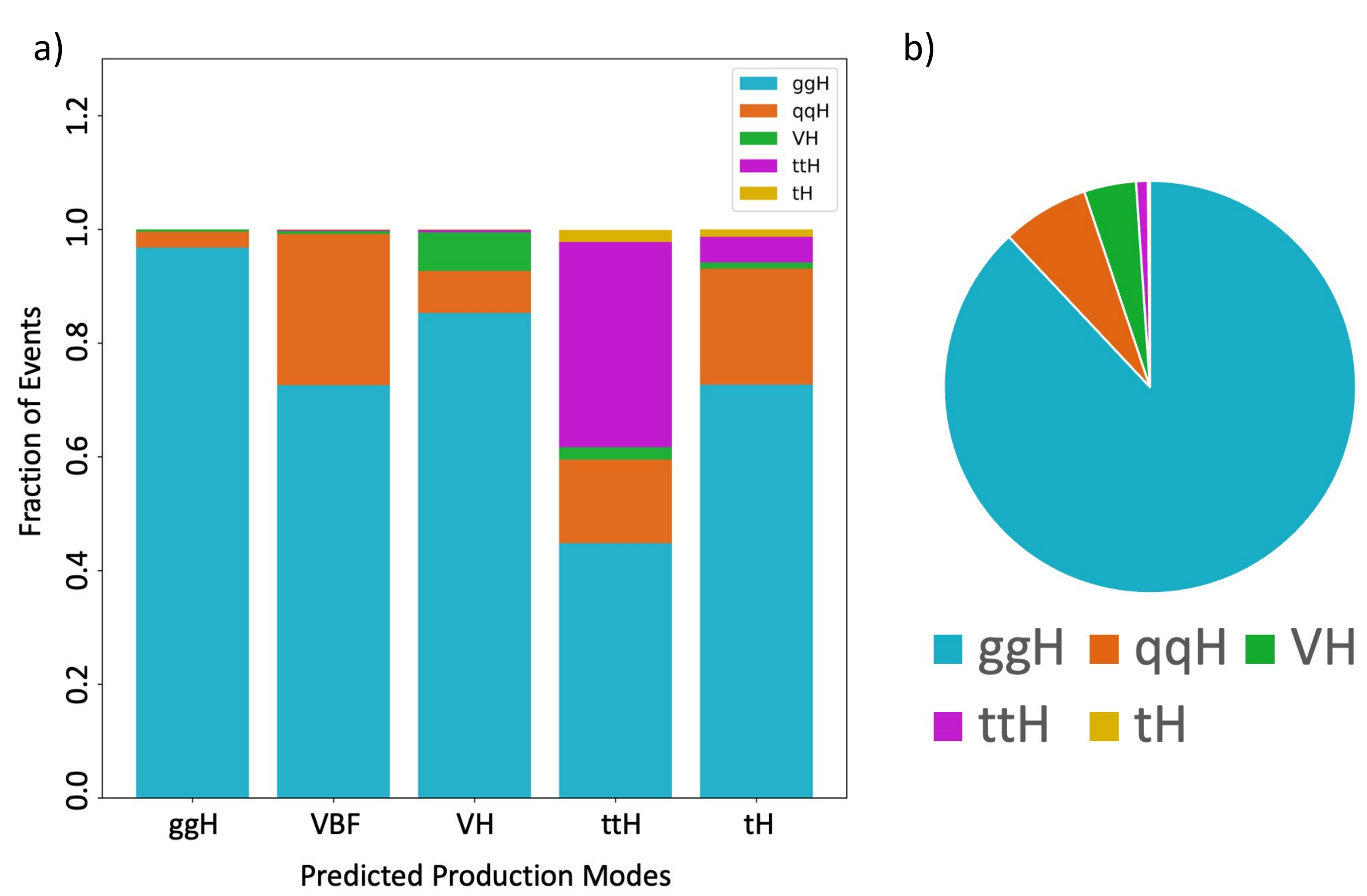

Fig. 6: a) Total normalized number of events for each predicted label b) Pie chart of relative production cross sections, explaining ggH contamination

- Because of its significantly higher cross section, ggH contaminates most of the other predicted classes
- This contamination motivates the need for further classification, for example by building binary BDTs to remove contamination from the other production modes (and background events) in each predicted class

4. Conclusion & Outlook

- The novel unified approach for the classification of the Higgs boson's production modes in the diphoton decay channel proves to be promising
- ML techniques are preferable over the traditional cut-based approach since they offer a unified approach and increased performance
- Further investigation is required to determine the ML model most suitable for the analysis
- Next steps include further splits of the production modes according to different kinematic regions of phase space and further background rejection
- Ultimately, this analysis will unravel a deeper understanding of Higgs boson parameters in the SM and can provide cutting edge insights into physics BSM

5. References

[1] E. J. T. Scott. *Measurements of Higgs boson properties using the diphoton decay channel with the Compact Muon Solenoid experiment*. PhD thesis. Imperial College London; 2019.
[2] CMS Collaboration, *Measurements of higgs boson production cross sections and couplings in the diphoton decay channel at √s = 13 tev*, Journal of High Energy Physics, vol. 2021, no. 7, Jul 2021.
[3] CMS Collaboration, *Combined measurements of Higgs boson couplings in proton-proton collisions at s√= 13 TeV*, The European Physics Journal C, vol. 79, no. 5, May 2019.
[4] B. P. Roe et al. *Boosted decision trees as an alternative to artificial neural networks for particle identification*, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment. Volume 54. Issues 2–3; 2005.
[5] Judith E. Dayhoff. *Neural Network architectures: an Introduction.* Van Nostrand Reinhold Co., USA. 1990