Study of Higgs Boson decaying into pair of electrons using Gradient Boosted Decision tree

Arpan Maity[*] **Krishna Kant Parida**[†]
School of Computer Science
National Institute of Science Education and Research (NISER), Bhubaneswar
Homi Bhabha National Institute (HBNI)
Jatni, Khurda, Odisha-754028

Abstract

In this project an identification and classification method is presented for the search of Higgs boson decaying into a pair of electrons (e^+e^-) during proton proton collision events at $\sqrt{s} = 13$ TeV. The analysis is done using Gradient Boosted Decision Tree (GBDT) on a simulated dataset used to classify the event categories for Higgs Boson production via vector boson fusion (VBF). A comparative analysis was done using various learning rates, their implications along with necessary feature engineering applied to the datasets in order to obtain meaningful classification and their improvization.

1 Introduction

Higgs Boson was discovered by the ATLAS and CMS Collaborations in 2012. Since then, measurements based on interactions of the Higgs Bosons with various standard model (SM) particles have been made. Based on the prediction of standard model, the coupling strength is directly proportional to the mass of the interacting fermions (spin -½ particles). The evidences for decay of Higgs boson into a pair of τ leptons are well recorded, but to that of decay into a pair of muons are highly contested. Hence, getting evidences of Higgs bosons coupling with electrons (or first generation fermions), owing to predicted branching fraction of 3.0×10^{-4}, is a challenging task and hence yet to be confirmed experimentally.

The base paper have briefly introduced the functioning of the CMS detector and the global event reconstruction of individual particles using particle flow (PF) algorithm. Their analysis strategy, adapted from one used for Higgs decay into photons, is based on classifying dielectron trigger signals using various parametric thresholds on selected features like transverse momentum p_T, angle ϕ and pseudorapidity η. Other than that, the thresholds were also chosen according to the characteristics of the CMS detector for obtaining both signal and the background.

A simulated sample dataset along with background were generated using Monte Carlo (MC) Techniques corresponding to various signal processes for Higgs production. Loose event preselection added with one targeting VBF events (VBF preselection procedure) were applied for ensuring consistency with the required decay process detection. Event categorization primarily uses a boosted decision tree in order to discriminate the VBF based Higgs production as signal from background events. Categories are defined using the output scores of the classifying BDT, which are non-overlapping by construction.

Hence, each category to be analyzed are defined as per selections on a multivariate (MVA) discriminant. And the final model of an MVA-based classifier was trained to distinguish the signal events

[*] arpan.maity@niser.ac.in

[†] krishna.parida@niser.ac.in

Preprint. Under review.

from the dominant Drell-Yan (DY) background signals. Finally, in each category, dielectron invariant mass distribution (m_{ee}) were fitted from which, an attempt to extract an upper limit on the branching fraction was made.

1.1 Related Works

Since identification and classification of Higgs boson decaying into pair of electrons, as predicted by the standard model with the upper cap in the branching fraction to be at the most stringent to date, analysis from other experimental collaborations with different detector designs have also been attempted. Although CMS collaboration were the first to claim about the evidence for the decay of Higgs boson to muons (second generation fermions), similar attempts were also made by ATLAS collaboration in order to validate the findings by their counterpart.

Recently, ATLAS collaboration found evidence of a rare Dalitz (three body) decay of Higgs bosons into two leptons (muon or electron pairs) and a photon. This was done using a categorization of a unique experimental signature of overlapping lepton pairs. Other than that both CMS and ATLAS collaborations are looking for various other lepton flavor violating decays of the Higgs Boson, which can be enhanced in beyond standard model (BSM) physics.

More machine learning based applications are made using classification and regression studies for observing Higgs decay into a pair of Bottom quarks by implementing both MVA-based BDT and Deep Neutral Network (DNN). Very recently, there are proposals for study Higgs boson decay into b and c jet were made under LHCb collaboration using Quantum Machine Learning (QML) by constructing a quantum classifier and running the algorithms on quantum simulators adn finally in real quantum computers to perform the measurements.

2 Dataset and Baseline Algorithm

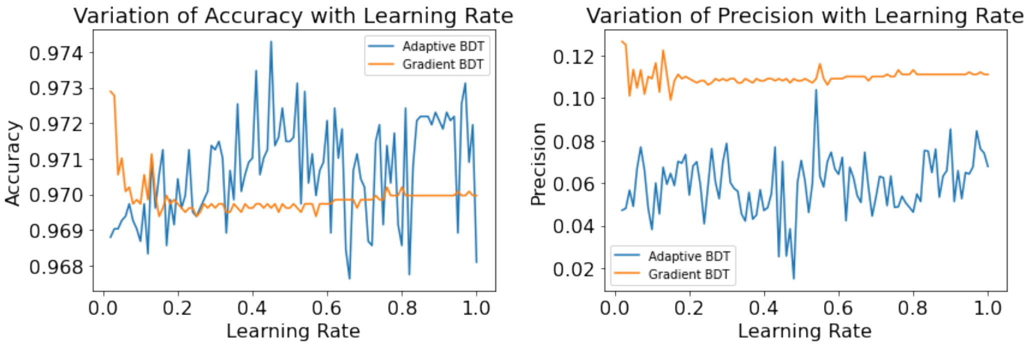
The baseline is built upon classification of Higgs decay into pair of electrons using a Gradient Boosting Decision Tree (GBDT), trained using a dataset simulated by Monte Carlo techniques using `PYTHIA8`. Specific root files generated by the simulator consisting of both signal and background datasets were obtained. These root files consists of directories containing various trees accessible by `ROOT` packages. The trees of the dataset contain identified leptonic and hadronic jets which are in practice are identified from the raw detector data using PF algorithm.

Further machine learning based analysis were to be done on this dataset by selecting specific feature branches from the trees representing VBF produced Higgs. A VBF preselection for identifying VBF produced Higgs bosons were currently deemed to be optional, subjecting to time constraints. The root file also contains the CMS detector response which was simulated using `GEANT4` package, which completes the overall procedure of simulation of event generation to detection.

Data belonging to specific features, mainly representing the transverse momentum p_T, angle ϕ and pseudorapidity η of each of the electrons were taken into account and are collected into an analyzable format like .csv file. Based on the categorization conditions on the feature values, the GBDT (XGBoost classifier) was trained with around 70% of the overall dataset, where 15% of the dataset was used for validation, and final 15% was used for testing purpose. The section of data to be tested was chosen randomly from the overall dataset which ensures the test data to remain beyond observation bounds.

The redundant features containing no significant data (single value for the fearture throughout the dataset) were removed and conditions for event categorization were applied. The model was trained repeatedly over variation of learning rate, which upon each successive training was tested using the test data from the dataset. The measures of accuracy, precision, recall and BDT score were obtained from from each testing procedure at different learning rates.

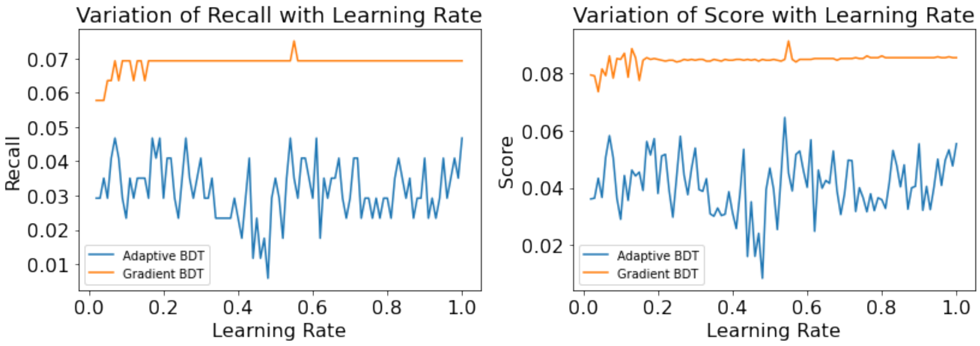
Finally, the baseline analysis was done by comparision with adaptive boosting based BDT (adaBoost classifier) in terms of their accuracy and BDT score. This aids in setting the learning rate of the GBDT and in search of possible tradeoffs associated with the learning rate. Similar comparision based analysis can also be done for various other hyperparameters like tree depth etc.



(a) Accuracy

(b) Precision

Figure 1: Accuracy and Precision comparision between Gradient Boosted and Adaptive Boosted Decision Tree



(a) Recall

(b) Score

Figure 2: Recall and Score comparision between Gradient Boosted and Adaptive Boosted Decision Tree

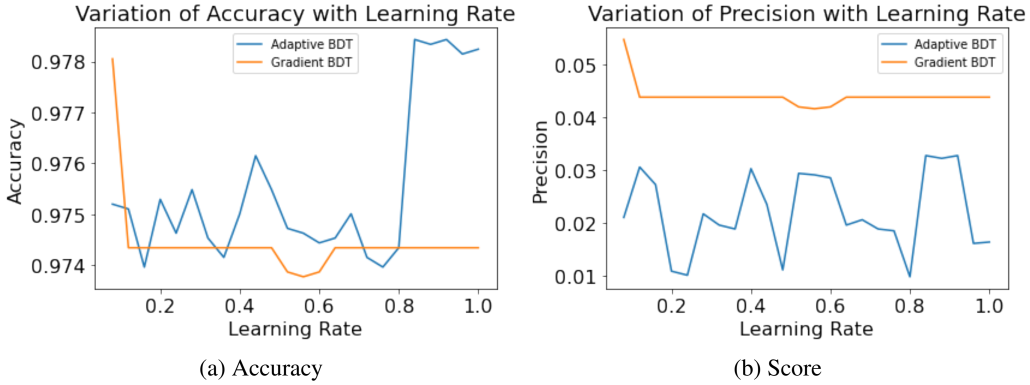
3 Experimental Analysis

The required tree extraction was done using C++ program running in `ROOT` software, from which the features branches were extracted into .csv file format using `PyROOT` module and `Pandas` framework. Finally, the BDT classifier with gradient boosting and adaptive boosting were trained using `scikit-learn` packages. The full code can be found in the provided GITHUB repository.

Figure 1a and figure 1b provide the comparision of Accuracy and Precision obtained by differently boosted BDT models. We can find that although accuracy levels for adaptive boosted BDT are higher in comparision with the gradient boosted counterparts, the precision of gradient boosted BDT is much higher in comparision to the adaptive boosted ones.

Additionally, figure 2a and figure 2b provide a comparision of Recall values and Score allotted to the categories between the differently boosted BDT models. Higher and consistent score implies higher likelihood of data belonging to the actual signal. Also, higher recall implies higher performance and better classification actual positive instances. In both cases, Gradient boosted BDT have made an edge over its Adaptive boosted counterpart.

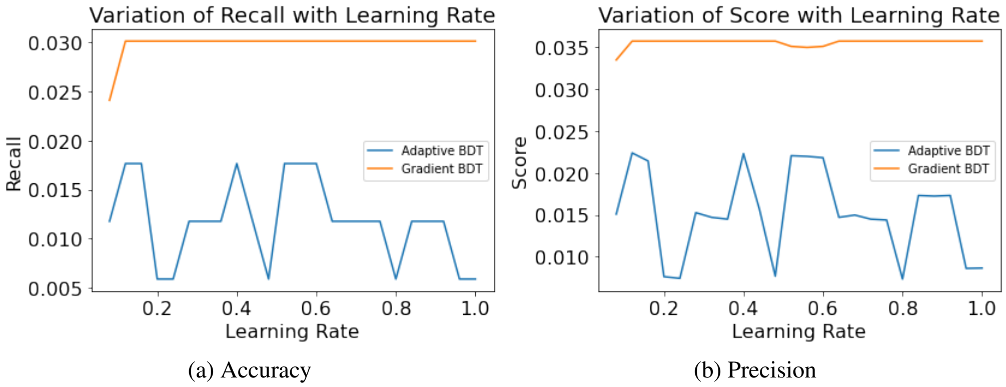
Figure 3a, figure 3b, figure 4a and figure 4b provide the similar comparision, but with larger background dataset. One can notice that although the overall accuracy have increased, the precision, recall and scores have been reduced nearly by a factor of half.



(a) Accuracy

(b) Score

Figure 3: Accuracy and Precision comparision between Gradient Boosted and Adaptive Boosted Decision Tree



(a) Accuracy

(b) Precision

Figure 4: Accuracy and Precision comparision between Gradient Boosted and Adaptive Boosted Decision Tree with larger background dataset

4 Conclusion and upcoming prospectives

Therefore, we have obtained the comparitive analysis of the Gradient Boosted Decision Tree with that of its Adaptive boosted counterpart. Since the Gradient Boosted BDT have worked at par with the Adaptive boosted one, we can improvise it further by looking beyond being MVA-based classifier. One can use stochastic based modelling of the GBDT, or directly into a gradient boosted distributed decision tree. Based on the analysis of the base paper, one can even continue to obtain the dielectron invariant mass distribution for each of the classified events and make an estimate of the branching fraction obtained from the improvised code.

Also, we removed redundant features manually, this can further be improvised using an autoencoder based unsupervised learning algorithm. Also, the performance of the prepared BDT classifier can be compared with neural network based algorithms such as Recurrent Neural Network (RNN). Such comparisons can also be included with the time complexity of the classifier algorithms, for more comprehensive distinctions.

References

[1] CMS Collaboration (2022) *A search for the Higgs boson decay to a pair of electrons in p-p collisions at* $\sqrt{s} = 13$ *TeV* CMS-HIG-21-015, `CERN-EP-2022-131`

[2] ATLAS Collaboration (2019) *A search for the dimuon decay of the Standard Model Higgs boson in p-p collisions at* $\sqrt{s} = 13$ *TeV with ATLAS detector.* ATLAS CONF Note, `ATLAS-CONF-2019-028`

[3] ATLAS Collaboration (2021) *Evidence for Higgs boson decays to a low-mass dilepton system and a photon in p p collisions at* $\sqrt{s} = 13$ *TeV with theATLAS detector.* ATLAS CONF Note, `ATLAS-CONF-2021-002`

[4] ATLAS Collaboration (2021) *Searches for lepton-flavour-violating decays of the Higgs boson in* $\sqrt{s} = 13$ *TeV p-p collisions with the ATLAS detector.* Phys. Lett. B 800 (2020) 135069

[5] Cagnotta, A. ; Carnevali, F. ; De Iorio A. (2022) Machine Learning Applications for Jet Tagging in the CMS Experiment. *Appl. Sci.* **2022**, 12, 10574.