UNIVERSITY of CALIFORNIA
SANTA CRUZ

PROSPECTS FOR DISCOVERY OF $t\bar{t}HH$ PRODUCTION AT THE HL-LHC USING BOOSTED DECISION TREES

A thesis submitted in partial satisfaction of the
requirements for the degree of

BACHELOR OF SCIENCE

in

PHYSICS

by

Jonathan O. Tellechea

June 2020

The thesis of Jonathan O. Tellechea is approved by:

Professor Michael Hance
Advisor

Professor Michael Dine
Chair, Department of Physics

Abstract

Prospects for discovery of $t\bar{t}HH$ production at the HL-LHC using Boosted Decision Trees

by

Jonathan O. Tellechea

A measurement of the Higgs boson self-coupling is crucial step in probing the Higgs potential. One way of probing the Higgs self coupling is through measurements of events with two Higgs bosons. This thesis studies the $t\bar{t}HH$ production mode in proton-proton collisions at $\sqrt{s} = 14$ TeV assuming the use of an upgraded ATLAS detector at the High-Luminosity LHC. The study attempts to improve on a cut-based analysis method by using machine learning via a Boosted Decision Tree. Considering events with at least one prompt lepton, we find a maximum significance of 0.35 σ with no systematic uncertainties on the background. While the significance of this channel remains modest, in combination with other production modes it will contribute towards measurements of the Higgs self-coupling at the HL-LHC.

Contents

To my parents,

Oswaldo and Rosa Tellechea

To my brother,

Samuel Tellechea

Acknowledgements

I have had many friends, colleagues, and mentors who have been by my side on my academic journey for which I am grateful. Jayesh Bhakta my professor, mentor and boss at LACC STEM Pathways was very helpful he advised me in physics and life and gave me the opportunity to hone my knowledge as a tutor. Marcos M. Alvarez and Glen Baghdasarian my professors, research advisors, friends, and mentors, have taught me a lot, have introduced me to research and with their guidance I was able to publish my first research paper. They were crucial to my transition into a 4 year university. I would like to thank my friends Edna Santos, Ramsey Issa, Lina Yi and Victor Gomez who have helped me stay on track and been positive influences in my life. I would like to thank my thesis advisor Michael Hance for the research opportunity, sharing of his knowledge, friendship, advice and mentoring while at UCSC which without him this thesis would not be possible.

1

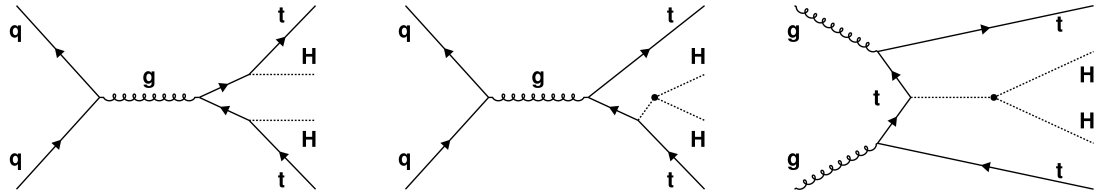
Introduction

The Standard Model (SM) of particle physics is the foundation that helps us understand high energy physics. The SM characterizes the elementary particles that make up matter and the forces that mediate interactions between them. The SM is a theory that explains most of our universe thus allowing for accurate predictions. Currently the SM can not tell us anything about gravity or dark matter. Quarks and leptons are half-integer spin fermions with interactions that are mediated by integer-spin bosons [9]. The Higgs boson (H), currently has not been fully probed which motivates us to study its properties [1]. The Higgs field is able to couple with itself, this is referred to as Higgs self coupling. The self-coupling strength, often denoted by λ can tell us more about the nature of the SM Higgs boson[7]. If the coupling strength is different than that predicted by the SM, this could be due to new physics, describing phenomena beyond the SM. A percise measurement of the Higgs self coupling strength can help us understand the ElectroWeak Phase Transitions (EWPT) in the early universe [7].

To better understand the Higgs self couplings we use graphical representations called Feynman diagrams which give us the physics and mathematical expressions needed to calculate the strength of coupling at vertices λ, cross section σ, and decay widths Γ. To calculate these quantities the sum over all possible Feynman diagrams are needed, similar to a Taylor series, where the higher-order terms contribute the least [10]. We can use the LHC, which has proton-proton (pp) collisions at $\sqrt{s} = 14$ TeV, to produce a sample of events with two Higgs bosons with a cross section that depends on λ_{HHH}, the trilinear Higgs self coupling.

The channel studied in this thesis is $pp \rightarrow t\bar{t}HH$, where t is for a top quark and $\bar{t}$ is for an anti-top quark. The $H \rightarrow b\bar{b}$ decay is considered, given its large branching ratio of 58%, where b is for a bottom quark and $\bar{b}$ is for an anti-bottom quark [2]. The leading order Feynman diagrams for $t\bar{t}HH$ are shown in Figure 1. The two diagrams shown in Figures 1b and 1c are self coupling Higgs boson processes. These processes start with an off shell Higgs boson which decays into two Higgs bosons, and then the remaining t or $\bar{t}$, and H decay into b or $\bar{b}$. The diagram in Figure 1a does not contain a λ_{HHH} vertex, therefore this Feynman diagram interferes with the other diagrams in the calculation of the scattering amplitude.

The $t\bar{t}HH$ channel is promising and previous attempts to observe di-Higgs production in the $t\bar{t}HH$ production mode have estimated only modest sensitivity at the HL-LHC [2]. The interference observed in Figure 1a does not seem to contribute much in calculating the matrix elements, making this channel less sensitive to the value of λ than other production modes. The number of events in $t\bar{t}HH(HH \rightarrow b\bar{b}b\bar{b})$ are small



(a) Uncoupled Higgs (b) Coupled Higgs (c) Coupled Higgs

Figure 1: Leading order Feynman diagrams for $t\bar{t}HH$ production.

compared to the backgrounds. The backgrounds considered here include $t\bar{t}b\bar{b} + jets$, $t\bar{t}H(H \to b\bar{b}) + jets$, and $t\bar{t}Z(Z \to b\bar{b}) + jets$ where Z is the Z boson and jets are any fermions.

The number of events N depends on the luminosity L, cross section σ, and a filter efficiency given by Equation (1).

$$N = \sigma \cdot L \tag{1}$$

$t\bar{t}HH$ production at $\sqrt{s} = 14$ TeV has a cross section of 0.981 fb. ATLAS plans to collect an integrated luminosity of 3000 fb^{-1} at the CERN HL-LHC.

Using a cut-based analysis one can attempt to isolate the background events from the $t\bar{t}HH$ events [2]. This can be done by looking for event shape variables that are used to discriminate between $t\bar{t}HH$ signal and background. This strategy was pursued in [3] and removes a lot of signal events. The number of signal events is calculated using Equation (1) by normalising with the branching ratio; 0.981 fb $\cdot\, 0.58^2 \cdot$ 3000 fb^{-1} = 990. While the backgrounds are 100 to 1000 times bigger. Therefore cutting the signal has

to minimised. With a plethora of possible event shape variables which have not been exhausted, we switch over to machine learning algorithms. The analysis presented in this thesis uses a Boosted Decision Tree (BDT).

The BDT uses a multivariate algorithm to apply a cut on the data which reduces the efficiency of the background and increases the efficiency of the $t\bar{t}HH$ signal; this is represented by the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve goes from 0 to 1 in both axes; therefore the maximum would be zero background and all signal which would result in an area of 1.00 (unitless). To improve the signal verses the background certain event shape variables and four vector components are fed to the BDT to train and give back a higher area under the ROC curve.

Some bottlenecks in using the BDT are computing power and/or time. Worse is that these constraints are inversely proportional. These constraints prevent us from feeding the BDT all the data components. With these constraints we need to be creative to select certain event shape variables that improve the area under the ROC curve without the need to use all four-vector quantities. A trial and error method is used to reduce the BDT's time to calculate the ROC curve. Feeding the BDT event shape variables that have merit as possible identifiers of $t\bar{t}HH$ production will need to be selected and tested to see if these variable optimize the BDT. However with the huge dimensionality of the data it is not straightforward to distinguish these variables.

Using a BDT can find regions in phase-space that can identify signal events. Looking at the ROC curve in Figure 18 the goal is to optimize the separation of the $t\bar{t}HH$ signal from background. The ROC curve gives us the flexibility to move along

the curve to find the spot where we can maximize the signal and reject the background.

2

Data Generation

The data used in the analysis was simulated using the conditions in [3], and summarised in Table 1. The signal and background samples are generated by using Monte Carlo generators. The signal is generated using MadGraph for the matrix element and Pythia8 for the parton shower, while the backgrounds use Sherpa. All samples are generated at leading order in QCD. The background samples are filtered by requiring at least one charged lepton with p_T >20 GeV. The number of events N at a luminosity of L, cross section σ, and one-lepton filter is given by equation below:

$$N = \sigma \cdot Filter \cdot L \tag{2}$$

The number of background events are calculated using the given cross section, luminosity, and one-lepton filter efficiency in Table 1. The signal has no filter and its cross section is only reduced by the $H \to b\bar{b}$ branching ratio.

Sample	Generator	σ (fb)	Filter	Events in 3 ab^{-1}	Events Generated
$t\bar{t}HH$ ($HH\rightarrow b\bar{b}b\bar{b}$)	MadGraph/Pythia8	0.33	-	990	20,000
$t\bar{t}b\bar{b}$ + jets	Sherpa	3750	0.52	5,850,000	6,000,000
$t\bar{t}H$ ($H\rightarrow b\bar{b}$) + jets	Sherpa	371	0.55	612,000	600,000
$t\bar{t}Z$ ($Z\rightarrow b\bar{b}$) + jets	Sherpa	163	0.55	269,000	300,000

Table 1: Summary of the signal and background samples used in cut based analysis. A charged lepton filter was applied to the background samples for the cut based analysis. The same filter was applied to the signal sample for the BDT analysis to prevent any bias.

3

Analysis

This analysis uses the semi-leptonic final state, which is represented by the leading order Feynman diagrams for $t\bar{t}HH$ production as seen in Figure 1. The background samples semi-leptonic final state consist of an electron or muon and at least 4 b quarks [3]. The data used in this paper is constructed using the parameters used in [3]. The raw data consist of energy-momentum four vectors, type and number of fundamental particle. Event shape variables are used to create a cut which can minimized the background. The variables that have some discrimination power include :

The average separation in pseudorapidity (srap) between two b-tagged jets

$$\langle \eta(b_i, b_j) \rangle = \frac{1}{N^2 - N} \sum_{i=1}^{N} \sum_{j>i}^{N} |\eta_i - \eta_j|. \tag{3}$$

Higgs boson candidate mass M_{bb}, $p^\mu = (E_i, \boldsymbol{p}_i)$ and $q_\mu = (E_j, \boldsymbol{p}_j)$ [10]

$$M_{bb} = \sqrt{p^\mu q_\mu} \tag{4}$$

Centrality which is the scalar sum of p_T for all jets, divided by the energy sum of all jets.

$$\begin{aligned} E_{Total} &= \sum_{i=1}^{N} E_i \\ Centrality &= \frac{1}{E_{Total}} \sum_{i=1}^{N} p_{T_i} \end{aligned} \tag{5}$$

The scalar sum of pT for b-tagged jets, H_B

$$H_B = \sum_{i=1}^{N} p_{T_i} \tag{6}$$

The cut-based analysis presented in [3] is defined by choosing requirements on the variables defined above that enhance the predicted signal yield relative to the predicted background yield. The variables described above are shown in Table 2. Choosing events with at least 6 jets, 4 b-tagged jets $N_b = 4$, and a $\langle \eta(b_i, b_j) \rangle < 1.25$ reduces the background by a factor of $7.5 \cdot 10^{-5}$ with a signal efficiency of $6.0 \cdot 10^{-3}$ as seen in Figures 2 and 3. No cut selection offered significant improvement in background rejection or signal efficiency. Testing these regions and or introducing new variables is a way to optimize the event selection. Some new event shape variables not use in [3] but used on the BDT are seen in Equations (7) and (8). The ΔR is used as the radius of a cone where a jet and a lepton are enclosed. A top quark decaying into two leptons via the W boson and a b-tagged jet; a small ΔR could correlate to a b-tagged jet as opposed to the other b-tagged jets originating from the λ_{HH} vertex.

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} \tag{7}$$

Transverse momentum is the momentum perpendicular to the beam of protons. Prior to the collision transverse momentum is zero. Therefore after the collision using conservation laws we can find the missing transverse momentum m_T and at angle ϕ it exited.

$$m_T = \sqrt{2E^l E^\nu (1 - Cos(\Delta\phi))} \tag{8}$$

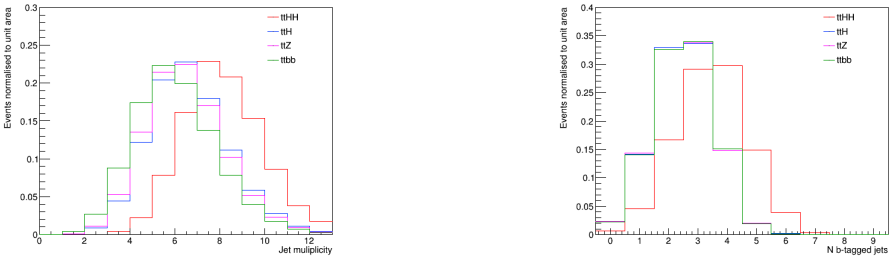


Figure 2: Left: jet multiplicity, number of jets. Right: N_b, number of jets being reconstructed from a bottom quark.

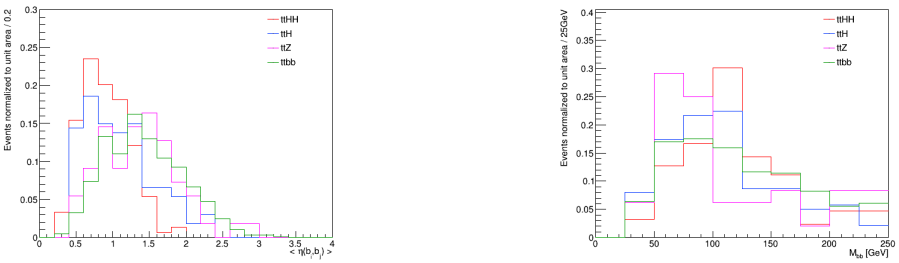


Figure 3: Left: $\langle \eta(b_i, b_j) \rangle$, average separation in pseudorapidity between two b-tagged jets calculated using Equation (3). Right: M_{bb}. Calculated using Equation (4).

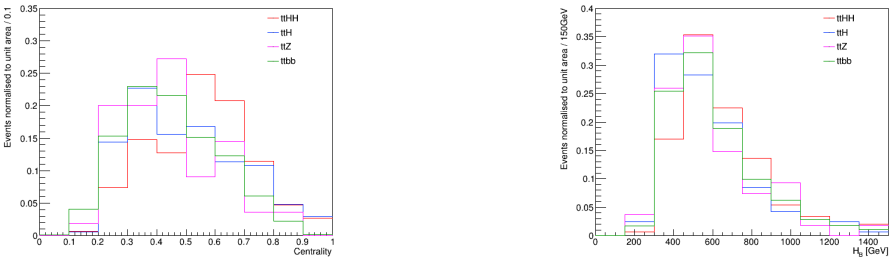


Figure 4: Left: Centrality; Right: H_B. Both calculated using Equations (5) and (6).

A large set of measured objects are required to describe each event which makes optimizing a cut-based search challenging. An alternative method is the use of a machine learning algorithm that can maximize the significance of the signal.

As we turn to Machine learning to assist us, we will use a BDT implementation from a Python library called Scikit-learn [8]. The BDT Classifier used is AdaBoost with a Stagewise Additive Modeling using a Multi-class Exponential loss function (SAMME) algorithm [6]. The adaBoost classifier uses N weak classifiers and weights the data by $w_i = 1/N$ where i is the iterated index and N is the number of classifiers. If the training point fails the match its weight is increased hence boosted, and if the match is correct then the weight is decreased[6]. This is iterated over 500 to 1000 times, the classifiers are then combined and a score is given, the BDT score. This BDT score distribution in Figure 18 can show us where to apply a cut with given BDT score.

The data is split, 10% is stored for future evaluation at random to not cause any bias. The remaining 90% is split further, half for training and half for testing. The data that is trained are the event shape variables that show the most promise. We have chosen to train the data in phases, with each new phase containing the variables

from the previous plus a few more. Phase 1 we used jet multiplicity as the event shape variable, this was to test the BDT. Phase 2 we added N_b, and $\langle\eta(b_i, b_j)\rangle$. Phase 3 we added Centrality, H_B, and M_{bb}. Phase 4 we added ΔRi, and m_Ti where (i = 1,2,3 for the 3 leptons). With each phase fed to the BDT we obtain BDT plots similar to Figure 18, and Figure 5. Figure 5 shows witch variables were more important. Since we are limited by computation power and time, Figure 5 can help us pick variables that contribute the most. Even though we can not know how other variables will contribute with new variables. Scanning through the ROC curve in Figure 18 an optimized BDT score threshold is chosen to display the cuts on the event shape variables for Figures 6, 9 and 12.

4

Results

Using the cut based analysis leaves very few events as summarized in Table 2. This does not come as a surprise as the cross section is small for $t\bar{t}HH$ production. It is challenging to find a region in the event shape variables that will optimise the signal.

Sample	No cuts	Trigger	One lepton	≥ 7 jets	≥ 5 b-tags	$\eta(b_i, b_j)$	≥ 6 b-tags
$t\bar{t}HH$	990	301	258	169	25	6	6
$t\bar{t}H$	612000	351668	296206	94202	1914	113	113
$t\bar{t}Z$	269000	152516	128580	37185	750	21	21
$t\bar{t}b\bar{b}$	5850000	3487500	2935502	646765	12481	368	368
Total background	6731000	3991684	3360288	778152	15145	502	502

Table 2: Summary of cuts applied to signal and background event. $\eta(b_i, b_j)$ column are for cuts with a $\langle \eta(b_i, b_j) \rangle < 1.25$.

Using a BDT to train on new event shape variables which can shed light on how

which variables work. As we explore, the variables that are important remain consistent through the phases. Due to this consistency we will focus on phase 4, The importance plot is part of the BDT which gives a score to each variable as seen in Figure 5. We can see the top three variables are the N_b, $\langle \eta(b_i, b_j) \rangle$, and Centrality.

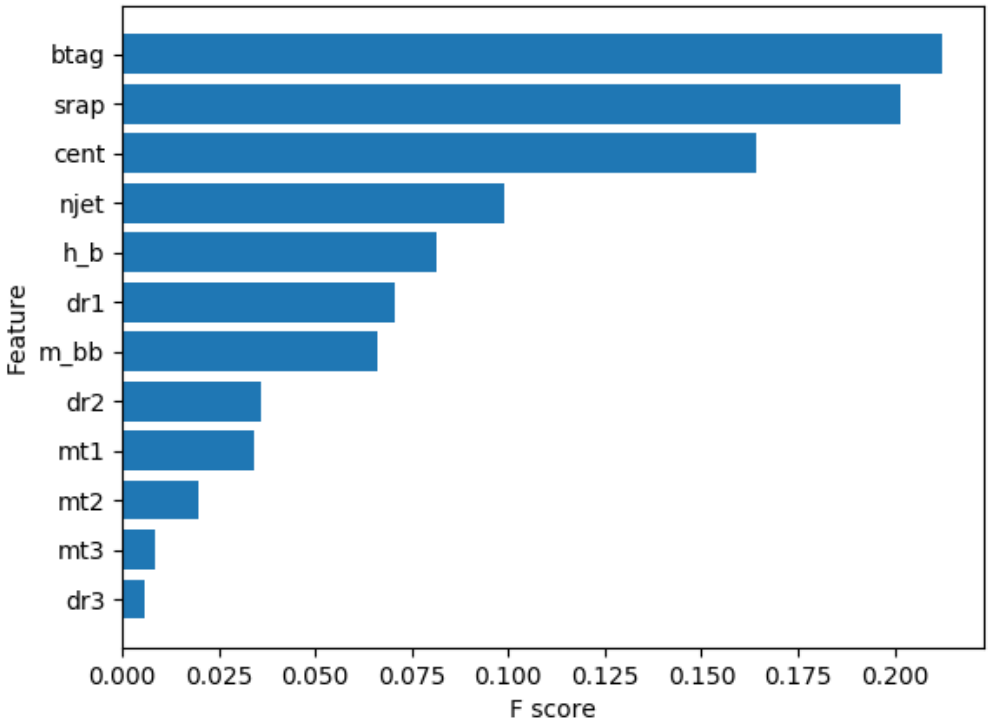


Figure 5: Importance; This plot ranks variables after training and testing and ranks them by giving an F score, the higher the F score the more important the variable is hence the name.

First, looking at N_b in Figure 6 we can see at $N_b \geq 5$ the signal does not change much while the background gets reduced. While for $N_b < 5$ both signal and background

are reduced. Since the events are plotted on a log scale this reduction is an important feature. We can see that signal events are favored for $N_b \geq 5$. Next, looking at $\langle \eta(b_i, b_j) \rangle$ in Figure 6 the $\langle \eta(b_i, b_j) \rangle < 4$ is favored for the signal events. The use of the log scale on the plots in Figure 6 shows that even with these distinction the background events are still dominant. The 2D heatmap in Figures 7 and 8 gives us all the BDT scores for signal and background respectively. Now looking at Figure 9 the plots do not show any

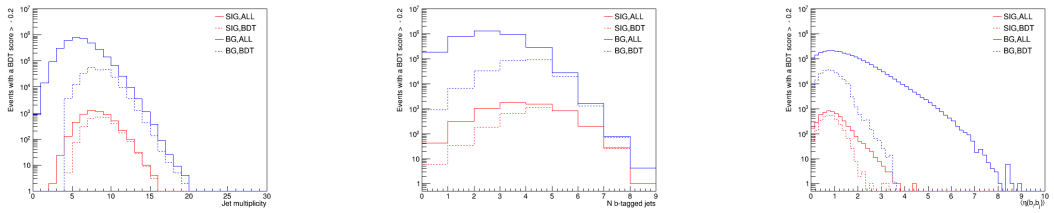


Figure 6: Left: jet multiplicity, number of jets. Middle: N_b, number of jets being reconstructed from a bottom quark. Right: $\langle \eta(b_i, b_j) \rangle$, average separation in pseudorapidity between two b-tagged jets calculated using Equation (3). Solid lines represent full data while dashed lines represent data with a BDT score greater than -0.2.

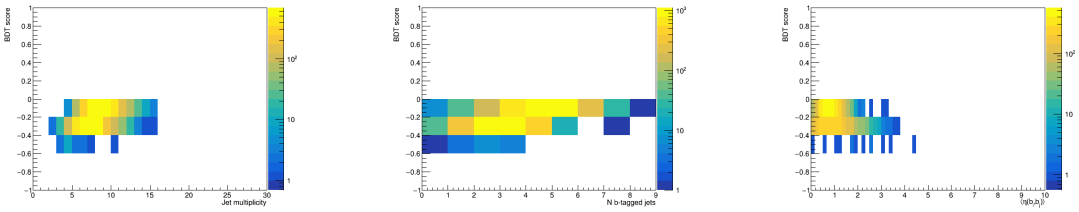


Figure 7: A 2D heatmap with BDT scores on the Y-axis; Same X-axis as in Figure 6.; logarithmic color scale for number of events. Number of signal events and BDT score for Phase 4.

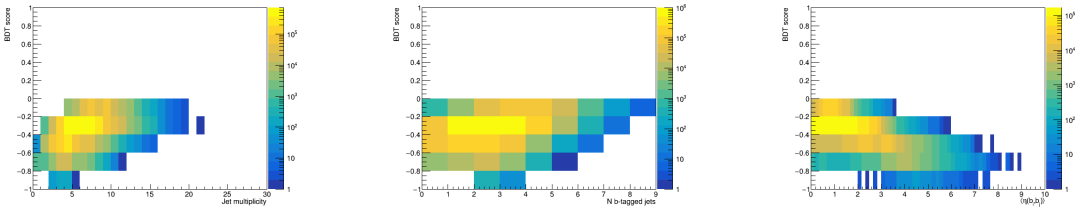


Figure 8: A 2D heatmap with BDT scores on the Y-axis; Same X-axis as in Figure 6.; logarithmic color scale for number of events. Number of background events and BDT score for Phase 4.

obvious distinction and the background is still dominant. The BDT makes it possible to find connections between event shape variables that we cannot see by looking at 1D plots of event shape variables. This can be seen in Figure 5 where the Centrality variable is crucial, yet looking at Figures 9 to 11 it is not obvious. Again, the Figure 12 shows

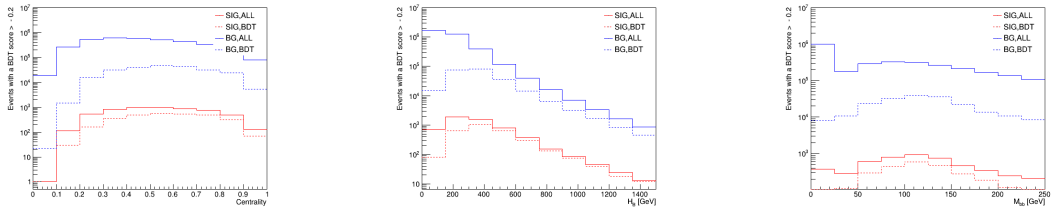


Figure 9: Left: Centrality; Middle: H_B; Right: M_{bb}. All calculated using Equations (4) to (6). Solid lines represent full data while dashed lines represent data with a BDT score greater than -0.2.

no distinction and these event shape variables are not crucial as seen in Figure 5. The only feature that stands out is majority of events have a small ΔR as seen in Figures 12 to 14. Lastly, Figures 15 to 17 show no distinctions and are ranked low Figure 5.

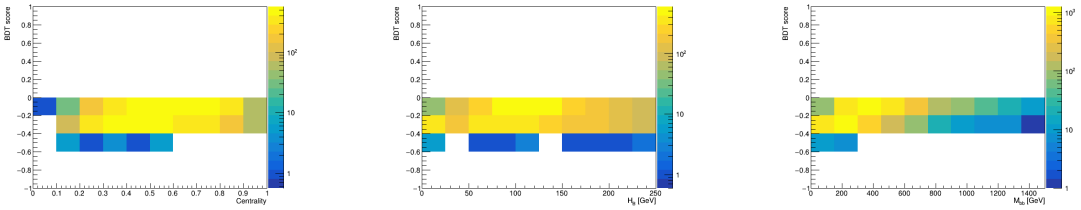


Figure 10: A 2D heatmap with BDT scores on the Y-axis; Same X-axis as in Figure 9.; logarithmic color scale for number of events. Number of signal events and BDT score for Phase 4.

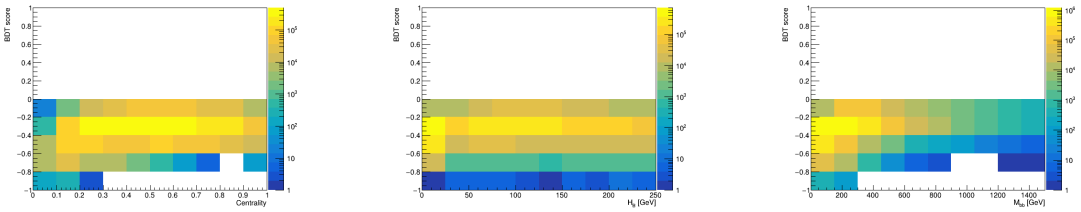


Figure 11: A 2D heatmap with BDT scores on the Y-axis; Same X-axis as in Figure 9.; logarithmic color scale for number of events. Number of background events and BDT score for Phase 4.

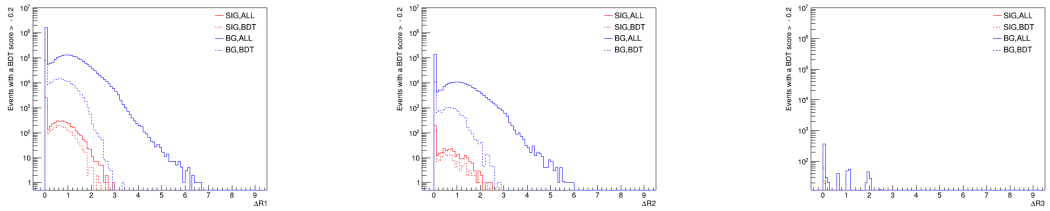


Figure 12: Left: $\Delta R1$; Middle: $\Delta R2$; Right: $\Delta R3$. All calculated using Equation (7). Solid lines represent full data while dashed lines represent data with a BDT score greater than -0.2.

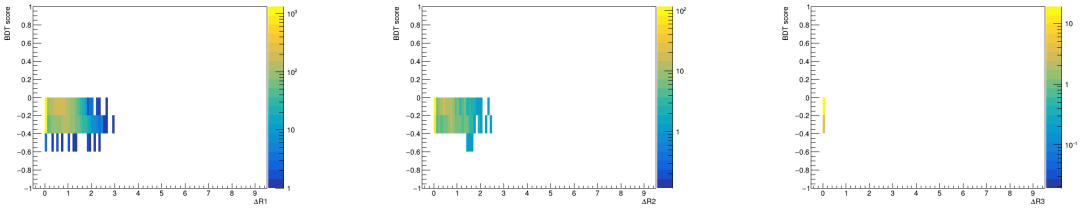


Figure 13: A 2D heatmap with BDT scores on the Y-axis; Same X-axis as in Figure 12.; logarithmic color scale for number of events. Number of signal events and BDT score for Phase 4.

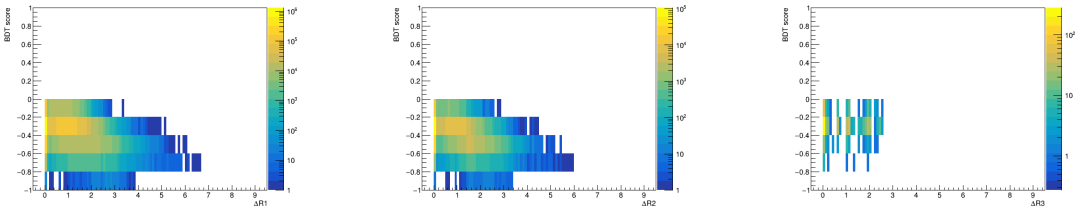


Figure 14: A 2D heatmap with BDT scores on the Y-axis; Same X-axis as in Figure 12.; logarithmic color scale for number of events. Number of background events and BDT score for Phase 4.

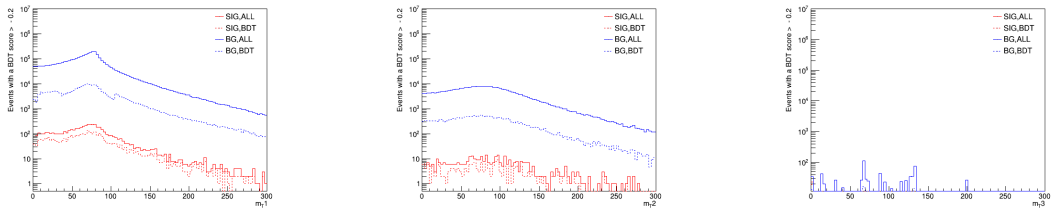


Figure 15: Left: m_T1; Middle: m_T2; Right: m_T3. All calculated using Equation (8). Solid lines represent full data while dashed lines represent data with a BDT score greater than -0.2.

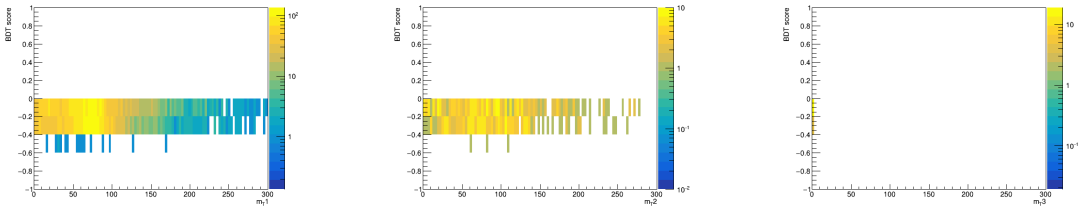


Figure 16: A 2D heatmap with BDT scores on the Y-axis; Same X-axis as in Figure 15.; logarithmic color scale for number of events. Number of signal events and BDT score for Phase 4.

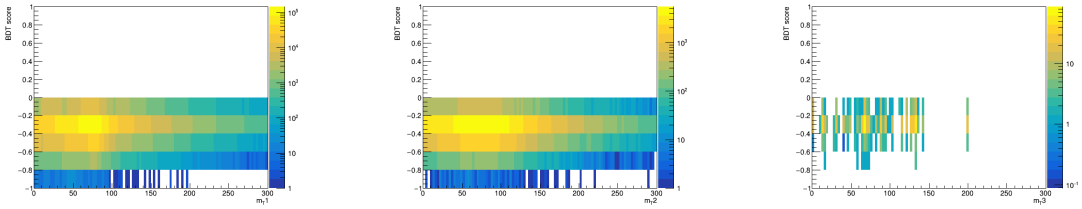


Figure 17: A 2D heatmap with BDT scores on the Y-axis; Same X-axis as in Figure 15.; logarithmic color scale for number of events. Number of background events and BDT score for Phase 4.

Even though some of these figures do not show any distinctions, the expected behavior of the figures can assure us that there are no artifacts that are biasing the analysis. The ROC curve in Figure 18 has an area of 0.86 (unit-less). Compared to [3] at the same background efficiency we find a signal efficiency that is larger by a factor of 1.003, which tells us that the BDT and the cut based analysis at the same background efficiency are effectively identical. The significance of the cut based analysis reported in the [3] is 0.26 σ. Scanning the ROC curve to find the an estimated max significance, which include no systematic uncertainties on the background, we find a maximum significance of 0.35 σ.

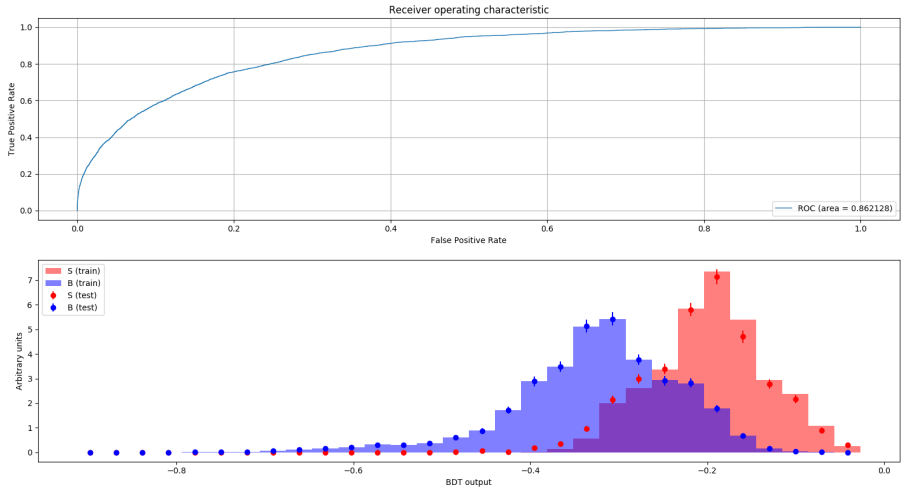


Figure 18: Trained and tested using Phase 4. Top: ROC curve; True positive rate, False positive rate are the efficiencies of the signal and background, respectively. Bottom: BDT output distribution used for over-training checks.

The computation and time restraints prevent us from using all low level vari-

ables and some high level variables. Using the event shape variables that have shown promise in the cut base analysis are fed to the BDT to hopefully optimise time and or area under the ROC curve in Figure 18.

5

Conclusion

The cut-based analysis has been shown to be limited on event shape variables that have a clear discrimination feature. Improving this analysis by creating new event shape variables is an inefficient method alone. However, with the help of machine learning we can discover the importance of some event shape variables. The cut-based analysis alone was optimised for a statistical significance of $t\bar{t}HH$ production using $S/\sqrt{B}$, which resulted in a 0.26 σ. With the use of a BDT the statistical maximum significance of $t\bar{t}HH$ production resulted in a 0.35 σ with no systematic uncertainties on the background. With the limited event shape variables fed to the BDT, we find that a BDT can beat a cut-based analysis. The $t\bar{t}HH$ production mode looks promising in contributing with measurements of the Higgs self-coupling at the HL-LHC.

Although only a selected amount of variables were used in this thesis adding both low and high level variables should continue to improve the sensitivity. Testing other algorithms and classifiers along with testing and training hyperparameters can be

studied to converge more quickly on an optimal BDT configuration. This could then allow for more high level event shape variables to be used. In this thesis a BDT was used for machine learning, other methods like deep learning have shown to have promise with particle physics [5, 4].

Bibliography

[1] ATLAS Collaboration. Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC. *Phys. Lett. B.* 716 (2012). DOI: 10.1016/j.physletb.2012.08.020.

[2] ATLAS Collaboration. Optimisation of the ATLAS b-tagging Performance for the 2016 LHC Run. ATLAS-PHYS-PUB-2016-012. 2016. URL: http://cds.cern.ch/record/2160731.

[3] ATLAS Collaboration. Prospects for Observing $t\bar{t}HH$ Production with the ATLAS Experiment at the HL-LHC. ATL-PHYS-PUB-2016-023. 2016. URL: http://cds.cern.ch/record/2220969.

[4] P. Baldi, P. Sadowski, and D. Whiteson. Enhanced Higgs Boson to $\tau^+\tau^-$ Search with Deep Learning. *Phys. Rev. Lett.* 114.11 (2015). DOI: 10.1103/PhysRevLett.114.111801.

[5] Dan Guest, Kyle Cranmer, and Daniel Whiteson. Deep Learning and Its Application to LHC Physics. *Annu. Rev. Nucl. Part. Sci.* 68.1 (2018). DOI: 10.1146/annurev-nucl-101917-021019.

[6] Trevor J. Hastie et al. Multi-class AdaBoost. *Statistics and Its Interface* 2 (2009). DOI: 10.4310/SII.2009.v2.n3.a8.

[7] Tao Liu and Hao Zhang. Measuring Di-Higgs Physics via the $t\bar{t}hh \to t\bar{t}b\bar{b}b\bar{b}$ Channel. arXiv: 1410.1855 [hep-ph]. 2014.

[8] F. Pedregosa et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12 (2011).

[9] Donald H Perkins. Particle Astrophysics. Oxford University Press, 2009.

[10] Michael E Peskin and Daniel V. Schroeder. An Introduction to Quantum Field Theory. CRC press, 2018.