Optimization Studies for the H $\rightarrow$ WW Boosted Decision Tree Analysis

Jessica Strickland

October 18, 2014

Abstract

The aim of this project was to follow the ATLAS $H \rightarrow WW$ boosted decision tree (BDT) analysis and try to optimize training variables, pre-selection cuts, and training parameters [1, 2]. Machine learning was done with Monte Carlo samples for the $H \rightarrow W^+W^- \rightarrow e\mu\nu_e\nu_\mu$ channel. A multivariate analysis was executed by way of boosted decision tree in an attempt to improve the original ATLAS $H \rightarrow WW$ BDT analysis. The goal of the BDT is to separate the Higgs signal from the continuous WW background only. Once an optimal set-up was found and used for training, the weights produced from the BDT output were used categorize an unknown data set. Finally, region cuts were made to the final BDT output to observe the performance of the training, and it appeared to perform well.

Contents

Acknowledgments

I would like to extend my gratitude to Prof. Michel Lefebvre for his supervision and guidance and to Dr. Manuela Venturi as well during my time at CERN. The time they dedicated is truely appreciated. I would also like to thank Ewan Hill and Tony Kwan for their assistance. Finally, I would like to thank the Institute of Particle Physics (IPP) and the National Science and Engineering Research Council of Canada (NSERC).

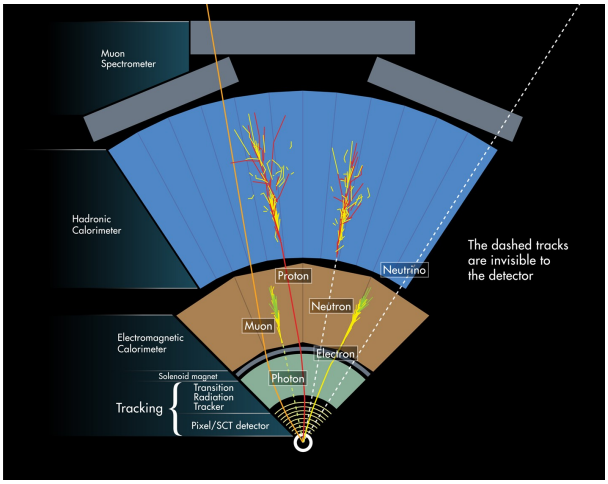


Figure 1: How the sections of the ATLAS experiment detect particles.

1 Introduction

1.1 The LHC and ATLAS

To begin I will introduce ATLAS, the experiment which I am doing this project for. The Large Hadron Collider (LHC) has 4 main experiments, one of which is ATLAS (A Toroidal LHC Apparatus), a general purpose dector. The LHC is the world's largest and highest energy particle accelerator. It operated at $\sqrt{s} = 8$ TeV in 2012, and is aspiring to 14 TeV during the second run starting in June, 2015. Proton beams are accelerated by the LHC and subsequently collided at the ATLAS interaction point. Given the overwhelming amount of data produced in the collision, two levels of pre-selection triggers are applied, followed by an event filter stage which is able to reduce the rate to 200 events/s. To detect a wide variety of phenomena, good resolution for leptons, photons, and jets is required. The detector also includes a very accurate muon sub-detector. Such an excellent resolution is crucial for measuring the missing energy of neutrinos which is also used in this project. ATLAS has multiple layered components to detect different particles (see Figure 1):

- Inner Detector: the tracker is used to accurately measure the trajectory of charged particles. It uses a 2 T solenoid magnetic field to bend the particles in the plane perpendicular to the beam axis (transverse).
- Calorimetry: there is an electromagnetic calorimeter and a hadronic one. The information they provide is crucial to measure the energy of electrons, photons, jets, and to calculate the missing energy in the transverse plane.
- Muon Spectrometer: it is located in the outer part of the ATLAS detector to absorb and measure the momentum of muons.

1.2 The $H \to W^+W^- \to e\mu\nu_e\nu_\mu$ Channel

In this project I worked with the channel $H \to W^+W^- \to e\mu\nu_e\nu_\mu$ as it is one of the most important channels for the decay of the recently discovered, SM Higgs boson. The signature of the $H \to WW$ channel is characterized by two oppositely charged leptons and their corresponding neutrinos, as shown in Figure 2. For this project, only the eμ final state with no accompanying jets was considered, since it is the most sensitive. As shown in Figure 3, the main background is the continuous WW production [1]. Other backgrounds of this final state include the top and anti-top quark, W^+ jets, and diboson processes. Due to its dominance, the project focused only on the continuous WW production [1] in the qq and gg channels and did not take other backgrounds into account. In contrast, the standard analysis focuses on all backgrounds concurrently.

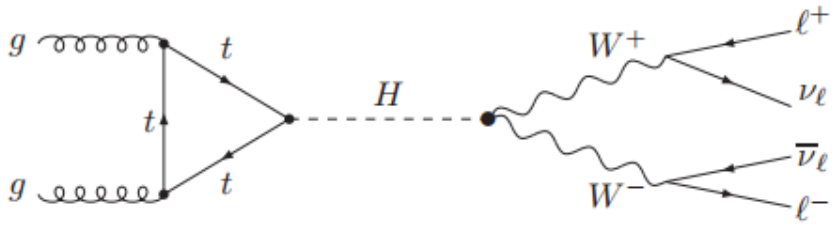


Figure 2: A Feynman diagram depicting the production of the Higgs boson via gluon fusion, then decaying to two W bosons, and lastly to two leptons and their corresponding neutrinos.

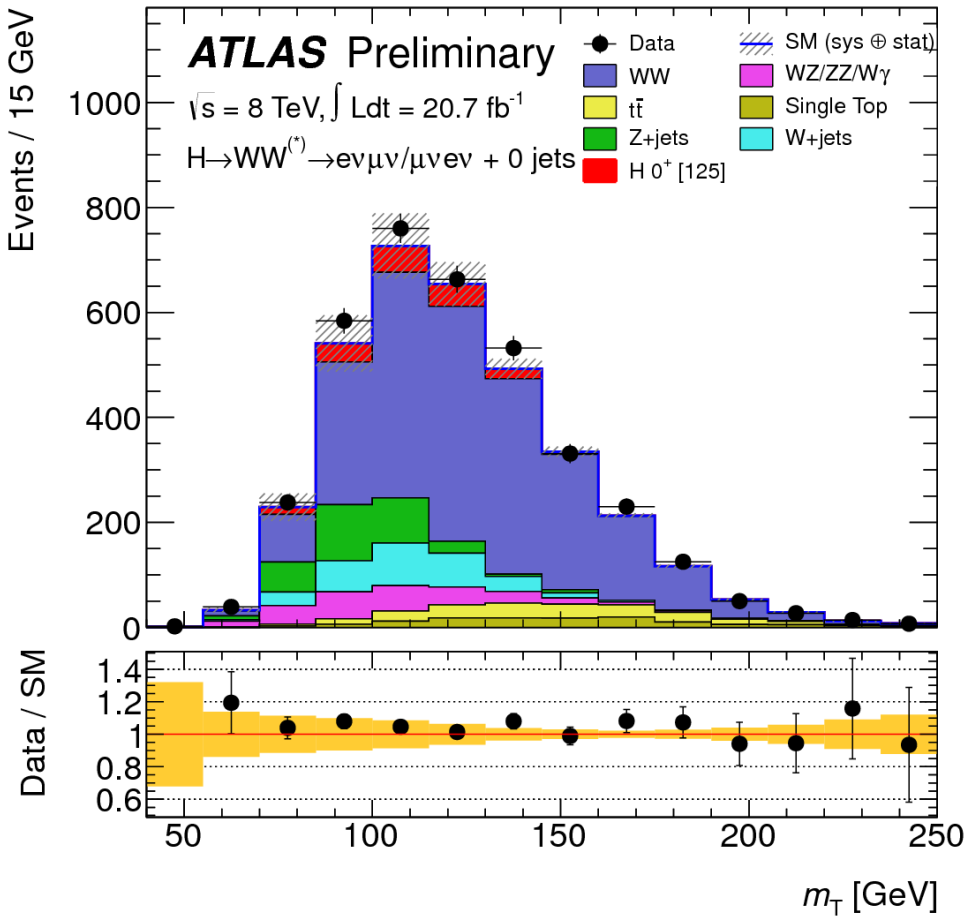


Figure 3: Distributions of the transverse mass for the $H \to WW$ channel. The excess shown in red is the presence of the Higgs boson as signal.

2 Analysis

2.1 TMVA and BDTs

Following the original analysis, this project employed Multivariate Analysis and Boosted Decision Trees. "TMVA provides a ROOT-integrated environment for the processing, parallel evaluation and application of multivariate classification and regression techniques" [3]. In particular, the technique used was the decision tree, which was boosted in order to reduce the impact of statistical fluctuations. A BDT is a progression of binary decisions involving one variable at a time until a stop criterion is satisfied. The trees are finally combined into a single classifier which is given by a (weighted) average of the individual decision trees (training) [3]. Once the machine has *learned* through by training with a known set of signal and background, it is used to classify an unknown set.

2.2 Standard Four Variables

The original ATLAS $H \to WW$ BDT analysis used four variables during the training, listed below. The distributions for each can be seen in Figure 4.

- m_T: this variable is the transverse mass and is very important in this channel. Essentially, the final state mass cannot be fully reconstructed due to the presence of the two neutrinos, however, m_T is a good approximation. m_T is the mass in the x and y plane, where $m_T < m_H$. You can see in Figure 7 that the signal peaks a little below the Higgs mass (~ 125 GeV).

$$m_T = \sqrt{E^2_{T,\ell\ell+\nu\nu} - |\overrightarrow{p}_{T,\ell\ell+\nu\nu}|^2} = \sqrt{(E_{T,\ell\ell+\nu\nu} - E_T^{miss})^2 - |\overrightarrow{p}_{T,\ell\ell+\nu\nu} + \overrightarrow{p}_T^{miss}|^2}$$

- $m_{\ell\ell}$: the invariant mass of the two lepton system, e and μ.
- $\Delta\phi_{\ell\ell}$: the azimuthal angle between the two leptons, e and μ.
- $p_{T,\ell\ell}$: the combined transverse momentum of the two lepton system, e and μ.

$$p_{T,\ell\ell} = \sqrt{(p_{x,e} + p_{x,\mu})^2 + (p_{y,e} + p_{y,\mu})^2}$$

2.3 BDT Output and ROC Curves

There are methods of interpreting the effectiveness of the training, such as examining the BDT output and a receiver operating characteristic (ROC) curve. BDT training is a machine learning algorithm which learns how to discriminate between signal and background events using the variables provided. It is seen, in Figure 5, that the trained BDT did a fairly good job separating the signal from the background. Most, but not all, of the background events have a different BDT output value than the signal, concentrated near negative one.

In order to evaluate the performance, the metric for optimization is the ROC curve: a representation of the signal efficiency at a given background rejection (see Figure 6). The points that make up the ROC curve are the corresponding BDT output values that each have a ratio of signal to background. By maximizing the area under the ROC curve, we maximize the efficiency, which implies a higher separation power of the BDT.

3 Results

3.1 Optimization Adjustments

3.1.1 Optimum Combination of Variables

Many different combinations of variables were tested to find a set that optimized the training. More variables were added to the basic set mentioned in section 2.2, and the BDTs were then compared using ROC curves. At most, 7 variables were used because increasing the number of input variables further

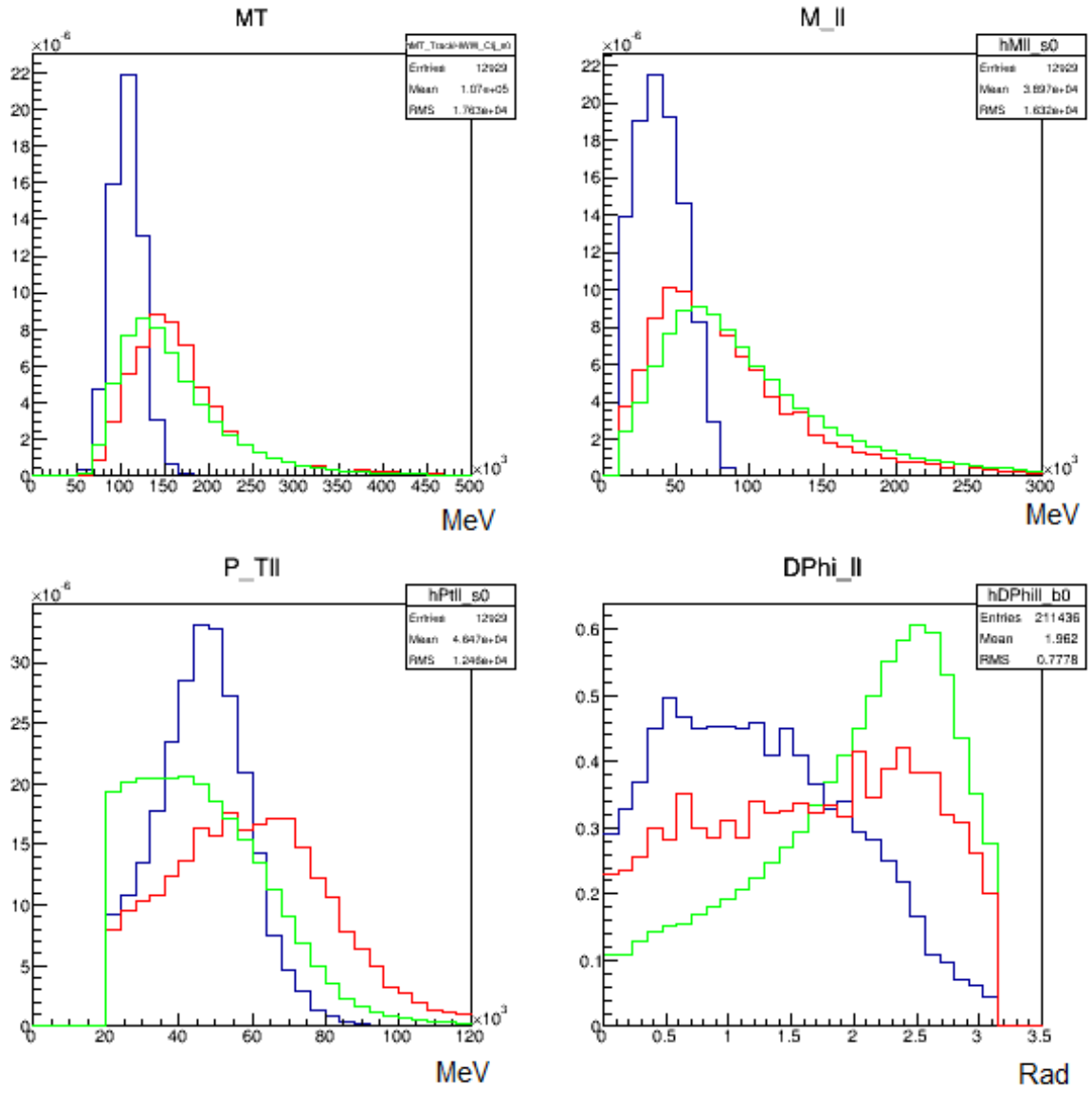


Figure 4: Distributions of the 4 standard variables for the $H \to W^+W^-$ channel displaying the signal in blue, ggWW background in green, and qqWW background in red.

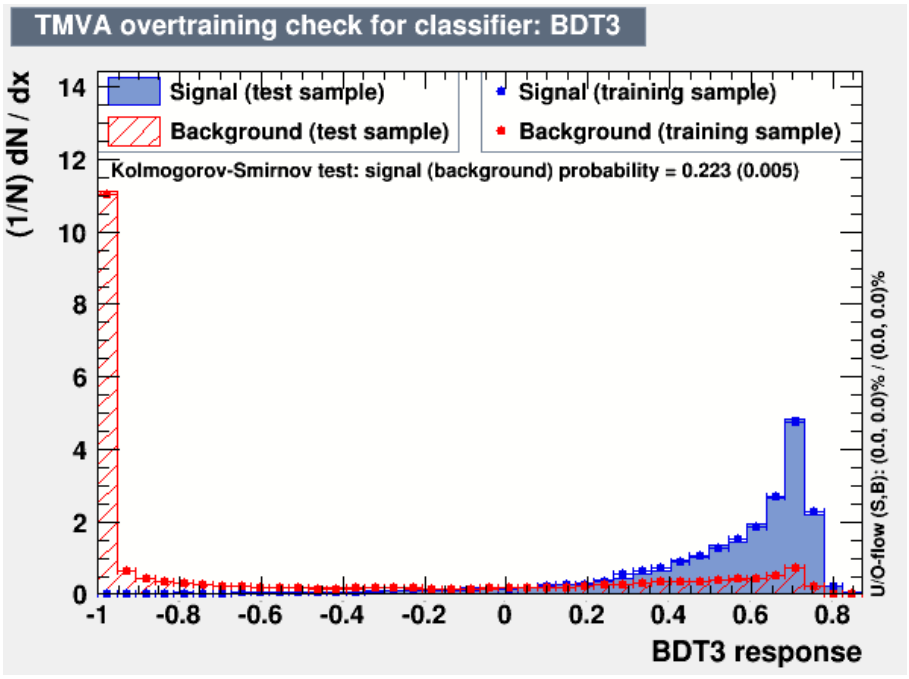


Figure 5: Given signal and background Monte Carlo samples, this is the distribution of the BDT output for the training with optimal combination of 7 variables that will be later discussed in section 3.1.1. The signal is shown in blue and the background in red.

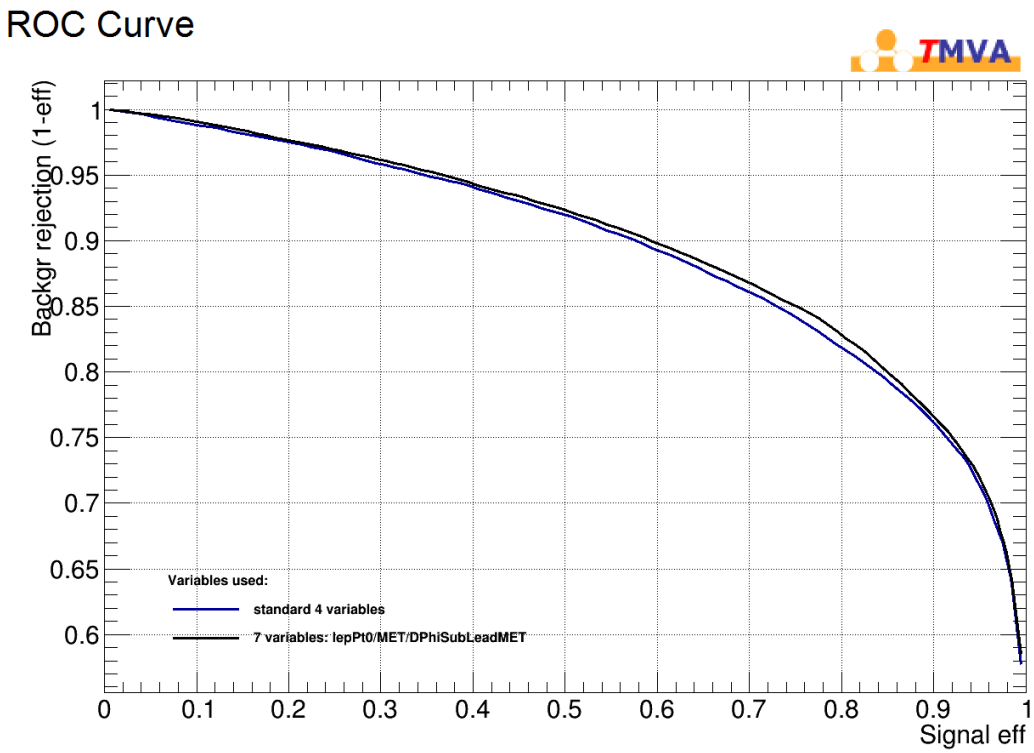


Figure 6: Two Receiver Operating Characteristic curves are displayed corresponding to the BDT output for training with the standard 4 variables and the optimal combination of 7 variables that will be later discussed in section 3.1.1.

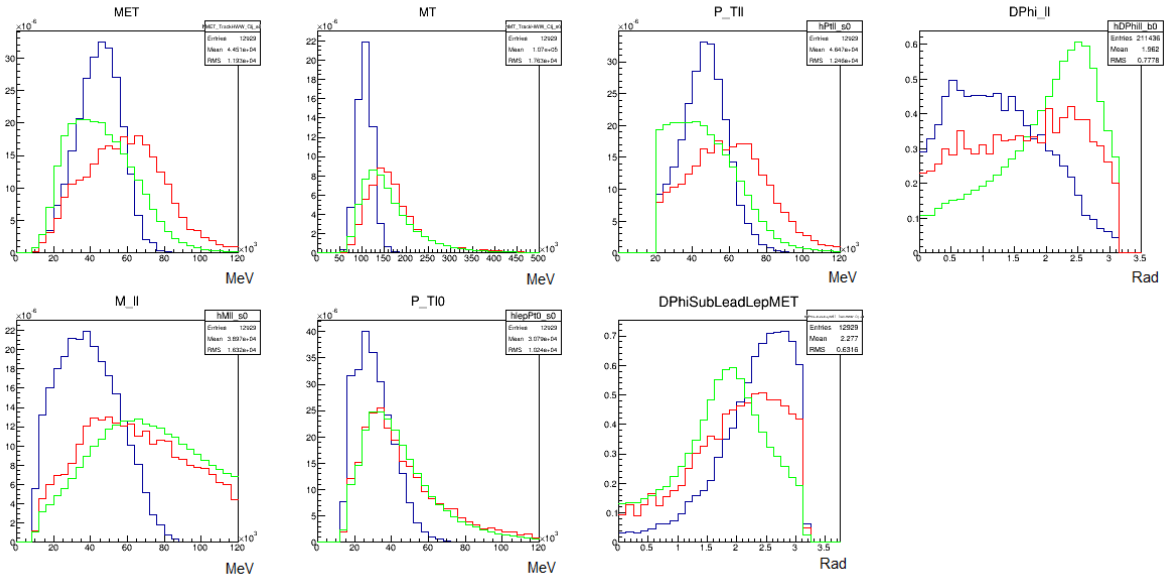


Figure 7: Distributions of the 7 variables of the optimal combination found for the $H \to W^+W^-$ channel displaying the signal in blue, the ggWW background in green, and the qqWW background in red.

does not result in an increase in the separation power as it ignores non-discriminating variables. The combination found to perform the best is (ranked during training with regards to variable importance):

1. m_T : transverse mass.
2. $\Delta\phi_{\ell\ell}$: azimuthal angle between the two leptons, e and μ.
3. $m_{\ell\ell}$: invariant mass of the two lepton system, e and μ.
4. MET: this variable is the missing transverse energy. The negative sum of the transverse momentum of all detected particles is equivalent to the sum of the transverse momentum of the neutrinos.
5. $\Delta\phi_{\ell 1,\mathrm{MET}}$: azimuthal angle between the sub-leading lepton and the missing transverse energy.
6. $p_{T,\ell\ell}$: transverse momentum of the two lepton system, e and μ.
7. $p_{T,\ell 0}$: transverse momentum of the leading lepton.

This combination gave the greatest separation power when compared to the performance of the standard 4 variables, as seen in the comparative ROC curves in Figure 6. All 7 variable distributions are shown in Figure 7, and it is clear that some variables have more contrasting shapes and values when comparing signal to the backgrounds. As shown above, MET and $\Delta\phi_{\ell 1,\mathrm{MET}}$ had more variable importance than the standard variables $p_{T,\ell\ell}$. In addition, if the number of variables used during training was reduced below 4, drastic decreases in separation power were observed. It was especially sensitive to the variable m_T. After many variations, this combination was then kept constant so other parameters could be manipulated.

3.1.2 Depth

Another training parameter that could be adjusted was the depth. The depth is the number of times that there is splitting at a node, or how many decisions that will be allowed during the training. A depth of 3 was used in the standard analysis to prevent overtraining of the BDT. If the depth is too large, the training can be too specific and will be according to the statistical fluctuations of that set. For this project, it was theorized that increasing the depth could allow for better separation power, as more variables were added. However, it was seen that increasing the depth up to 6 did not lead to a significant

Variable (GeV)	Standard Cut	Cut 1	Cut 2	Cut 3
MET	> 20	< 80	N/A	< 80
$m_{\ell\ell}$	> 10	N/A	< 100	< 90
$p_{T,\ell\ell}$	> 20	< 90	N/A	< 80

Table 1: In this table the standard cuts are shown for the three variables for which the thresholds were adjusted, MET, $m_{\ell\ell}$, and $p_{T,\ell\ell}$, in GeV. The cuts made were in addition to the standard cut. For example, Cut 2, as shown, entails MET > 20 GeV, $p_{T,\ell\ell} > 20$ GeV, and 10GeV $< m_{\ell\ell} < 100$ GeV.

enough improvement relative to the risk found of overtraining. It was concluded that the depth would remain at 3.

3.1.3 Pre-selection Cuts

Adjustments were also made to the pre-selection cuts. Pre-selection cuts are made to optimize BDT efficiency before the training. When manipulating other parameters, the standard pre-selection cuts were used:
$p_{T,\ell\ell} > 20$ GeV
$m_{\ell\ell} > 10$ GeV
$p_{T,\ell 0} > 22$ GeV
$p_{T,\ell 1} > 15$ GeV
MET > 20 GeV

Next, the variable combination and depth remained constant while changes were made to the thresholds in order to focus better on the signal and WW topologies [1, 2]. For example, the distribution for $m_{\ell\ell}$ in Figure 4, shows there is no signal above 100 GeV. Essentially, if there is data where $m_{\ell\ell}$ is greater than 100 GeV, the BDT does not need to try to determine whether or not it is signal, as it is definitely background. Because it is so distinctive, a cut can be made prior to training.

Different cuts were attempted and can be seen in Table 1. In all cases the standard pre-selection cuts produced better results. For example, in Figure 8, the ROC curves with the standard cut, for either the same set, or even the standard 4 variables performed better than Cut 2. Therefore, it was decided that the best training would occur if the standard pre-selection cuts were used.

3.2 Application and BDT Output

Once the BDT is trained it can be used to classify a new set. The weights produced when it *learned* from one half of the sample were then applied to the other half of the sample. The BDT variable that was generated was plotted for the signal and both background samples without any region cuts, as seen in Figure 9. These plots are stacked and scaled to event size in Figure 10. Sharp peaks show that the BDT has done a good job at separating the two. It is observed that most of the signal sample was classified as signal, and most of the background was interpreted as background. The signal is concentrated around 0.7, while the background is concentrated around –1. Next, region cuts were made to these plots in both the signal and WW control region.

3.3 Region Cuts and Purity

To further examine the effectivness of the BDT, we took the output and made region cuts. This means that cuts were made when producing the BDT output plots to focus on a certain region, much like zooming in on a particular region. This allowed us to investigate how much signal and background was present in the regions where you would expect them. One numerical representation of this is the purity.

$$P = \frac{S}{S+B}$$

The purity in the signal region, shown in Figure 11, is 0.135. To compare, the purity in the WW control region, shown in Figure 12, is 0.009 which is distinctly less. These results follow the expected outcome.

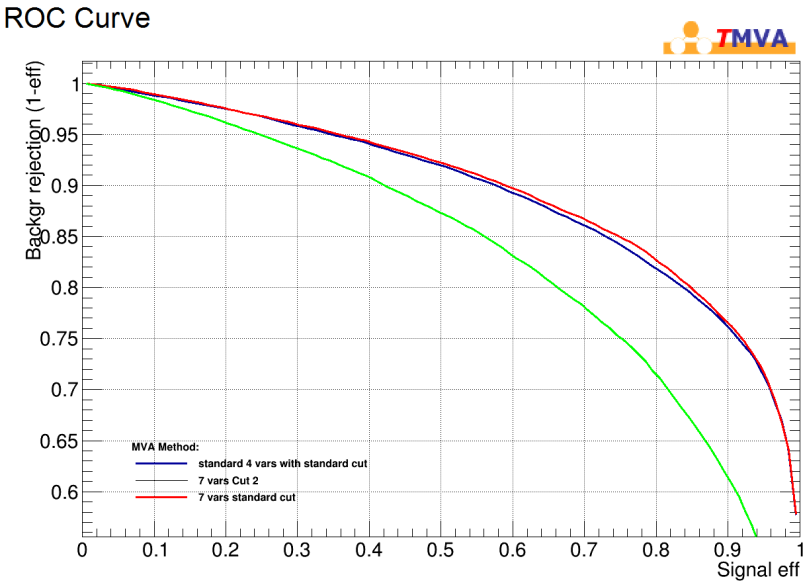
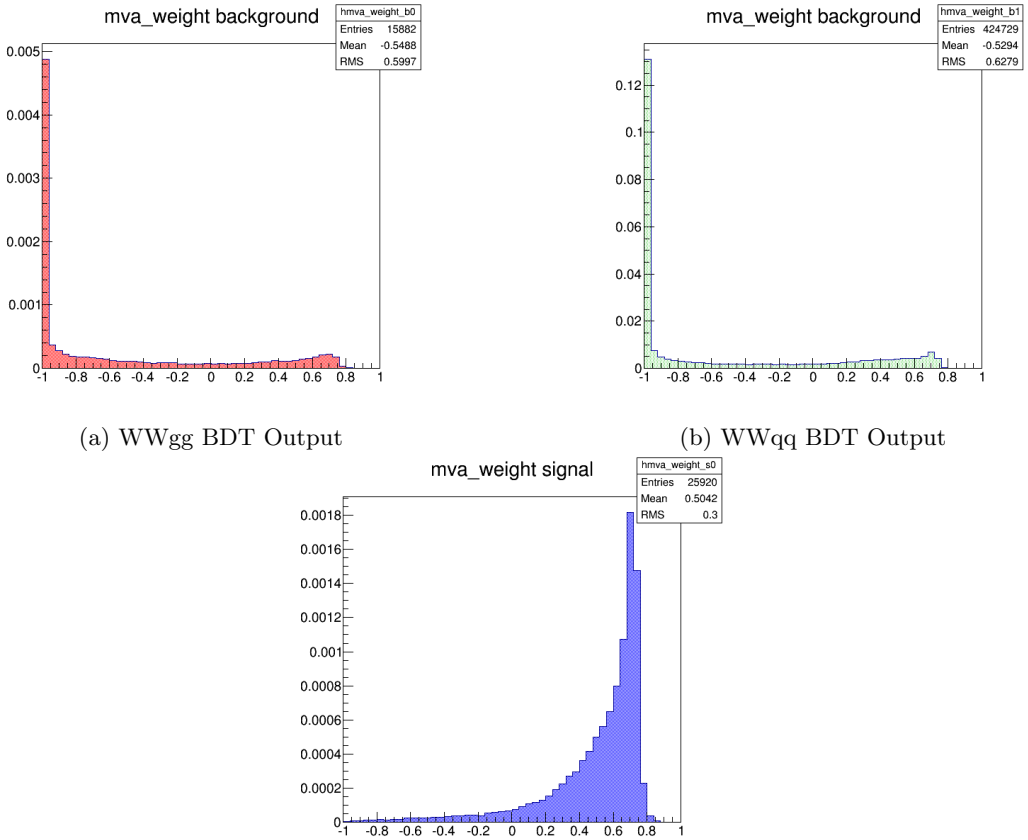


Figure 8: ROC curves of the standard 4 variables at a depth of 3, with standard cuts (blue), the 7 variables of the optimal combination listed in section 3.1.1, at a depth of 3, with both standard cuts (red), and Cut 2 (green).



(a) WWgg BDT Output

(b) WWqq BDT Output

(c) ggH BDT Output

Figure 9: The BDT Output after applying the training to the other half of the sample.

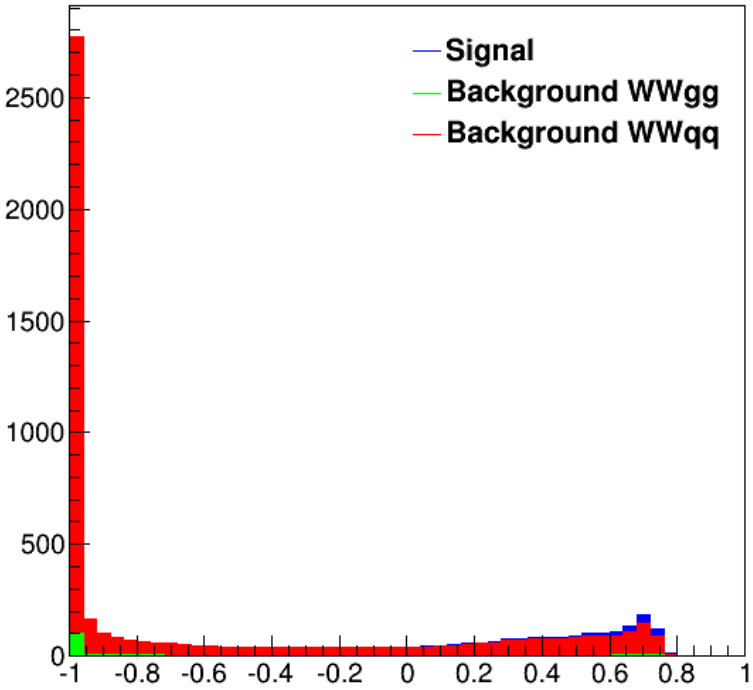


Figure 10: Stacked distributions of the BDT output for the signal and background for the $H \to W^+W^-$ channel scaled to event size, without any region cuts. This application was done after training with the optimal combination listed in 3.1.1, the standard pre-selection cuts, and a depth of 3.

It is seen that the BDT was successful as there is clearly an excess of signal in the signal region, and very little signal in the control region.

4 Conclusion

During the training I adjusted the variables used, pre-selection cuts, and depth to optimize the ability of the BDT to separate signal from background in the $H \to W^+W^- \to e\mu\nu_e\nu_\mu$ channel. ROC Curves and the BDT output were used to determine the most efficient setting of the training parameters. The best performance was obtained by training at a depth of 3, with the standard pre-selection cuts ($p_{T,\ell\ell} > 20$ GeV, $m_{\ell\ell} > 10$ GeV, $p_{T,\ell 0} > 22$ GeV, $p_{T,\ell 1} > 15$ GeV, MET > 20 GeV) and the following 7 variables: m_T, $\Delta\phi_{\ell\ell}$, $m_{\ell\ell}$, MET, $\Delta\phi_{\ell 1,\mathrm{MET}}$, $p_{T,\ell\ell}$, $p_{T,\ell 0}$.

The weights produced while training one half of the Monte Carlo sample were used to classify the other half of the set. Then, I plotted the final BDT output and also made region cuts in both the signal and WW control region to observe the performance of the BDT. Sharp peaks in the BDT output show that the BDT has done a good job at separating the two and applying it to another sample. As expected, it was observed that there was an excess of signal in the signal region, and a deficit in the control region. The purity was 0.135 in the signal region and 0.009 in the WW control region.

Optimization is critical because machine learning is an efficient method used by ATLAS to explore important properties of particles, such as the Higgs boson. The goal of this project was to manipulate certain training parameters in order to improve the training and that has been achieved.

References

[1] The ATLAS Collaboration. Study of the spin properties of the Higgs-like particle in the $H \to WW^{(*)} \to e\mu\nu_e\nu_\mu$ channel with 21 $fb^{\neg 1}$ of $\sqrt{s} = 8$ TeV data collected with the ATLAS detector. ATLAS-CONF-2013-031, 2013.

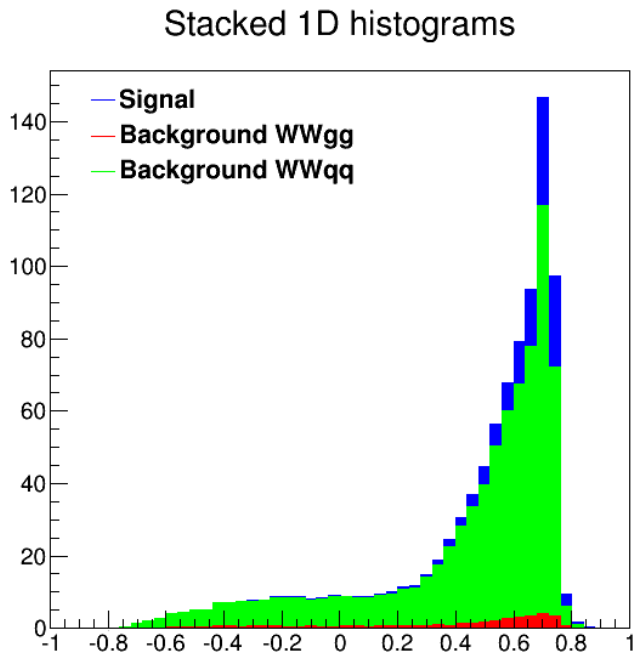


Figure 11: Just as figure 10, stacked distributions of the BDT output for the signal and background for the $H \to W^+W^-$ channel scaled to event size. This application was done after training with the 7 variables of the optimal combination found, the standard pre-selection cuts and a depth of 3. However, it is now plotted within the signal control region. **Signal Region:** $p_{T,ll} > 20$ GeV, $m_{ll} > 55$ GeV, $p_{T,l0} > 22$ GeV, $p_{T,l1} > 15$ GeV, MET> 20GeV, $\Delta\phi_{ll} > 1.8$ rad, 93.75 GeV $< m_T < 150$ GeV.

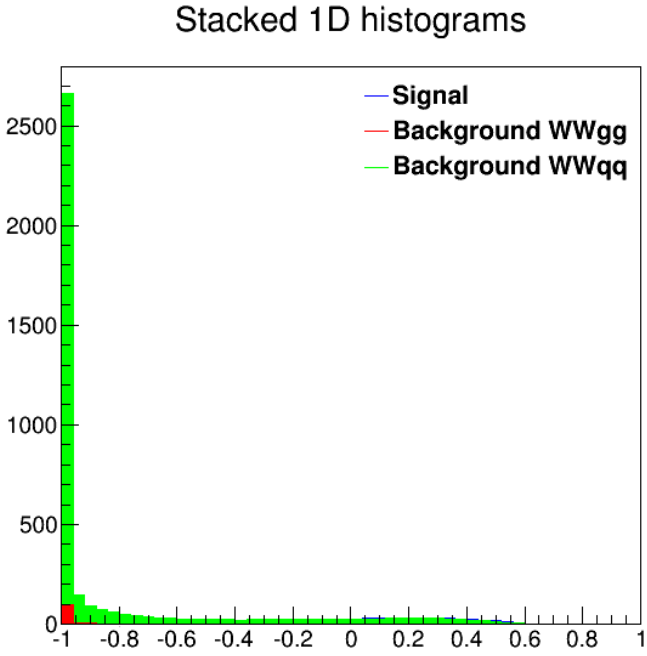


Figure 12: Just as figure 10, stacked distributions of the BDT output for the signal and background for the $H \to W^+W^-$ channel scaled to event size. This application was done after training with the 7 variables of the optimal combination found, the standard pre-selection cuts and a depth of 3. However, it is now plotted within the WW control region. **WW Control Region:** $p_{T,ll} > 20$ GeV, 10 GeV $< m_{ll} < 55$ GeV, $p_{T,l0} > 22$ GeV, $p_{T,l1} > 15$ GeV, MET > 20 GeV.

[2] The ATLAS Collaboration. Evidence for the spin-0 nature of the Higgs boson using ATLAS data. Phys.Lett., B726:120144, 2013.

[3] A. Hoecker and others. TMVA 4: Toolkit for Multivariate Data Analysis with ROOT. arXiv:physics/0703039v5,2009.