Search Optimization using Boosted Decision Trees for a Vector Boson Fusion SUSY Signature

Emily Thompson*

Advisor: Dr. Christian Sander

DESY - Hamburg, Germany

September 7, 2017

Abstract

The purpose of this project is to investigate the use of boosted decision trees in the search for charginos and neutralinos originating from a vector-boson fusion (VBF) process in the ATLAS detector. Charginos produced in the VBF process are assumed to decay with a branching fraction of 100% to staus and tau sneutrinos. An iterative removal algorithm is used to determine a ranking of the optimal variables to train the boosted decision trees. Two reference points in compressed SUSY scenarios, both with a mass splitting of Δm = 50 GeV, are investigated. The optimal analysis for the reference point $m_{\tilde{\chi}_1^0}$ = 100 GeV, $m_{\tilde{\chi}_1^\pm}$ =$m_{\tilde{\chi}_2^0}$ = 150 GeV is found to have a significance of 3.65, and the optimal analysis for the reference point $m_{\tilde{\chi}_1^0}$ = 150 GeV, $m_{\tilde{\chi}_1^\pm}$ =$m_{\tilde{\chi}_2^0}$ = 200 GeV is found to have a significance of 2.03.

*s6emthom@uni-bonn.de

Contents

1 Introduction

In the following subsections, the necessary theoretical background is summarized. First, the Minimal Supersymmetric Standard Model (MSSM) is briefly motivated and explained. Then, the motivation for this specific search analysis involving vector boson fusion is discussed. Finally, the main analysis method used in this project, boosted decision trees, is described.

i The Minimal Supersymmetric Standard Model

Supersymmetry (SUSY) is an extension of the Standard Model (SM), in which the space-time Poincaré symmetry is extended by a symmetry that transforms fermions into bosons, and vice versa. SUSY is motivated by the fact that it can provide a framework for the unification of particle physics and gravity, as well as an explanation for the hierarchy problem arising from the large difference between the electroweak symmetry breaking scale (~ 100 GeV) and the Planck scale ($\sim 10^{19}$ GeV) [1]. If SUSY was an exact symmetry of nature, then all superparticles would have the same mass as their SM counterparts. As no superparticles have been observed yet, SUSY must be a broken symmetry, allowing for SUSY particles to be much heavier than their SM counterparts. While there are many versions of SUSY, this analysis is constrained to consider the Minimal Supersymmetric

Standard Model, which is the simplest extension of the SM. The particle content of the MSSM is briefly summarized as follows:

- The gauge supermultiplets consist of gluons and their *gluino* fermionic superpartners, as well as SU(2)$\times$U(1) gauge bosons that are accompanied by their *gaugino* fermionic superpartners. The *winos* are the superpartner particles of the W bosons ($W^{\pm}$ and W_0^3), while the *bino* is the superpartner of the weak hypercharge gauge boson [2].
- The matter supermultiplets consist of 3 generations of quarks and leptons and their superparticles, *squarks* and *sleptons*, along with their corresponding antiparticles [2].
- The MSSM contains two complex Higgs doublets, resulting in 5 Higgs bosons total: one CP-odd and two CP-even neutral Higgs bosons and two charged Higgs bosons, called *higgsinos* [2].
- The charged gauginos and the neutral gauginos mix with the higgsinos to create physical states of definite mass, namely the *charginos* ($\tilde{\chi}_i^{\pm}$, $i = 1, 2$ in order of increasing masses) and *neutralinos* ($\tilde{\chi}_j^0$, $j = 1, 2, 3, 4$ in order of increasing mass), respectively [2].

The MSSM Lagrangian is constructed using all possible 4-dimensional supersymmetric interaction terms that satisfy SU(3)$\times$SU(2)$\times$U(1) gauge invariance and $B-L$ conservation, where B is baryon number and L is lepton number. As a consequence of this $B - L$ invariance, one can also include another multiplicative quantum number, called R-parity, defined as $R = (-1)^{3(B-L)+2S}$, for a particle with spin S. Thus, the SM particles have $R = +1$ and their superpartner particles have $R = -1$. By requiring that R-parity is a symmetry, one can deduce that a superparticle must decay into an odd number of other superparticles, along with any number of SM particles. Therefore, the lightest supersymmetric particle (LSP) must be stable, and is thus a good candidate for dark matter. A candidate for the LSP in the MSSM is the lightest neutralino ($\tilde{\chi}_1^0$), which is a mixture of the bino, the neutral wino, and the neutral higgsinos [2]. The LSP is also considered to be only weakly interacting. Thus, if any LSPs are created at the Large Hadron Collider, they will traverse the ATLAS detector without leaving a signal.

ii Motivation for Analysis

This analysis targets the production of charginos and neutralinos in vector boson fusion processes. To date, the masses of the strongly produced gluinos and first- and second-generation squarks have been excluded up to 1 TeV [1]. However, the mass limits on the weakly produced charginos and neutralinos are considerably lower, because these processes have smaller cross sections at the Large Hadron Collider. The mass limits for charginos and neutralinos are summarized in Fig. 1.

These mass limits are especially low in the compressed mass spectrum scenarios of SUSY, where the mass of the LSP is only slightly lighter than the other SUSY particles. These scenarios have mass points close to the line labeled $\mathrm{m}_{\tilde{\chi}_2^0} = \mathrm{m}_{\tilde{\chi}_1^0}$ in Fig. 1. Previous searches for charginos and neutralinos have mainly focused on direct production processes, where the final states are characterized by a large missing transverse energy from the escaping LSP and leptons with high transverse momenta that arise from the decay chains of the

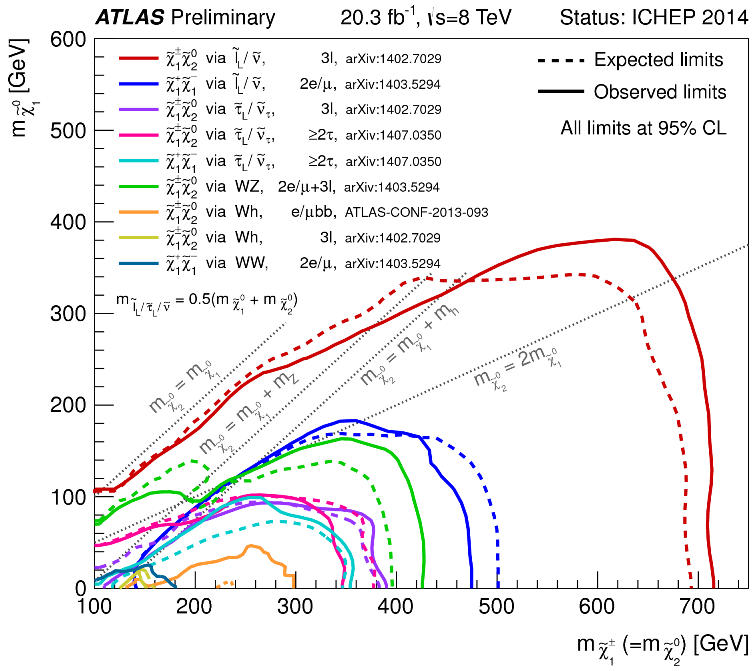


Figure 1: Summary of the ATLAS search for electroweak production of charginos and neutralinos. Each color corresponds to a different topology and decay channel scenario. The dashed and solid lines show the expected and observed limits, respectively.

superparticles. This search, however, is only sensitive to scenarios where the mass splitting between the LSP and the other supersymmetric particles is relatively large. If this mass splitting is small, then the leptons from the decay chain will have low transverse momenta, making the triggering of these events difficult.

A solution to this is to search for the production of charginos and neutralinos in vector boson fusion. Vector boson fusion offers a unique signature of two forward jets in opposite hemispheres of the detector, thus resulting in a high dijet invariant mass. These features make VBF production of charginos and neutralinos a promising probe for compressed SUSY scenarios that would otherwise be too experimentally difficult to explore via direct production. While there are multiple different VBF SUSY signatures, the specific topology investigated in this analysis is illustrated in Fig. 2.

In this topology, two W bosons fuse to create two charginos of either same or opposite sign. The most probable decay of a chargino is currently unknown. In this search we assume that the chargino decays with a branching fraction of 100% to staus or tau sneutrinos. This is theoretically motivated because SUSY scenarios with a small mass splitting and a light stau are favored in coannihilation processes that set the dark matter relic density to be consistent with experimental observations [3]. The mass limits for this decay process are depicted by the pink and purple lines in Fig. 1. The two reference points investigated in this report are as follows:

- Reference point 1 (RP1): $m_{\tilde{\chi}_1^0} = 100$ GeV, $m_{\tilde{\chi}_1^{\pm}} = m_{\tilde{\chi}_2^0} = 150$ GeV
- Reference point 2 (RP2): $m_{\tilde{\chi}_1^0} = 150$ GeV, $m_{\tilde{\chi}_1^{\pm}} = m_{\tilde{\chi}_2^0} = 200$ GeV

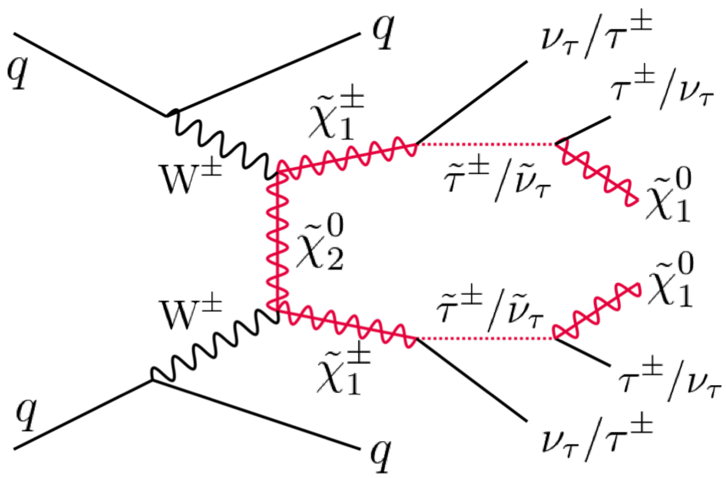


Figure 2: Feynman diagram of example VBF production studied in this analysis

Both reference points have a mass splitting of Δm = 50 GeV and are thus classified as compressed SUSY scenarios.

iii Multivariate Analysis: Boosted Decision Trees

In this analysis, the method of Boosted Decision Trees within the Toolkit for Multivariate Analysis (TMVA) is used. TMVA is a ROOT-integrated analysis toolkit that hosts a large variety of multivariate classification algorithms [4]. One such algorithm is boosted decision trees (BDT). A BDT is a binary tree structured classifier, that splits the phase space into many regions that are eventually classified as signal or background.

The TMVA BDT algorithm works as follows. The data is first split into training and testing subsets. Training starts at the root node, where the algorithm scans through all kinematic variables provided to it, and chooses the variable and cut that maximizes the separation between signal and background in the training data subset. The separation of each proposed cut is evaluated using the *Gini Index*, defined by $p \cdot (1-p)$, where the purity, p, is the ratio of signal events to all events in the node. The sample is then divided according to the optimal cut criterion into two new nodes, where one is signal-like and the other is background-like. This process continues until either a node has reached a pre-specified minimum number of events, or a maximum depth in the tree. These limits are set to help prevent the tree from *overtraining*, which occurs when the tree has learned statistical fluctuations in the training sample. After one of these limits has been reached, all of the events in the *leaf* node are then characterized as signal or background. The final collection of nodes is called a decision tree.

In order to stabilize the classification performance with respect to statistical fluctuations in the training sample, a method called *boosting* is used. After the first tree is trained, the events are re-weighted such that the signal events that were misclassified as background (and vice versa) receive a higher weight. Then, another tree is grown using these new event weights. This continues iteratively until a *forest* of trees is created.

Based on the output of all trees in the forest, each event is assigned a BDT response value ranging from -1 to +1. Background-like events will have a BDT response value shifted towards -1, and signal-like events will have a response value shifted towards +1. The BDT classifier then creates a *receiver operating characteristic* curve (ROC) by applying sequential cuts to the BDT response value and calculating the respective signal efficiencies and background rejections. The integral of the ROC curve indicates the level of performance of the classifier; the closer the integral is to 1, the better the classifier has performed. While there are other performance indicators provided by the TMVA BDT output, the ROC integration is the only indicator used in this analysis.

2 Analysis Procedure

In this section, the exact steps of the analysis are detailed. First the generation of the signal and background sample events are briefly outlined. Then, the kinematic variables considered in this analysis are discussed. Finally, the method used to determine the most discriminating variables, iterative removal, is described.

i Sample Generation

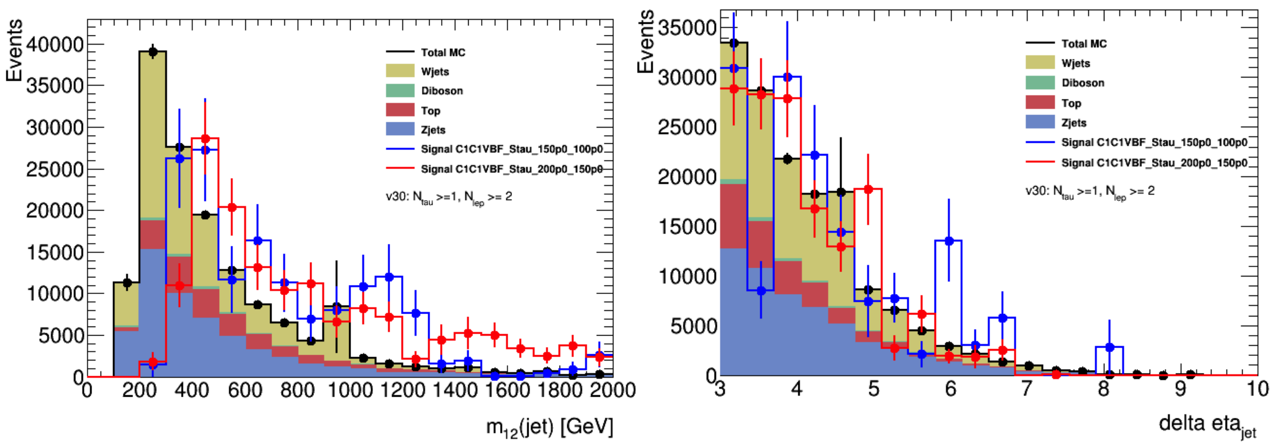
Monte Carlo (MC) samples are generated such that the signal events have at least one hadronically decaying tau lepton, at least two leptons total, and at least two jets with high separation. The signal samples are created to enrich the number of events with highly separated and highly energetic jets. The most prominent sources of background arise from the production of W or Z bosons in association with jets (W/Z+jets), top quark production, and diboson production. The W+jets background arises predominantly from when a W boson decays leptonically, and one or more jets are misidentified as a lepton. The top background contains both $t\bar{t}$ and single top production.

ii Input variables and precuts

In total, 19 variables are considered in this analysis. A summary of these variables is presented in Tab. 1. More complete descriptions of m_{T2}, meff, and drtt are listed in Appendix A. To remove these unphysical differences between signal and background due to the MC event generation, precuts are made before beginning the analysis. The precuts executed are: $\Delta\eta(\text{jet}) > 3.0$, $p_{\mathrm{T},1}(\text{jet}) > 30$ GeV, $p_{\mathrm{T},2}(\text{jet}) > 30$ GeV, $p_{\mathrm{T},1}(\text{lep}) > 20$ GeV and $p_{\mathrm{T},2}(\text{lep}) > 20$ GeV. All event variables are first plotted to visually analyze their distributions. Two of these event histograms, showing the dijet invariant mass and $\Delta\eta$ between the leading jets, are presented in Fig. 3(a) and Fig. 3(b), respectively. As expected, the signal events are significantly shifted towards higher values of the dijet invariant mass. Though less pronounced, the signal events also have slightly higher jet separation in pseudorapidity.

Variable	Description
MET	Missing transverse energy (magnitude)
m_{T2}*	Stransverse mass
m_{12}(lep)	Invariant mass of leading leptons
m_{12}(jet)	Invariant mass of leading jets
mt12	Sum of transverse mass of leading leptons
meff *	Effective mass
drtt *	ΔR of leptons
$p_{\mathrm{T},1}$(lep) $p_{\mathrm{T},2}$(lep)	Transverse momentum of leading (1) and subleading (2) leptons
$p_{\mathrm{T},1}$(jet) $p_{\mathrm{T},2}$(jet)	Transverse momentum of leading (1) and subleading (2) jets
η_1(jet) η_2(jet)	Pseudorapidity of leading (1) and subleading (2) jet
$\Delta\eta$(jet)	Separation in η of leading jets
$\eta_1\eta_2$(jet)	$\eta_{1,\mathrm{jet}} \cdot \eta_{2,\mathrm{jet}}$
$\mid \Delta\phi(\mathrm{jet}) \mid$	Separation in ϕ of jets, where ϕ is the azimuthal angle
$m_{\mathrm{T},1}$(jet) $m_{\mathrm{T},2}$(jet)	Transverse mass of leading (1) and subleading (2) jet
n_{jet}	Number of jets

Table 1: Brief descriptions of the event variables considered in this analysis. Starred (*) variables are described in more detail in Appendix A.



(a) $m_{12,\ \mathrm{jet}}$ distribution

(b) $\Delta\eta_{\ \mathrm{jet}}$ distribution

Figure 3: Distributions of the invariant mass of the jets (a) and difference in pseudorapidity of the jets (b) after the precuts have been applied. The signal area is normalized to the total MC background to show the distribution shapes. Reference point 1 is shown in blue and Reference point 2 is shown in red.

Additional vetos to be used in combination with the precuts are also analyzed. The three types of vetos investigated are a b-jet veto, a m_{ll} veto, and a Z boson veto. An event passes the b-jet veto ('bVeto') if there are no b-tagged jets. This cut significantly reduces the top background. An event passes the m_{ll} veto ('mllVeto') if the combination of leptons in the event with the maximum invariant mass yields $m_{12} > 12$ GeV. This cut removes taus that are produced via low mass resonance decays. Finally, the Z boson veto involves both a cut on the invariant mass of the leptons, as well as a cut on the 'qualEle' variable. The 'qualEle' variable is based on a BDT for tau selection that is used to remove overlaps with electrons. To ensure a tau has been selected, 'qualEle' must be 0. Additionally, the invariant mass requirement to pass the Z veto ('zVeto') is $\mid m_{12} - m_Z \mid > 10$ GeV. Here, the mass of the Z boson in this cut is set to be $m_Z = 70$ GeV. This is because the Z boson mass in di-tau decays is reconstructed lower than its actual mass, due to the neutrinos resulting from the tau decays [5]. This shift of the Z mass peak can be seen in the m_{12} distribution shown in Fig. 4.

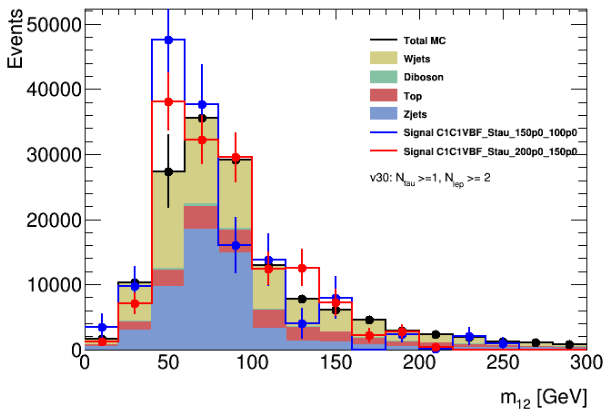


Figure 4: Distribution of m_{12} showing the shifted Z mass peak after the precuts have been applied. The signal area is normalized to the total MC background to show the distribution shapes. Reference point 1 is shown in blue and Reference point 2 is shown in red.

The following combinations of cuts are analyzed:

- cutA = bVeto + precuts
- cutB = bVeto, mllVeto + precuts
- cutC = bVeto, mllVeto, zVeto + precuts

The total event counts after applying these cuts are summarized in Tab. 2. cutA is excluded from further analysis in this project, because cutB is able to exclude an appreciable amount more of background events with negligible loss of signal when compared to cutA. Thus, the following analysis steps are carried out 3 times, once with only the precuts, once with cutB, and once with cutC.

iii Iterative Removal Analysis

Iterative removal (IR) is a procedure used to determine the optimal variables to train the BDT classifier. The optimal variables are the ones that have the most discriminating power, or the highest ability to separate signal from background. When choosing the

Cut	Signal events (RP1)	Signal events (RP2)	Background events	(S/B)$\times 10^3$ (RP1)	(S/B)$\times 10^3$ (RP2)
No precuts	136.27	50.23	2528000	0.0539	0.020
Precuts	43.58	27.66	148597	0.293	0.186
cutA	42.84	27.30	119019	0.359	0.229
cutB	42.84	27.23	118858	0.360	0.229
cutC	22.47	17.47	65555	0.342	0.266

Table 2: Event counts after applying cuts specified in Sec. 3.ii. The last two columns show the ratio of signal to background events, scaled by 1000.

optimal variables, one must also take into account correlations. If two variables both have high separation power, but are highly correlated, then likely only one of these is needed to train the BDT classifier. The iterative removal process is able to take this type of correlation into account when selecting the optimal variables. In the IR process, one first begins with the entire variable list. Each variable is removed one at a time, and the BDT classifier is trained and tested without this variable in the list. The variable that, when removed, decreases the ROC integration the least, is deemed the least important variable, and discarded permanently. The process continues until an entire ranking of the variables, from least important to most important, is created.

One alternative to this method is the 'brute force' method. In this method, one tries every possible combination of the desired final number of variables to find the best performing combination. In this project, where 19 variables are considered, using the brute force method is unfeasible to accomplish because of the number of times one would have to train and test the BDT classifier. For example, if one wanted to determine the best 5 variables with the brute force method, the BDT classifier would have to be trained and tested $\binom{19}{5} = 11,628$ times. On the contrary, the iterative removal process only requires $19 + 18 + ... + 3 + 2 = 189$ iterations for a complete variable ranking. Thus, IR offers a much more efficient way to create a variable ranking. For details on the IR algorithm, refer to [6].

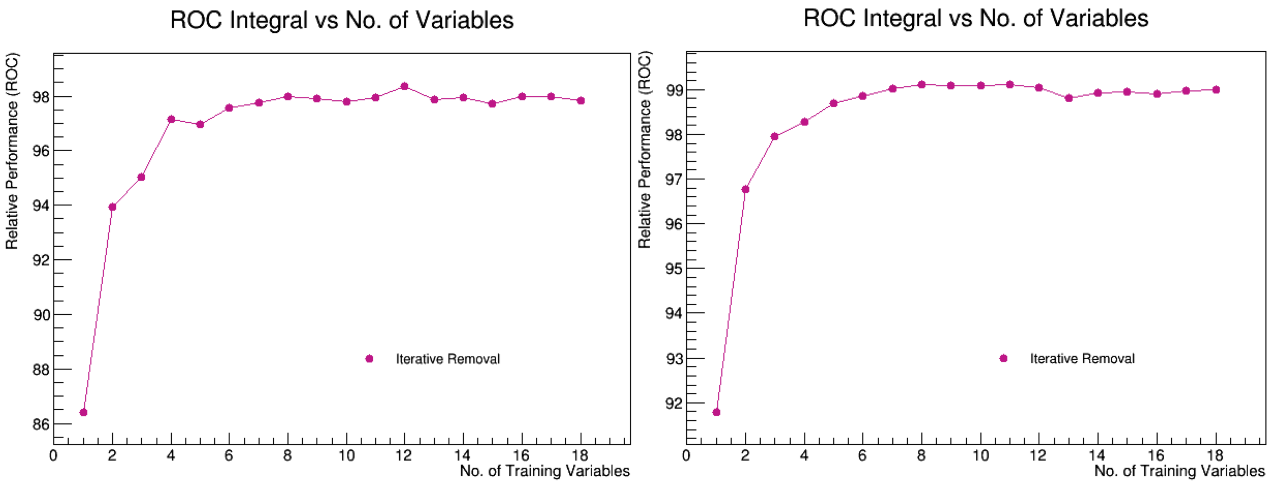
A summary of the BDT classifier parameters used in this analysis are shown in Table 3. A description of these parameters can be found in the TMVA users guide [4]. In addition to the event variables, the event weights are also sent to the BDT such that the events are properly weighted for training and testing. To shorten the computation time of the iterative removal analyses, the number of trees is reduced to 200 from the amount used to produce the final results (800).

3 Results

For each set of cuts on each reference point, the iterative removal analysis is executed to determine a variable ranking, and a plot of the ROC integration as a function of the number of variables is made. An example of this plot for both reference points with the precuts applied is shown in Fig. 5. Similar plots for both reference points with cutB and cutC applied can be found in Appendix B. For each analysis, the first local maximum of

BDT Parameter	Value
NTrees (IR)	200
NTrees (Results)	800
MaxDepth	3
MinNodeSize	5%
nCuts	20
BoostType	AdaBoost
NegWeightTreatment	IgnoreNegWeightsInTraining

Table 3: Summary of important BDT training parameters. Details can be found in [4].



(a) Reference point 1 with precuts

(b) Reference point 2 with precuts

Figure 5: ROC integration scaled to 100 as a function of the number of variables in the IR analysis

the ROC integral occurring with more than 5 variables is chosen as the optimal number of variables to use for that analysis. This criteria allows the ROC integration to be maximized while the number of variables is still kept minimal. A summary of the best variables chosen for each run and reference point is listed in Tab. 4. For each reference point and each set of precuts, the best set of variables is sent to a BDT classifier to be trained with 800 trees. The BDT response value, scaled with the event weights, for both reference points with precuts applied is shown in Fig. 6. The BDT response histograms for the cutB and cutC runs are shown in Appendix C.

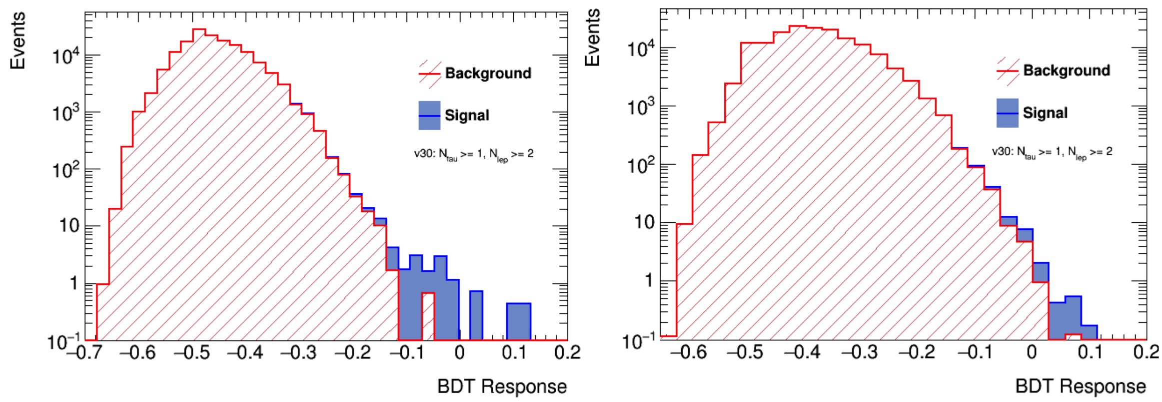
The final step in the analysis is to find the optimal cut on the BDT response value such that the significance is maximized. The significance, S, is estimated using:

$$\mathrm{S} = \frac{S}{\sqrt{S+B}}, \quad (1)$$

where S and B are the number of signal and background events after the cut has been applied, respectively. Sequential cuts are applied to the BDT response histogram of each run, and the significance of each cut is calculated. The results for both reference points with precuts applied are shown in Fig. 7. The results for both reference points with cutB

	Reference 1			Reference 2		
	precuts only	cutB	cutC	precuts only	cutB	cutC
ROC integral	0.975	0.968	0.980	0.990	0.993	0.990
Nvar	8	7	8	8	8	9
# 1	$\mathrm{n_{jet}}$	$\mathrm{n_{jet}}$	$\mathrm{n_{jet}}$	MET	MET	MET
# 2	$\mathrm{m_{T,2}(jet)}$	$\mathrm{m_{T,1}(jet)}$	mT2	$\mathrm{n_{jet}}$	$\mathrm{n_{jet}}$	$\mathrm{n_{jet}}$
# 3	drtt	$\mathrm{m_{12}(jet)}$	drtt	drtt	$\mathrm{m_{12}(jet)}$	$\mathrm{m_{T2}}$
# 4	$\mathrm{m_{12}(jet)}$	drtt	$\mathrm{m_{T,2}(jet)}$	$\mathrm{m_{T2}}$	drtt	drtt
# 5	$\mathrm{m_{12}}$	$\eta_2\mathrm{(jet)}$	$\mid \Delta\phi\mathrm{(jet)} \mid$	$\mid \Delta\phi\mathrm{(jet)} \mid$	$\mathrm{m_{12}}$	$p_{\mathrm{T},1}\mathrm{(jet)}$
# 6	mt12	$\mathrm{m_{T2}}$	MET	$\mathrm{m_{12}}$	$\mathrm{m_{T2}}$	$\mathrm{m_{12}}$
# 7	$\eta_2\mathrm{(jet)}$	MET	$\mathrm{m_{12}(jet)}$	$\eta_1\eta_2\mathrm{(jet)}$	$\eta_1\mathrm{(jet)}$	$\eta_2\mathrm{(jet)}$
# 8	$\Delta\eta\mathrm{(jet)}$		$p_{\mathrm{T},1}\mathrm{(lep)}$	$\Delta\eta\mathrm{(jet)}$	$\eta_1\eta_2\mathrm{(jet)}$	$\mid \Delta\phi\mathrm{(jet)} \mid$
# 9						meff

Table 4: Variable ranking according to iterative removal analysis. Included is the final number of variables selected, Nvar, and the optimal ROC integral value.

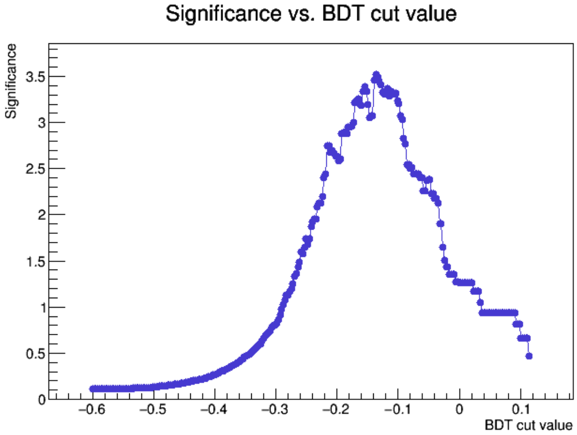


(a) Reference point 1 with precuts (b) Reference point 2 with precuts

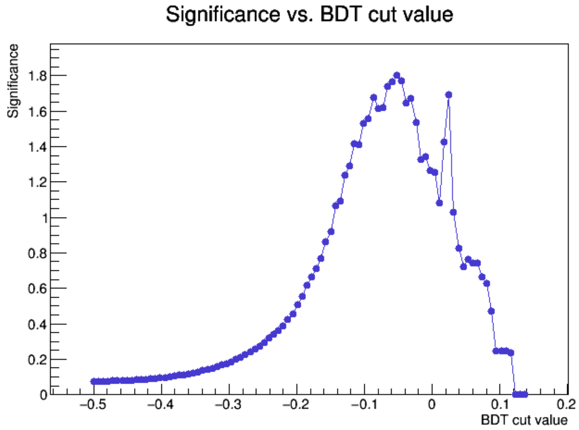
Figure 6: BDT response value distributions for signal and background events, scaled with the event weights. Shown are the precut-only runs for both Reference points.

and cutC are shown in Appendix D. The maximal significances and BDT cut values for all runs are listed in Tab. 5.

In the cutB and cutC runs for Reference point 2 (See Appendix C Fig. 9(c) and (d)), the events with negative weights are excluded from the calculation of the significance. This is because at some cut values, there are more background events with significant negative event weights than signal events, and the significance can not be calculated.



(a) Reference point 1 with precuts



(b) Reference point 2 with precuts

Figure 7: Estimated significance as a function of the BDT response cut value for each reference point with precuts applied.

Reference point 1

Cut	Maximum Significance	Signal	Background	BDT cut
precuts	3.52	13.9	1.8	> -0.136
cutB	3.65	18.1	6.5	> -0.148
cutC	3.37	14.8	4.4	> -0.285

Reference point 2

Cut	Maximum Significance	Signal	Background	BDT cut
precuts	1.88	7.9	9.6	> -0.048
cutB	2.03	10.3	15.7	> -0.059
cutC	1.89	7.3	7.7	> -0.103

Table 5: Maximum significance using cuts on the BDT response values.

4 Interpretation

As shown in Tab. 5, the most sensitive analysis for Reference point 1 and Reference point 2 had a significance of 3.65, and 2.03, respectively. Both maximal significances occurred when the precuts were applied in combination with a bVeto and an mllVeto. Thus, neither analysis is sensitive enough to make a discovery, and only the analysis for Reference point 1 is sensitive enough to claim evidence. These analyses are therefore in need of more optimization before they can be applied to a search with real ATLAS data.

While the optimal variable list for Reference point 1 only includes 7 variables, the list for Reference point 2 includes 9 variables. The plots in Fig. 5 and Fig. 8 demonstrate that supplying the BDT classifier with more variables for training does not always lead to better results. A method like iterative removal is always needed to identify the best variables to maximize the performance of the BDT classifier.

Both optimal variable lists include n_{jet}, MET, drtt, η_2(jet), and m_{T2}. The top variable for every set of cuts with Reference point 1 is n_{jet}, while the top variable for each run with Reference point 2 is MET. This is likely because the mass of the escaping LSP in Reference point 2 (150 GeV) is larger than that in Reference point 1 (100 GeV), and therefore the MET in Reference point 2 events will be shifted to larger values that further separate it from background. It is unclear why the number of jets is an important variable in many variable lists. However, the possibility that this is due to the way these sample events are generated has not been excluded. Thus, further investigation into the exact mechanism for creating these MC events is required to confirm these results and ensure that no biases are present in the MC event data.

Negative event weights are ignored by the BDT classifier when it is in the training stage. Nevertheless, negative event weights cause a problem for the analysis of Reference point 2 when cutB and cutC are applied. In the calculation of the significance, often times there are more negative background events than positive signal events when cutting on BDT response values in the tails of the background distributions, resulting in an imaginary significance. However, this issue only arose in the cutB and cutC runs of Reference point 2, which is likely due to the low signal event counts in these runs, as seen in Tab. 2.

5 Conclusion and Future Work

While using BDTs to search for the production of charginos and neutralinos via VBF in compressed SUSY scenarios is a promising approach, this analysis has shown that more optimization is needed before this technique can actually be applied.

One of the challenges with this analysis is the requirement to apply precuts. Although these precuts significantly reduce the background event count, it also reduces the signal event count by roughly 68% for Reference point 1 and 46% for Reference point 2 (see Tab. 2). These precuts are on the 'safe-side', such that it is certain that no biases are present in the $\Delta\eta$, p_T(lep), or p_T(jet) distributions. However, it should be investigated whether these precuts can be relaxed without introducing biases, such that less signal events are cut.

This analysis can be optimized in a plethora of ways. One could experiment with trying new variables, such as the recently developed 'razor variables' [7]. Additionally, one can run an analysis to determine the best parameters for the training of the BDT classifier, such that the signal and background separation is optimized while the trees still remain safe from overtraining.

Once this analysis has been further optimized, it can be compared to other analyses such as the rectangular cuts method in TMVA. Using rectangular cuts to define signal regions is a more traditional approach to data analysis in ATLAS, and was very common in the analysis of data from ATLAS during Run 1 of the LHC. However, with better modeling of data and the development of TMVA, the TMVA methods such as BDTs are likely to be more abundantly used for the analysis of data from Run 2 and future runs of the LHC.

References

[1] C. Patrignani et al. Review of Particle Physics. *Chin. Phys.*, C40(10):100001, 2016.

[2] Howard E. Haber. Supersymmetry, part i (theory). *Particle Data Group*, September 2015.

[3] Bhaskar Dutta, Alfredo Gurrola, Will Johns, Teruki Kamon, Paul Sheldon, and Kuver Sinha. Vector Boson Fusion Processes as a Probe of Supersymmetric Electroweak Sectors at the LHC. *Phys. Rev.*, D87(3):035029, 2013.

[4] J. Stelzer J. Therhaag E. von Toerne H. Voss A. Hoecker, P. Speckmayer. Tmva 4 users guide. *arXiv:physics/0703039*, October 4, 2013.

[5] G. Aad et al. Search for the electroweak production of supersymmetric particles in $\sqrt{s}$ = 8 tev pp collisions with the atlas detector. *Physical Review*, D 93(052002), 2016.

[6] Sitong An. Optimising variable selection for machine learning analysis in atlas tth search. DESY Summer Student Program 2017, September 2017.

[7] Chatrchyan et al. Search for supersymmetry with razor variables in pp collisions at $\sqrt{s}$ = 7 TeV. *Phys. Rev. D*, 90(CMS-SUS-12-005. CMS-SUS-12-005. CERN-PH-EP-2014-057):112001. 40 p, May 2014.

A Event variable descriptions

$\mathrm{m_{T2}}$: The stransverse mass is defined as:

$$\mathrm{m_{T2}} = \mathrm{min}_{\vec{q}_{\mathrm{T}}}[\ \mathrm{max}(\mathrm{m_T}(\ \vec{p}_{\mathrm{T},1}\ , \vec{q}_{\mathrm{T}}), \mathrm{m_T}(\ \vec{p}_{\mathrm{T},2}, E_{\mathrm{T}}^{\mathrm{miss}} - \vec{q}_{\mathrm{T}})\ )\ ],$$

where $\vec{p}_{\mathrm{T},1}$ and $\vec{p}_{\mathrm{T},2}$ are the transverse momenta of the leading and subleading leptons, respectively. $\vec{q}_{\mathrm{T}}$ is a transverse vector that minimizes the larger of the two transverse masses, m_{T}. The stransverse mass distribution has a kinematic endpoint for events where two massive pair produced particles each decay semi-visibly [5].

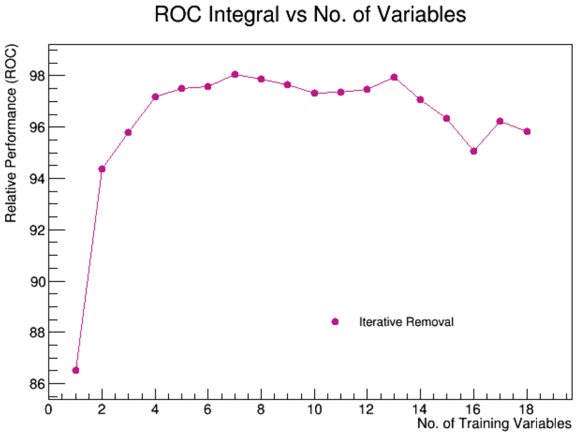
meff: The effective mass is the scalar sum of the transverse momenta of the leptons, jets, and $E_{\mathrm{T}}^{\mathrm{miss}}$ in the event:

$$\mathrm{meff} = E_{\mathrm{T}}^{\mathrm{miss}} + \sum p_{\mathrm{T}}^{\mathrm{leptons}} + \sum p_{\mathrm{T}}^{\mathrm{jets}}$$

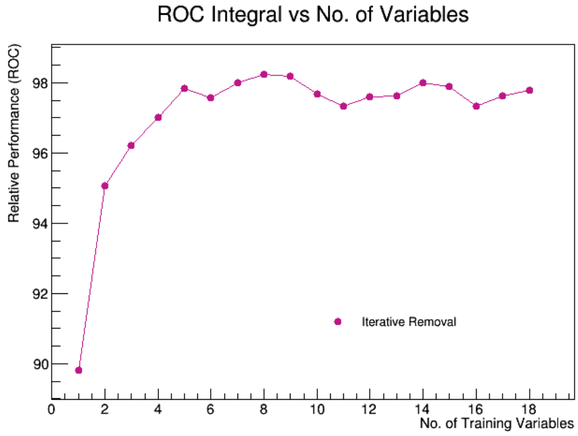
drtt: The separation of the leptons, drtt, is characterized by ΔR, which is defined: as:

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

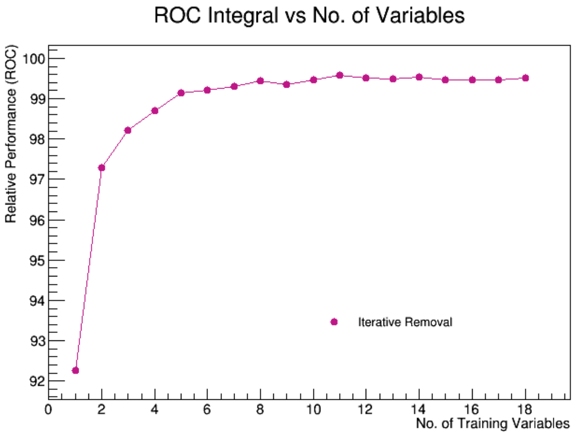
B ROC integral plots



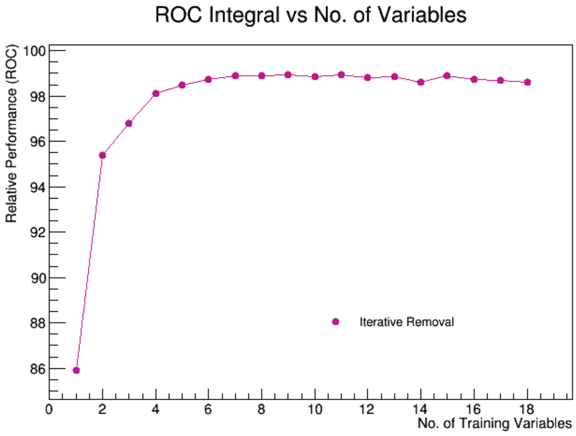
(a) Reference point 1 with cutB



(b) Reference point 1 with cutC



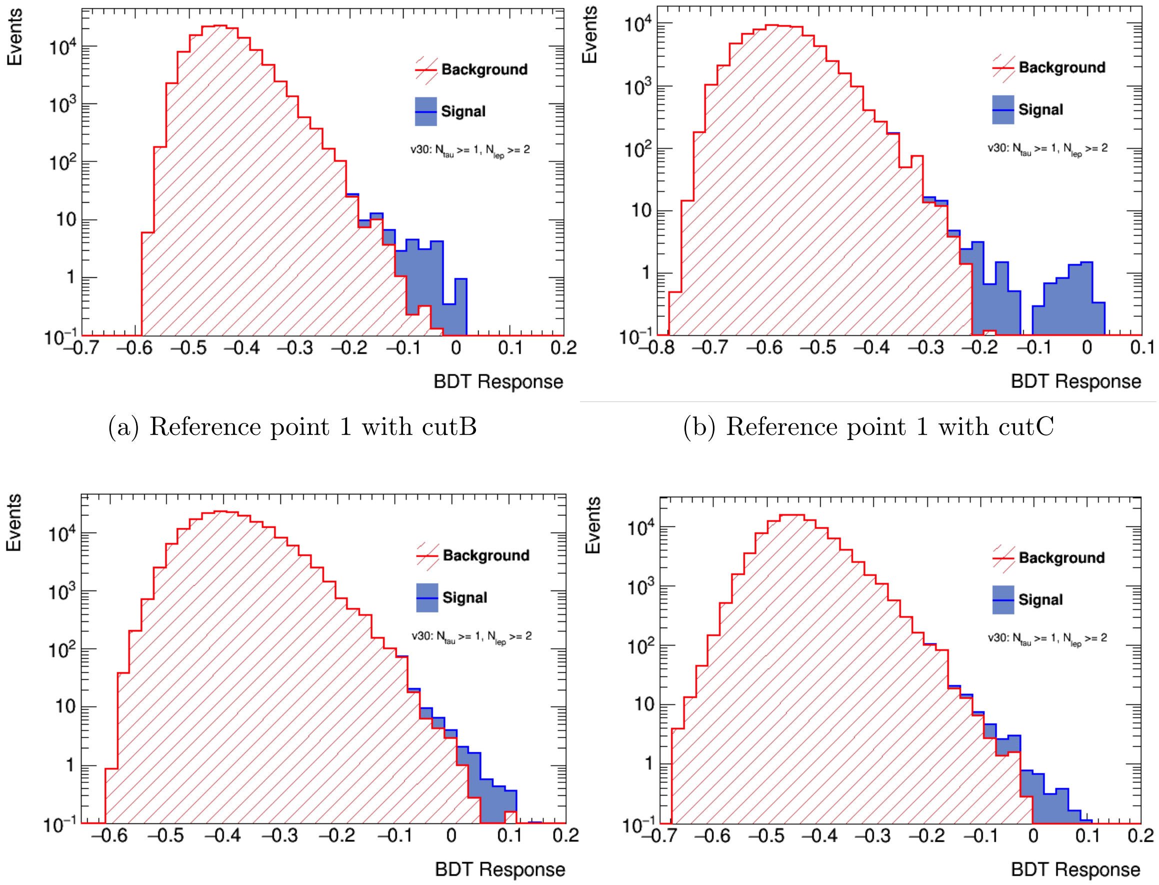
(c) Reference point 2 with cutB



(d) Reference point 2 with cutC

Figure 8: ROC integration scaled to 100 as a function of the number of variables in the iterative removal analysis, for cutB and cutC of each reference point.

C BDT response histograms



(a) Reference point 1 with cutB

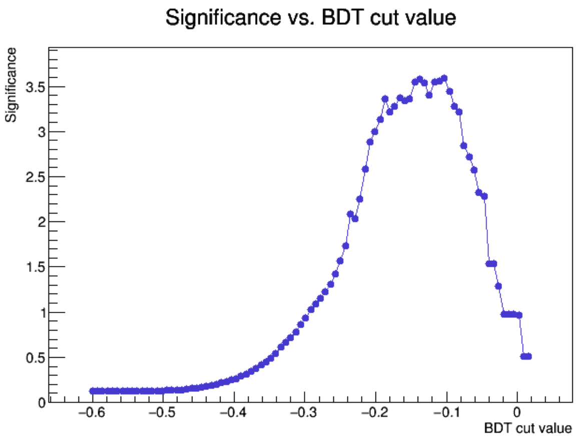
(b) Reference point 1 with cutC

(c) Reference point 2 with cutB

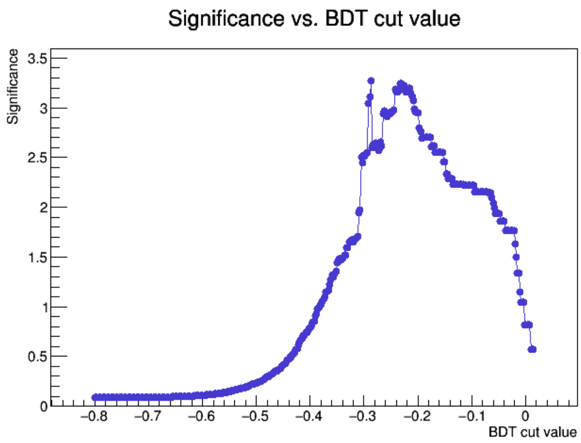
(d) Reference point 2 with cutC

Figure 9: BDT response value distribution for signal and background events, scaled with the event weights. For (c) and (d), the negative event weights are excluded, as discussed in Sec. 3.

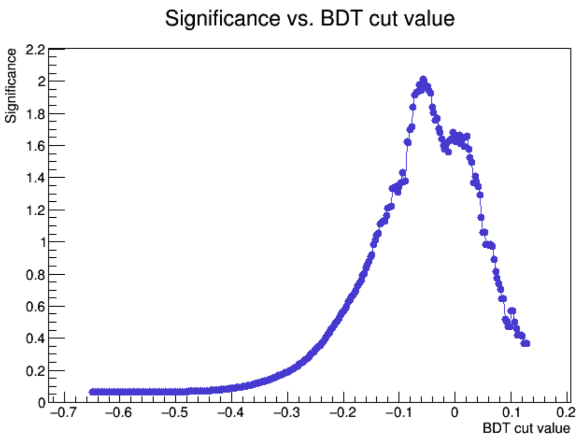
D BDT cut optimization plots



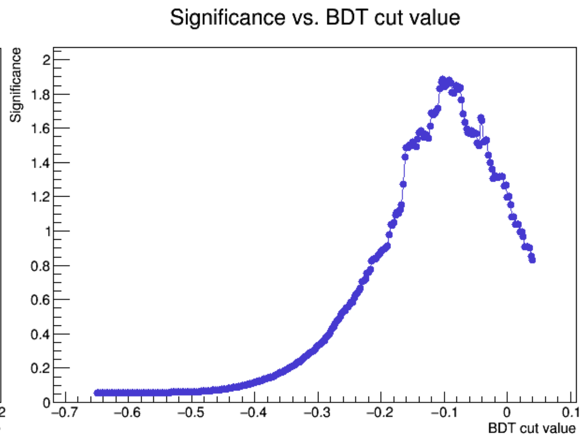
(a) Reference point 1 with cutB



(b) Reference point 1 with cutC



(c) Reference point 2 with cutB



(d) Reference point 2 with cutC

Figure 10: Estimated significance as a function of BDT response cut value for each reference point with cutB and cutC applied.