Searching for Supersymmetric Top Quarks at the LHC

Onkur Sen
Rice University
onkur@rice.edu

Dr. Paul Padley
Rice University
padley@rice.edu

April 22, 2013

Contents

1 Introduction

1.1 Motivation and background

Experiments at the LHC have posited the existence of a Higgs-like particle in the 125 GeV region. However, the precision of these calculations has been brought into question because in the Standard Model, higher-loop corrections to this mass are quadratically divergent.

The most commonly-studied method that would cancel these divergences is supersymmetry, which posits that for every fermion/boson, there exists a corresponding boson/fermion "superpartner" with the same mass and quantum numbers except spin. Originating from string theory (although it stands alone as an idea), supersymmetry has been widely favored to hold true, and some have already constructed supersymmetric extensions of the Standard Model. However, the existence of a supersymmetric particle has not yet been experimentally confirmed.

With the recent success of the Higgs experiments, however, a supersymmetric particle that has come to the forefront is the superpartner of the top quark known as the **stop quark** or **squark**. In particular, the decay hypothesized is:

$$gg \to \tilde{t}\bar{\tilde{t}}$$

From here, the supersymmetric particle decays as follows:

$$\tilde{t} \to t + \tilde{\chi}_1^0,$$

where $\tilde{\chi}_1^0$ is the lightest supersymmetric particle that is hypothesized within a given parameter space. Furthermore, in a collision, we expect a further decay of the top quark into a bottom quark and a W boson:

$$t \to b + W,$$

and finally, the W boson decays into two jets that are observed as a final result of the collision:

$$t \to b + jj,$$

See Figure 1 for a Feynman diagram summarizing the top quark decay sequence.

Dutta et al. [7] provide a theoretical framework for searching for stop quarks that is based on multiple successive "cuts" that iteratively shrink the set of data in order to locate the desired decay mode. In this thesis, we instead attempt to use machine learning (in particular, boosted decision trees) to holistically consider multiple aspects of the events and then classify

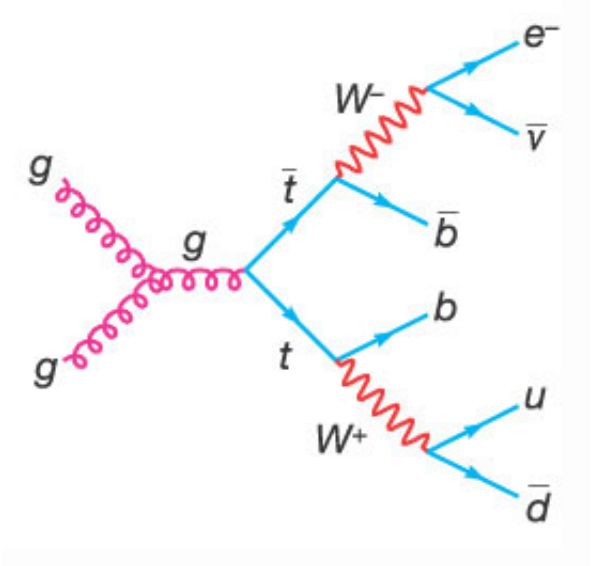

Figure 1: A Feynman diagram illustrating the full path of the top quark decay. The resultant particles emanating out of the W bosons are treated as untagged jets in our simulation. [3]

them as signal (stop quark decays) or background (top-top, or ttj, collisions). We hope that this will be a more effective approach in searching for the decay of stop quarks.

1.2 The machine learning problem: classification

1.2.1 Problem description and related work

As we are trying to separate one kind of decay from another, a machine learning perspective lends itself well to this problem. In particular, we are attempting to solve the **classification problem**: given a set of categories and a training set of data with the category known for each instance, can we accurately identify to which category a new instance belongs?

Note that classification only refers to the task of using a finite set of discrete labels to classify any given instance; for continuous input, the task is referred to as *regression*. In our case specifically, the classification is binary: we aim to distinguish between **signal** (i.e., supersymmetric decays) and **background** (i.e., top-top decays).

So far, various machine learning approaches have been attempted [4]:

- Support vector machines (separate data into two classes using a hyperplane that maximizes the "margin")
- Linear discriminant analysis (find linear combination of features that best separates data)
- Artificial neural networks (use different layers of connected "neurons," beginning with an input layer and ending with an output layer)

However, decision trees have been particularly effective in previous CMS analyses, such as $H \to W^+W^-$ [2].

1.2.2 Decision trees

A **decision tree** is a full binary tree with three characteristics that appeal to intuition:

- An internal node is a "question" which can be answered using the variables in consideration.
- The two edges emanating from an internal node represent the paths taken where the condition is and is not satisfied (i.e, the answer to the question is either yes or no).
- The leaf nodes contain the category which this observation belongs to given the previous responses given along the path to the node; additionally, a confidence level is given regarding this answer.

Thus, a decision tree resembles a flowchart in that by answering a sequence of yes/no questions, one can arrive at a classification of an observation. More explicitly, the internal nodes of a decision tree concern threshold values for the variables in question (only one variable is considered at each node), and the leaf nodes represent a class to which the event can be assigned.

But how are these decision trees actually *constructed*? In particular, how are the threshold established for making an effective classifier? As Bishop points out, the ideal way to do this would be to minimize a general, multivariate sum-of-squares error function; however, this is computationally infeasible given both the number of threshold values that must be computed as well as the number of events over which this decision tree will be running. Instead, the approach used is a greedy algorithm that works as follows:

Suppose there are D input variables to split the decision tree on. By exhaustive search, the optimal place to split on a variable d_i is at the local

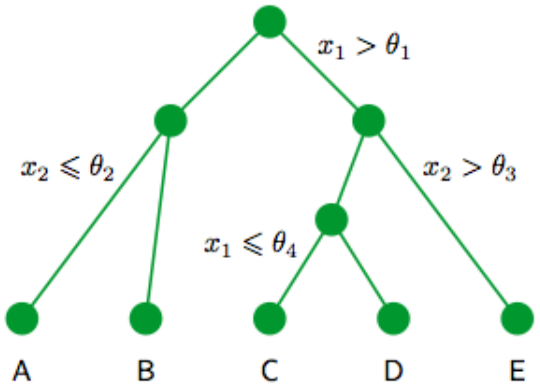

Figure 2: An example decision tree with threshold cutoffs for the feature variables at the internal nodes. The unlabeled edges correspond to the negations of the corresponding labeled edges. In this case, the leaf nodes A, B, C, D, and E are all different classes. [5]

average of the data. This is done for all choices of each variable d_i to be split, and the value giving the smallest sum-of-square error function is used as the threshold.

Furthermore, in many cases, decision trees are grown out to a certain fixed size and then *pruned* by removing nodes so that the resultant tree is an even more effective classifier. However, we chose not to use pruning in our experiments.

The advantages of using decision trees include [11]:

- an appeal to intuition
- robustness in handling both numeric and string attributes
- an ability to represent any discrete-value classifier, and
- an ability to handle datasets with errors or missing values

However, a big pitfall of decision trees is that since they are inherently based on the notion of "divide and conquer", they can perform extremely poorly if the data is not characterized by a few highly relevant attributes, but instead by complex interactions between many less relevant factors.

1.2.3 Boosting

Decision trees are often used in conjunction with a machine learning technique called **boosting** which aims to answer the question [12]: can a combination of weak learners, each only slightly better than a random classifier, be strategically combined to make a single strong classifier? An example of boosting in everyday decision making would be as follows: given a specifically oriented set of data, we may be able to come up with a "rule of thumb" for classification. However, this scheme cannot be extended to a general data set. In this case, boosting would involve a set of many rules of thumb for different cases in order to account for the generality of the expected input [8].

Note that boosting is defined in relation to **supervised learning algorithms** (including decision trees), i.e., those that train on a set of data for which the classification label is already known before attempting to classify a new event.

The most well-known boosting algorithm is known as **AdaBoost**, which stands for adaptive boosting (Algorithm 1). We are given a set of training data $\{(x_j, y_j)\}_{j=1}^{m}$, where each x_j represents an input (in this case, an event for classification), and y_j represents the corresponding classification label. For a two-class scheme like in our case with signal and background, a common assumption is $y_j \in \{+1, -1\}$. AdaBoost works over a number of iterations T, and at each time step aims to find a weak hypothesis h_t that fits the domain space X to the classification labels according to a distribution D_t, which determines how much each training instances is weighted at time step t. The goodness of fit of h_t is determined by the error:

$$\epsilon_t = \sum_{i:h_t(x_i)\neq y_i} D_t(i).$$

After the error is calculated, the distribution is reweighted so that the misclassified instances have higher weights, which forces the hypothesis to account more for them in the following iteration.

In our approach to classifying decay of stop quarks, we use **boosted decision trees** (BDTs). These have been particularly successful in previous CMS analyses because the philosophy behind boosting perfectly complements the major weakness of decision trees. More explicitly, although an individual decision tree performs poorly when connections between many different factors are present, the collective nature of boosting allows for each decision tree to be focused on one or a few factors (consequently specializing in the data and epitomizing a "weak learner"). Combining these weak classifiers together makes a strong, versatile classifier that is more robust

Given: $\{(x_j, y_j)\}_{j=1}^m$, where $x_j \in X$, $y_j \in Y = \{+1, -1\}$
Initialize $D_t(i) = \frac{1}{m}$
for $t = 1 \to T$ **do**
    Train weak learner with distribution D_t
    Get weak hypothesis $h_t : X \to \{+1, -1\}$ with error
    $\epsilon_t = \sum_{i:h_t(x_i) \neq y_i} D_t(i)$
    Update the distribution:

$$D_{t+1}(i) = \frac{1}{Z_t} D_t(i) \exp\left(-\alpha_t y_t h_t(x_i)\right),$$

    where Z_t is a normalization factor and $\alpha_t = \frac{1}{2} \ln(\frac{1-\epsilon_t}{\epsilon_t})$.
end
Return the final hypothesis:

$$H(x) = \operatorname{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right).$$

Algorithm 1: Algorithm for AdaBoost [8].

on a new dataset. Furthermore, BDTs are generally less susceptible than other methods to **overtraining** [4], the phenomenon where the classifier is to attuned to the training data (which may have been biased to begin with) to perform well in a general setting. This speaks to the robustness and versatility of the approach.

1.2.4 The physics context

Recall that the data that we are training on comes in the form of **jets**, which provide a tagged representation of the final result of a particle collision. Beginning with a pair of particles (e.g., two protons, two gluons), the collision results in a stream of different particles emanating in various directions. Each jet corresponds to one of these final particles.

In particular, since are interested in stop quark decays and top-top background, we will be focusing on the bottom-tagged and untagged jets in the collisions. Our goal is to recreate a top quark system from a trijet combination of one bottom-tagged jet and two untagged jets (bjj). The justification for this particular emphasis comes from the mechanism for the decay of the top quark:

$$t \to b + W$$
$$W \to jj$$

Thus, the two untagged jets represent the W boson, and all three jets combined represent the top quark.

2 Experimental Approach

2.1 Implementing theory

2.1.1 Preliminaries

Since we are not sure whether this decay mode is indeed viable, we are using simulated data to see how feasible the separation of signal and background can be. Thus, we first obtained phenomenological data from Dutta et al. [7], who have been working on providing the theoretical mechanisms for searching for this mode. The phenomenological data was generated and processed using PYTHIA [13] and returns many different objects per event, where each object corresponds to a final particle (e.g., jet, electron, muon, tau).

Our first step was to isolate the jets in these events and plot the energy distribution. See Figure 3 for energy distributions for sample SUSY and ttj event runs.

Next, we applied some **baseline cuts** to remove instances which can be easily classified as background to prevent bias in our training set towards those instances. The baseline cuts applied were [7]:

- Reject any event with fewer than 4 untagged jets or fewer than 2 bottom-tagged jets.
- Within the space $|\eta| \leq 2.5$ (where η denotes the pseudo rapidity), reject events where the leading jet has $p_T < 100$ GeV or where any other jet has $p_T < 30$ GeV.
- Within the space $|\eta| \leq 2.5$, reject events containing isolated electrons and muons with $p_T > 10$ GeV.
- Within the space $|\eta| \leq 2.1$, reject events containing taus with $p_T > 20$ GeV. Unlike the authors of [7], we crudely assume perfect tau identification.
- Reject any event where the missing transverse energy $\not{E}_T < 100$ GeV.

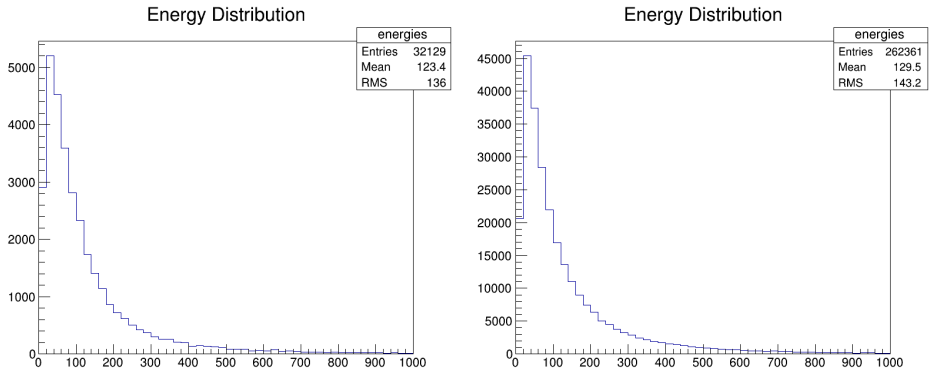

Figure 3: Energy distributions for jets in collision event runs. Left: supersymmetry space with $m(\tilde{t}) = 400$ GeV and $m(\tilde{\chi}_1^0) = 100$ GeV. Right: top-top background with 6 jets. The distributions are remarkably similar, indicating the difficulty of the task of separating signal from background *based on energy alone*.

2.1.2 Isolating two top quark systems

Next, we applied the theoretical approach of Dutta et al. [7] to separate each event into two top quark systems (see Figure 4). Broadly speaking, we attempt to find trijet bjj combinations that:

- maximize vecorially summed p_T,
- approximate the mass of the top quark well, and
- approximate the mass of the W boson well.

This was done as follows:

1. Select only the jets in the event, and separate them into those that are bottom-tagged (a set B) and those that are not (a set J).
2. Calculate the vectorial transverse momentum of all combinations of one bottom-tagged jet and two untagged jets. More explicitly, for all $b \in B$ and $j_1, j_2 \in J$ such that $j_1 \neq j_2$, calculate:
$$p_T(b, j_1, j_2) = p_T(b) + p_T(j_1) + p_T(j_2).$$
3. Choose the two trijet combinations $(b_1, j_{1,1}, j_{2,1})$ and $(b_1, j_{1,2}, j_{2,2})$ with the highest $||p_T||$.

4. Calculate $M3$, the invariant mass of the entire trijet combination, for both systems:
$$M3 = invariant\_mass(b, j_1, j_2)$$

5. Calculate $M2$, the invariant mass of the two untagged jets only, for both systems:
$$M2 = invariant\_mass(j_1, j_2)$$

6. Calculate the χ^2 error from the top quark and W boson masses:
$$error(b, j_1, j_2) = (m_{top} - M3)^2 + (m_W - M2)^2$$

7. The system with the lower error (b^*, j_1^*, j_2^*) is labeled as top quark A.

8. Remove b^* from B, and j_1^*, j_2^* from J.

9. Repeat steps 2-6. The system with the lower error is labeled as top quark B.

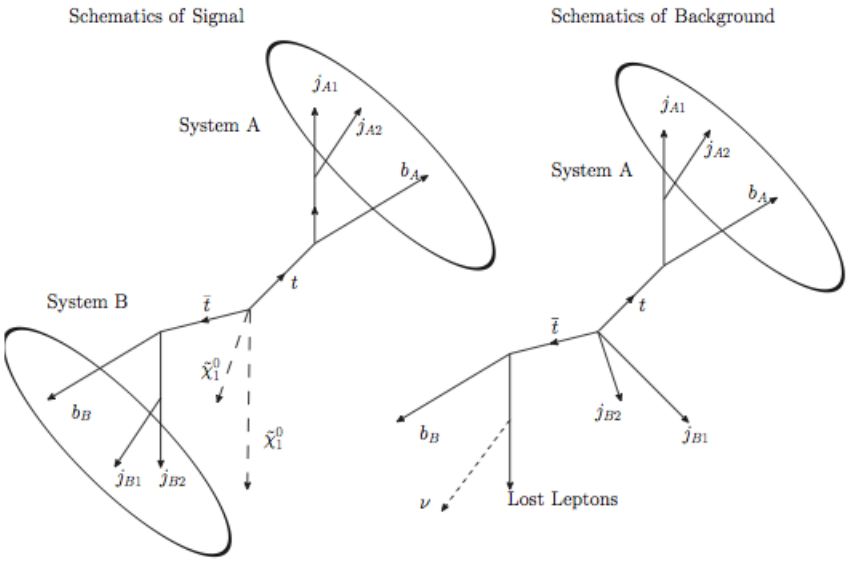

Figure 4: A schematic for the top quark systems obtained from each event in the cases of both signal and background [7].

A justification for the theoretical approach here: The use of $M3$ has been shown to be effective to separate signal from background in previous top quark studies [6, 14] and is most successful in events with at least 4 jets (our data has up to 6). Furthermore, transverse energy analysis has been effective in separating QCD events and is most effective with 3 jets; this is why we construct two top quark systems consisting of 3 jets each from every event.

2.2 Defining and customizing the BDT

2.2.1 Parameters considered

After identifying top quark systems A and B, Dutta et al. identify certain parameters as being useful in classifying events as signal or background [7]. Instead of performing iterative cuts, we include them as parameters for input into a boosted decision tree. We include both a mathematical notation for the parameter as well as the variable name in our code in parentheses.

- $\not{E}_T$ (`missing_eT`): the magnitude of missing transverse energy.
- $M3_A$ (`m3a`): $M3$ of top quark A (see Figure 5 for distribution).
- $M2_A$ (`m2a`): $M2$ of top quark A (see Figure 5 for distribution).
- $M3_B$ (`m3b`): $M3$ of top quark B.
- $M2_B$ (`m2b`): $M2$ of top quark B.
- $\Delta\phi(b_A, \not{E}_T)$ (`angles_b_a`): azimuthal angle between bottom quark of system A and missing traverse energy.
- $\Delta\phi(j_{1A}, \not{E}_T)$ (`angles_j1_a`): azimuthal angle between first untagged jet of system A and missing traverse energy.
- $\Delta\phi(j_{2A}, \not{E}_T)$ (`angles_j2_a`): azimuthal angle between second untagged jet of system A and missing traverse energy.
- $\Delta\phi(b_B, \not{E}_T)$ (`angles_b_b`): azimuthal angle between bottom quark of system B and missing traverse energy.
- $\Delta\phi(j_{2A}, \not{E}_T)$ (`angles_j1_b`): azimuthal angle between first untagged jet of system B and missing traverse energy.
- $\Delta\phi(j_{2B}, \not{E}_T)$ (`angles_j2_b`): azimuthal angle between second untagged jet of system B and missing traverse energy.

- $M_T(b_B, \not{E}_T)$ (`mT_b`): invariant transverse mass of bottom quark B and missing transverse energy.

A note on the raison d'etre of the azimuthal angles: theoretically the sum of all transverse momenta in the collision should be 0 since both particles are colliding in line with the beam axis. However, this is not the case experimentally; therefore, a missing transverse energy term arises. Because the kinematics of the resultant neutrinos are different for the supersymmetric and top-top cases, we can use that information to separate signal from background in the BDT.

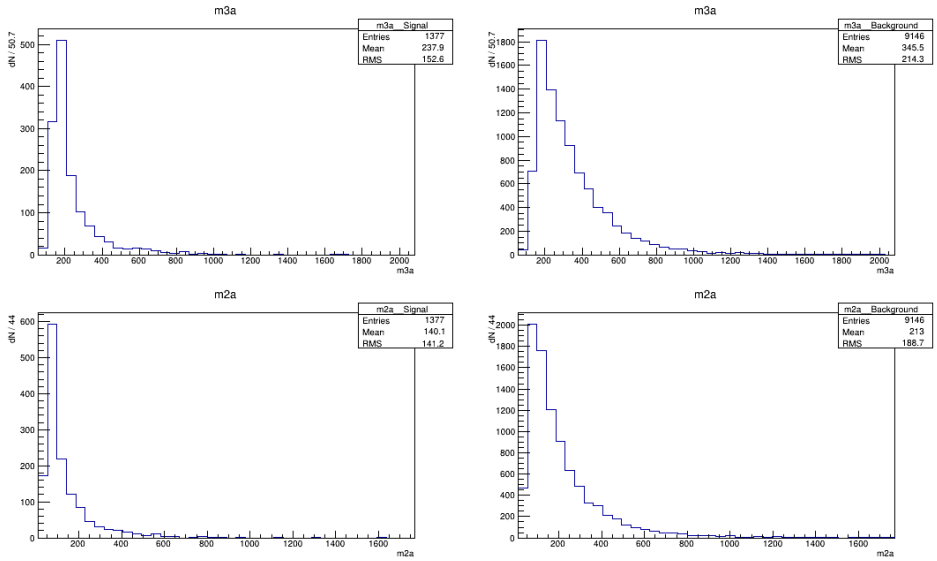

Figure 5: Distributions of $M3_A$ (top) and $M2_A$ (bottom) for jets in collision event runs. Left: supersymmetry space with $m(\tilde{t}) = 400$ GeV and $m(\tilde{\chi}_1^0) = 100$ GeV. Right: top-top background with up to 6 jets.

2.2.2 Implementation details

Software framework: We use boosted decision trees through PyROOT, a Python wrapper of the popular C++ data analysis framework ROOT [1]. In particular, we used the library TMVA (Toolkit for Multivariate Analysis), which provides machine learning methods for analyzing multivariate data [10]. Our previous experiments have used the Java framework Weka [9].

Boosting: As mentioned before, we used AdaBoost in particular for boosting our decision trees.

Supersymmetric parameter space: The authors of [7] show that out of the many supersymmetric spaces, the one where $m(\tilde{t}) = 400$ GeV and $m(\tilde{\chi}_1^0) = 100$ GeV provides the highest significance and throughput of signal and the is more likely to be the correct one. This is our primary test domain although we experiment with others as well.

Data sets: We split our original dataset into training and test sets of equal size. Furthermore, both the training and test sets contain data that is half signal and half background. We use a cross section of 10 fb^{-1} for the background (ttj) events with 0-6 jets.

Number of trees: In accordance with common practice, we choose the number of trees to be approximately the square root of the number of training/test instances.

3 Results and Discussion

3.1 BDT Classification and Overtraining Check

Figure 6 shows the classification values provided by the trained BDT over the test set. In general, a value of $+1$ corresponds to signal while a value of -1 corresponds to background. Thus, we see that the signal values tend towards the expected values while the background values are relatively scattered across the spectrum. However, these two distributions still provide a viable high-level mechanism for separating signal for background, especially since the classification of background for values close to $+1$ is sparse.

Furthermore, it is important to consider how much background we can reject while retaining as much signal as possible. This is represented in Figure 7. Ideally the curve would be similar to half of a square wave: a flat line of amplitude 1 for all values to 1 and then instantly dropping to 0. This indicates that we are rejecting the entire background while retaining any desired level of signal efficiency. However, as we see in our experimental results, the actual curve for background rejection begins at 1 and then slopes down towards 0 as signal efficiency increases.

We stated earlier that BDTs comprise a robust learning method that can be applied to a wide variety of data sets and that in particular, they tend not to overtrain on data. This allows them to classify new instances of data independently of any bias that could have been provided, induced,

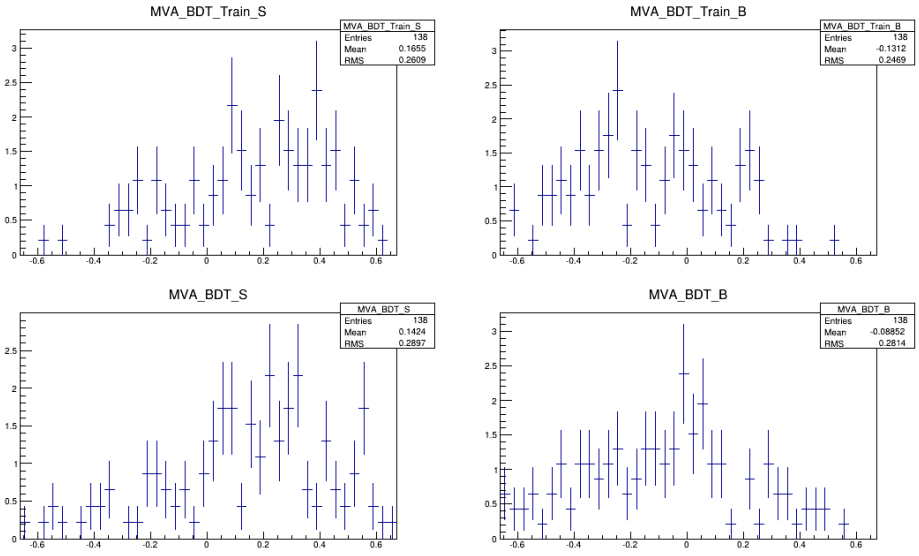

Figure 6: BDT classification values over the training (top) and test (bottom) data sets as signal (left) and background (right).

or otherwise present in the training data set. However, in our analyses, we saw that the signal efficiencies at different background efficiencies for the training set was significantly less than those for the training set (see Table 1). Therefore, we can see that our data has fallen prey to overtraining, a result we suspect stems from our small data set. With larger data sets, we should be able to build a more robust classifier.

Background Eff.	**Signal Eff. (Training)**	**Signal Eff. (Test)**
0.01	0.199	0.119
0.10	0.455	0.303
0.30	0.746	0.670

Table 1: Signal efficiency vs. background efficiency for training and test data sets being processed by our BDT.

3.2 Significance: Background Rejection vs. Signal Efficiency

In the "cut and count" experiments performed by Dutta et al. [7], the background is separated into three categories: ≤ 2 jets, ≥ 3 jets, and "other."

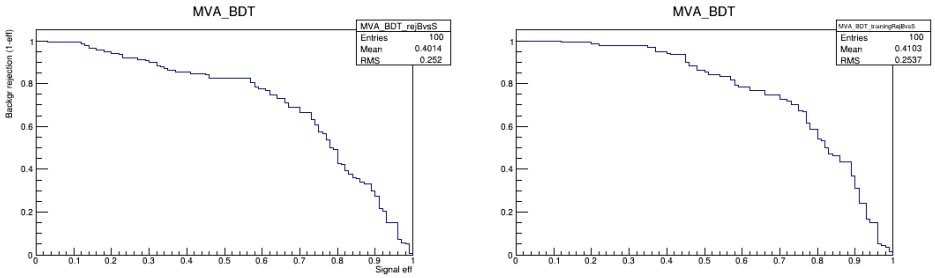

Figure 7: Rejection of background vs. signal efficiency for both test (left) and training (right) data sets.

The background efficiencies obtained in the supersymmetric parameter space which we confine ourselves to ($m(\tilde{t}) = 400$ GeV and $m(\tilde{\chi}_1^0) = 100$ GeV) are 0.26, 0.28, and 0.14, respectively. The signal efficiency is 0.20. They scale their significance to a cross section of 50 fb^{-1} and obtain:

$$\text{Significance} = \frac{S}{\sqrt{B}} = \frac{50 \times 0.20}{\sqrt{50 \times (0.26 + 0.28 + 0.14)}} = \mathbf{1.71}. \tag{1}$$

However, since we only look at the rejection of top quark backgrounds, we define a **figure of merit** (fom):

$$\text{fom} = \frac{S}{\sqrt{B_{top}}}, \tag{2}$$

where B_{top} is the number of top quark background events. Thus, the fom for [7] is:

$$\text{fom} = \frac{S}{\sqrt{B_{top}}} = \frac{50 \times 0.20}{\sqrt{50 \times (0.26 + 0.28)}} = \boxed{\mathbf{1.92}}. \tag{3}$$

In the calculations for our signal and background efficiencies, we had to scale our numbers by their expected baseline ratios. They reported a scaled signal ratio of 5.55 after baseline cuts and background ratios of 192 for ≤ 2 jets and 147 for ≥ 3 jets. Therefore:

$$\text{Signal Efficiency} = S = 50 \times 5.55 \times 0.119 = \mathbf{33.02} \tag{4}$$

$$\text{Top Background Efficiency} = B_{top} = 50 \times (192 + 147) \times 0.01 = \mathbf{169.5} \tag{5}$$

Thus, we obtain a significance of:

$$\text{fom} = \frac{S}{\sqrt{B_{top}}} = \frac{33.02}{\sqrt{169.5}} = \boxed{\mathbf{2.54}}, \tag{6}$$

indicating that our BDT approach is more effective than the "cut and count" approach in distinguishing signal from background.

4 Conclusions and Future Work

We have provided a machine learning approach using boosted decision trees to search for stop quarks based on the theoretical framework set forth by Dutta et al. [7] We are able to successfully separate supersymmetric signal from top-top background in a more sophisticated approach than the iterative "cut and count."

Although we have nuggets of promise in our initial experiments, there are a few things we can do to improve the quality of our results:

- Currently we using a limited data set that represents a cross-section of 10 fb^{-1}. We would like to run these experiments on larger data sets and cross sections corresponding to around 100 fb^{-1}.

- We also do not consider the backgrounds classified as "Other." A simple improvement would be to train on the existing backgrounds but test on the "Other" backgrounds and see how well the classifier performs.

- From a software engineering perspective, we began with highly inefficient programs that mined the data primitively and ended with a more robust approach that was nearly 10 times faster. However, there are tweaks in efficiency yet to be implemented that will be particularly useful when scaling our methods to larger data sets.

- We focused on the supersymmetric parameter space that was most favored by Dutta et al. [7]; however, we would like to use the holistic nature of the BDT approach to examine multiple supersymmetric parameter spaces. This is likely to be particularly effective in evaluating the viability of those parameter spaces since our initial results seem to provide a more sophisticated search mechanism for the stop quark.

- We would like to examine the individual decision trees more precisely and see if we should adjust the construction of the decision trees. For instance, we could let the decision trees have different heights, or we could experiment with different pruning approaches to fine-tune the trees after the initial stage of construction.

The potential impact of this project after the above improvements have been implemented lies in our ability to strongly suggest the evidence of a supersymmetric particle that is likely to exist given recent experimental discoveries and solid theoretical groundwork. Furthermore, we would be able to detect it *now* in the current parameter space. This decreases our dependence on physical and mechanical constraints, particularly those of the collider where the "real data" would be obtained from (as opposed to the simulated data that we are using).

5 Acknowledgements

First and foremost, I would like to thank Dr. Paul Padley for advising me on this research project. This work would have been impossible without his guidance, suggestions, stories, and lighthearted humor.

I would also like to thank the group of Bhaskar Dutta for providing us with both the theoretical basis for this work as well as the phenomenological data which made our experiments possible. In particular, I would like to thank Kuver Sinha for giving a talk about the theory behind their work and for meeting with us to discuss research possibilities and nuances.

In addition, Robert Brockman simultaneously worked on the same data using a neural network and facilitated the genesis and cross-fertilization of many ideas that arose during the course of this research.

Lastly, my heartfelt thanks go to Vinita Israni and my parents Sandip and Debjani Sen for constantly supporting me in my endeavors and pushing me to be better each day.

References

[1] The ROOT System Home Page, 2008.

[2] Search for the standard model higgs boson decaying to a w pair in the fully leptonic final state in pp collisions at sqrt(s) = 8 tev. Technical Report CMS-PAS-HIG-12-017, CERN, Geneva, 2012.

[3] V.M. Abazov et al. A precision measurement of the mass of the top quark. *Nature*, 429:638–642, 2004.

[4] Pushpalatha C. Bhat. Multivariate analysis methods in particle physics*. *Annual Review of Nuclear and Particle Science*, 61(1):281–309, 2011.

[5] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[6] Serguei Chatrchyan et al. Measurement of the Top-antitop Production Cross Section in *pp* Collisions at $\sqrt{s} = 7$ TeV using the Kinematic Properties of Events with Leptons and Jets. *Eur.Phys.J.*, C71:1721, 2011.

[7] Bhaskar Dutta, Teruki Kamon, Nikolay Kolev, Kuver Sinha, and Kechen Wang. Searching for top squarks at the lhc in fully hadronic final state. *Phys. Rev. D*, 86:075004, Oct 2012.

[8] Yoav Freund and Robert E. Schapire. A short introduction to boosting. In *In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1401–1406. Morgan Kaufmann, 1999.

[9] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

[10] Andreas Hoecker, Peter Speckmayer, Joerg Stelzer, Jan Therhaag, Eckhard von Toerne, and Helge Voss. TMVA: Toolkit for Multivariate Data Analysis. *PoS*, ACAT:040, 2007.

[11] Oded Maimon and Lior Rokach, editors. *The Data Mining and Knowledge Discovery Handbook*. Springer, 2005.

[12] Robert E. Schapire. The strength of weak learnability. *Mach. Learn.*, 5(2):197–227, July 1990.

[13] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. PYTHIA 6.4 Physics and Manual. *JHEP*, 0605:026, 2006.

[14] Zhenyu Ye. Top Quark Mass Measurements at the Tevatron. 2011.