Bachelor's Thesis Astronomy

Applications of Boosted Decision Trees for Cosmic Ray Particle Identification

Aakash Bhattacharya S3767663

Supervisors: Dr Manuela Vecchi, Marta Borchiellini

Abstract

The AMS-02 is at the forefront of cosmic ray identification, being the first detector with the potential to detect cosmic antiparticles. This relies heavily on accurate velocity measurements in order for proper mass reconstruction and particle identification. This is particularly challenging for single charge particles whose signal is easily disrupted by noise, negatively affecting the velocity reconstruction. This project aims to identify variables to which can be used to train BDTs to effectively discriminate background and clean data. The BDT was successful in cleaning the data and producing more clear mass peaks. 9 variables were identified to help clean the data.

Contents

Abstract.. 1
Acknowledgements..1
Contents.. 3
Chapter 1: Introduction..4
Chapter 2: Theory.. 5
2.1) Cosmic rays...5
2.2) AMS-02..5
2.2.1) RICH...7
Cherenkov radiation...8
2.3) Particle identification...10
Chapter 3: Methodology...11
Elementary cuts.. 11
Variable selection...12
BDTs.. 14
Chapter 4: Results...16
Chapter 5: Conclusion.. 23
Bibliography..24
Appendix.. 25

Chapter 1: Introduction

Cosmic rays (CRs) are high energy charged particles which originate from outside the solar system. They were discovered in the early 1900s, following the discovery of X-rays, radioactivity and the electron[1]. Viktor Franz Hess was able to measure an extraterrestrial source of ionizing radiation which increased with altitude, by attaching electroscopes to balloons. He later went on to win a Nobel prize for his discovery[2]. As studies on CRs progressed, from ground based to space-based spectrometry, much has been learned about the composition of CRs. Through particle identification we know that the most abundant species are hydrogen (H) and helium (He), and that protons dominate about 90% of the CR spectrum[3].

This paper focuses on the AMS-02 experiment, a high precision particle detector aboard the international space station (ISS). It is able to make precise measurements of charge, velocity, and momentum of the particle, which allows the reconstruction of mass and particle identification. It is also able to detect the sign of the charge of a particle, giving it the potential to detect the first CR anti nuclei[4]. This relies heavily on the ability to reconstruct a precise velocity measurement, which is difficult for single charge isotopes, whose background noise easily disrupts velocity reconstruction.

This study aims to employ boosted decision trees (BDTs) to clean data in order to obtain more accurate velocity reconstructions which are used for reconstructing mass distributions of CR events. Mass distributions calculated using AMS data taken over a month will be compared, using standard selection cuts, and employing a BDT to make further cuts. Chapter 2 will discuss the relevant theory, covering the AMS-02 experiment and relevant concepts for particle detection. Chapter 3 will be an explanation of the methodology used, with a brief section explaining the use of BDTs. Chapter 4 will present the final results and discuss their implications and finally Chapter 5 will present the general conclusions of this paper.

The broader motivation for this work lies in the search for dark matter. It was first discovered in the early 20th century, where by comparing rotation curves and velocities of objects in the sky, it was found that the mass of galaxies and clusters are much higher than what can be observed in visible matter. In recent cosmology, it has been concluded that less than 20% of the universe's matter is baryonic, and that the rest must be dark matter[5]. However the remaining dark matter mass has not been observed yet. The neutralino (χ) has been discussed as a potential dark matter candidate. It is predicted that its annihilation could produce a low energy antideuteron flux. Studies have been conducted using monte carlo simulations to potentially identify this antideuteron flux, should the predictions for the neutralino model be true[4]. This is a truly interesting theory and field which could lead to new revelations about dark matter and stress the general relevance of this study on noise reduction.

Chapter 2: Theory

This chapter includes all the relevant theory in order to understand the meaning of this work. It will cover a description of the AMS experiment, focussing on the RICH detector explaining Cherenkov radiation and finally covering particle identification .

AMS-02

This study utilized data from the AMS-02. At the forefront of precision particle detection in space, the AMS has been taking data for over 12 years, having detected over 220 billion CR events[7]. As aforementioned, it is able to make precise measurements of the velocity, momentum, charge and charge sign and rigidity of particles, which allows for particle identification.

The AMS consists of 7 sub-detectors and a magnet, as illustrated in figure 1. The Transition Radiator Detector (TRD) separates electrons and positrons by transition radiation, and is also able to determine the charge of nuclei by measuring the energy loss (dE/dx). The Time of Flight (TOF) detector consists of four scintillator planes, two above and two below the magnet. The planes consist of 8-10 paddles, equipped with 2 or 3 photomultiplier tubes (PMT) on the end. By detecting the time taken Δt for a particle to traverse the length of the magnet as well as the length of said path L, the tracker is able to measure $\beta = \Delta t / cL$ with a resolution of about 4% for particles with Z=1 and $\beta \approx 1$[6]. The Anti-Coincidence Counter (ACC) rejects particles with high incidence angle and the ECAL is used for lepton–hadron separation and measuring the energy of the particle[8].

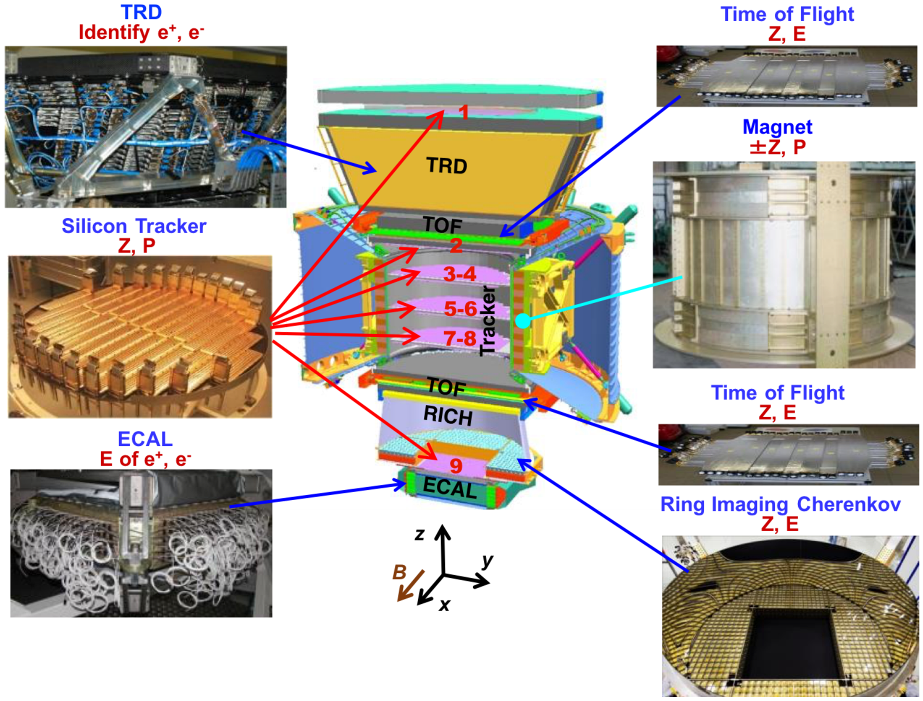


Figure 1: Schematic drawing with real images of the AMS detectors[7].

The silicon tracker, magnet and RICH are relatively more important for this work. The silicon tracker consists of 9 layers, one above the TRD, one above the ECAL hole and 7 layers inside the magnet which make up the inner tracker. The magnet has a strength of 0.15T and works together with the tracker to help identify particles. Each of the tracker layers detects the two-dimensional coordinates of the particle, which can then allow for the reconstruction of the trajectory and curvature of the path. As shown in figure 2, this allows a geometric determination of the momentum given by equation 1, where z is the charge and e is the charge of an electron[2].

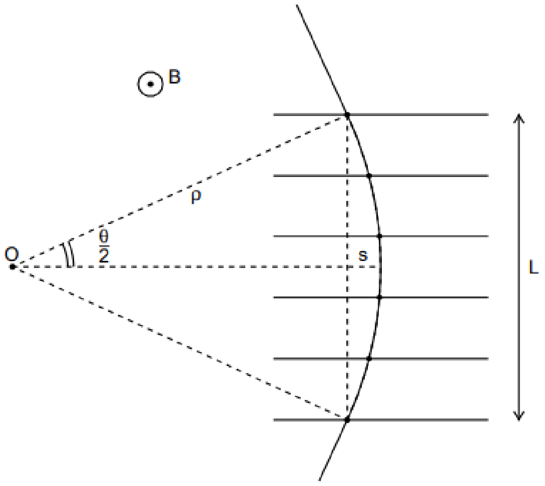


FIG. 41: Sketch of the momentum measurement in a uniform magnetic field (in the bending view). Here and in the following of this section L is the height of the region permeated by the magnetic field, ρ is the radius of curvature of the track, θ_{B} is the deflection angle, and s is the sagitta of the track.

Figure 2: geometrical values used to determine momentum from the particle track in the detector[2]

$$p = \rho z e B \tag{1}$$

Once the momentum and charge of the particle are known, the rigidity can be defined and determined as

$$R = \frac{pc}{ze} \tag{2}$$

This constant is very useful as it is proportional to the mass, velocity $(p = mv)$ and charge of the particle which can all be used for particle identification. However, it does not allow for convenient mass reconstruction without an accurate velocity measurement, which is made in the RICH and will be the focus of this paper.

RICH

The RICH is a ring imaging cherenkov detector. As shown in figure 3, the detector consists of a radiator plane, a conical reflector and a photodetection plane. The reflectors allow for more signals to be saved, as they reflect events with high incidence angles back to the photodetection pane. The photodetection plane is made of 680 PMTs with a hole in the middle of 64 × 64 cm2 which corresponds to the effective area of the ECAL below it.

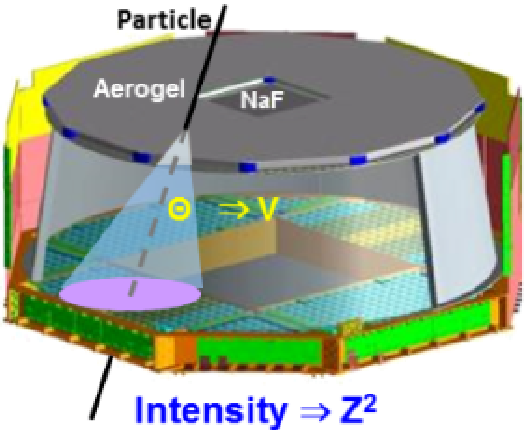


Figure 3: A diagram of the AMS-02 RICH detector[7]

Cherenkov radiation

The RICH detector is able to measure the velocity of particles through a phenomenon known as Cherenkov radiation. Cherenkov radiation occurs when a charged particle passes through a dielectric medium at a speed greater than the speed of light in the medium. This causes a light cone to be formed, as shown in figure 4[1]. Due to the fact that the particle must travel faster than the speed of light in the medium, the threshold for cherenkov radiation occurs if

$$v > \frac{c}{n}, \; \beta \geq \frac{1}{n} \tag{3}$$

Where v is the velocity of the particle, c is the speed of light and n is the refractive index of the medium.

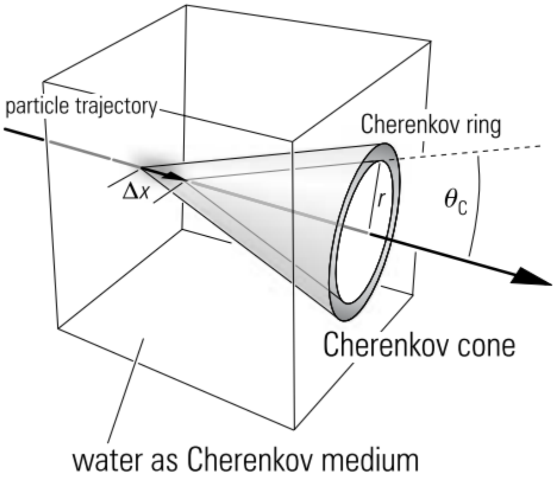


Figure 4: Diagram illustrating a three dimensional view of Cherenkov cone formation in water[1] (Note the AMS uses Aerogel and Sodium Fluoride)

The angle formed inside the cone is known as the cherenkov angle θ_c and is given by[1]:

$$cos(\theta_c) = \frac{1}{\beta n} \tag{4}$$

The RICH is able to distinguish events by detecting scattered noise hits from the Cherenkov ring, followed by an eventual strong hit near the center of the region when the particle crosses the PMT plane itself, as shown in figure 6. Figures 5 and 6 show an example of a simulated proton event from the side and from the top of the detector respectively. Figure 5 shows the particle going through the whole detector whereas figure 6 shows only the radiator and photodetection plane. This is also a good example as it shows the use of the conical reflector saving the edge of the cherenkov ring. The AMS detector then has two methods of reconstructing the charge and velocity. The first method is a geometrical reconstruction based on the valid hits recorded. The second method considers all hits recorded, and uses a maximum of a likelihood function to determine a reconstruction[4].

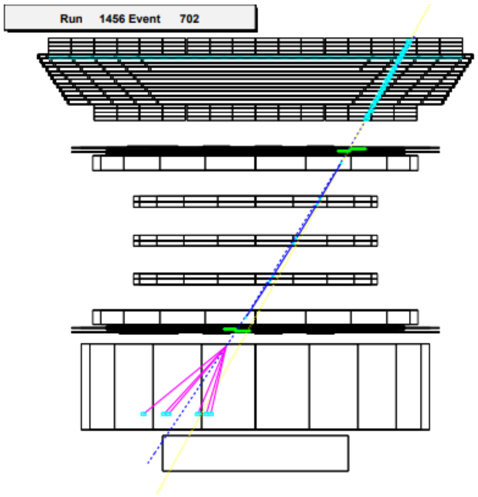


Figure 5: A simulated proton event in the AMS-02[4]

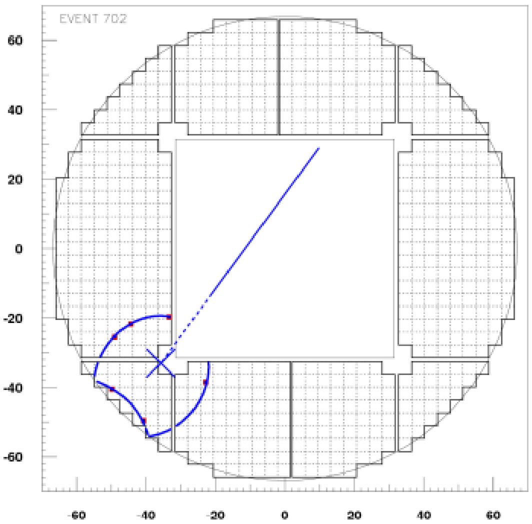


Figure 6: The same event from figure 5 as seen in the RICH[4]

The radiator plane is made of panels of sodium fluoride (NaF) and silica aerogel (AGL). The central panels consist of 16 NaF tiles of 85 × 85 × 5 mm^3. Surrounding these tiles are 92 113 × 113 × 25 mm^3 tiles made of the AGL. From equation 1 we can calculate threshold beta values for which the AGL is able to detect particles are $0.96 < \beta < 0.996$ with an accuracy of $\Delta\beta/\beta\,(\beta = 1,\ Z = 1) \approx 0.12\%$. The NAF is only able to detect particles with $0.75 < \beta < 0.97$, to an accuracy of $\Delta\beta/\beta\,(\beta = 1,\ Z = 1) \approx 0.35\%$. The Naf has a refractive index of n = 1.33, whereas the AGL has one of n = 1.05. Although the NaF has a lower resolution, these tiles are necessary to ensure that the light cones of particles passing through the area will become large enough to avoid the ECAL hole in the bottom of the detector[6]. This is because the NaF has a higher refractive index than the AGL and will therefore form larger light cones to avoid the ECAL since equation 4 shows that a higher cherenkov angle implies a higher refractive index at a given velocity. This also makes it better suited for detecting lower velocity particles, since equation 3 states that a lower beta and higher refractive index will result in the same size cone.

Particle identification

Now combining measurements from both the tracker and the RICH, the mass can be reconstructed using equation 5. The error propagation is shown in equation 6[8]. The fourth power dependence of the resolution on the Lorentz factor shows the importance of an accurate velocity measurement as it improves as beta →1.

$$m = \frac{RZe}{\beta\gamma} = \frac{RZ\sqrt{1-\beta^2}}{\beta} \tag{5}$$

$$\left(\frac{\Delta m}{m}\right)^2 = \left(\frac{\Delta R}{R}\right)^2 + \gamma^4\left(\frac{\Delta\beta}{\beta}\right) \tag{6}$$

This stresses the general importance of this study, as although this paper only focuses on general noise reduction, advances in the area could lead to better noise reduction and as mentioned in the introduction, identification of cosmological antimatter.

Chapter 3: Methodology

This chapter will discuss the methodology employed in this study. It can be broken into three elementary steps. The first step consisted of plotting a simple mass distribution with basic cuts which ensured that the events had a clean track through the detector. These cuts also split the data between events that passed through the NaF radiator versus those that passed through the AGL. This gives a baseline to compare the results of the BDT training. The second step was to look for variables which could suggest where background noise comes from as well as be used to help identify signals from the background. Plots of background and signal hits were made to determine that there was a clear correlation to signal and background for each variable. Finally the BDTs were trained using the TMVA framework and a cut was applied using the BDT classifier and compared to the selection cut plots to look for noise reduction.

Elementary cuts

The following tables show a list of all the basic cuts made, along with explanations on their functions[6]. The masses were reconstructed using equation 5, using the rigidity found in the tracker and the beta factor found in the RICH.

Table 1: Cuts to ensure a good trigger for the detectors

Physics trigger	A part of the AMS called the trigger processor collects fast inputs from the TOF, ACC, and ECAL, checks them according to the goal of the detection, and then sends out a trigger signal for the detectors to start their readout cycles. In this case, the four TOF planes are checked for signals, along with making sure that the ACC did not receive too many rejected signals. This releases a trigger for the measurement of a Z=1 particle[10].

Table 2: Cuts made based on path reconstructions

Inner tracker track	The particle passed though the inner track
Good Inner tracker chisq	The inner tracker has a good χ^2 value
Single track	A single track was detected in the tracker

Good tracker fiducial volume	A good fiducial volume from the first TOF plane to the PMT plane on the RICH.
Good L1 reconstruction	A good velocity construction in the first plane of the tracker
Has Track in L1 FV	The particle has a good track in the fiducial volume of the first tracker plane

Table 3: Cuts based on Charge reconstructions

Good unbiased L1 Z	Removes outliers, such that the largest change found between L1 and L9 are 0.8<q<1.6
Good L1 Z	L1 of the tracker has a good charge between 0.8 and 1.6
Good TOF Z	TOF charge is between 0.75 and 1.25
Good inner tracker Z	The inner tracker charge (L2-L8) are between 0.75 and 1.5

Table 4: Cuts used to separate AGL from NaF events

Track in AGL	Geometric interpolation of the path from the tracker shows the particle passed through the AGL radiator plane
AGL beta within threshold	0.96 < beta<0.996
Track in NaF	Geometric interpolation of the path from the tracker shows the particle passed through the NaF radiator plane
NaF beta within threshold	0.75 < beta<0.97

Variable selection

The next step was plotting signal vs background distributions for different variables were plotted in order to assess whether they would be appropriate use for training the BDT. The plots were made assuming a mass resolution of about 10%[6] , therefore signal-like events were considered with a mass reconstruction within 2σ from the proton mass, being $0.75 < m < 1.12\, GeV/c^2$. Events were considered background-like with a mass reconstruction greater than 4σ from the proton mass. Variables 8 and 9 are extensions to the literature used for this study.

1) RICH charge → The RICH defines the charge Z as $Z = \sqrt{N_{pe}/N_{exp}}$ where N_{pe} is the number of photoelectrons (PE) collected in the ring and N_{exp} is the number of PEs a Z=1 particle could emit, based on the reconstruction in the detector for the event.
2) Kolmogorov probability → This variable compares the distribution of the PEs in the ring with a uniform distribution as they should be uniformly distributed. A high probability indicates a high similarity between the distributions.
3) Fraction of PEs in the ring→ This compares the number of PEs that were considered in the reconstruction to the total number of PEs detected.
4) Number of unused hits → The number of unused hits when reconstructing the ring. Events with more unused hits could be more likely to be noisy or have a faulty reconstruction.
5) Number of PMTs → Events where there are a large number of PEs and a small amount of PMTs tend to be noisy. Furthermore events with large numbers of PMTs tend to also be noisy, as a low number of PEs are expected due to the single charge and the relation in equation 3.
6) Radiator impact point (x, y): → Indirect hits on the radiator panels tend to lead to poor reconstructions due to fewer Cherenkov photons being detected. This could be due to hitting the edge of the radiators or the borders between the tiles.
7) RICH/TOF velocity ratio: → The particle should have a consistent velocity in the detector. Should the velocities measured by the RICH and TOF vary, this suggests something wrong in the reconstruction and will lead to noisy events.
8) Reflected hits → Events which are reconstructed using reflected hits, such as the example in figure 7 could tend to lead to more noisy reconstructions due to the added interactions in the detector as well as relying on more computing in order to reconstruct the ring accordingly.
9) Impact angle → The impact angle could relate to noise, as higher impact angles could be related to more reflected hits, as well as more chance that the particles have gone through interactions earlier in the detector such to disturb the path and reconstruction. Furthermore they could suggest something about the path taken through the detector and in combination with reflected hits and the impact point could be used to determine if there are potential biases in the instrument. Finally this parameter also scales with the total distance through the radiator as well as the RICH fiducial volume which further suggest it could be a candidate for cleaning data.

In order to assess whether a variable would be appropriate to use for training the BDT, we looked for variables that showed differences between the distribution for signal vs background. This means that the BDT can use them to classify events better. Plots of all the variables are included in the appendix, however Figure 7 shows an example of the charge measured by the RICH for AGL events. This plot shows that the RICH charge has a strong relation to both signal and background events. By looking at the graphs we could make intuitive cuts at $0.6 < Z < 1.4$, however this still leaves a large overlap of background and signal events, showing the room for BDT training. There is a clear distinction between the two distributions, making it a good variable to use. Figure 8 shows the distribution for the radiator impact plane in the y axis for the NaF. It is much less clear where a cut could be made, and while the distributions do look very similar, one can see that the signals tend to be clearer closer to the center, whereas the background events tend to spike at the edges.

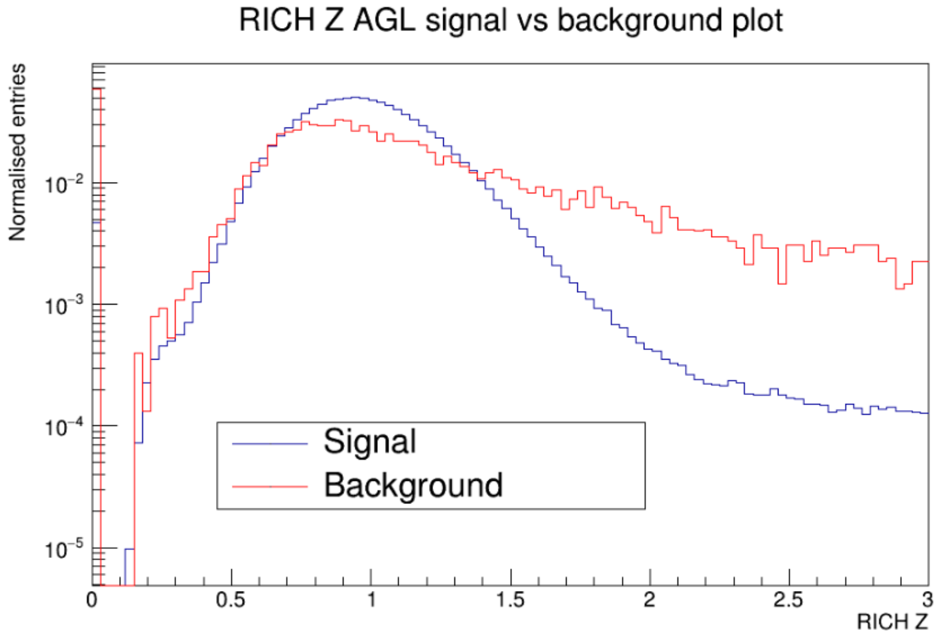


Figure 7: Signal vs background separation for the charge measured by RICH

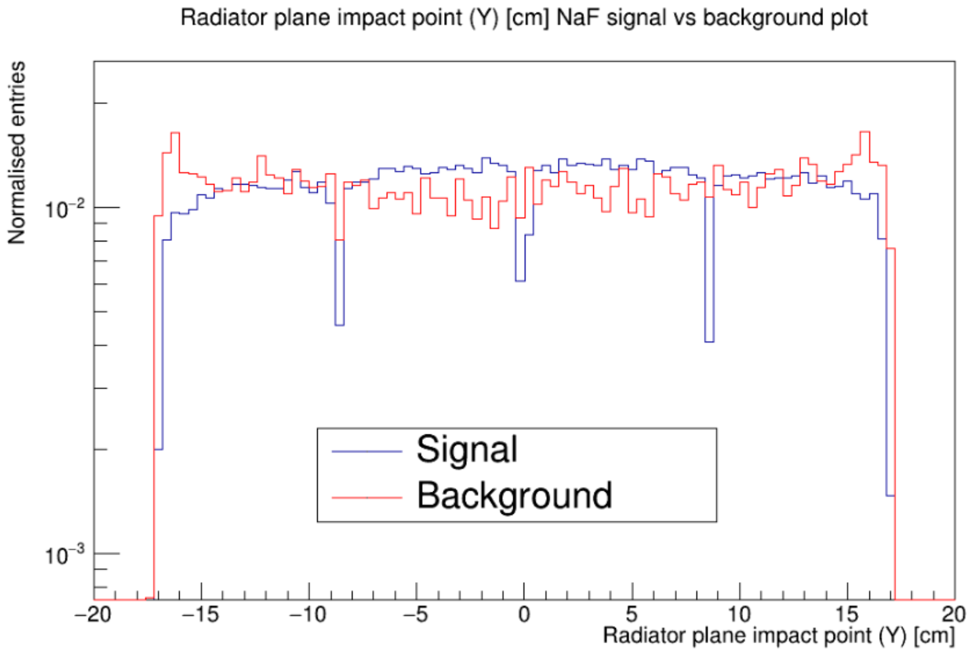


Figure 8: Signal vs background separation for the Y coordinate of the impact point in the NAF radiator

BDTs

The final step was to set up and train the BDTs using the variables selected in the previous step. Boosted decision trees are an application of machine learning which are an extension of decision trees. This project will use them to help identify signal vs background events and they will be explained in this context.

Decision trees work by using a cut-based multivariate technique. They can be visualized as trees, where conditions are called nodes which extend through branches as shown in figure 9. Final nodes represent a decision and are called leafs. Decision trees follow the structure from top to bottom, checking each decision node until a leaf is reached. A limitation to binary trees is that in our case, not all signal events may satisfy all of the conditions set to identify them as such. For example, if when using a tree which makes a cut that all events with more than 4 reflected hits are background, a large amount of events are lost which could potentially be signal. This is comparable to using a cut based selection of events, as was done in the first step.

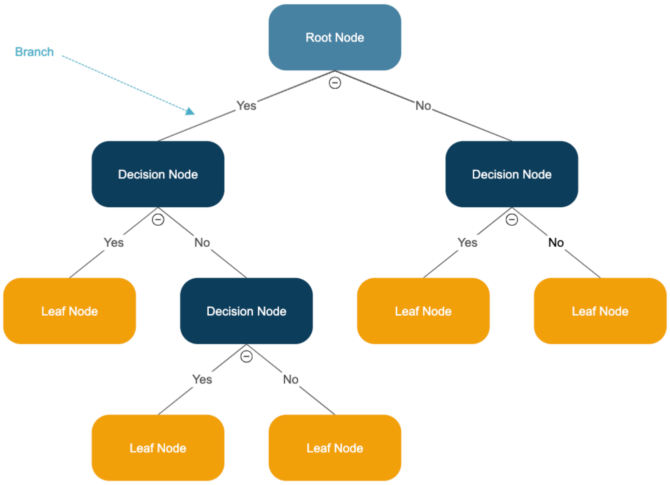


Figure 9: Explanation of decision trees[11]

BDTs are, as the name suggests, an even more powerful form which uses boosted trees. For this project the TMVA framework[10] was used, which allowed for easy manipulation of the data which was in ROOT files. ROOT is a framework which allows for object oriented programming in order to handle the data. The BDTs were set up with the Adaboost algorithm and used the Gini index to weight variables as will be explained.

The Adaboost algorithm trains to use the variables given as decision nodes called classifiers. This algorithm trains each classifier taking into account the abilities of the previous classier. This is done through utilizing stumps, which are decision trees for variables known as weak classifiers. These are very short decision trees, usually consisting of a node and two leafs, in this case classification of either signal or background identification.

The algorithm first makes classifications with each of the stumps using a dataset and measures the amount of true and false positives and negatives. In this case the amount of correctly vs incorrectly classified signal vs background events as shown in table 5. All the events are given an equal weight at first.

Table 5: Nomenclature for classifications

	True (correctly classified)	False (incorrectly classified)
Positive (signal)	Correctly classified signal	Incorrectly classified signal
Negative (background)	Correctly classified background	Incorrectly classified background

From here the first stump is chosen by comparing the Gini indices of the different variables. The Gini index is defined as the probability of misclassification, so the number of false positives and negatives over the total number of classifications. Once the first stump is chosen, the sample weights of incorrectly classified events must be increased in order to allow the BDT to focus on these more, and the weights of correctly classified events must be decreased. This is done using the following equations:

$$Classifier\ weight\ =\ \frac{1}{2}log(\frac{1-total\ error}{total\ error}) \tag{7}$$

$$New\ incorrect\ sample\ weight\ =\ old\ weight\ \times\ e^{classifier\ weight} \tag{8}$$

$$New\ correct\ sample\ weight\ =\ old\ weight\ \times\ e^{-classifier\ weight} \tag{9}$$

Where the total error is the sum of the weights of the misclassified events.

These new weights are now normalized and the next stump is selected using a weighted Gini index, where the number of events are counted up in terms of their weights. Finally to make a classification once the BDT is trained, the sum of each classifier's decision and weight is added up for an event and the decision (signal vs background) is made based on the leaf with the most weight.

Training the BDT required the data to be split into two sets, one for training and eventually one to test it on. For this reason when training the BDTs, the data was split into two sets of odd vs even numbered events. Furthermore they were split between AGL and NaF events, as with the rest of the work. The dataset of odd events was then used to train the BDT and then tested on the even events, and vice versa after which the two sets were combined again. This ensured that no data was wasted and reduced the chance of overtraining. Overtraining occurs due to the fact that not all the data accurately represents the underlying signal that is being identified. This may cause the BDT to target noise or irrelevant features of the data, which makes the cleaning less generalizable[12].

The results section will discuss interpreting the output of the BDTs in order to receive the final mass distributions.

Chapter 4: Results

This section will present and explain the results found through this study. Figure 10 shows the normalized mass distributions after making the elementary cuts as discussed in step 1 of the methodology. Traces of bumps around Z = 1 and Z=2 can be seen, and it is also clearly visible that the AGL has much less noise towards the tail end. This is due to the AGL having more surface area, therefore collecting more events which statistically makes it tend more towards the distribution we are trying to model.

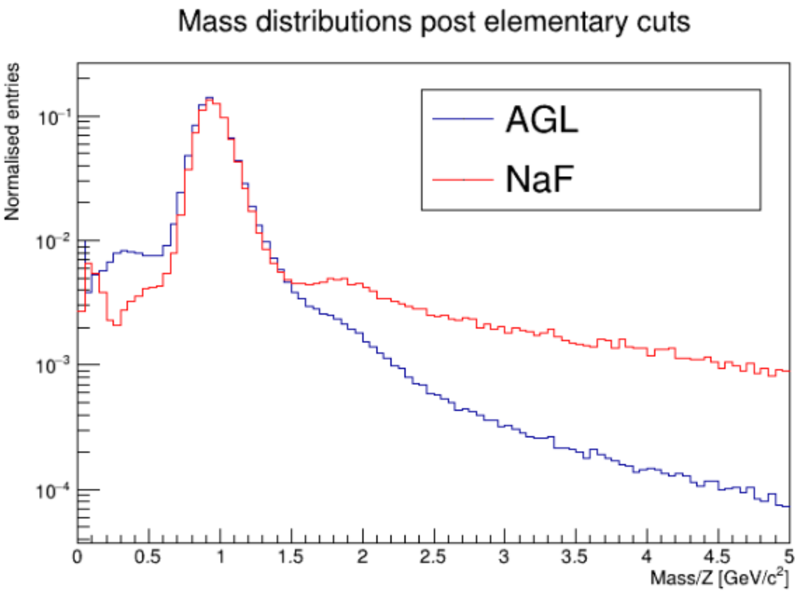


Figure 10: Elementary mass distributions found after applying basic selection cuts to the data

For the second step of variable selection, the following list discusses each of the variables patterns and motivation to train along with some supporting figures. In general, all the variables plotted were found to have clear enough distinctions between signal and background distributions. All the plots are included in the appendix.

1. RICH charge → As previously mentioned, by looking at figure 7 an intuitive cut can be made at 0.6<Z<1.4. The more the value varies from 1, the more chance of a mis-reconstruction.

2. Kolmogorov probability → Figure 11 shows the example for agl, where clearly there is a correlation with a higher probability showing less background and more signal.

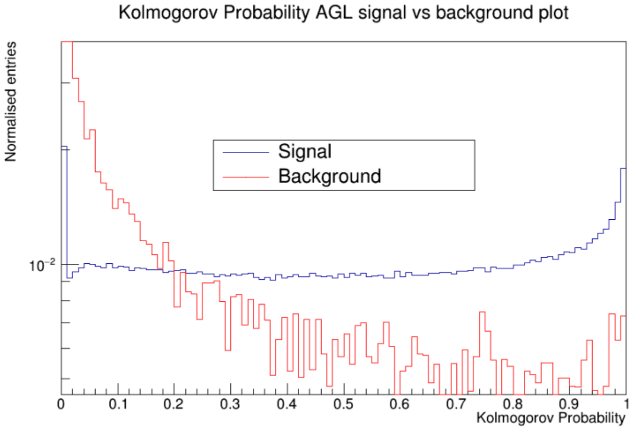


Figure 11: Kolmogorov probability Signal vs Background for AGL

3. Fraction of PEs in the ring→ Figure 12 shows when more PEs are used, a better reconstruction occurs.

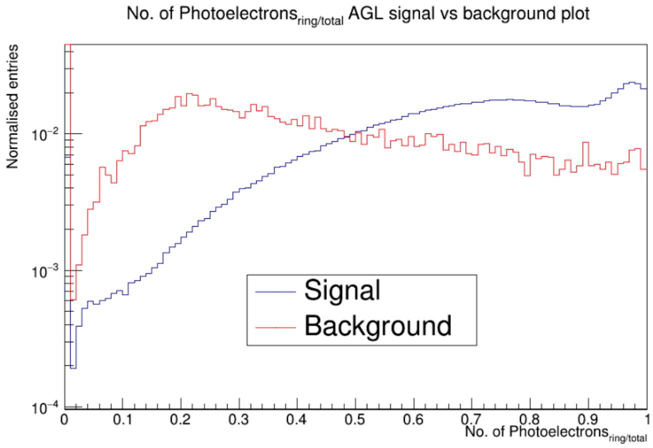


Figure 12: Fraction of PEs used Signal vs Background for AGL

4. Number of unused hits → Figure 13 shows a clear pattern, where the more unused hits tend to signal more poorly reconstructed events.

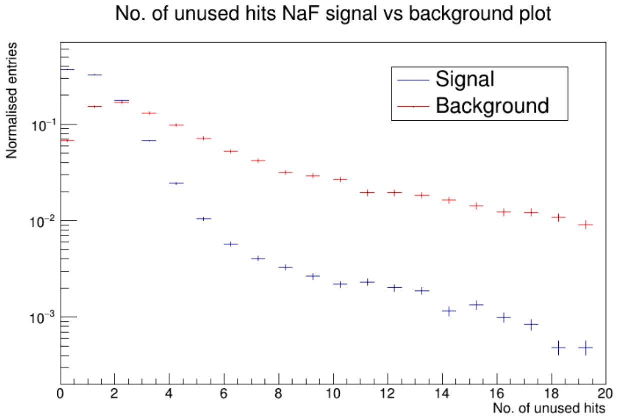


Figure 13: number of unused hits Signal vs Background for NaF

5. Number of PMTs → Figure 14 shows indeed that pmt<2 events have more chances of being poor reconstructions. Furthermore when too many are activated this tends to noise, which makes sense since there should not be too many PMTs being activated due to equation 3.

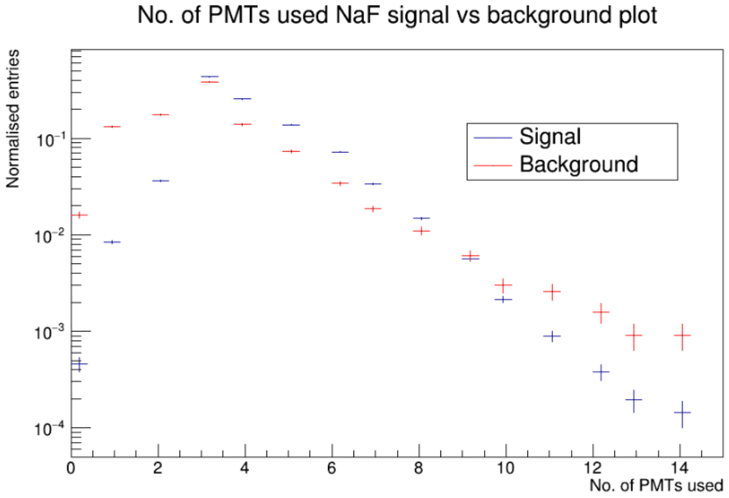


Figure 14: Number of PMTs Signal vs Background for NaF

6. Radiator impact point (x, y): → As previously discussed, figure 8 shows potential for background cleaning.
7. TOF/RICH velocity ratio: → Figure 15 shows an example, as the value varies from 1, there is more noise produced. Seemingly more so if the TOF records a higher velocity.

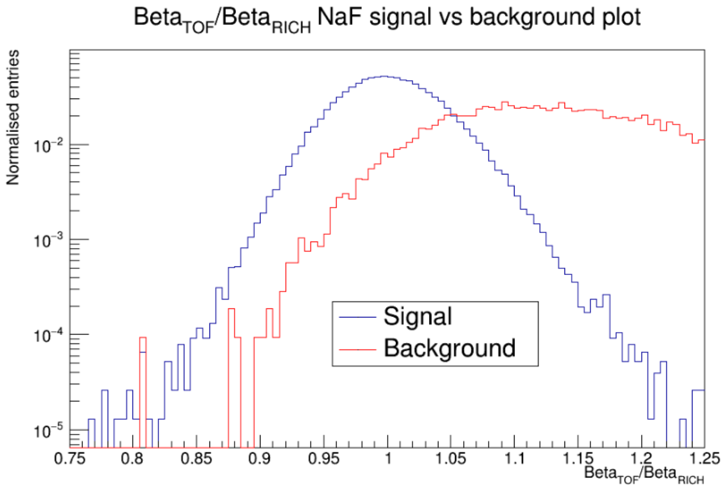


Figure 15: TOF/RICH Signal vs Background for NaF

8. Reflected hits → As expected the noise increased with the number of reflected hits used, as shown in figure 16.

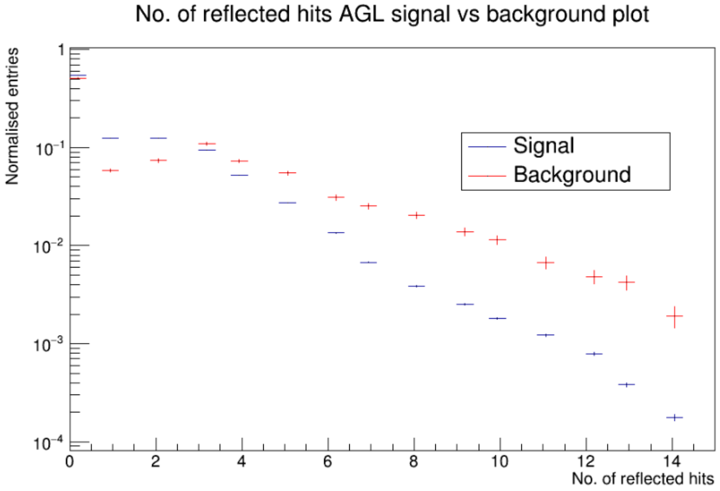


Figure 16: Reflected hits Signal vs Background for AGL

9. Impact angle → This variable also showed a high contribution to the background noise. Furthermore it seems there was some kind of bias in the AGL, figure 17, such that impacts around -pi/2 seemed to cause a peak in background events. This could be a random occurrence that is characteristic of the data being used, further tests would have to be done to confirm it but it does leave interesting room for the bdt to train on.

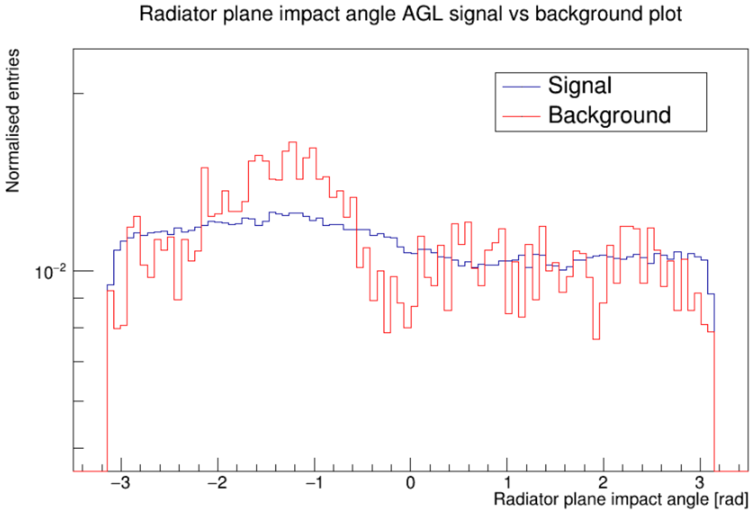


Figure 17: Impact angle Signal vs Background for AGL

Finally after training the BDTs the following forms of output were analyzed in order to make cuts on the data. Firstly the receiver operating characteristics (ROC) curves were checked in order to assess and confirm the training of the BDT. As shown in figure 18, ROC plots the true positive rate against the false positive rate, in this case signal efficiency against the background rejection. The closer the curve is to the top right corner, as shown in the example, the better. A failed classifier that simply randomly chooses a classification would be the same line but running through the center of the graph diagonally with no curve[12]. Furthermore the separation output distributions, as shown in figure 19 were used to ensure the BDTS were able to properly separate the two. As usual, the rest of the curves are included in the appendix.

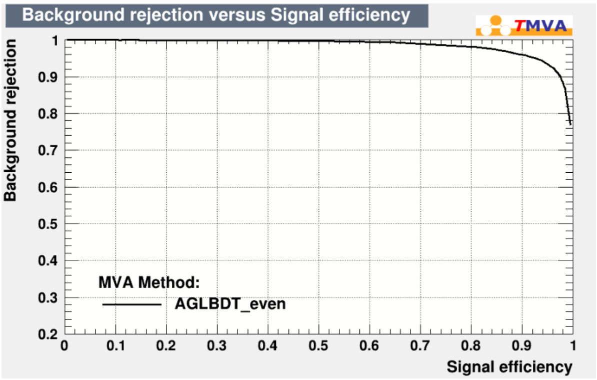


Figure 18: Example ROC curve for the even AGL data set

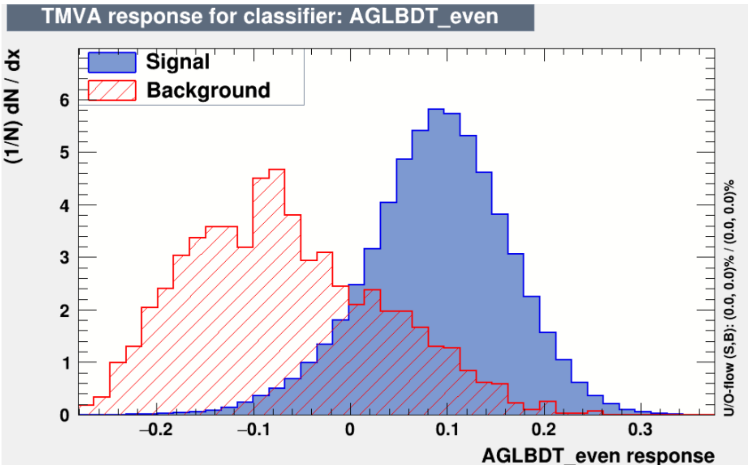


Figure 19: Example BDT output distribution for background vs signal

From here in order to select a cut value for the BDT classifier, cut efficiency graphs were analyzed, as shown in figure 20. This plot shows the signal efficiency (blue), background efficiency (red) and the cross section significance (green), which is defined as $S/\sqrt{(S + B)}$ and its maximum value predicts the best value to make the BDT cut at. The TMVA framework allows for the classifiers to be tested based on the amount of signal and background events, so these were plugged in and a number was selected from the plot, as illustrated in figure 20. Table 5 shows the number of events used for signal and background, AGL, NaF and even or odd events as well as the final values used to make the BDT cuts.

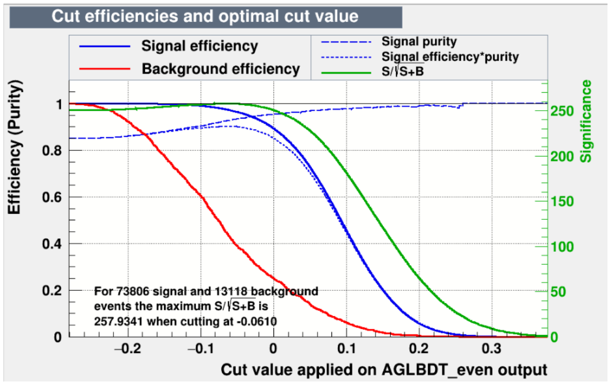


Figure 20: Example cut efficiency graph of signal efficiency (blue), background efficiency (red) and cross section significance (green).

Table 5: Number of events used for each dataset with final BDT values chosen

	AGL even	AGL odd	Naf even	NaF odd
Signal	73806	74035	18898	18838
Background	13118	13128	1221	1295
BDT cut value	-0.0610	-0.0601	0.0911	-0.1087

These cuts were then implemented and the final mass distributions were reconstructed, recombining the even and odd numbered events. The final plots shown in figure 21 and 22 show a clear success in cleaning the signal. The noisy high mass tails are significantly reduced and the BDT was in turn able to successfully identify more true signal events. This is also evident in the fact that the normalized distribution has a higher peak after the BDT training, showing that relatively more true positives were identified. This shows that the use of BDTs can in fact help with cleaning data and confirms the objective of this study.

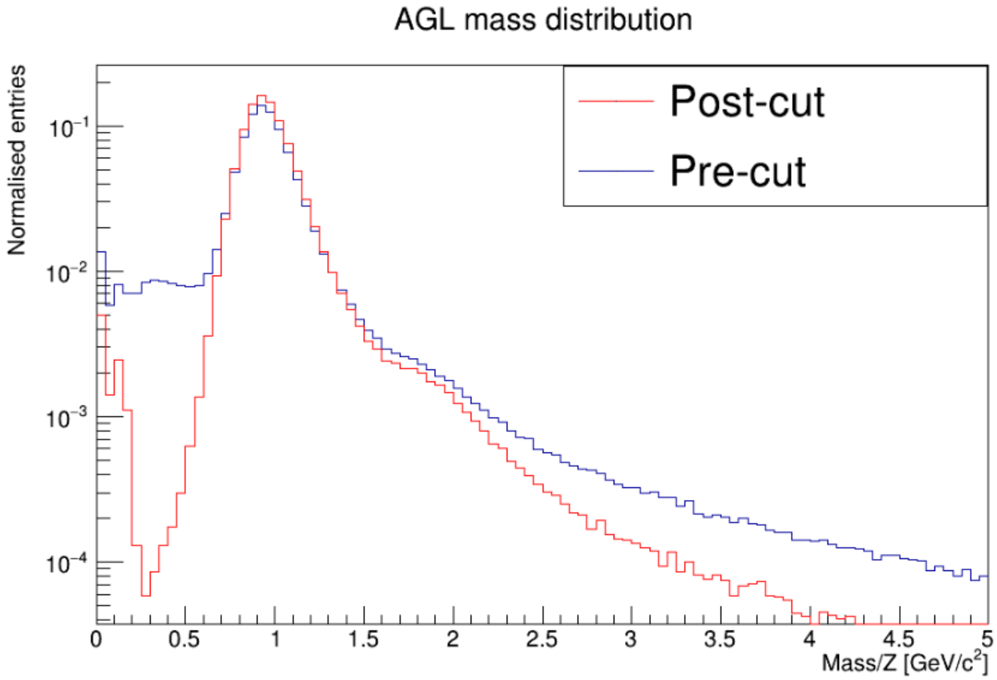


Figure 21: AGL mass distributions, before and after creating the BDT cut.

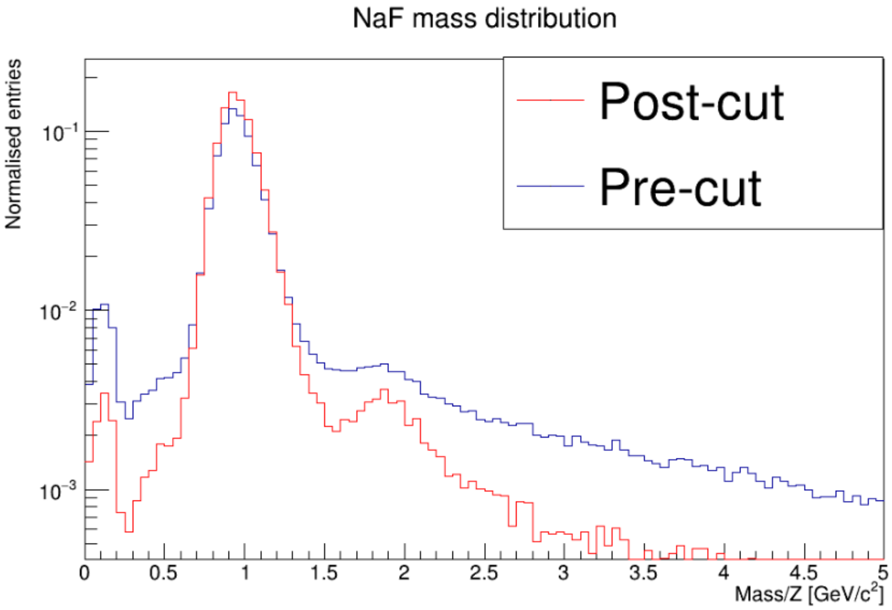


Figure 22: NaF mass distributions, before and after creating the BDT cut.

It is important to note further extensions to this study, the main one lying in potential optimisation of the BDTs. This project, being more so focussed on the astrophysical aspect rather than mastering machine learning, relied on basic adaboost separation using the Gini index. The main focus was between identifying potential sources of background to study and testing whether the BDTs help in doing so. However there are potentially more powerful machine learning techniques which could be employed. Furthermore, as mentioned in the introduction, the true potential in this study lies in the potential to apply these techniques in order to look for dark matter traces in the future[4].

Chapter 5: Conclusion

The aim of this paper was to investigate the potential for using BDTs to identify and clean background noise within the AMS-02 detector in order to obtain cleaner velocity and mass reconstructions for single charge particle identification.

The 9 variables tested were the RICH charge, Kolmogorov probability, the number of photoelectrons used in reconstruction, the number of unused hit, the number of PMTs activated, the radiator impact point, the ratio between beta measured by the RICH and TOF, the number of reflected hits used and the impact angle of the event. They were all found to have significant noise contributions and were used to train a BDT to clean data.

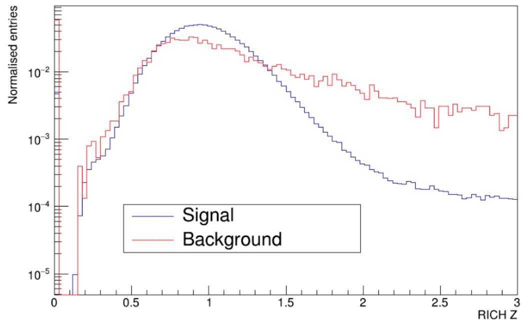
The adaboost algorithm was used with the Gini index for separation, and it was found to be effective in reducing noise and better identifying signal events. Further extensions to this research lie in both optimizing the use of BDTs or other machine learning algorithms to clean the data. The larger scope of this study lies in the potential to extend its findings to the observation of antimatter.

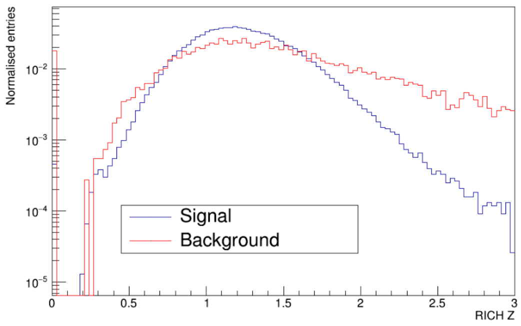
Bibliography

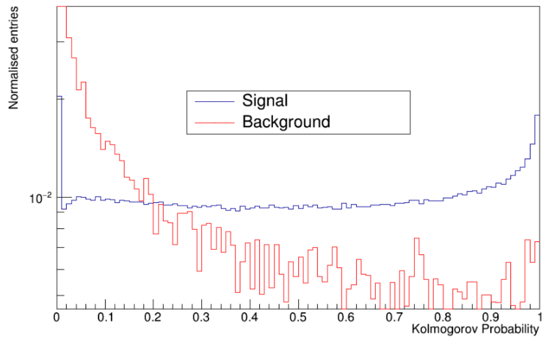
1) Grupen, Claus, and Tilo Stroh. *Astroparticle Physics*. Springer, 2020.
2) Baldini, Luca. "Space-Based Cosmic-Ray and Gamma-Ray Detectors: a Review." arXiv preprint arXiv:1407.7631 (2014).
3) Gomez-Coral, Diego Mauricio, et al. "Current status and new perspectives on cosmic ray deuterons." arXiv preprint arXiv:2303.09775 (2023).
4) Arruda, Luísa, Fernando Barao, and Rui Pereira. "Particle identification with the AMS-02 RICH detector: search for dark matter with antideuterons." arXiv preprint arXiv:0710.0993 (2007).
5) Bertone, Gianfranco, and Dan Hooper. "History of dark matter." Reviews of Modern Physics 90, no. 4 (2018): 045002.
6) Aguilar, M., et al. "The Alpha Magnetic Spectrometer (AMS) on the international space station: Part II—Results from the first seven years." Physics reports 894 (2021): 1-116.
7) “02: The Alpha Magnetic Spectrometer Experiment.” *AMS*, ams02.space/.
8) Bueno, E. F., F. Barão, and M. Vecchi. "A parametric approach for the identification of single-charged isotopes with AMS-02." Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 1031 (2022): 166564.
9) Giovacchini, F., et al. "The AMS-02 RICH detector: Status and physics results." Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 952 (2020): 161797.
10) Andreas Hocker et al. “TMVA - Toolkit for Multivariate Data Analysis”. Ch. 8.13 (2007). arXiv:`physics/0703039`.
11) “Decision Tree.” Decision Tree - Learn Everything About Decision Trees, www.smartdraw.com/decision-tree/
12) Coadou, Yann. "Boosted decision trees." Artificial Intelligence for High Energy Physics. 2022. 9-58.
13) Alpat, B., et al. "Charge determination of nuclei with the AMS-02 silicon tracker." Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 540.1 (2005): 121-130.
14) Li, Zi-Yuan, et al. "Antiproton identification below threshold with the AMS-02 RICH detector." Chinese Physics C 41.5 (2017): 056001.
15) Carnini, Marco, and Alessandro Pastore. "Trees and forests in nuclear physics." Journal of Physics G: Nuclear and Particle Physics 47.8 (2020): 082001.

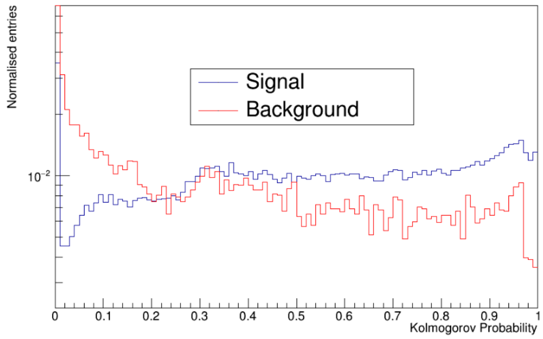
Appendix

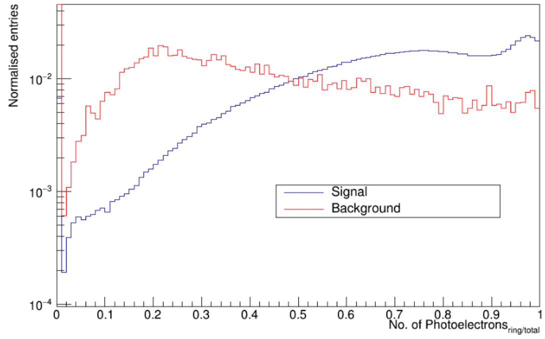
Variable separation plots:

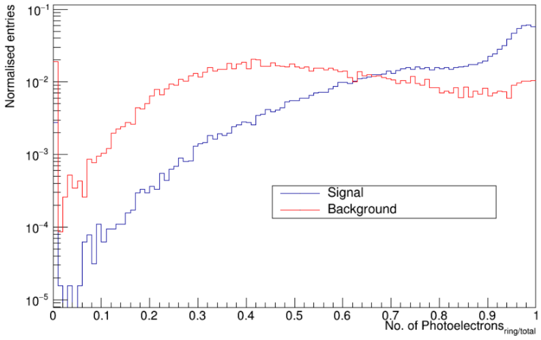

RICH Z AGL signal vs background plot

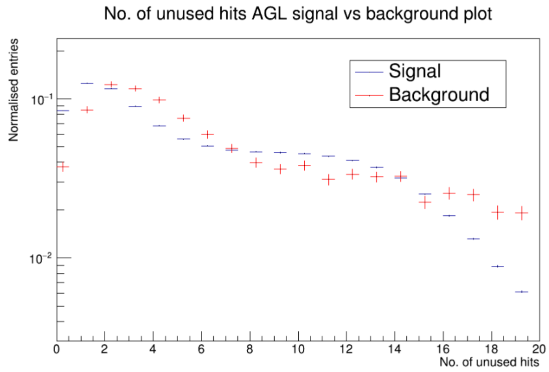

RICH Z NaF signal vs background plot

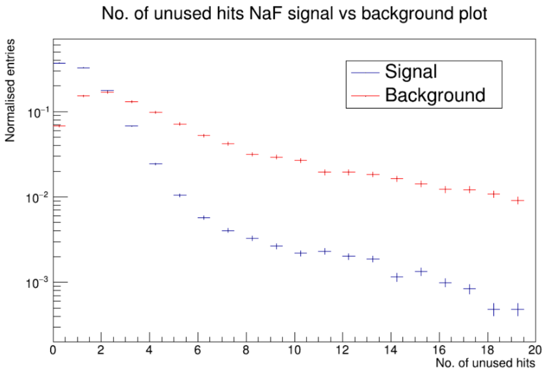

Kolmogorov Probability AGL signal vs background plot

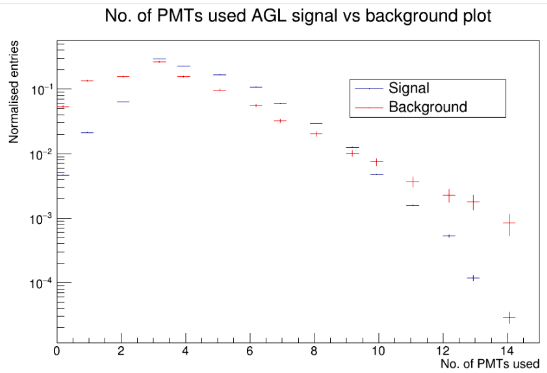

Kolmogorov Probability NaF signal vs background plot

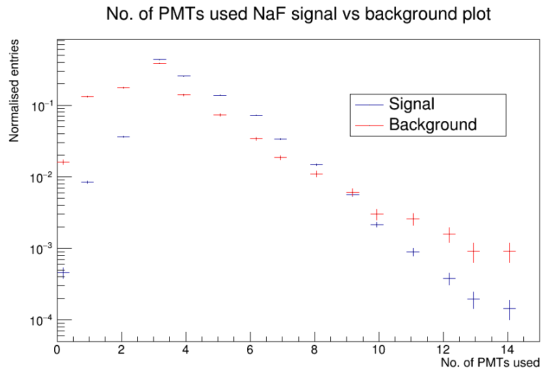

No. of Photoelectrons$_{\mathrm{ring/total}}$ AGL signal vs background plot

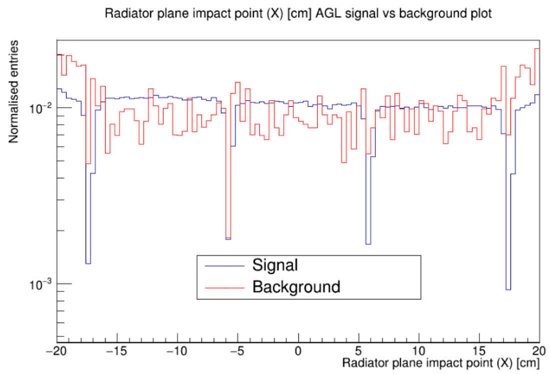

No. of Photoelectrons$_{\mathrm{ring/total}}$ NaF signal vs background plot

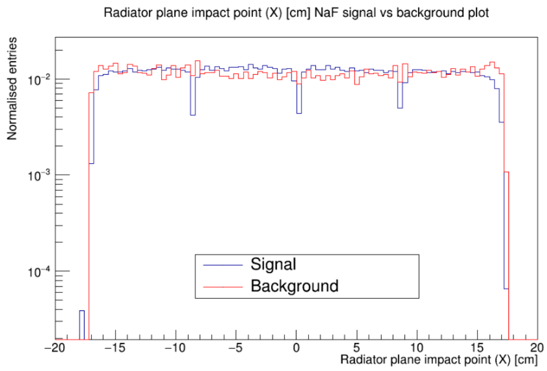

No. of unused hits AGL signal vs background plot
Normalised entries
10^{-1}
10^{-2}
0
2
4
6
8
10
12
14
16
18
20
No. of unused hits
Signal
Background

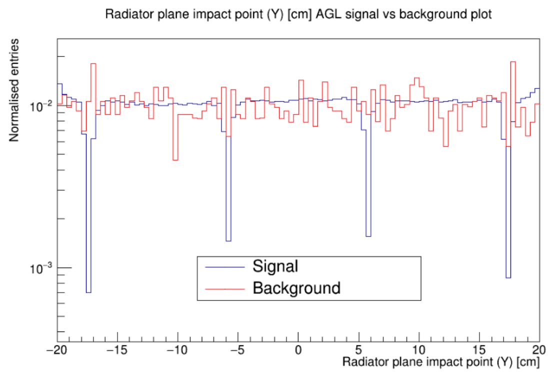

No. of unused hits NaF signal vs background plot
Normalised entries
10^{-1}
10^{-2}
10^{-3}
0
2
4
6
8
10
12
14
16
18
20
No. of unused hits
Signal
Background

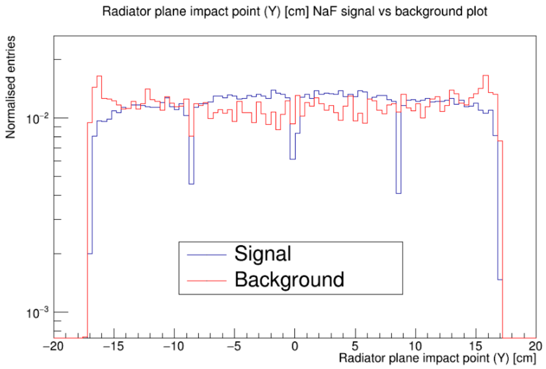

No. of PMTs used AGL signal vs background plot
Normalised entries
10^{-1}
10^{-2}
10^{-3}
10^{-4}
0
2
4
6
8
10
12
14
No. of PMTs used
Signal
Background


No. of PMTs used NaF signal vs background plot
Normalised entries
10^{-1}
10^{-2}
10^{-3}
10^{-4}
0
2
4
6
8
10
12
14
No. of PMTs used
Signal
Background

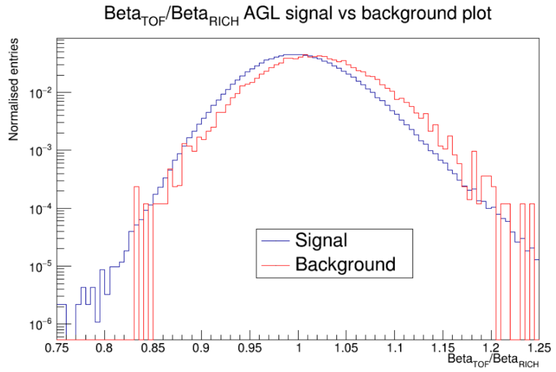

Radiator plane impact point (X) [cm] AGL signal vs background plot
Normalised entries
10^{-2}
10^{-3}
−20
−15
−10
−5
0
5
10
15
20
Radiator plane impact point (X) [cm]
Signal
Background

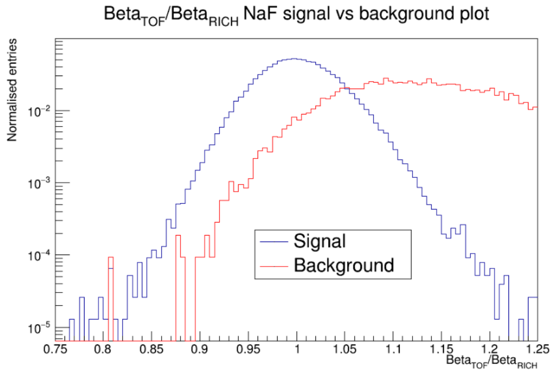

Radiator plane impact point (X) [cm] NaF signal vs background plot
Normalised entries
10^{-2}
10^{-3}
10^{-4}
−20
−15
−10
−5
0
5
10
15
20
Radiator plane impact point (X) [cm]
Signal
Background

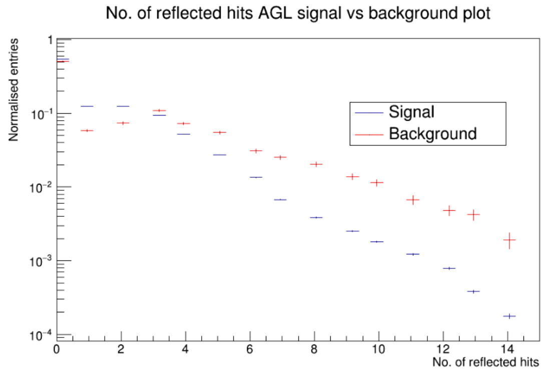

Radiator plane impact point (Y) [cm] AGL signal vs background plot
Normalised entries
10^{-2}
10^{-3}
−20
−15
−10
−5
0
5
10
15
20
Radiator plane impact point (Y) [cm]
Signal
Background

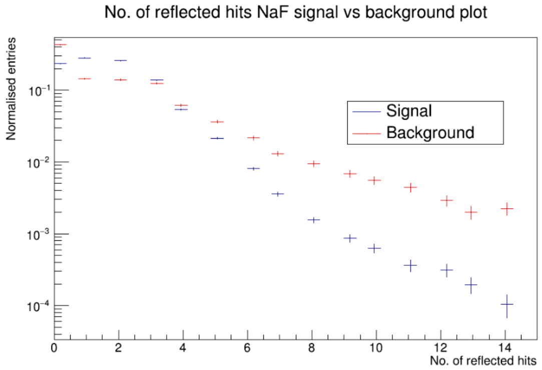

Radiator plane impact point (Y) [cm] NaF signal vs background plot
Normalised entries
10^{-2}
10^{-3}
−20
−15
−10
−5
0
5
10
15
20
Radiator plane impact point (Y) [cm]
Signal
Background

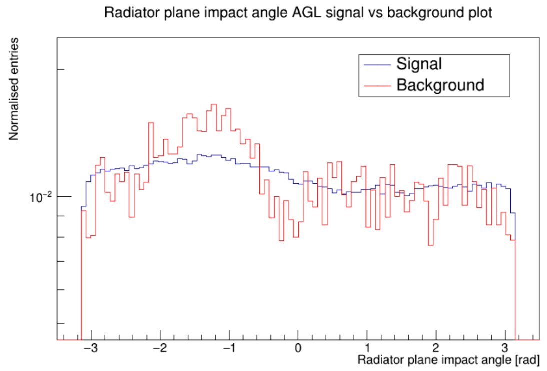
e

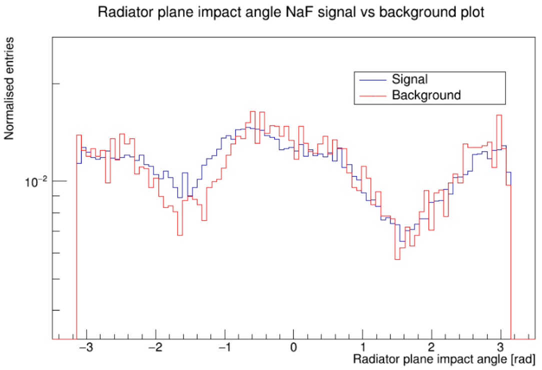

Beta$_{TOF}$/Beta$_{RICH}$ AGL signal vs background plot


Beta$_{TOF}$/Beta$_{RICH}$ NaF signal vs background plot

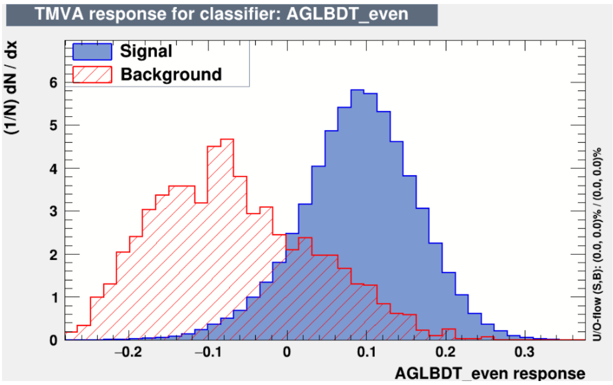

No. of reflected hits AGL signal vs background plot

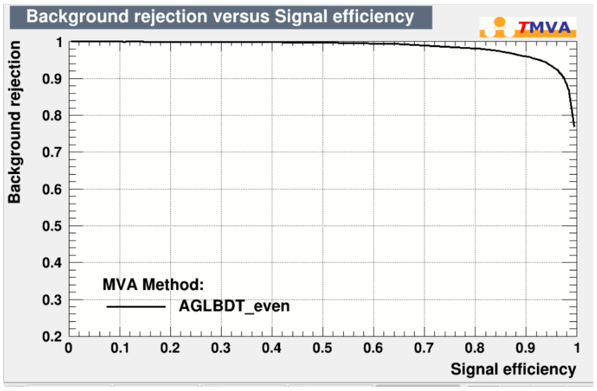

No. of reflected hits NaF signal vs background plot

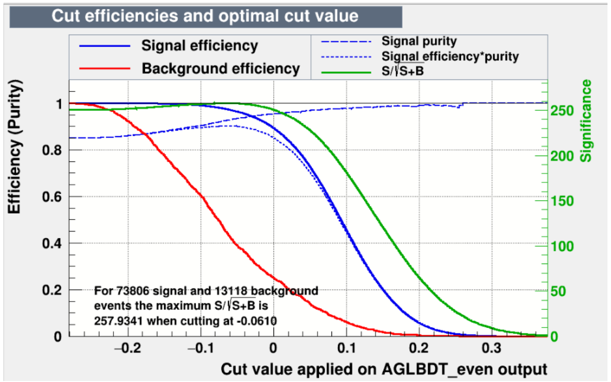

Radiator plane impact angle AGL signal vs background plot


Radiator plane impact angle NaF signal vs background plot

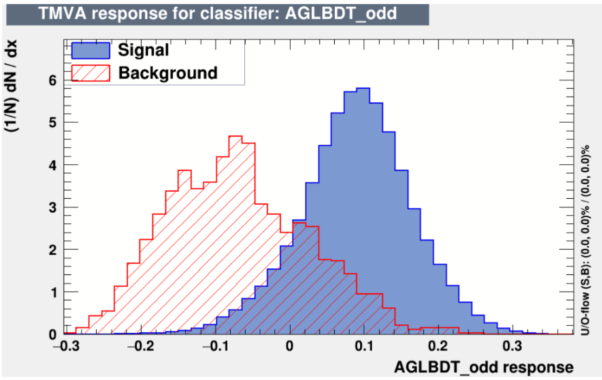
BDT output, ROC & cut efficiency plots

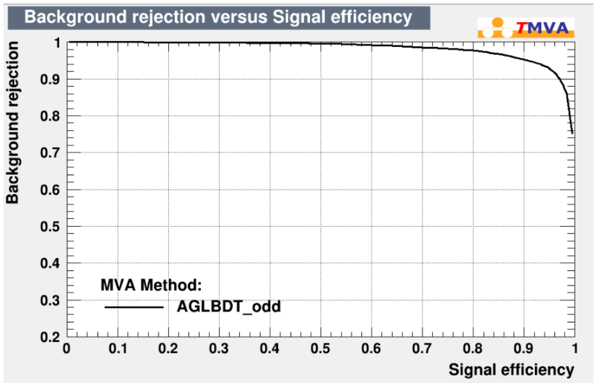


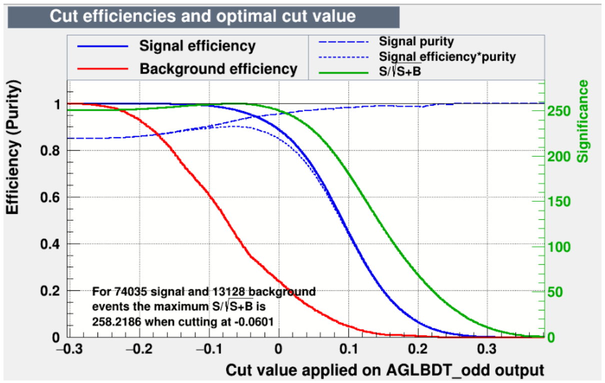




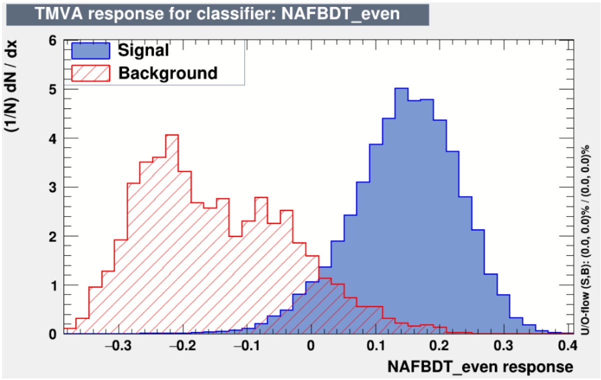
For 73806 signal and 13118 background events the maximum S/√S+B is 257.9341 when cutting at -0.0610

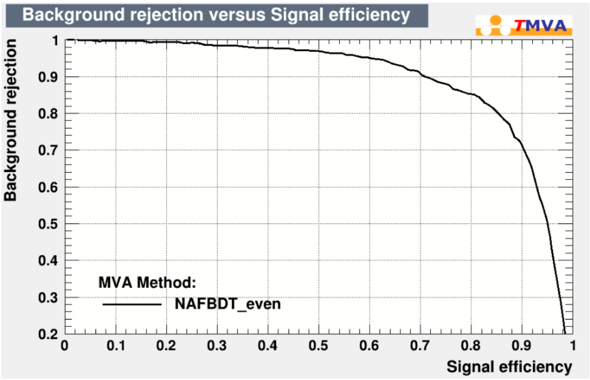


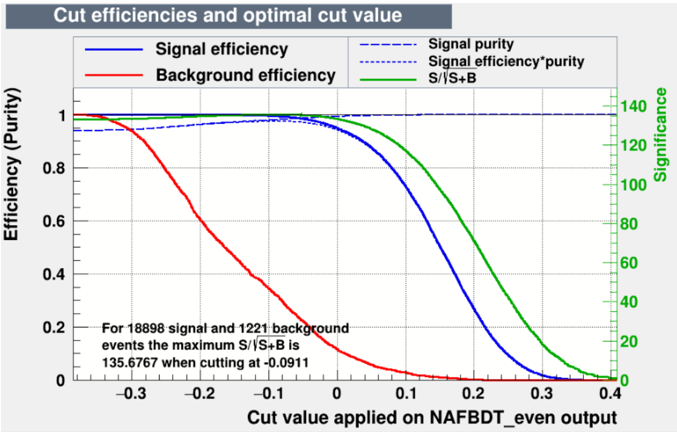


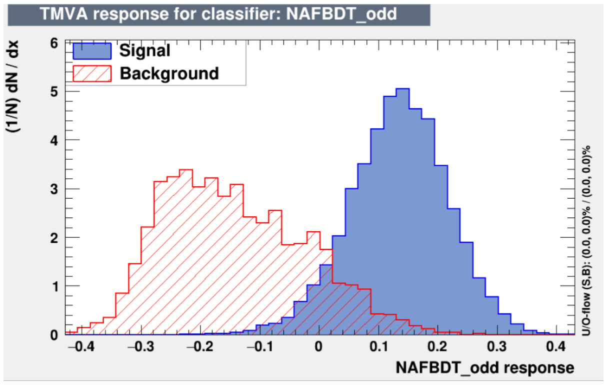


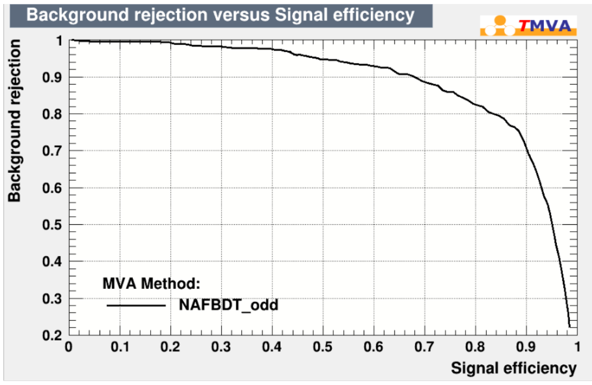
For 74035 signal and 13128 background events the maximum S/√S+B is 258.2186 when cutting at -0.0601

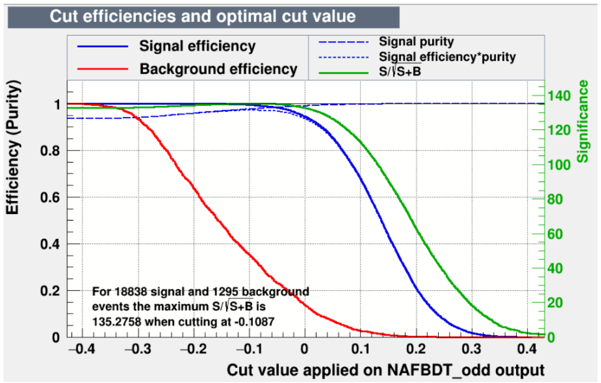

TMVA response for classifier: NAFBDT_even
Signal
Background
(1/N) dN / dx
NAFBDT_even response
U/O-flow (S,B): (0.0, 0.0)% / (0.0, 0.0)%
0
1
2
3
4
5
6
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4


Background rejection versus Signal efficiency
TMVA
Background rejection
Signal efficiency
MVA Method:
NAFBDT_even
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Cut efficiencies and optimal cut value
Signal efficiency
Background efficiency
Signal purity
Signal efficiency*purity
S/√S+B
Efficiency (Purity)
Significance
Cut value applied on NAFBDT_even output
For 18898 signal and 1221 background
events the maximum S/√S+B is
135.6767 when cutting at -0.0911
0
0.2
0.4
0.6
0.8
1
0
20
40
60
80
100
120
140
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4


TMVA response for classifier: NAFBDT_odd
Signal
Background
(1/N) dN / dx
NAFBDT_odd response
U/O-flow (S,B): (0.0, 0.0)% / (0.0, 0.0)%
0
1
2
3
4
5
6
−0.4
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4


Background rejection versus Signal efficiency
TMVA
Background rejection
Signal efficiency
MVA Method:
NAFBDT_odd
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Cut efficiencies and optimal cut value
Signal efficiency
Background efficiency
Signal purity
Signal efficiency*purity
S/√S+B
Efficiency (Purity)
Significance
Cut value applied on NAFBDT_odd output
For 18838 signal and 1295 background
events the maximum S/√S+B is
135.2758 when cutting at -0.1087
0
0.2
0.4
0.6
0.8
1
0
20
40
60
80
100
120
140
−0.4
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4