Gamma/Hadron Separation Methods with VERITAS

Maryn K. Askew

McGill University Department of Physics

August 13th, 2023

Abstract

A fundamental challenge faced by Imaging atmospheric Cherenkov telescopes is the suppression of cosmic-ray background events which constitute the vast majority of detections, while keeping good sensitivity to gamma-ray events. With previous success in many fields of physics, machine learning has emerged as a viable approach to discriminate between signal and background events. This is achieved by characterizing air showers based on differences in their shower image properties, thereby determining whether the primary particle is more likely to be a gamma-ray or a hadronic particle. In this study, machine learning techniques are applied to simulated data and subsequently to real Crab data from the Very Energetic Radiation Imaging Telescope Array System (VERITAS) to compare with VERITAS's standard moderate cuts for Crab-like sources.

1 Introduction

Advancements in technology have enabled thorough investigation of the most intricate aspects of the universe. Ground-based detectors incorporating imaging atmospheric Cherenkov telescopes (IACTs) aim to study the most energetic phenomena, searching for astrophysical sources that emit gamma-ray radiation. However, a significant challenge faced by the high-energy physics (HEP) community lies in discriminating between gamma-ray and hadronic events, as cosmic-ray showers largely dominate the detections made by these ground-based detectors.

By discerning the distinctions between between gamma-ray showers and hadronic events, one can characterize the air shower and separate the two. Upon the arrival and interaction of a primary particle with the Earth's atmosphere, a cascade of secondary particles called and extensive air shower occurs, which can then be parameterized by data collected by a ground-based detector [1]. Gamma-ray showers are mainly composed of electron-positron pairs, whereas hadronic showers are predominantly composed of muons resulting from the decay from charged pions and kaons. The presence of muons tend to cause hadronic showers to be more spread out, while gamma-ray showers tend to be more compact and focused.

In addition to the effective method of employing "box" cuts to separate gamma/hadron events, machine learning techniques such as Boosted Decision Trees (BDTs) and Neural Networks (NNs) have gained increasing traction within the realm of IACTs like VERITAS [2], HESS [3], and MAGIC [4], and others. Machine learning techniques leverage the differences to build models that can discriminate between gamma-ray and hadronic events based on features extracted from the Cherenkov images captured by the telescopes. By learning from simulated data and real observations, these techniques enhance the accuracy of separation, enabling a more precise identification of gamma-ray sources amidst the background of cosmic-ray events.

1.1 The VERITAS Observatory

The Very Energetic Radiation Imaging Telescope Array System (VERITAS) is a ground-based observatory designed to detect very-high-energy (VHE) gamma rays from celestial

sources with the highest sensitivity in the range of about 100 GeV to 10 TeV [5]. It is located at the Fred Lawrence Whipple Observatory in southern Arizona, USA and operates using an array of four 12m imaging atmospheric Cherenkov telescopes, each equipped with a large mirror and a camera system of 499 individual photomultiplier tubes to achieve a 3.5 degree field of view. Because the majority of detections made by VERITAS are cosmic-rays, gamma/hadron separation is a necessity to study gamma-ray astronomy. The main method of gamma/hadron separation that VERITAS uses are standard "box" cuts which involves applying cuts on relevant air shower parameters to isolate and retain as many gamma events as possible while rejecting hadron events.

1.2 VERITAS Air Shower Features

One way IACTs differentiate between particle air showers is by using Hillas parameters [6]. These parameters describe the elliptical images captured by the telescopes. The width and length of such images depend on the primary particle. The combined weighted averages of these measurements, derived from multiple telescopes, yield two vital parameters for air shower classification: the mean reduced scaled width (MCSW) and the mean reduced scaled length (MSCL). It is expected that the MSCW and MSCL distributions are centered around zero for gamma-ray showers while hadronic showers are longer and wider. Another distinguishing parameter is the second brightest pixel, labeled as SizeSecondMax. Typically, gamma-ray showers are expected to be brighter than hadronic showers on average, resulting in a larger value for SizeSecondMax. The emission height and shower core are also characterized by the primary particle. Lastly, the distributions for energy and emission height χ^2 values as well as the dES are different between gamma-ray showers and hadronic showers, so they can also be valuable for classification tasks. To summarize, the air shower parameters used to determine between a gamma-ray or hadronic event are: MSCW, MSCL, EmissionHeight, EmissionHeightChi2, EChi2S, SizeSecondMax, Core, and dES.

2 Data Sets

2.1 Monte Carlo Simulations

For the training and validation of the machine learning models, Monte Carlo (MC) simulations of VERITAS–like data were generated by two different software packages. Extensive air showers were simulated by the program CORSIKA (COsmic Ray SImulations for KAscade) [7] while the light from those showers were processed using GrOptics [8]. GrOptics simulates the response of the telescopes and produces a similar read out that one would obtain by collecting data from the actual VERITAS array. For gamma/hadron separation, two particle air showers were simulated. One simulation of approximately 400,000 events with electrons as the primary particle were treated as gamma-ray air showers since after the first interaction the showers are essentially identical. Then, hadronic air showers were simulated with protons as the primary particle, which consisted of 1.7 million events that were treated as background. Appendix A shows the distributions of each of the relevant air shower parameters for electron and proton showers.

2.2 Real VERITAS Crab Data

In order to test whether our classifiers truly work, data available from VERITAS on the Crab nebula between 2013 to 2018 ($\sim$17 hours) were selected. The Crab holds significant importance as it is the remnant of the historical supernova event SN 1054 which was documented by Chinese astronomers in the year 1054 [9]. The Crab is considered a standard in the VHE field for its brightness and steady flux at TeV energies. It is one of the brightest sources of gamma-ray emissions in the sky which has made it a natural laboratory for studying high energy particle interactions and calibrating instruments.

3 Gamma/Hadron Separation Techniques

3.1 Machine Learning Techniques

Machine Learning Techniques (MLTs) can either be supervised or unsupervised. Supervised MLTs are used to take known data to train and fit a model that can later be used to predict unknown data. Classification is an example of supervised machine learning. In contrast, unsupervised machine learning is applied to previously unknown data where the machine works by itself to figure out patterns in the data. Two machine learning models were used with the goal of classifying gamma-rays and hadrons in order to successfully separate the events while keeping most of the gamma-ray events and rejecting hadron events. Due to their previous success in the VHE field, the two models used are Boosted Decision Trees (BDTs) and Neural Networks (NNs). The general consensus for both techniques is that BDTs outperform NNs, however a BDT is very prone to overfitting such that altering a parameter by a tiny bit drastically affects the performance of the BDT. For this classification task, gamma-ray events are assigned with the number 1 while hadron events are assigned with 0 for both the BDT and NN. After the models are trained and fit to the simulated data, they output a probability of whether the event is a gamma-ray or hadron.

3.2 Boosted Decision Trees

Boosted decision trees are powerful ensemble learning techniques used for both classification and regression tasks, combining the strengths of traditional decision trees and boosting algorithms. A decision tree is a flowchart-like structure where each internal node represents a feature or attribute, each branch represents a decision rule based on that attribute, and each leaf node represents an outcome or prediction. The pros of decision trees are that they are easy to interpret, but they are very prone to latching onto peculiarities in data, leading to the concept of overfitting. Boosting is an iterative ensemble learning technique that combines multiple weak learners into a strong learner. It aims to improve the model's performance by focusing on the mistakes made by the previous weak learners. Boosting assigns higher weights to the instances that were misclassified and lower weights to the correctly classified

instances. In each iteration, a new weak learner is trained to correct the mistakes of the previous ensemble.

Here, the xgboost python package [10] in combination with the scikit-learn python package [11] is used to implement a gradient boosting algorithm to create a BDT model. To preprocess the data, scikit-learn's StandardScaler was used to rescale the training and test data to improve the model's performance. To tune the model's hyperparameters with the purpose of conserving computation power, scikit-learn's RandomizedSearchCV was used with the result of 500 trees, a learning rate of 0.1, a max_depth of 5 nodes, a min_child_weight of 3, a min_split_loss set to 1, a subsample of 60%, and a colsample_by set to 0.7 were used to prevent overfitting. In addition, early stopping was implemented to further prevent overfitting. To check for overfitting, the consistency of the output responses of the BDT for the training and and testing set were observed, as seen in Figure 1.

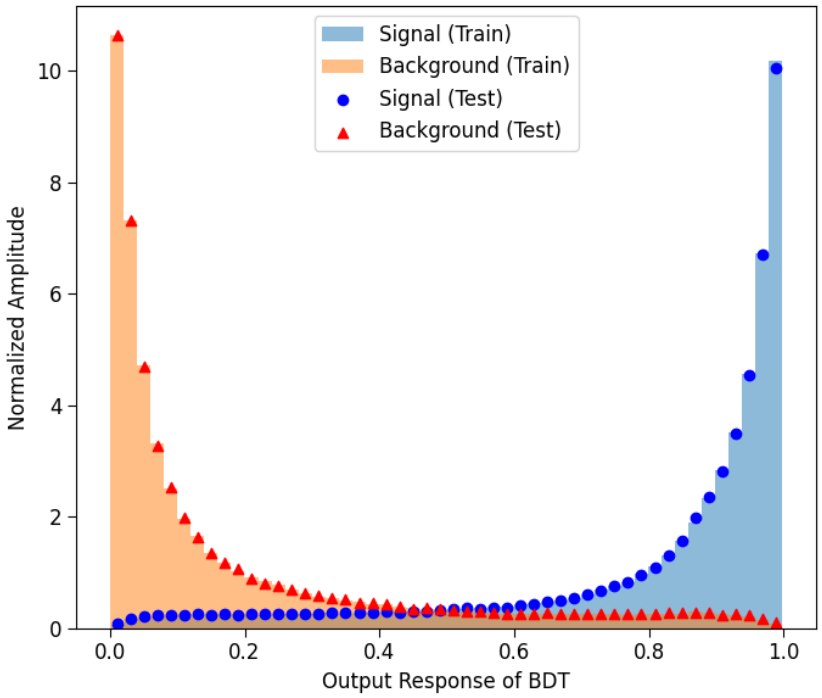

Figure 1: The probability distributions of the output response for the BDT with normalized amplitude on the MC signal and background data.

subsectionNeural Networks

Neural Networks (NNs) are a class of machine learning models inspired by the structure and functioning of the human brain. They excel at capturing complex relationships and

patterns in data. The structure of a neural network is composed of layers called "neurons", and starts with an input layer which receives the raw input data features. Then, hidden layers transform the input using learned weights and biases, which are initially assigned randomly. The final layer is the output layer which provides predictions or probabilities of the ouput. Each layer or "neuron" computes a weighted sum of its inputs, adds a bias term, and then passes the result through an activation function, much like a synapse.

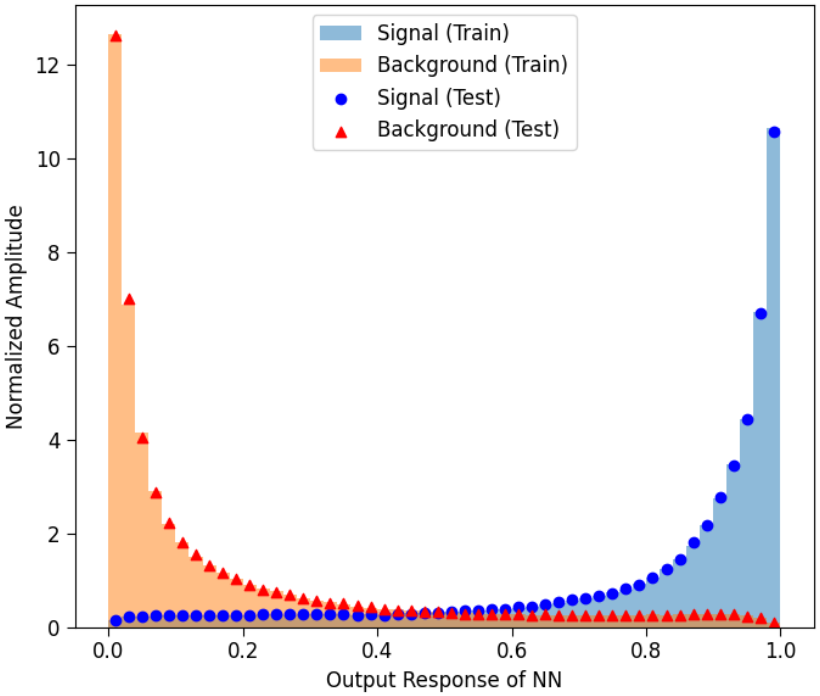

Figure 2: The probability distributions of the output response for the NN with normalized amplitude on the MC signal and background data.

The activation function introduces non-linearity which allows the network to catch complexities in the data. For the NN model, data was preprocessed the same way as for the BDT. For this work, Tensorflow's Keras [12] was used to create a simple NN with a total of four layers: an input layer with one neuron per input parameter, two hidden layers of 16 neurons each and a relu activation function, and an output layer with one neuron and a sigmoid activation function, giving the probability of that an event is a gamma-ray. Early stopping was also introduced to the NN to prevent overfitting. The output responses were once again observed but this time for the NN, as shown in Figure 2.

3.3 VERITAS Standard Cuts

By applying "box" cuts on individual parameters, VERITAS has been able to successfully separate the majority of gamma/hadron events [13]. Moderate cuts that are generally used for Crab-like sources are: 400 < SizeSecondMax, 6. < EmissionHeight < 1.e10, -1.2 < MSCW < 0.3, and -1.2 < MSCL < 0.5. Using these cuts, comparisons with the performances of the MLTs can be made with Standard Cuts (SC) that are similar to what VERITAS uses.

4 Results

4.1 MLT Performance on MC Data

The sample of events used from the simulated air showers consisted of a 1:1 ratio of signal to background. This sample was used to train and test both models. For training and testing purposes, 60% of the MC dataset was allocated to training, while the remaining 40% was reserved for testing. The ROC curves for the BDT and NN were obtained using sci-kit learn's roc_curve method, as seen in Figures 3 and 4.

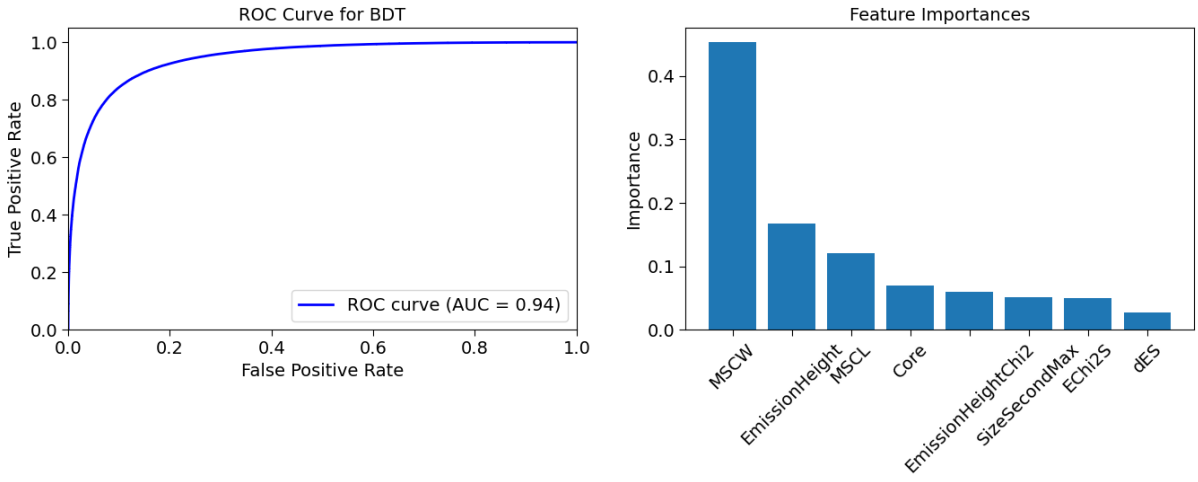

Figure 3: The ROC curve of the BDT (left) shows the discrimination power of the BDT with an AUC score of 0.94, while the feature importances (right) are displayed to analyze which parameters the BDT relies on to classify the air showers for the MC data.

The importances of the air shower parameters for the BDT in Figure 3 are shown where the MSCW emerges as the most discriminating feature. By studying the feature importances, we can tell if the model has potentially been overfit if it tends to latch on to a certain feature

that would not make sense for classifying gamma/hadron events. In Figure 4, the NN's validation loss on the simulated test data improves by approximately 2%. However, considering that further epochs might not yield significant improvements, early stopping was implemented, ending the training at 20 epochs.

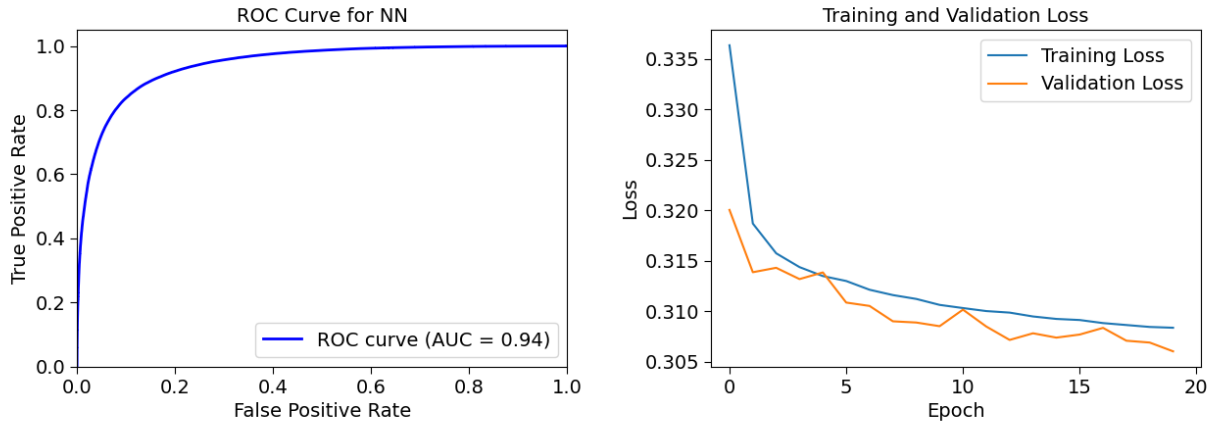

Figure 4: The ROC curve of the NN (left) shows the discrimination power of the NN with an AUC score of 0.94 with the evolution of the training vs. validation loss (right) over 20 epochs for the MC data.

4.2 Testing on VERITAS Crab Data

To assess the performance of each classifier, sky maps at the Crab region for each gamma/hadron separation technique were calculated with gammapy's WStatCountsStatistic [14, 15] with an exlusion region at the Crab. One thing to note, however, is that the inclusion of Wobble backgrounds during the map calculations introduced "holes" in the map, where significance calculations couldn't be conducted. Nevertheless, the BDT and NN models underwent a similar evaluation process: event probability thresholds were set to separate gamma-ray events from background noise while retaining as much signal as possible. This threshold was determined empirically by maximizing the summed significance of the Crab over the duration of 17 hours of runs. Figure 5 shows how the significance of the Crab is affected by the probability cut for the BDT model with a resulting sky map with 0.96 being the threshold for kept events. The same was done with the NN but with a slightly higher probability cut-off of 0.98, as displayed in Figure 6. Table 1 summarizes the results for the performances of the BDT, the NN, and SC on real VERITAS Crab data. The individual significance of each of

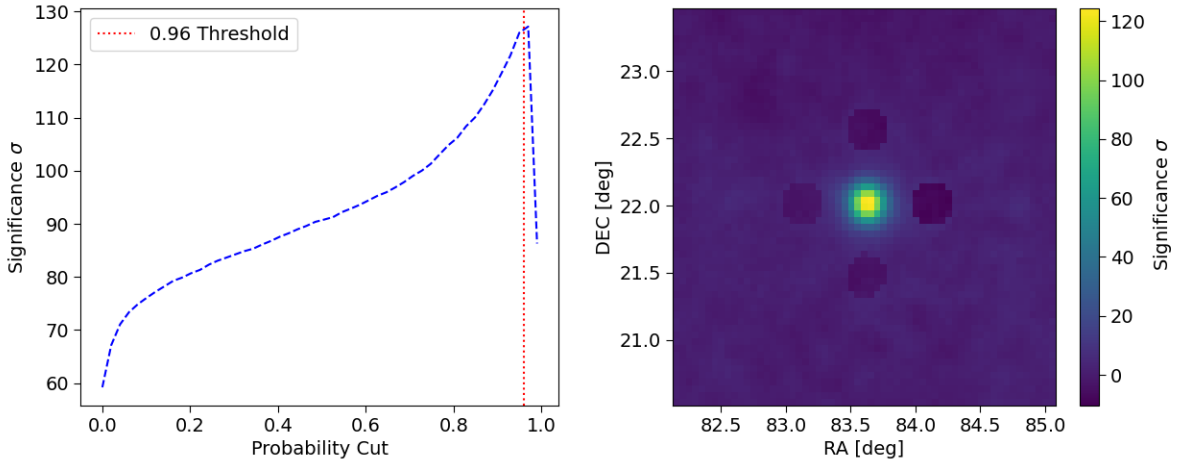

Figure 5: The effects of applying different thresholds on the Crab significance for the BDT model (left). A threshold of 0.96 for gamma/hadron separation was applied to maximize significance as shown in the sky map of the Crab Nebula (right)

the 35 runs are listed with the significance of the summed signal. In addition, the percent differences between the performance of the BDT and SC, and the NN and SC are listed with positives representing improvements to VERITAS's SC and negatives indicating a worse

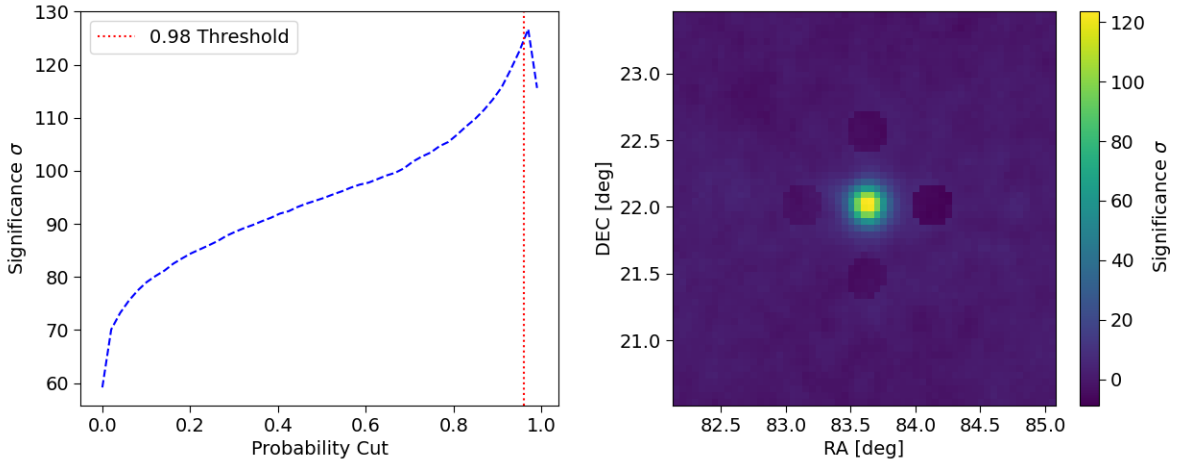

Figure 6: The effects of applying different thresholds on the Crab significance for the NN model (left). A threshold of 0.98 for gamma/hadron separation was applied to maximize significance as shown in the sky map of the Crab Nebula (right)

performance. Notably, the BDT model performs the best with the highest overall significance at 127.6σ, slightly outperforming the NN and VERITAS's moderate cuts, both achieving

127.0σ.

	Signifcance			% Difference	
Run	SC	BDT	NN	SC & BDT	SC & NN
1	25.8	26.0	25.9	+0.8	+0.4
2	23.8	22.7	22.5	-4.7	-5.6
3	23.2	24.8	25.1	+6.7	+7.9
4	22.5	21.2	21.1	-5.9	-6.4
5	25.0	25.8	26.6	+3.1	+6.2
6	23.5	23.3	22.3	-0.9	-5.2
7	20.7	19.3	20.5	-7.0	-1.0
8	23.2	24.9	26.3	+7.1	+12.5
9	21.8	23.3	23.6	+6.7	+7.9
10	21.0	20.3	21.0	-3.4	+0.0
11	22.6	23.7	22.9	+4.8	+1.3
12	21.1	21.4	21.0	+1.4	-0.5
13	21.5	23.0	23.3	+6.7	+8.0
14	21.7	22.0	21.0	+1.4	-3.3
15	21.6	21.6	22.0	+0.0	+1.8
16	22.6	26.0	24.0	+14.0	+6.0
17	21.1	22.0	23.3	+4.2	+9.9
18	22.2	22.3	21.9	+0.4	-1.4
19	21.2	20.9	22.1	-1.4	+4.2
20	14.6	13.7	14.1	-6.4	-3.5
21	20.8	22.0	21.9	+5.6	+5.2
22	22.6	20.4	21.0	-10.2	-7.3
23	21.9	22.0	20.9	+0.5	-4.7
24	21.7	19.1	18.2	-12.7	-17.5
25	18.9	16.6	18.4	-13.0	-2.7
26	20.6	20.0	18.6	-3.0	-10.2
27	20.0	21.0	20.7	+4.9	+3.4
28	18.2	19.5	19.8	+6.9	+8.4
29	21.8	22.0	21.2	+0.9	-2.8
30	18.5	19.3	18.5	+4.2	+0.0
31	19.0	19.7	19.3	+3.6	+1.6
32	18.7	19.5	17.9	+4.2	-4.4
33	20.3	21.0	20.6	+3.4	+1.5
34	20.8	19.8	19.0	-4.9	-9.0
35	22.4	21.5	21.3	-4.1	-5.03
Sum	127.0	127.6	127.0	+0.5	+0.0

Table 1: The Crab significance obtained from sky maps of each individual 30 minute run for each gamma/hadron separation method, including the percent difference (positive being an improvement). The last row is the summed signal of all runs ($\sim$17 hours).

5 Discussion

Both the BDT and NN performed well on the MC data as observed in Figure 3 and 4 by the ROC curves. These curves highlight that both models exhibit a high true positive rate while effectively maintaining a low false positive rate. This signifies the ability of the models to accurately distinguish between gamma and hadron events. Their performance is further supported by the Area Under the Curve (AUC) score of 0.94, closely approaching the optimal value of 1 that signifies excellence for both the BDT and NN models. However, solely concentrating on these scores might overlook nuances when comparing to the models' performance on real Crab data. Although it can be seen that the BDT leads in terms of overall significance in Table 1, the differences between the BDT and NN, as well as the SC, are relatively marginal. Additionally, we notice that the spread of the significance for both the BDT and NN is much larger than the SC. This a hint that while the machine learning models provide the advantage of potentially better discrimination, they can also show more variability in their performance across different runs. For instance, the NN's performance varies by 17% below SC in run 24 but improves by 12.5% in run 8. Similarly, the BDT model shows a performance decrease of 13.0% in run 25 and an increase of 14.0% above SC in run 16. Considering the percentage differences between the machine learning models and VERITAS's SC, the BDT performs better than SC in about 60% of the runs, while the NN surpasses SC for approximately 50% of the runs.

An air shower parameter that is traditionally a highly important distinguishing factor is SizeSecondMax, as found in other gamma/hadron separation machine learning experiments done at VERITAS [13]. However, by the distributions in Appendix A and the feature importances in Figure 3 we notice that this parameter is not the largest discriminating factor. This may be due to the fact that only electron and proton showers were used to simulate gamma-ray and hadronic events, whereas including a simulation of, for example, helium showers as part of the hadronic events may demonstrate higher discrimination between all relevant parameters, including SizeSecondMax. Such inclusion could potentially enhance the BDT and NN models' performance on real Crab data.

Furthermore, given the extensive energy range covered in the datasets, partitioning it

into multiple energy bins could yield more information about the strengths and weaknesses of MLTs in different areas. Splitting the data into energy bins could reveal potential trends that might not be apparent when analyzing the entire dataset as a whole. If the MLTs exhibit variability in performance across different energy intervals, it could signal the need for different hyperparameter tuning or feature engineering tailored to those specific energy ranges. For example, some air shower features are better discriminators at different energies, which would be reflected in the feature importances plot for the BDT. If MLTs consistently surpass standard cuts in certain bins, it suggests that MLTs have certain advantages in those energy ranges which can be studied further.

6 Conclusions

The BDT and NN models exhibit strong discriminatory capabilities on simulated data, as evident by their ROC curves and AUC scores. They also perform well in evaluating real Crab data, with the BDT model generally outperforming the NN and VERITAS's moderate cuts in terms of overall significance. However, the models also exhibit more varied performance compared to VERITAS's moderate cuts, emphasizing the importance of considering different methods such as binning the data into energy bins or adding more data to obtain a comprehensive understanding of their capabilities.

Acknowledgements

The author would like to thank Ken Ragan, Ste O'Brien, Matt Lundy, and Samantha Wong for providing insight and guidance as well as VERITAS for providing the necessary data to conduct this project.

References

[1] S. De, W. Maitra, V. Rentala, and A. M. Thalapillil, "Deep learning techniques for imaging air cherenkov telescopes," Apr. 2023, [Online; accessed 19-August-2023].

[Online]. Available: https://doi.org/10.1103%2Fphysrevd.107.083026 1

[2] M. Krause, E. Pueschel, and G. Maier, "Improved γ /hadron separation for the detection of faint γ -ray sources using boosted decision trees," *Astroparticle Physics*, vol. 89, pp. 1–9, mar 2017, [Online; accessed 19-August-2023]. [Online]. Available: https://doi.org/10.1016%2Fj.astropartphys.2017.01.004 1

[3] S. Ohm, C. van Eldik, and K. Egberts, "γ/hadron separation in very-high-energy γ-ray astronomy using a multivariate analysis method," *Astroparticle Physics*, vol. 31, no. 5, pp. 383–391, jun 2009, [Online; accessed 19-August-2023]. [Online]. Available: https://doi.org/10.1016%2Fj.astropartphys.2009.04.001 1

[4] J. A. et. al, "Implementation of the random forest method for the imaging atmospheric cherenkov telescope MAGIC," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 588, no. 3, pp. 424–432, apr 2008, [Online; accessed 19-August-2023]. [Online]. Available: https://doi.org/10.1016%2Fj.nima.2007.11.068 1

[5] GammaWiki, "Main Page — GammaWiki," 2022, [Online; accessed 19-August-2023]. [Online]. Available: https://veritas.sao.arizona.edu/GammaWiki/index.php?title=Main_Page&oldid=160574 2

[6] A. M. Hillas, "Cerenkov Light Images of EAS Produced by Primary Gamma Rays and by Nuclei," p. 445, Aug. 1985, [Online; accessed 19-August-2023]. 2

[7] KIT, "CORSIKA (COsmic Ray SImulations for KAscade)," Apr. 2023, [Online; accessed 19-August-2023]. [Online]. Available: https://www.iap.kit.edu/corsika/index.php 3

[8] "GrOptics," May 2019, [Online; accessed 19-August-2023]. [Online]. Available: https://github.com/groptics/GrOptics 3

[9] R. Garner, "Messier 1 (The Crab Nebula)," 2017, [Online; accessed 19-August-2023]. [Online]. Available: https://www.nasa.gov/feature/goddard/2017/messier-1-the-crab-nebula 3

[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794, [Online; accessed 19-August-2023]. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939785 5

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011, [Online; accessed 19-August-2023]. 5

[12] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, and J. Dean, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, [Online; accessed 19-August-2023]. [Online]. Available: https://www.tensorflow.org/ 6

[13] GammaWiki, "BDT Gamma Hadron Separation," 2021, [Online; accessed 19-August-2023]. [Online]. Available: https://veritas.sao.arizona.edu/GammaWiki/index.php?title=BDT_Gamma_Hadron_Separation&oldid=149929 7, 11

[14] M. de Naurois, "Very High Energy astronomy from H.E.S.S. to CTA. Opening of a new astronomical window on the non-thermal Universe," Ph.D. dissertation, Ecole Polytechnique, 2012. 8

[15] Keith Arnaud, Craig Gordon, Ben Dorman Kristin Rutkowski, "Statistics in XSPEC," Jul. 2023, [Online; accessed 19-August-2023]. [Online]. Available: https://heasarc.gsfc.nasa.gov/xanadu/xspec/manual/XSappendixStatistics.html 8

A Visualizing Air Shower Parameters

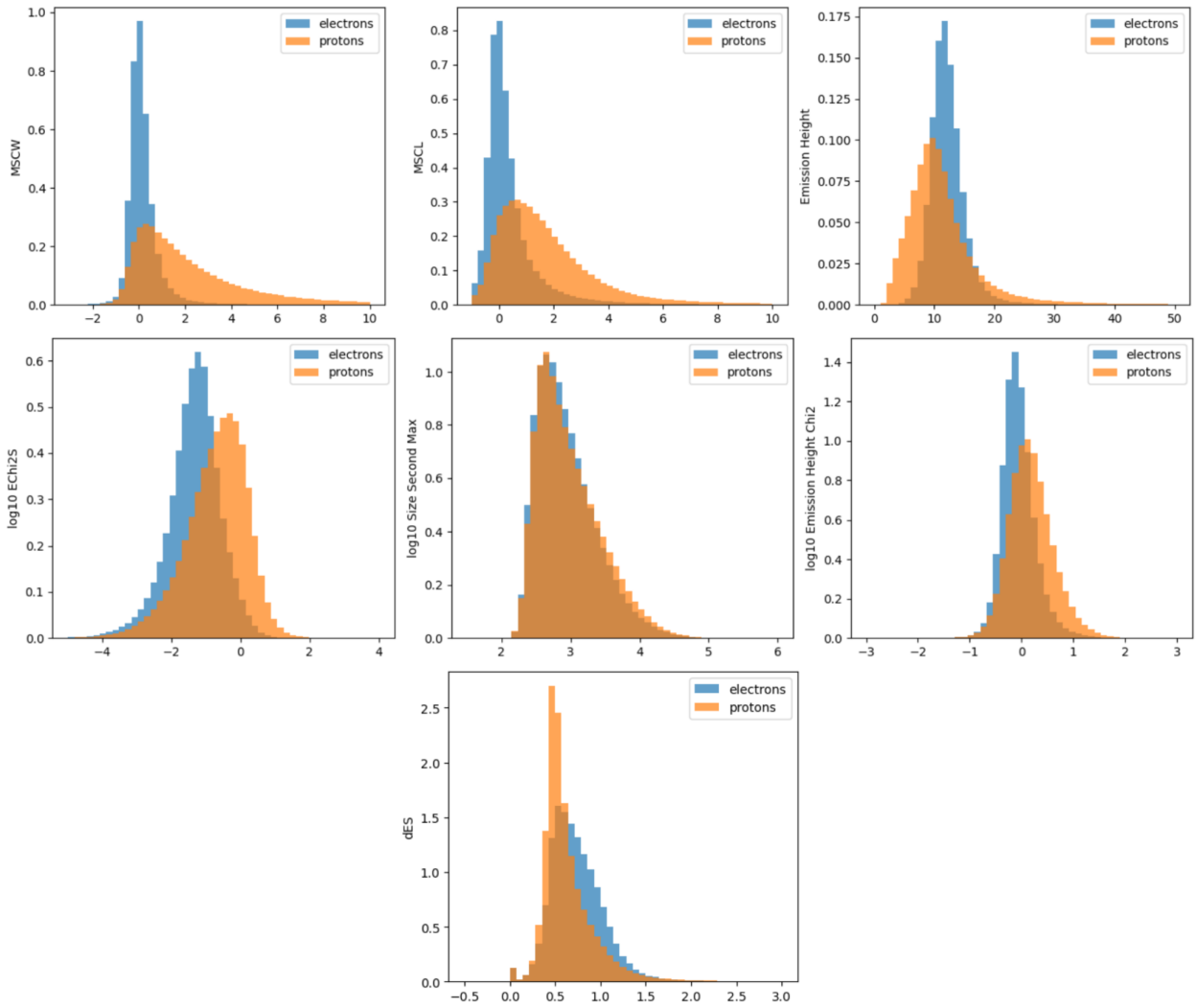

Figure 7: The normalized distributions of each air shower parameter: MSCW, MSCL, Emission Height, log10 ECHi2S, log10 SizeSecondMax, log10 Emission Height Chi2, and dES for electron and proton showers predicted to be relevant in the classification task of gamma/hadron events. SizeSecondMax does not seem to be as discriminating as the other six parameters however it was kept since it has previously been used as a highly important parameter in other literature.