Enhancing Lepton Identification in CLAS12 using Machine Learning Techniques

Mariana Tenorio-Pita

Abstract

In this study, we explore the application of machine learning techniques for lepton identification in CLAS12 physics experiments. Our specific goal is to minimize pion contamination, which is observed in both the experimental and simulated datasets. We develop, evaluate, and validate two machine learning models: Boosted Decision Trees and Multilayer Perceptron. Each model underwent training with two sets of variables, and rigorous validation was conducted using simulated and experimental data to ensure reliability. The results revealed the effectiveness of the applied models in mitigating background contamination. By retaining around 90% of leptons in the samples, the background can be reduced by a factor of 10. Notably, the Boosted Decision Tree exhibited superior performance in accomplishing this task.

I. MOTIVATION

In CLAS12, the Event Builder (EB) collects and organizes the responses from the different CLAS12 subsystems. In addition to outputting all the information necessary for physics analysis, it is also responsible for associating the responses to particles and executing particle identification (PID). The particle identification uses the Particle Data Group Montecarlo particle numbering scheme [1]. The EB assigns 11 and -11 for electrons and positrons, respectively. The conditions for assigning the PIDs to electrons and positrons are [2]:

- A minimum of 2.0 photoelectrons in the High Threshold Cherenkov Counter (HTCC) for particles with $P < 4.9$ GeV.
- A minimum deposition energy of 60 MeV in the pre-shower calorimeter (PCAL).
- A total measured sampling fraction (SF) within 5σ of the parameterized momentum-dependent sampling fraction (SF_P). The sampling fraction is defined as:

$$SF = \frac{E_{ECAL}}{P} \tag{1}$$

  E_{ECAL} is the total energy deposited in the Electromagnetic Calorimeter (ECAL), and P is the momentum measured in DC. The parameterization of the mean and sigma of the SF is given by [3]:

$$SF_P = p_1 * \left(p_2 + \frac{p_3}{E_{ECAL}} + \frac{p_4}{E_{ECAL}^2} \right) \tag{2}$$

  where p_i are run-dependent parameters calibrated for each sector.

At $P > 4.9$GeV, both leptons and charged pions produce a signal in HTCC, thus, in this region, the PID assigned by the EB is only based on the calorimeter responses.

A. Evidence of pion contamination

The evidence of π^+ contamination in the positron sample can be seen when selecting events with one electron, one positron, and one proton. In Figure 1, where positron polar angle vs. momentum is shown, we can observe a cluster of events above 4.9 GeV at forward angles ($\theta < 10^o$), indicating pion contamination.

Another piece of evidence can be observed in the exclusive reaction $ep \to e'\pi^+(n)$ when a low energy electron and a particle with $PID = -11$ with $P > 4.5$GeV are selected in the forward

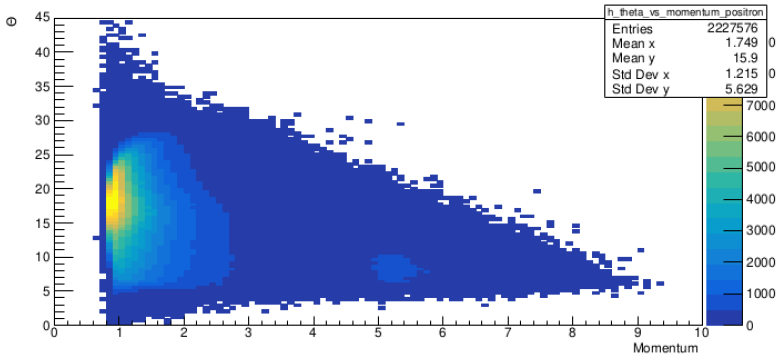


Figure 1: θ vs P distribution for positrons. The cluster of events above 4.9 GeV at forward angles indicates contamination from π^+s identified as e^+.

detector (FD). The particles with $PID = -11$ are assigned a mass of the pion, 0.1395 GeV, and the missing mass of the $(e'e^+(m_\pi))$ is calculated. Figure 2 shows the missing mass distribution where a clear peak at the neutron mass can be observed. This proves that the particle sample with ID$= -11$ contains π^+s.

Finally, pion contamination is also present in simulations. When we generate a sample of π^+, when looking at the difference between the generated vs the reconstructed sample, about 10% of π^+ with $P > 4.7$ GeV were identified as positrons.

II. TMVA ROOT FRAMEWORK

The Toolkit for Multivariate Analysis (TMVA) is an open-source ROOT-integrated environment for processing, evaluating, and applying multivariate classification and regression techniques [4]. All multivariate techniques in TMVA belong to the family of supervised machine learning algorithms.

With TMVA packages, we can use several variables, such as the sampling fractions, to distinguish between leptons (signal) and pions (background). For TMVA, the analysis consists of two phases: the training and the application. The multivariate method is trained, tested, and evaluated in the training phase using user-provided data with a clear separation of signal and background. The success of the model is dependent on the quality of the training data. A weight file containing the results is created at this stage. In the application phase, the weight file is used to evaluate the data set where the composition is unknown. For

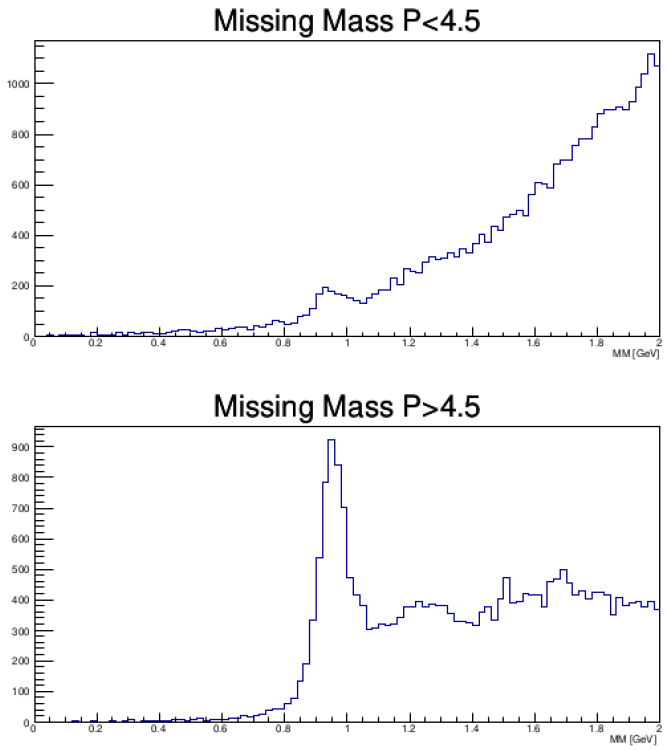


Figure 2: The missing mass distributions of reaction $ep \to e'e^+X$ when pion mass is used for the positron. The top plot is for positron momenta below the HTCC pion threshold, and the bottom plot is for the positrons with momenta greater than the pion threshold in HTCC.

each event, the variables used for the training are evaluated, and a single output is obtained as a score. From this output, we can apply a cut to maximize the background suppression while maintaining good signal efficiency.

Two multivariate methods were tested for this work: Boosted Decision Trees (BDT) and Multilayer Perceptron (MLP), which is a feedforward artificial neural network. A total of 6 classifiers were trained: two for each RGA experimental configuration (Fall 18 Inbending, Fall 18 Outbending, Spring 19 Inbending), where one was to identify electrons while the other was to identify positrons. The classifiers were trained using simulation samples, with the signal sample consisting of $e^-(e^+)$ while the background consisted of $\pi^-(\pi^+)$ that had been assigned a PID= $11(-11)$ by the EB.

A. 6 and 9-Variable models

In the first approximation, a total of 9 variables were considered: Momentum (P), the polar (θ) and azimuthal (ϕ) angles, the PCAL, ECIN, and ECOUT sampling fractions (SF), and the sum of the 2^{nd}-moments (m2) for PCAL, ECIN, and ECOUT, defined as:

$$m2 = \frac{m2u + m2v + m2w}{3} \tag{3}$$

where $m2u$, $m2v$, and $m2w$ are the 2^{nd} moments of the shower on each readout side of the calorimeter (u, v, and w). The distributions for these variables can be observed in Figure 3, where the blue distributions correspond to the signal (lepton) and the red to the background (pion). To eliminate any possible bias from using P, θ, and ϕ, a second model was studied where we kept only the sampling fractions and the second moments.

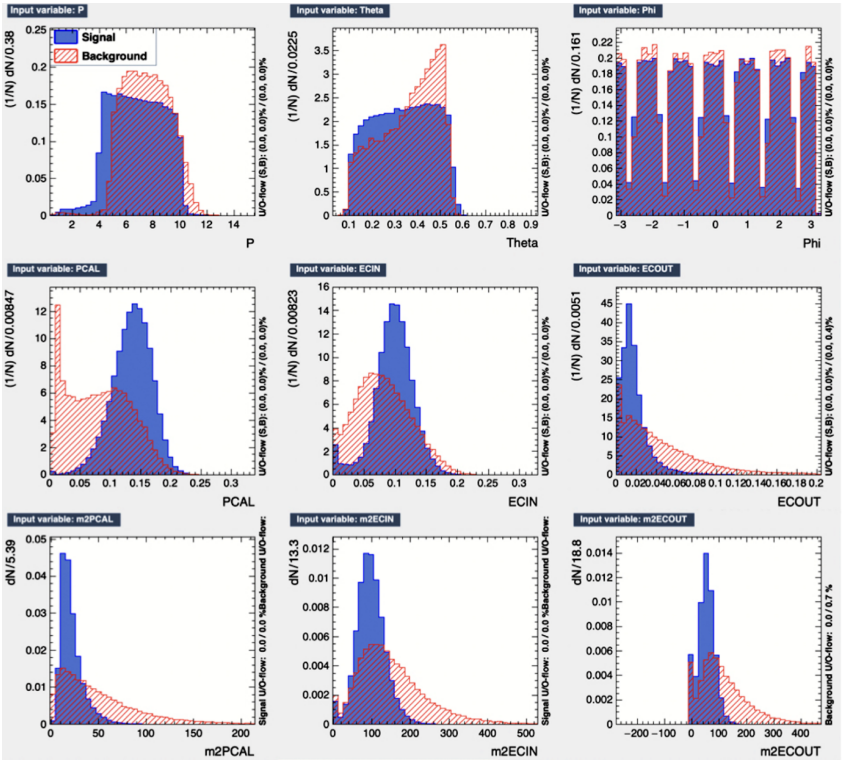


Figure 3: Distributions of signal (blue) and background (red) events over the nine variables used to train the algorithm.

III. TRAINING AND TESTING OF ALGORITHMS

The training was done on BDT and MLP using the previously mentioned variables. The input data sets are divided into two sub-sets. One dedicated to training and the other to testing.

This ensures that the evaluation of the algorithms is statistically independent from the training sample. Figure 4 shows the results produced from this testing. The MVA responses from the training and test samples are superimposed. As we can see, both distributions match, meaning the methods perform well.

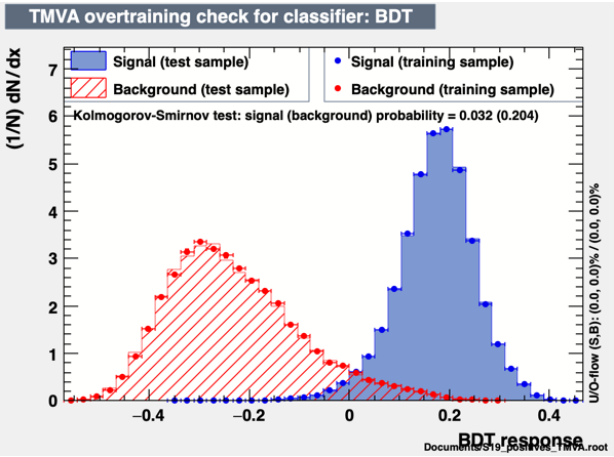


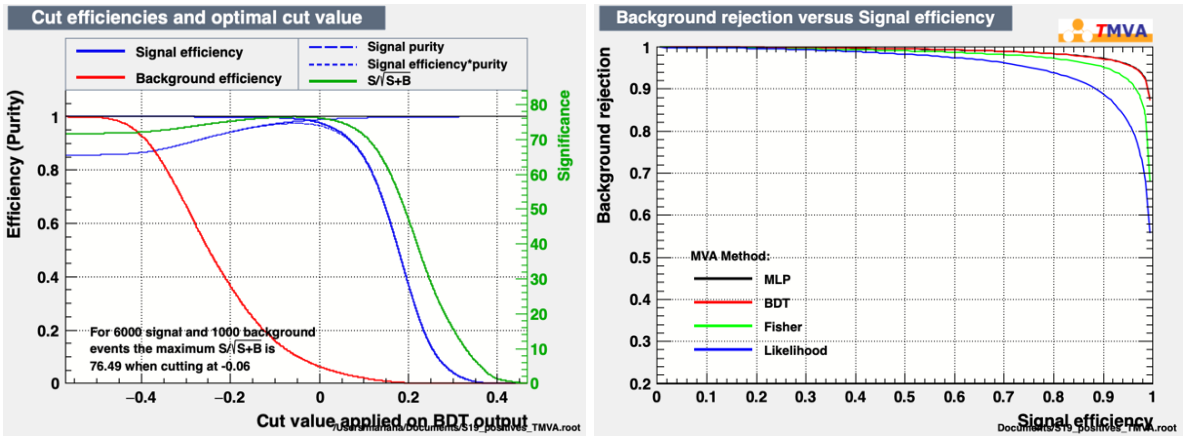
Figure 4: Distributions of signal and background events as a function of the classifier value.

The cut efficiencies and the optimal cut value are obtained for each evaluated method. In Figure 5(a), the solid blue line represents the signal efficiency, the solid red line represents the background efficiency, and the green solid line is the significance, $s = S/\sqrt{S+B}$. The cut that will give us the maximum significance value is the optimal cut, which has the most real leptons while reducing most of the pions. This was done for every MVA (MLP and BDT) and model (6 and 9 variable models) for each RGA configuration.

Finally, we can compare the efficiency of each method by comparing the Receiver Operating Characteristic (ROC) curves. The ROC curve represents the signal efficiency versus the background rejection, that is, the fraction of the signal kept vs. the fraction of the background removed. The more robust method tends to be in the upper right corner. In Figure 5(b), we can observe the resulting ROC curves of four different MVA, where BDT and MLP are the most efficient.

IV. VALIDATION

Up to this point, we have trained all the models and obtained the optimal cut efficiency values for each configuration provided by TMVA. The next step is to validate these models in simulations and data, select the most efficient model, and establish the optimal cut value.



(a) (b)

Figure 5: The results of the TMVA analysis. (a) efficiencies of the signal identification (blue) and the background rejection (red) as a function of classifier value. The green line is the significance of the signal over the whole sample. (b) ROC curves for various MVA methods. The best-performing methods are MLP (black) and BDT (red).

A. Validation of models on Simulations

The model was validated in simulations using an independent Montecarlo data set. The response for each event was obtained and stored. Based on this, we found that we have four possible outcomes:

- True Positive: The particle is identified as a lepton, and it was a lepton.
- False Positive: The particle is identified as a lepton, but it was a pion.
- True Negative: The particle is identified as a pion, and it was a pion.
- False Negative: The particle is identified as a pion, but it was a lepton.

We can calculate the signal efficiency and background rejection rate using this information. Figure 6 shows the distribution of signal (top) and background (bottom) events as a function of P, θ, and ϕ for initial samples, solid-line histograms, and after the cut on the classifier of the 9-variable model, 0.0, the blue and orange points. The ratio of the number of events after the cut over the number of events in the initial sample is shown with black points under each distribution. As can be seen, the true positive rate (TPR) for this case is $\sim 97.8\%$, while the false positive rate (FPR) is $\sim 6\%$.

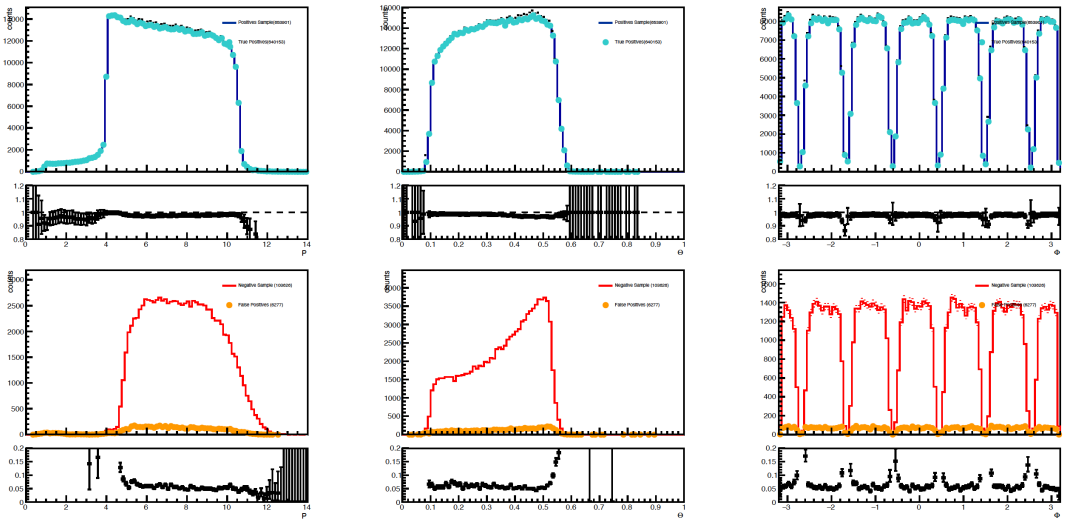


Figure 6: Distribution of signal (top) and background (bottom) events as a function of P, θ and ϕ for initial samples, solid-line histograms, and after the cut on the classifier of the 9-variable model, the blue and orange points. The ratio of the number of events after the cut over the number of events in the initial sample is shown with black points under each distribution.

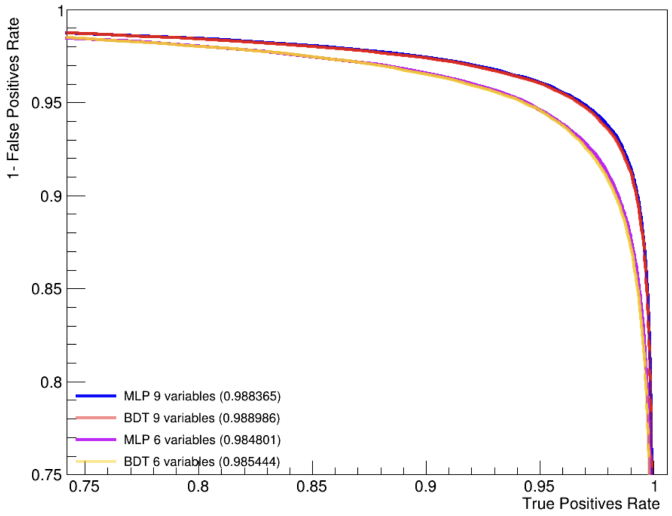


Figure 7: ROC curve obtained from simulations for different models. The 9-variable models (blue and red) perform better than the 6-variable models (yellow and purple).

We can reconstruct the ROC curve for the simulations by applying a series of cuts in the $[-0.6, 0.6]$ range. The objective is to a) Validate the results obtained from the TMVA analysis, b) Compare the performance of the 9 and 6 variables models, and c) look for the optimal cut to apply. Figure 7 shows the ROC curve obtained from this analysis. In this figure, MLP and

BDT 9-variable models perform better than the 6-variable model versions.

B. Validation of models on Data

Clean signal and background samples are necessary to test the model's performance with data. We use the radiation of photons from leptons to obtain the signal sample. It has been established that as electrons and positrons propagate from the target, they lose energy by radiating photons that are then detected in the ECAL. Therefore, we can obtain the lepton sample by looking at the reaction $e^- p \rightarrow e^{-(+)} \gamma X$. The photons emitted by the outgoing electron can be identified by looking at the $\Delta\theta = \theta_\gamma - \theta_l$ distribution, where θ_γ is the photon polar angle, and θ_l is the lepton polar angle. In Figure 8, $\Delta\phi$ vs $\Delta\theta$ is plotted for electrons (left) and positrons (right) from the Spring 2019 pass2 dataset. A sharp peak at $\Delta\theta \approx 0$ corresponds to leptons and radiated photon pairs.

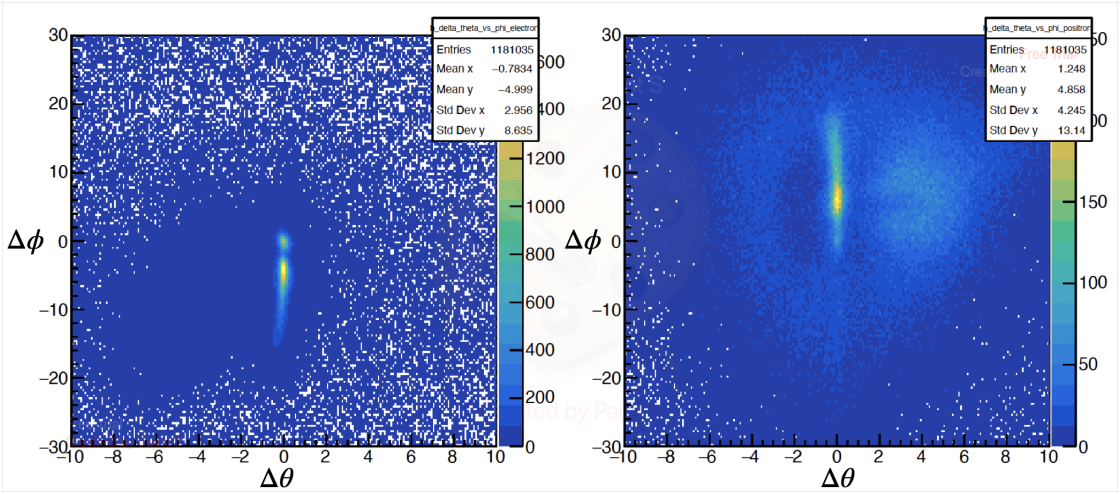


Figure 8: $\Delta\phi$ vs $\Delta\theta$ for electrons (left) and positrons (right) for Spring 2019, pass2 dataset. The concentration of events at $\Delta\theta \sim 0$ are the events with the true leptons.

The signal sample was obtained using RGA *InclusivePositron* and *resIncl* skims. We select events that contain one $e^+(e^-)$ with $P > 4.5$ GeV and one photon. When looking at the $\Delta\theta$ distribution, Figure 9, the events at $\Delta\theta \approx 0$ correspond to true lepton events. One thing to notice is that the signal is cleaner in the case of electrons than in the positron case. We can obtain the number of true electrons in the sample by fitting the distribution with a Gaussian plus a polynomial function. Doing so with $\Delta\theta$ distributions for various classifier cut values, we can obtain signal efficiency as a function of the cut. The procedure is the same for both electrons and positrons.

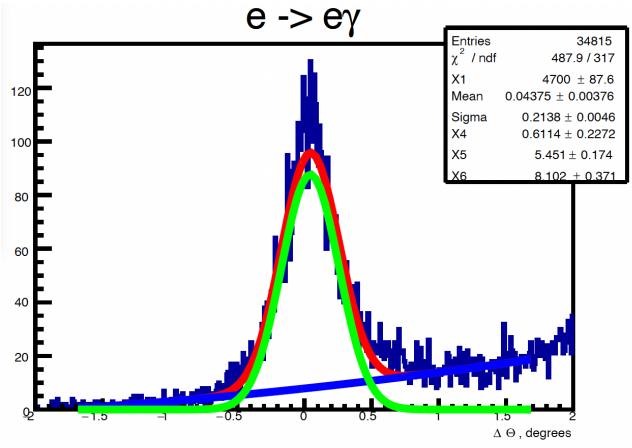


Figure 9: $\Delta\theta$ distribution for the positrons case. The peak represents the events with the true leptons. The red line is the total fit of the signal (green) plus the background (blue).

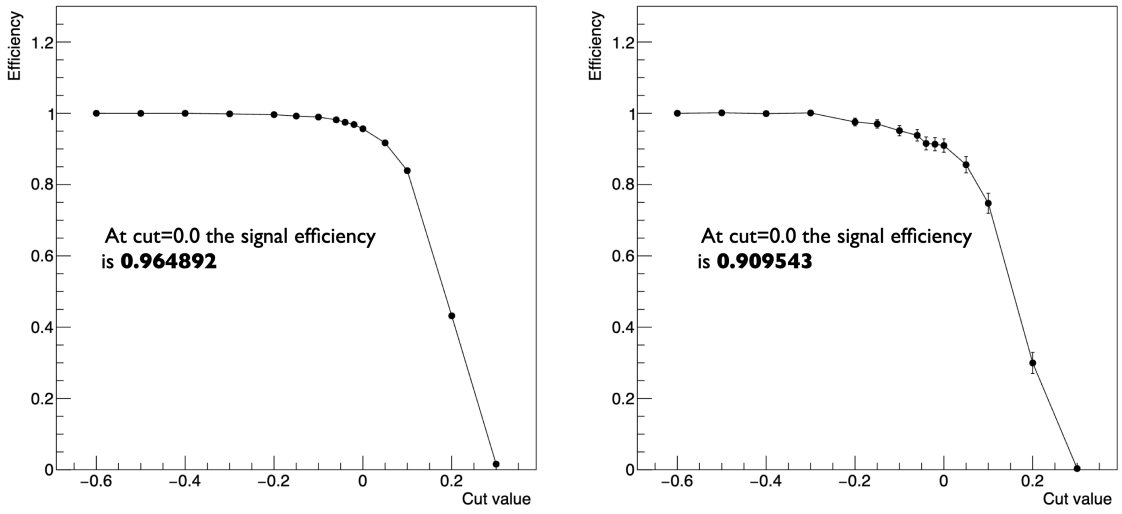
At the cut on the classifier, 0.0, we can observe that the TPR is around 95% for the inbending data sets with the BDT-6 models for both the positron and electron classifiers. On the other hand, the percentage is around 89% for the outbending configuration in all models. The results of all the configurations are displayed in Table I and Table II. From these results, we can conclude that the 6-variable model performs better in the data sets. These results are compatible with the results obtained from the validation with simulations described before. The dependence of TPR on the classifier cut in the region from -0.6 to 0.4 is shown in Figure 10. This dependence and the Background suppression curve will allow us to construct the ROC curve and find the optimal cut for each configuration and model.

Model	Configuration	TPR
BDT-6	Spring 2019	96.47% ±0.27%
	Fall 2018 Inbending	95.65% ±0.18%
	Fall 2018 Outbending	90.95% ±1.88%
BDT-9	Spring 2019	92.50% ±0.37%
	Fall 2018 Inbending	94.60% ±0.20%
	Fall 2018 Outbending	89.45% ±2.01%

Table I: TPR of positron identification at 0.0 cut.

Model	Configuration	TPR
BDT-6	Spring 2019	94.30% ±0.02%
	Fall 2018 Inbending	94.94% ±0.02%
	Fall 2018 Outbending	89.86% ±0.02%
BDT-9	Spring 2019	91.65% ±0.02%
	Fall 2018 Inbending	92.33% ±0.03%
	Fall 2018 Outbending	88.17% ±0.03%

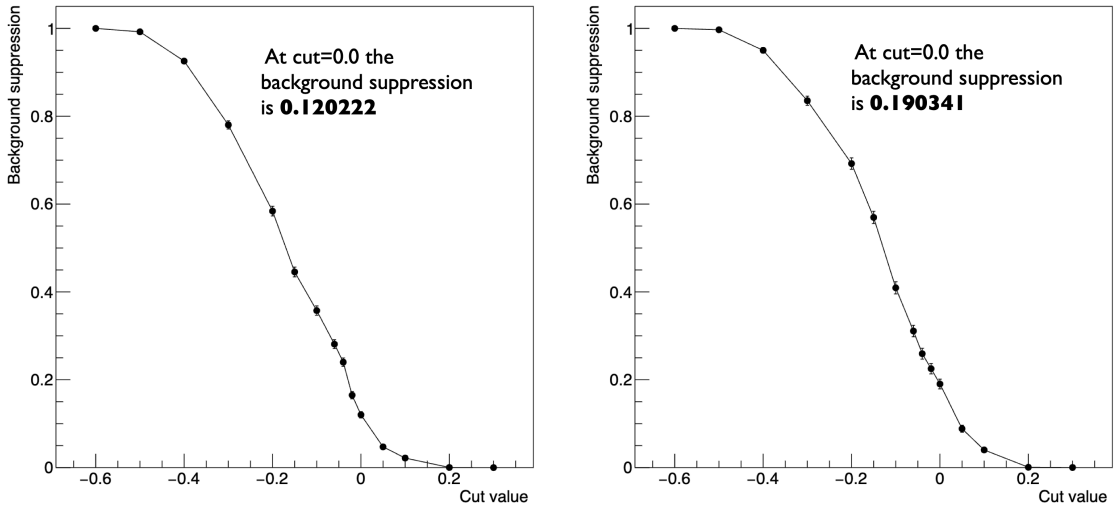
Table II: TPR of electron identification at 0.0 cut.



(a) (b)

Figure 10: Signal efficiency curves for the positron BDT 6 model for (a) inbending and (b) outbending datasets. We can see that around the cuts at 0, the efficiency is well over 90%.

Regarding background suppression, the sample acquisition presents a slightly more intricate process. For positrons, the reaction described in Section I A, $ep \rightarrow e'\pi^+_{PID=-11}X$, is employed, where X would correspond to the missing neutron and $\pi^+_{PID=-11}$ refers to a particle that has been identified as a positron, $PID = -11$, but we used the mass of a pion in the kinematic calculation. The background sample for positrons was obtained using the RGA *jpsitcs* skim. As a first step, we select events with one electron and one particle with PID$= -11$, both in the forward detector, and assign the mass of the pion, $m = m_{\pi^+}$, to the identified positron. Then, we keep only the events where the particle identified by the EB as a positron has $P > 4.5GeV$. Similarly, as with the signal, we fit the peak at the neutron mass to extract the number of misidentified pions in the sample. At the cut value of the classifier 0.0, the FPR is 4% for BDT 9 and 10% for the 6 variable model for the inbending data sets. We obtain the background suppression curve by applying the series of cuts on the classifier, Figure 11. These numbers are consistent with the results obtained from the validation in simulations that had an FPR of $\sim 6\%$. As for the background sample for electrons, we use the outbending positrons to validate the background suppression rate in the inbending data set and vice versa.



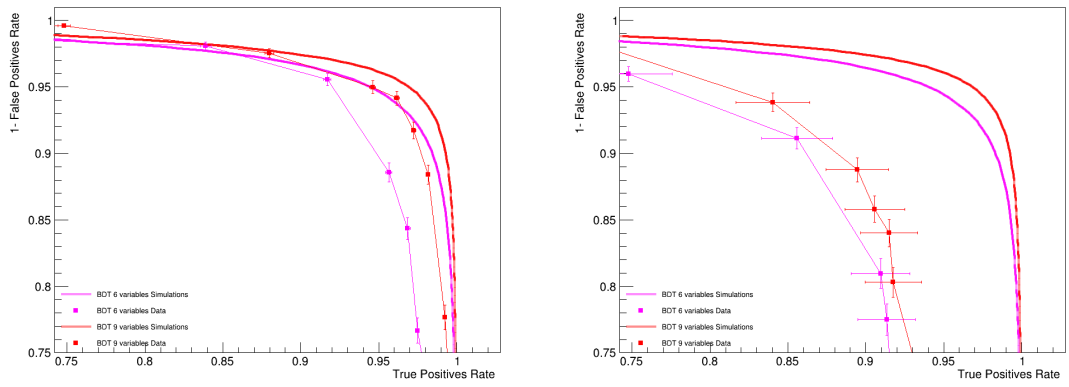
(a) (b)

Figure 11: Background suppression curve for the positron BDT 9 model for (a) inbending and (b) outbending datasets. This curve indicates how much of the background is left after each cut on the classifier. We can observe that at the cuts around 0, the background left in the sample is less than 20% of the original amount.

V. COMPARISON BETWEEN DATA AND SIMULATIONS

By combining the information obtained from signal efficiency and the background suppression, we can reconstruct the ROC curve for data and compare it with Figure 7 of simulations. The comparison is shown in Figure 12. As can be seen, the results obtained from the validation in data are consistent with those from the validation using simulations, the general efficiency of the BDT 9 variable model is larger than the BDT 6 variable model. But, overall, the efficiency of the cuts in the data is lower than what is obtained in simulations.

In addition, we are interested in obtaining the ratio of efficiency in simulations and data as a function of the cut in the response. As it can be seen from Figure 13, the efficiency ratio, i.e., the discrepancy between MC and data around the classifier cut at zero, is less than 25%. Interestingly enough, while the efficiency performance of the BDT 9 variable model is better than that of others, the discrepancy between MC and data is much smaller for the 6 variable BDT model.



(a) (b)

Figure 12: ROC curve comparing the results of the validation on Data (square points) and simulation (solid line) for BDT 6 and BDT 9 variable models for (a) inbending and (b) outbending datasets. In both cases, the results from the data are lower than those in MC, but the trend is consistent with simulation results.

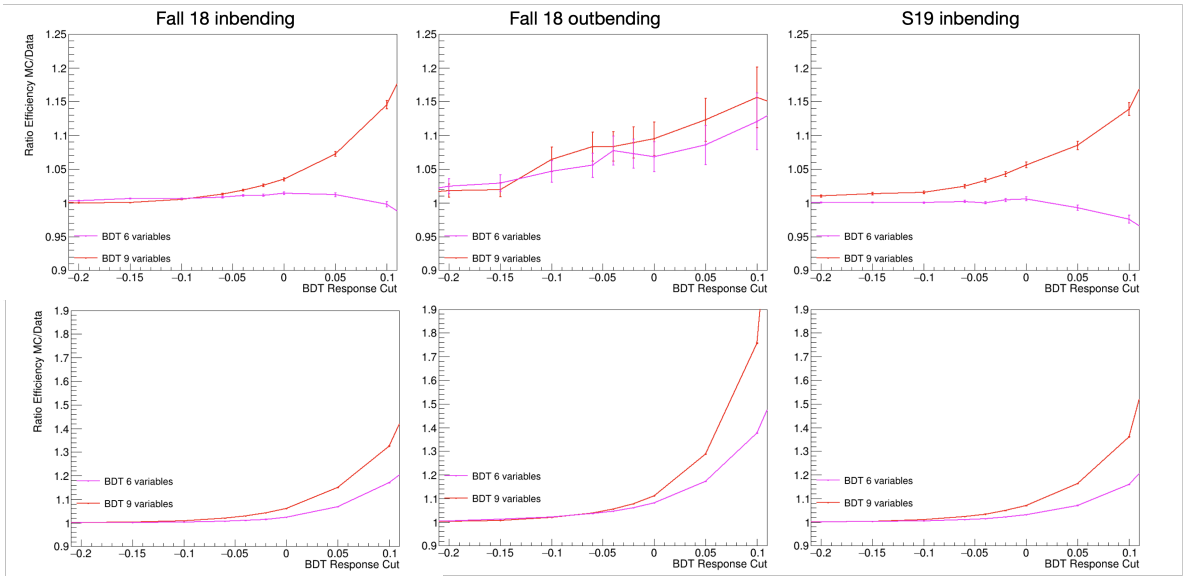


Figure 13: Ratios of efficiencies from MC to data for positron (upper) and positron (lower) identification classifiers for each data set as a function of the cut in the response.

VI. CONCLUSION

The machine learning methods can applied to discriminate between the leptons and the pions to reduce pion contamination in the electron and positron samples both in data and MC. In this study, we developed, rigorously evaluated, and validated two machine learning

models: Boosted Decision Trees and Multilayer Perceptron. Each model was meticulously trained using two distinct sets of variables. To ensure the robustness and reliability of our models, extensive validation procedures were conducted using both simulations and experimental data. The results from the simulation validation highlight the models' overall efficiency, where the 9-variables model is more efficient. On the other hand, the validations on data, where the results obtained are compatible with the results from simulations, showed the model's capabilities to maintain a good percentage of the signal while removing most of the background. Furthermore, one interesting thing to notice is that even though the 9-variable model seems to have better overall efficiency, the efficiency ratio between data and Montecarlo showed a better performance with the 6-variable model.

Overall, the results revealed the effectiveness of the applied models in mitigating background contamination. For both the 6- and 9-variable models, the results show that in each data set, by retaining around 90% of leptons in the samples, the background can be reduced by a factor of 10. The next step would be to implement this work on Iguana to make this project a tool available for collaboration.

Acknowledgements

I extend my sincere gratitude to Pierre Chatagnon for their invaluable guidance and support throughout this analysis. Special thanks to Stepan Stepanyan and the di-lepton analysis group for their valuable feedback.

[1] Particle Data Group. Review of Particle Physics. *Progress of Theoretical and Experimental Physics*, 2020:083C01, 2020.

[2] V. Ziegler, N. Baltzell, F. Bossù, D. Carman, P. Chatagnon, M. Contalbrigo, R. De Vita, M. Defurne, G. Gavalian, G. Gilfoyle, D. Glazier, Y. Gotra, V. Gyurjyan, N. Harrison, D. Heddle, A. Hobart, S. Joosten, A. Kim, N. Markov, and M. Ungaro. The CLAS12 software framework and event reconstruction. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 959:163472, 2020.

[3] G. Asryan, S. Chandavar, T. Chetry, N. Compton, A. Daniel, N. Dashyan, N. Gevorgyan, Y. Ghandilyan, K. Giovanetti, K. Griffioen, K. Hicks, D. Kashy, G. Khachatryan, M. Khandaker, W. Phelps, J. Riso, A. Simonyan, C. Salgado, C. Smith, and M. Yurov. The CLAS12 forward electromagnetic calorimeter. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 959:163425, 2020.

[4] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. Von Toerne, H. Voss, M. Backes, T. Carli, O. Cohen, A. Christov, D. Dannheim, K. Danielowski, M. Jachowski, K. Kraszewski, M. Kruk, Y. Mahalalel, R. Ospanov, X. Prudent, A. Robert, and A. Zemla. TMVA - toolkit for multivariate data analysis. 2009.