PROCEEDINGS OF SCIENCE

Optimization of the proton background rejection in the measurement of the electron flux at high energies with CALET on the International Space Station

Sandro Gonzi,[a,b,c,*] **Eugenio Berti**[b,c] **and Lorenzo Pacini**[b,c] **for the CALET collaboration**

[a] *University of Florence, Department of Physics and Astronomy,*
Via Giovanni Sansone 1, 50019 Sesto Fiorentino, Italy

[b] *National Institute for Nuclear Physics INFN, Division of Florence,*
Via Bruno Rossi 1, 50019 Sesto Fiorentino, Italy

[c] *National Research Council CNR, Institute of Applied Physics IFAC,*
Via Madonna del Piano 10, 50019 Sesto Fiorentino, Italy

E-mail: sandro.gonzi@unifi.it

The Calorimetric Electron Telescope (CALET) is a cosmic-ray observatory operating since October 2015 onboard the International Space Station (ISS). The data processed since the beginning of the mission has made it possible to measure with high precision the inclusive flux of cosmic electrons and positrons (all-electron) in the multi-TeV region. The appearance of any structures in this energy region can potentially be connected to the presence of nearby astrophysical sources or dark matter. The CALET detector, consisting of a charge detector, an imaging calorimeter and a total absorption calorimeter has a total vertical thickness of about 30 radiation lengths. The construction characteristics of the instrument allow to obtain an energy resolution better than 2% for electrons and a proton rejection power of about 10^5. However, the exploration of the multi-TeV region involves dealing with a limited statistical sample and a large proton background. As a consequence, a complex multivariate analysis based on variables connected to the shower development has been adopted. In this contribution, we summarize the results of a study conducted on different multivariate analysis techniques in order to optimize the proton rejection at high energies in the all-electron flux measurement. In particular, we discuss the features of the different methods, the tuning of their parameters and the overall strategy to increase the separation between electrons and protons, avoiding the phenomenon of overfitting.

38th International Cosmic Ray Conference (ICRC2023)
26 July - 3 August, 2023
Nagoya, Japan

ICRC2023
38th International Cosmic Ray Conference
The Astroparticle Physics Conference

[*] Speaker

https://pos.sissa.it/

1. Introduction

The study of cosmic-ray electrons and positrons in the high-energy range provides a unique probe of nearby cosmic accelerators: due to the intense energy loss during diffusion, the observed electrons above 1 TeV can only be produced by sources within 1 kpc and therefore only few supernova remnants or pulsars located in the proximity of the Solar System can be deemed as their astrophysical sources. In addition, the apparent increase of the positron fraction over 10 GeV established by the Payload for Antimatter Matter Exploration and Light nuclei Astrophysics (PAMELA) [1] and the Alpha Magnetic Spectrometer (AMS-02) [2] experiments could involve the existence of some supposed positron sources with astrophysical or exotic origin as, respectively, nearby pulsars or dark matter. A precise measurement of the inclusive spectrum of cosmic electrons and positrons (all-electron) in the TeV region might thus reveal some peculiar spectral features which, compared to the ones expected by the numerous theoretical models available, can lead to a better understanding of the neighboring region of the Galaxy or to an improvement of existing cosmological models.

The CALorimetric Electron Telescope (CALET) [3] is a space experiment operating onboard the International Space Station (ISS) since October 2015 for long term observations of cosmic rays. Over the past few years the CALET Collaboration had performed a direct measurement of several cosmic-ray spectra, specifically of electron and positron (up to 4.8 TeV) [4, 5], proton (up to 60 TeV) [6, 7], Helium (up to 250 TeV) [8], Boron (up to 3.8 TeV/n) [9], Carbon and Oxygen (up to 2.2 TeV/n) [10] and Iron (up to 2.0 TeV/n) [11].

This paper describes the results of a study, concerning the all-electron spectrum measurement in the TeV region that has been conducted on different multivariate analysis techniques based on variables connected to the development of the shower in the detector. This approach allows optimizing the results when the analysis is dealing with a limited statistical sample of the electron signal and a large proton background. In detail: in section 2 the CALET detector is described, in section 3 the features of the different multivariate analysis selected methods are introduced and the procedure is explained, in section 4 the results of the parameters optimization of the selected methods are presented and discussed.

2. CALET detector

The CALET detector is an all-calorimetric instrument with a total vertical thickness equivalent to 30 radiation lengths (X_0) and 1.3 proton interaction lengths (λ_I), for particles at normal incidence.

The total instrument has a field of view of about 45° from zenith and a geometrical factor of about 1040 cm^2 sr for high-energy electrons. A charge detector (CHD), comprised of a pair of plastic scintillator hodoscopes arranged in two orthogonal layers, is placed at the top of the instrument in order to reconstruct the charge of the incident particle. The energy measurement relies on two independent calorimeters: a fine-grained preshower imaging calorimeter (IMC) followed by a total absorption calorimeter (TASC). The IMC is a sampling calorimeter alternating thin layers of Tungsten absorber, optimized in thickness and position, with layers of scintillating fibers read-out individually. The TASC is a tightly packed lead-tungstate ($PbWO_4$; PWO) segmented calorimeter, capable of almost complete absorption of the TeV-electron showers. In addition, a combination of the calorimetric observations with the ones performed with a dedicated gamma-ray burst monitor (CGBM) allows the CALET Collaboration to carry out gamma-ray astronomy.

The CALET design allows to reach an electromagnetic shower energy resolution of about 2% above 20 GeV and a protons rejection factor of about 10^5, making possible to extend a well established and understood analysis procedure to the high-energy region.

3. Multivariate analysis

Monte Carlo (MC) simulations of electrons and protons, performed with the EPICS [12] framework, were used to evaluate event selection and event reconstruction efficiencies, energy correction factors and the background contamination.

A group of pre-selections [13] allowed to obtain a well reconstructed sample of electron candidates, removing events outside acceptance and particles with charge $Z > 1$.

After pre-selections, the residual proton background in the analysis has to be removed by using dedicated rejection algorithms. A simple two-parameter cut (K-cut) [13], despite its simplicity, has proven to be powerful and stable in the low energy range, below 500 GeV. Its simple application leads however to unsatisfactory results when it is applied in the high energy range, which is why a multivariate algorithm is prefered to perform a background rejection above 500 GeV. A classification of the available multivariate algorithms and an optimization of their parameters in the high energy region has then been carried out, with the target of lowering as much as possible the resultant contamination ratios of protons in the final electron sample, while keeping a constant high efficiency of 80% for electrons.

The algorithms used to perform this work are the standard ones developed in the Toolkit for Multivariate Analysis (TMVA) [14]: this tool provides a ROOT-integrated [15] environment for the processing, parallel evaluation and application of multivariate classification and multivariate regression techniques. All multivariate techniques in TMVA make use of training events, for which the desired output is known, to determine the mapping function that discribes a decision boundary (classification). Among the classification algorithms available in the TMVA toolkit, the application of the Boosted Decision Trees (BDT), Artificial Neural Network (ANN) and Deep Learning (DL) has been tested as background-rejection algorithm to the CALET electron analysis. In particular:

- BTD: the classical TMVA approach has been proposed, by fixing some parameters as the minimum percentage of training events required in a leaf node (2.5%), the number of grid points in variable range used in finding optimal cut in node splitting (20), the boosting type for the trees in the forest (*Adaptive Boost* with a Learning rate of 0.5%), the *bagging* resampling technique (with a sample fraction of 0.5%) and the separation criterion for node splitting (*GiniIndex*);
- ANN: the MultiLayer Perceptrons (MLP) approach has been proposed, as it is the fastest and recommended neural network to be used in the TMVA toolkit, along with its bayesian extension referred to as Multi Layer Perceptrons Bayesian Neural Network (MLPBNN). The parameters fixed in the MLP approach are the neuron activation function type (*tahh*), the variable transformations to be applied prior to the MVA training (normalization) and the test rate performed for overtraining (5).

The MLPBNN approach offers a means to allow for more complex network architectures while at the same time regularizing the model complexity adaptively to avoid unwanted overfitting effects. It implements the Broyden-Fletcher-Goldfarb-Shannon (BFGS) training method that allows to reduce the number of iterations by cutting down on the computation time. The parameters fixed in the MLPBNN approach are the same as the MLP one with, in addition, the application a regulator to avoid over-training when the BFGS algorithm is used;

- DL: the Deep Neural Network (DNN) approach that provides an optimized implementation of feed-forward multilayer perceptrons that can be efficiently trained on modern multi-core and GPU architecture, has been proposed with the standard parameters assignment proposed by the TMVA toolkit.

The TMVA estimator has been constructed, for each one of the selected methods, by using a sample of 13 variables related to the shower development inside the CALET detector [13].

4. Results and discussion

Before starting the parameter optimization for the selected TMVA algorithms, the Monte Carlo samples of electrons (signal) and protons (background) exiting from the pre-selection step have been splitted into training and test samples with a random seed, so that in each energy bin there were the same number of training and test events.

A different combination of the parameters that characterize the application of each one of the selected method has been carefully optimized in order to have excellent discrimination and stable performances. In particular, the parameters that have been changed are:

- the number of trees in the forest (t) and the maximum depth of the decision tree allowed (d) for the BDT algorithm;
- the number of training cycles (n) and the specification of hidden layer architecture (h) for the MLP and MLPBNN algorithms;
- the layout of the network, concerning the specific layers number (h), the number of neurons of each layer (n) and the network activation function (rs, which stands for *RELU* for the internal neurons and *SIGMOID* for the output one or ss, which stands for *SIGMOID* everywhere) for the DNN algorithm.

For each one of the four selected algorithm a series of test has been performed by changing the specific parameter and the respective performances have been evaluated. This made it possible to select, for each method, the parameters combination that provided the best performance:

- BDT ($t = 1000$, $d = 5$) and BDT ($t = 200$, $d = 10$), for the BDT algorithm;
- MLP ($n = 600$, $h = 10$), for the MLP algorithm;
- MLPBNN ($n = 60$, $h = 9$), for the MLPBNN algorithm;
- DNN ($n = 20$, $h = 1$, ss), for the DNN algorithm.

In figure 1 the Receiver Operating Characteristic (ROC) diagram, showing the background rejection versus the signal efficiency, has been reported for the four selected methods in the energy bin with $E \in [2899, 4594]$ GeV.

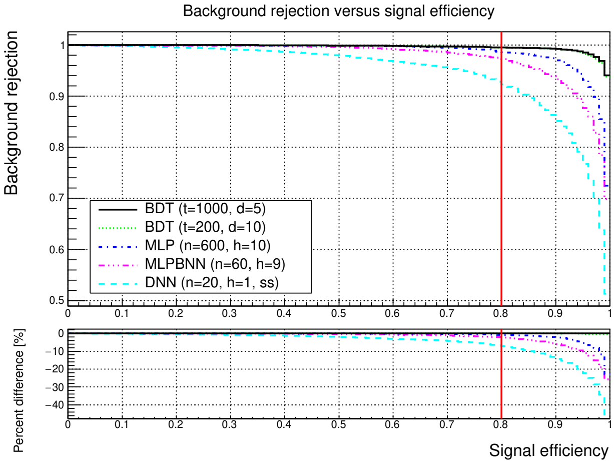


Figure 1: Receiver Operating Characteristic (ROC) diagram for the selected TMVA methods and percent difference with respect to the BDT ($t = 1000$ $d = 5$) one, chosen as the reference algorithm. The vertical line indicates the 80% signal efficiency fixed in the analysis for the electron selection.

It is evident that the BDT method is more performing with respect to the others.

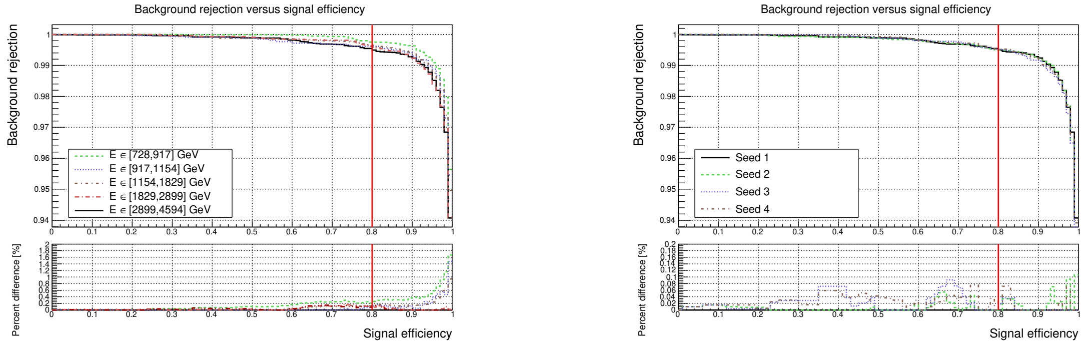
The BDT methods are confirmed to be the best ones by changing the tested energy bins in the high energy region (with fixed training-test splitting) or the seed in the training-test splitting (in a fixed energy bin in the high energy region): in both cases, at the fixed signal efficiency of 80%, the BDT methods are the best discriminating algorithms, with a percentage difference with respect to the better MLP method fixed under 2%.

Once the BDT ($t = 1000$ $d = 5$) has been identified as the better algorithm, we verified its performances as a function of the tested bin energy and of the training-test splitting. Results reported in figure 2 show that the performances obtained by changing that two parameters are very similar.

As a final test, the selected TMVA rejection algorithms have been applied on the CALET electron analysis and the measured fluxes have been compared in the high energy region. A sample of 2637 days of flight data, collected with a high-energy shower trigger and a consistently high live time fraction (~ 86%) in the full detector acceptance has been processed by following an analysis procedure similar to that used in the latest publication of the CALET Collaboration on this topic [5].

A first check has been performed by measuring the resultant proton contamination ratios of protons in the final electrons sample obtained with the the BDT ($t = 1000$ $d = 5$) algorithm. The measured contamination, which takes into account the real abundance of electron and protons in nature, is on average 10% above 500 GeV.

The stability of the analysis flux, obtained by changing the rejection algorithms has then been tested. The ratio of the fluxes obtained by using the TMVA selected methods with respect to the one obtained by using the BDT ($t = 1000$ $d = 5$) is reported in figure 3. The results are expected to



(a) performances as a function of the energy. The percent difference has been evaluated with respect to the energy bin with $E \in [2899, 4594]$ GeV.

(b) performances as a function of the training-test splitting. The percent difference has been evaluated with respect to the training-test sample splitted with seed 1.

Figure 2: performances of the BDT ($t = 1000$ $d = 5$) algorithm as a function of the energy and of the training-test splitting. The vertical line indicates the 80% signal efficiency fixed in the analysis for the electron selection.

be stable but methods that give a high proton contamination can lead to discrepant results because of the subtraction of a very high residual background.

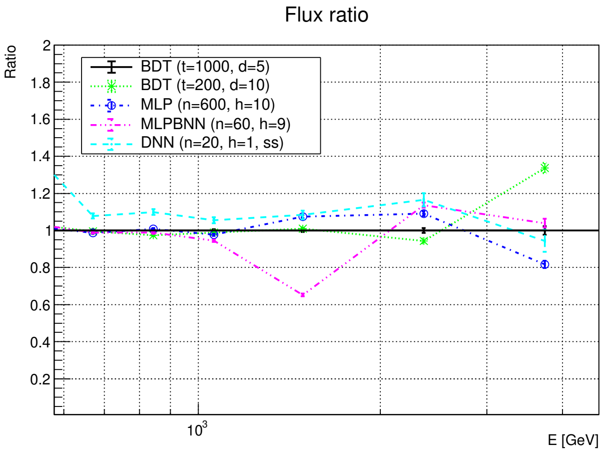


Figure 3: Flux ratio for the selected TMVA methods with respect to the BDT ($t = 1000$ $d = 5$) one. The error bars reported in the plot are only the statistical ones.

The result confirm that BDT ($t = 1000$ $d = 5$), the TMVA algorithm selected for the CALET electron analysis, turns out to be the best performing one. The MLP and MLPBNN methods result to be competitive with respect to the BDT in some energy bins but only the MLP turns out to be stable with the energy. The DNN method turns out to be unstable and more studies appear to be needed to deepen its understanding.

Acknowledgments

We gratefully acknowledge JAXA's contributions to the development of CALET and to the operations onboard the International Space Station. The CALET effort in Italy is supported by ASI under Agreement No. 2013- 018-R.0 and its amendments. The CALET effort in the United States is supported by NASA through Grants No. 80NSSC20K0397, No. 80NSSC20K0399, and No. NNH18ZDA001N-APRA18-0004. This work is supported in part by JSPS Grant-in- Aid for Scientific Research (S) Grant No. 19H05608 in Japan.

References

[1] O. Adriani *et al.* (PAMELA Collaboration), *An anomalous positron abundance in cosmic rays with energies 1.5-100 GeV*, *Nature* **458** (2009) 607 [astro-ph/0810.4995].

[2] L. Accardo *et al.* (AMS Collaboration), *High Statistics Measurement of the Positron Fraction in Primary Cosmic Rays of 0.5–500 GeV with the Alpha Magnetic Spectrometer on the International Space Station*, *Phys. Rev. Lett.* **113** (2014) 121101.

[3] S. Torii, P. S. Marrocchesi *et al.* (CALET Collaboration), *The CALorimetric Electron Telescope (CALET) on the International Space Station*, *Adv. Space Res.* **64** (2019) 12, 2531.

[4] O. Adriani *et al.* (CALET Collaboration), *Energy Spectrum of Cosmic-Ray Electron and Positron from 10 GeV to 3 TeV Observed with the Calorimetric Electron Telescope on the International Space Station*, *Phys. Rev. Lett.* **119** (2017) 181101 [astro-ph.HE/1712.01711].

[5] O. Adriani *et al.* (CALET Collaboration), *Extended Measurement of the Cosmic-Ray Electron and Positron Spectrum from 11 GeV to 4.8 TeV with the Calorimetric Electron Telescope on the International Space Station*, *Phys. Rev. Lett.* **120** (2018) 261102 [astro-ph.HE/1806.09728].

[6] O. Adriani *et al.* (CALET Collaboration), *Direct Measurement of the Cosmic-Ray Proton Spectrum from 50 GeV to 10 TeV with the Calorimetric Electron Telescope on the International Space Station*, *Phys. Rev. Lett.* **122** (2019) 181102 [astro-ph.HE/1905.04229].

[7] O. Adriani *et al.* (CALET Collaboration), *Observation of Spectral Structures in the Flux of Cosmic-Ray Protons from 50 GeV to 60 TeV with the Calorimetric Electron Telescope on the International Space Station*, *Phys. Rev. Lett.* **129** (2022) 101102 [astro-ph.HE/2209.01302].

[8] O. Adriani *et al.* (CALET Collaboration), *Direct Measurement of the Cosmic-Ray Helium Spectrum from 40 GeV to 250 TeV with the Calorimetric Electron Telescope on the International Space Station*, *Phys. Rev. Lett.* **130** (2023) 171002 [astro-ph.HE/2304.14699].

[9] O. Adriani *et al.* (CALET Collaboration), *Cosmic-Ray Boron Flux Measured from* $8.4\;\mathrm{GeV}/n$ *to* $3.8\;\mathrm{TeV}/n$ *with the Calorimetric Electron Telescope on the International Space Station*, *Phys. Rev. Lett.* **129** (2022) 251103 [astro-ph.HE/2212.07873].

[10] O. Adriani *et al.* (CALET Collaboration), *Direct Measurement of the Cosmic-Ray Carbon and Oxygen Spectra from* $10\;\mathrm{GeV}/n$ *to* $2.2\;\mathrm{TeV}/n$ *with the Calorimetric Electron Telescope on the International Space Station*, *Phys. Rev. Lett.* **125** (2020) 251102 [astro-ph.HE/2012.10319].

[11] O. Adriani *et al.* (CALET Collaboration), *Measurement of the Iron Spectrum in Cosmic Rays from* 10 GeV/n *to* 2.0 TeV/n *with the Calorimetric Electron Telescope on the International Space Station*, *Phys. Rev. Lett.* **126** (2021) 241101 [astro-ph.HE/2106.08036].

[12] K. Kasahara, *Introduction to Cosmos and some Relevance to Ultra High Energy Cosmic Ray Air Showers*, in proceedings of *24th International Cosmic Ray Conference (ICRC1995)*, Editrice Compositori, Bologna 1995.

[13] E. Berti *et al.* (CALET Collaboration), *The analysis strategy for the measurement of the electron flux with CALET on the International Space Station*, in proceedings of *37th International Cosmic Ray Conference (ICRC2021)*, PoS(ICRC2021)065 (2021).

[14] A. Hoecker *et al.*, *TMVA: Toolkit for Multivariate Data Analysis*, *CERN-OPEN-2007-007*, *arXiv:physics/0703039* [physics.data-an].

[15] R. Brun and F. Rademakers, *ROOT - An Object Oriented Data Analysis Framework*, in proceedings of *5th International Workshop on New Computing Techniques in Physics Research: Software Engineering, Neural Nets, Genetic Algorithms, Expert Systems, Symbolic Algebra, Automatic Calculations (AIHENP 96)*, *Nucl. Instrum. Meth. A* **389** (1997) 81.