*particles*

MDPI

*Review*

# Feature Selection Techniques for CR Isotope Identification with the AMS-02 Experiment in Space

**Marta Borchiellini [1,*], Leandro Mano [2], Fernando Barão [3,4] and Manuela Vecchi [1]**

[1] Kapteyn Astronomical Institute, University of Groningen Landleven 12, 9747 AD Groningen, The Netherlands; m.vecchi@rug.nl
[2] National Council of Scientific and Technological Development, SHIS Q1, Edifício Santos Dumont 203, Brasília 71605-001, Brazil; leandroyukiomano@gmail.com
[3] Laboratório de Instrumentação e Física Experimental de Partículas (LIP), 1649-003 Lisboa, Portugal; barao@lip.pt
[4] Departamento de Física, Instituto Superior Técnico—IST, Universidade de Lisboa—UL, Avenida Rovisco Pais 1, 1049-001 Lisboa, Portugal
* Correspondence: m.borchiellini@rug.nl

**Abstract:** Isotopic composition measurements of singly charged cosmic rays (CR) provide essential insights into CR transport in the Galaxy. The Alpha Magnetic Spectrometer (AMS-02) can identify singly charged isotopes up to about 10 GeV/n. However, their identification presents challenges due to the small abundance of CR deuterons compared to the proton background. In particular, a high accuracy for the velocity measured by a ring-imaging Cherenkov detector (RICH) is needed to achieve a good isotopic mass separation over a wide range of energies. The velocity measurement with the RICH is particularly challenging for $Z = 1$ isotopes due to the low number of photons produced in the Cherenkov rings. This faint signal is easily disrupted by noisy hits leading to a misreconstruction of the particles' ring. Hence, an efficient background reduction process is needed to ensure the quality of the reconstructed Cherenkov rings and provide a correct measurement of the particles' velocity. Machine learning methods, particularly boosted decision trees, are well suited for this task, but their performance relies on the choice of the features needed for their training phase. While physics-driven feature selection methods based on the knowledge of the detector are often used, machine learning algorithms for automated feature selection can provide a helpful alternative that optimises the classification method's performance. We compare five algorithms for selecting the feature samples for RICH background reduction, achieving the best results with the Random Forest method. We also test its performance against the physics-driven selection method, obtaining better results.

**Keywords:** feature selection; cosmic rays; isotope identification



**Citation:** Borchiellini, M.; Mano, L.; Barão, F.; Vecchi, M. Feature Selection Techniques for CR Isotope Identification with the AMS-02 Experiment in Space. *Particles* **2024**, *7*, 417–434. https://doi.org/10.3390/particles7020024

Academic Editor: Armen Sedrakian

Received: 20 December 2023
Revised: 29 March 2024
Accepted: 11 April 2024
Published: 20 April 2024



## 1. Introduction

Positive, singly charged nuclei dominate the galactic cosmic ray (CR) spectrum [1]. Cosmic rays can be divided into two main categories based on their production mechanism: primary cosmic rays, which are produced directly in stellar nucleosynthesis processes at the sources, and secondary cosmic rays, which originate from the nuclear interaction of primary CRs with the interstellar medium (ISM) during their propagation in the Galaxy [2]. Although protons dominate the isotopic composition of hydrogen in cosmic rays, a few percent of deuterons are also present. They are expected to be mostly of secondary origin since the primary deuterons produced in the first step of the proton–proton chain are depleted in the next step of the nucleosynthesis reaction [3]. Secondary deuterons are produced through inelastic interactions between CRs, mainly p, $^3$He, $^4$He, and the ISM. Thus, the identification of deuterons and the measure of their flux is essential for the study of cosmic ray propagation processes in the Galaxy. In particular, it is possible to factor

out the source contribution to the spectrum using secondary-to-primary ratios, such as deuteron-to-proton (d/p) and deuteron-to-helium-4 (d/$^4$He) [2] to constrain the parameters of the galactic propagation model.

Isotope identification for singly charged particles has already been performed by magnetic spectrometers such as PAMELA [4], IMAX [5], and CAPRICE [6] for CR energies up to a few GeV/n. The Alpha Magnetic Spectrometer (AMS-02) [7] will extend the energy range of isotopic composition measurement [8] to ∼10 GeV/n. However, the analysis presents challenges due to the intrinsic characteristics of the measurement. Isotopes are separated through their mass by combining the rigidity ($R = p\,c/Z\,e$, momentum per unit charge) and the particle's velocity, as follows:

$$m = \frac{R\,Z\,e}{\beta\,\gamma}\,, \tag{1}$$

where $Ze$ is the magnitude of the charge, $\beta = v/c$ is the velocity in speed of light units, and $\gamma$ is the Lorentz factor. The mass resolution can be derived from Equation (1):

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

Due to the dependence on the fourth power of the Lorentz factor, the velocity resolution's contribution to the total mass resolution dominates for $\beta \to 1$, which happens for energies higher than a few GeV/n. Therefore, efficiently cleaning the initial sample from events whose RICH velocities have not been accurately reconstructed is essential to identify singly charged isotopes and extend the energy range of the measurement.

Artificial intelligence (AI) methods, particularly machine learning (ML), are widely used for particle identification in particle and astroparticle physics [9–14]. In particular, boosted decision trees (BDTs) have been employed to classify events with misreconstructed RICH velocities and reject the background for the identification of deuterons [15]. For this type of classification task, complex data sets containing many variables and parameters are analysed, and data are often noisy and contain correlated information, which is thus redundant. Hence, a crucial step in the application of ML methods to classification tasks is selecting the variables (or features) that constitute the input required for model training. ML-driven feature selection techniques are emerging as suitable tools to optimise the performance of ML algorithms for classification tasks in particle [16] and astroparticle physics [17–19]. Furthermore, feature selection has already been used in cosmic ray identification for ground-based experiments in Herrera et al. [20] to rank the relevance of features involved in primary particle reconstruction from air shower simulations. The importance of feature selection lies in its ability to simplify the data analysis process. By identifying and selecting the most relevant features, it is possible to enhance the efficiency and accuracy of the classification algorithms, making the results more interpretable and robust. Moreover, feature selection aids in the avoidance of overfitting, a common pitfall in complex data sets. Focusing on essential features reduces the risk of models becoming overly tailored to the training data, thereby increasing their generalisation capabilities.

In this work, we study machine learning (ML) algorithms for feature selection in the context of singly charged cosmic ray isotope identification using the AMS-02 experiment and investigate whether these ML techniques could enhance the efficiency in reducing the RICH background compared to traditional physics-driven methods, such as Bueno et al. [15]. Five ML techniques are used to single out, from a sample of 130 features obtained by the detection and reconstruction of cosmic ray nuclei with the RICH detector of the AMS-02 experiment, the most promising features to identify the signal and reject the background. Furthermore, for comparison, the physics-driven approach proposed by Bueno et al. [15] is used to choose a set of features based on the knowledge of the detector and of the type of background to be reduced.

The paper is organised as follows: In Section 2, the AMS-02 detector is presented. The preparation of the data set and the algorithms used to perform feature selection are described in Section 3, along with the metrics used to evaluate the performance of the different methods. In Section 4, the results of the different models and their performance on the validation basis are followed by a discussion on the features selected by the best-performing model. We conclude in Section 5.

## 2. The AMS-02 RICH Detector

The Alpha Magnetic Spectrometer (AMS-02) is a cosmic ray detector operational aboard the International Space Station since May 2011 [7]. Its unique capabilities allow for the measurement of deuteron flux in previously unexplored energy ranges, extending it nowadays above a limit of 4 GeV/n.

AMS-02 consists of several subsystems as follows: A silicon tracker with nine layers, positioned from the top to the bottom of the detector, in conjunction with a permanent magnet of 0.15 T. This combination enables the measurement of the magnitude, the sign of the charge, and the rigidity of the particles; a transition radiation detector (TRD), designed to distinguish between leptons and hadrons; a time-of-flight (TOF) system, comprising two pairs of scintillators (upper TOF and lower TOF) located above and below the magnet, is responsible for measuring the velocity and charge $Z$ of the particles and serves as the primary trigger for the experiment; a ring-imaging Cherenkov detector (RICH), positioned below the lower TOF, is used to measure the particle velocity and charge $Z$; an anti-coincidence counter (ACC), which identifies and rejects particles with high-incidence angles; an electromagnetic calorimeter (ECAL), positioned below the RICH, is responsible for measuring particle energy and enabling differentiation between leptons and hadrons.

The RICH detector [21] plays a crucial role in measuring hydrogen isotope fluxes, allowing for their identification up to 10 GeV/n [22]. However, it is essential to emphasise that this task presents significant challenges due to the intrinsic nature of the Cherenkov effect. Because the intensity of the emitted radiation is proportional to $Z^2$ [23], singly charged isotopes generate a faint signal compared to higher $Z$ particles, making the velocity reconstruction vulnerable to background disruptions, especially near the threshold of each radiator.

The RICH detector [24–26] (see Figure 1) features a truncated conical shape with a 60 cm top radius, a 67 cm bottom radius, and an expansion height of 47 cm. The detector comprises a radiator plane, an expansion volume, and a photo-detection plane. The double radiator plane includes a central radiator consisting of 16 tiles of sodium fluoride (NaF) measuring $8.5 \times 8.5 \times 0.5$ cm$^3$, with a refraction index of 1.33. It is surrounded by 92 silica aerogel tiles measuring $11.5 \times 11.5 \times 2.5$ cm$^3$ and having a refraction index of 1.05.

The detection plane is equipped with an array of 680 photomultiplier tubes (PMTs) arranged in eight grids, four rectangular and four triangular. To minimise lateral losses of approximately 30% of the radiated Cherenkov photons, the expansion volume is surrounded by a highly reflective mirror that meets roughness specifications of better than 150 nm and exhibits a reflectivity exceeding 90% at $\lambda = 420$ nm.

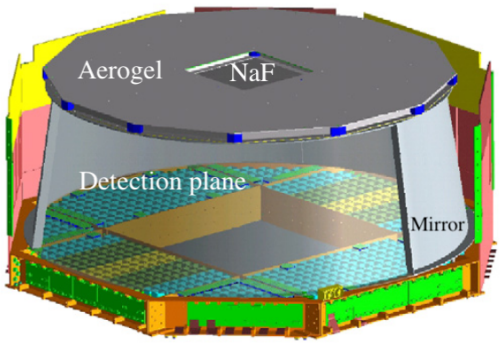


**Figure 1.** Sketch of the ring-imaging Cherenkov detector of AMS-02. Adapted with permission from Ref. [7]. Copyright 2011, AMS Collaboration.

## 3. Methodology

### *3.1. Database Description*

Six months of data collected by AMS-02 in space (from December 2015 to May 2016) were used in this work. While this particular choice is arbitrary, we believe it does not introduce any bias to our analysis, providing us with a statistically relevant sample. The performance of the AMS-02 detector has been extensively verified in the past years [27] and no specific time-dependent effect on performance was reported.

A first selection was applied to ensure the quality of the reconstructed tracker track and of the velocity measured by TOF for the used events. Singly charged isotopes were selected using TOF and silicon tracker charge measurements. The requirements on the tracker charge ($Z_{TRK}$) and upper and lower TOF charge ($Z_{UTOF}$, $Z_{LTOF}$) are $0.75 < Z_{TRK,UTOF} < 1.5$ and $0.75 < Z_{LTOF} < 1.3$, respectively. Then, the selected events were divided into two samples based on their reconstructed masses, following Bueno et al. [15]: events with a mass within $2\sigma$ from the proton mass ($0.75\ \mathrm{GeV/c^2} < m < 1.25\ \mathrm{GeV/c^2}$) or a mass above $4\sigma$ from the triton mass ($m > 4\ \mathrm{GeV/c^2}$) are considered to be signal-like or background-like, respectively. This preparation is necessary to have a labelled set of events to be used for the training of the classification method.

As was pointed out in Bueno et al [15], the residual background for the identification of cosmic ray deuterons consists mainly of events whose RICH velocity is poorly reconstructed due to noise disrupting the already weak signal produced by the $Z = 1$ particles. In particular, particles produced from the interactions of the incoming cosmic rays with the AMS-02 detector can induce spurious hits that are not related to the Cherenkov emission of cosmic rays. These ring-uncorrelated hits consist of additional photons generated as the particles produced by the interactions in the detector cross the aerogel radiator or while the same particles cross the PMT plane. These spurious events induce additional photon hits and consequently affect the reconstruction of the Cherenkov ring. Furthermore, interactions between particles happening in the region between the lower tracker and the RICH can modify the direction of the incoming cosmic ray, thus introducing a slight bias in the number of detected photon hits in the Cherenkov ring with respect to the reconstructed tracker track. These events whose velocities are poorly reconstructed constitute the background of this analysis and they are the ones to be rejected to improve mass resolution and correctly identify singly charged isotopes. They are mainly located in the high mass tail of the particles' mass distribution, as outlined in Bueno et al. [15]; for this reason, the background-like sample comprises events falling within that specific mass distribution region. Conversely, the mass range selected for the signal-like sample corresponds to the proton peak of the mass distribution and it is characterised by events whose velocity has been correctly reconstructed.

A set of 130 features was used for the analysis, including quantities directly measured by the RICH detector (e.g., the number of crossed PMTs and the number of hits), the reconstructed quantities (e.g., charge and velocity), and the expected values of the measured and reconstructed features computed using the reconstruction and input parameters of the event itself (e.g., expected number of photoelectrons associated with the Cherenkov ring). As shown in Figure 2, the used features can be divided into six classes: Charge, Track position, PMT number, Beta, Hit number, and Photoelectrons.

The class "Charge" contains seven features related to the magnitude of the reconstructed charge: these include two different estimates of the charge based on different Cherenkov ring reconstruction methods [28,29], the expected charge resolution and its mean square error, and the value for the Kolmogorov probability associated to the hypothesis of a uniform distribution of signal hits along the particle path [30]. The class "Track Position" contains 16 features related to the extrapolated tracker track of the particle inside the RICH, such as the coordinates of the impact point on the radiator plane and the angles that the reconstructed track forms with it, the distance of the impact point from the border of the radiator tile, and the information on the radiator tile crossed by the extrapolated track. The class "Beta" contains 40 features connected to the reconstructed velocity of

the particle: it includes, for example, the particles' velocity, reconstructed independently by two methods [28,29], the values of the velocity resolution, and the mean square error. The class "Hit number" delivers information about the number of photons detected in the photodetection plane by means of the number of photoelectron hits registered by the PMTs. Among these 26 features are included the total number of detected hits, and the number of hits inside and outside the Cherenkov ring. The class "Photoelectrons" includes 36 features incorporating different information about the measured number of photoelectrons (p.e.) deriving from the detection of the Cherenkov photons in the PMTs of the detection plane. For every event, the number of p.e. in the PMT with the highest number of p.e., the number of p.e. collected in and out of the ring, and the number of p.e. expected for a singly charged nucleus or an electron with the reconstruction and input parameters of the current event are available. Finally the class "PMT number" is related to the number of PMTs crossed in each event, and includes five features related to the measured and expected number of PMTs inside and outside of the Cherenkov ring. All the variables used in the analysis are listed and briefly described in Table A1.

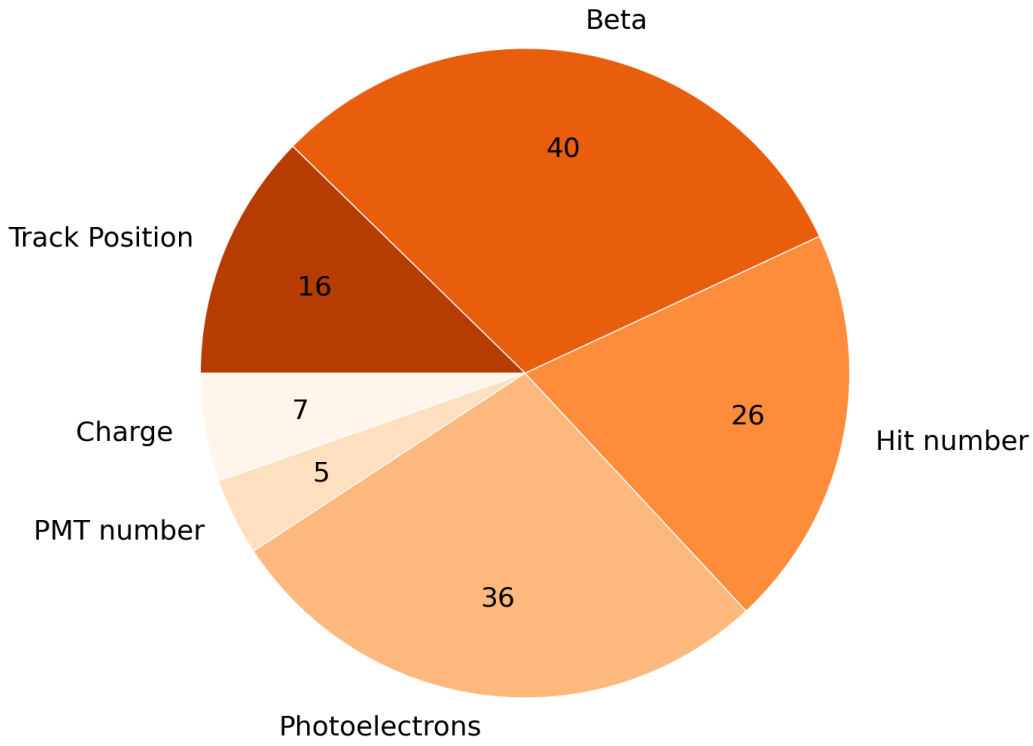


**Figure 2.** Pie chart showing the breakdown of the 130 features into six classes. The value in each slice represents the number of features in the corresponding class.

To illustrate the different behaviour that signal- and background-like events display, Figure 3 shows the distributions of the Kolmogorov probability and the magnitude of the charge, denoted by $Z$, for a signal-like (in blue) and a background-like (in red) sample of events. Both features show different distributions for the two samples; hence, they discriminate well between background-like and signal-like events. This conclusion derives from the physical phenomena that these features trace. The left plot shows the distribution of the Kolmogorov probability [30] that is obtained by performing for each event a Kolmogorov test on the azimuth distribution of emitted photons along the particle path, expected to be uniform for well-reconstructed events with a ring-like shape (i.e., signal-like events), and to be non-uniform for background events, whose rings include noisy hits. The Kolmogorov test was used to compare the expected cumulative distribution for the azimuth angle with the measured one. When the discrepancy between these two distributions is maximal the Kolmogorov probability scores very low values, justifying the behaviour of

the background-like sample in the left plot in Figure 3. On the other hand, the Kolmogorov probability is calculated such that it is almost uniformly distributed between 0 and 1 for ring-shaped events, as shown by the signal-like sample in the same plot. The distribution of the reconstructed charge is shown in the right panel of Figure 3. In the RICH detector, the square of the charge $Z$ is proportional to the number of photoelectrons associated with the Cherenkov ring (i.e., $Z^2 \propto N_{p.e.}$), with a scaling factor accounting for the ring acceptance and velocity dependence. We expect signal-like events to have a symmetric charge distribution peaked at $Z = 1$, as in Figure 3. Conversely, the background-like sample mostly contains events whose rings have spurious hits, leading to the characteristic high charge tail in the red distribution.

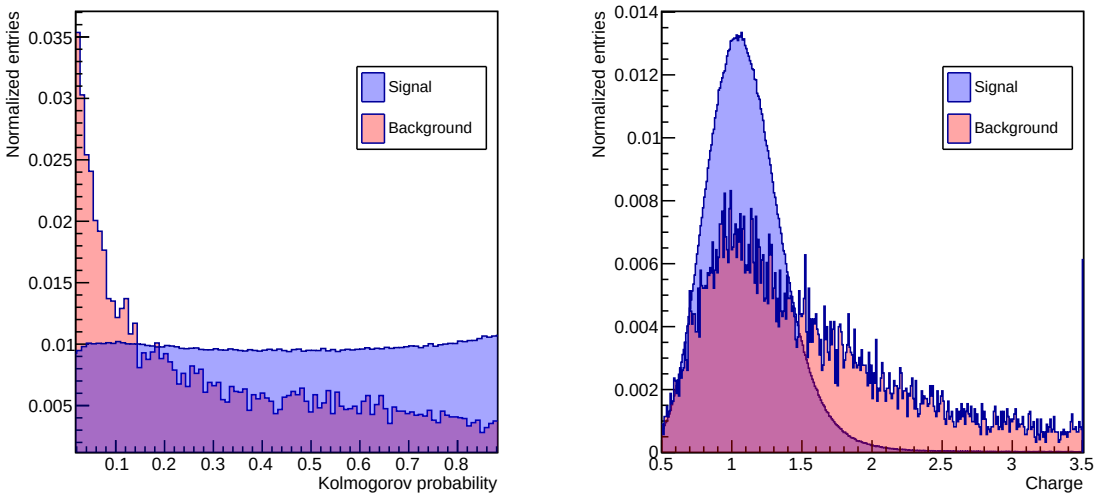


**Figure 3.** The distributions of reconstructed charge (**left**) and Kolmogorov probability (**right**) for the aerogel radiator for the signal-like (in blue) and background-like samples (in red). The purple regions represent the overlaps between these two samples.

### *3.2. Feature Selection Techniques*

Feature selection techniques play a crucial role in the data analysis and the modelling of ML algorithms. The importance of these techniques resides in the fact that not all attributes contribute equally to constructing an accurate and efficient model [31]. By selecting the relevant characteristics of the data set, it is possible to improve the generalisation capacity of the model, reducing the chance of overfitting [31–33]. Furthermore, other benefits obtained when applying feature selection are the reduction in noise and interference and the acceleration of the training time and computational efficiency, as it reduces processing resources [34,35]. Therefore, feature selection techniques are essential in optimising and improving ML models, allowing better interpretability, efficiency, and accuracy of data analysis.

The most straightforward approach in selecting the relevant features is to test every possible subset of features, finding the one that minimises the error rate. However, this is an exhaustive and computationally intractable search for real data sets. In this sense, the choice of the evaluation metric strongly influences the feature selection technique, and it is these evaluation metrics that distinguish between the three main categories of feature selection techniques [36,37], namely:

- Filter: Use a proxy measure instead of an error rate to score a subset of features;
- Wrapper: Use a predictive model to score subsets of features. Each new subset is used to train a model, which is tested on a validation set;
- Embedded: Is a comprehensive group of techniques that perform feature selection as part of the model-building process.

Thus, in order to achieve the project's objective of verifying the most relevant features in the database, we selected a technique for each category of feature selection techniques, namely: SelectKBest (filter), Random Forest - RF (wrapper), and linear regression (embedded). In addition to the mentioned techniques, Pearson's correlation is applied as a feature selection technique. We briefly describe the techniques used:

- Kbest: is an approach that selects the $k$ best attributes based on a statistical measure, such as the analysis of variance (ANOVA) used in this study. By defining a value for $k$, it is possible to choose the $k$ most significant features, which have a more relevant impact on the model's prediction [38,39];
- Random Forest (RF): is a technique that can be applied not only for building classification and regression models but also as a practical feature selection tool [40]. Random Forest performs several independent decision trees, each using different subsets of attributes and random samples from the data set. During this process, the algorithm calculates the importance of each attribute based on its contribution to the overall accuracy of the predictions [37];
- Linear Regression: the approach considers the coefficients of each attribute as a measure of the individual contribution in predicting the dependent variable [41]. The magnitude and sign of the coefficient indicate the impact on the target variable. Thus, features with higher and statistically significant coefficients are considered more important and can be selected as part of the feature selection process [37,41];
- Pearson's Correlation (CORR): the approach measures the strength and direction of the linear relationship between two continuous variables [42]. When calculating the Pearson correlation between each attribute and the target features, it is possible to obtain a value ranging from $-1$ to 1. A value close to $-1$ indicates a strong negative correlation. On the other hand, a value close to 1 indicates a strong positive correlation. A value close to 0 indicates a weak or no correlation. Based on the correlation values, it is possible to select attributes with a more significant correlation with the target variable [31,43].

Furthermore, for comparison purposes, the methodology proposed by Bueno et al. [15] is considered to select the last set of features. This method involves the detailed study of the events whose mass is incorrectly identified due to the interactions occurring within the AMS-02 detector and aims to identify the sources of interactions relevant to the RICH reconstruction background and mitigate this background efficiently. In this work, the features are chosen following a physics-driven approach based on the knowledge of the RICH detection mechanism and velocity reconstruction method and they can be combined to form more effective features. A multivariate estimator is subsequently used to complete the classification task.

### *3.3. Performance Evaluation Metrics*

Accurately evaluating the performance of the algorithms is essential to guide the choice of the best classification model [44]. Metrics such as accuracy, precision, F1-score, and recall play a central role in this evaluation, allowing the effectiveness and usefulness of the algorithms to be measured [45]. In the following, we will explore the meaning and calculation of these metrics, highlighting their relevance and providing a general description to ensure reliable and accurate AI systems:

- Accuracy: this metric provides a general measure of the model's ability to correctly predict classes. It is helpful in scenarios where all classes have similar importance. It considers true positives (correctly classified cases) and true negatives (correctly classified negative cases) concerning the total number of examples. Accuracy is defined as follows:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Examples}}; \tag{3}$$

- Precision: this metric focuses on the quality of the model's positive predictions. It is particularly relevant when false positives have a substantially more significant impact than false negatives. It provides deeper insight into the model's ability to avoid the erroneous classification of negative examples as positive. Precision is defined as follows:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}; \quad (4)$$

- Recall: this metric, also known as sensitivity, focuses on the model's ability to identify positive cases while effectively minimising false negatives. The recall metric is defined as follows:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}; \quad (5)$$

- F1-score: this metric combines the precision and recall metrics to provide a balanced measure of model performance. It is particularly relevant when the balance between accurately identifying positive cases and minimising false positives and false negatives is essential. The F1-score is defined as follows:

$$\text{F1-score} = 2 \times \left( \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right); \quad (6)$$

In summary, performance evaluation metrics, including accuracy, precision, recall, and F1-score, play an essential role in evaluating classification algorithms. They provide valuable insights into the quality and effectiveness of forecasts, adapting to different needs and contexts.

## 4. Experiments, Results, and Discussion

This section describes the experiments conducted in this study and provides a critical evaluation of the results obtained. Section 4.1 describes the steps involved in processing and balancing data to create an equitable and reliable data set. Next, Section 4.2 examines the strategies employed to identify the most informative features of data sets. Subsequently, Section 4.4 presents the results achieved through the developed models, highlighting their performance metrics and predictive capacity. Finally, Section 4.5 explores the role and contribution of the Random Forest algorithm in achieving the study objectives.

### *4.1. Preparation of the Database for the Experiments*

The data sample used in this work is made of almost 3 million well-reconstructed singly charged events crossing the RICH detector of AMS-02. This sample primarily comprises events whose mass is well reconstructed (signal), with a tiny fraction of about one percent of events whose mass is misreconstructed (background). The disproportion between the two samples can introduce significant biases in statistical analysis and modelling, undermining the effectiveness and reliability of ML algorithms. To solve this issue, we use the `RandomUnderSampler` technique available on the `Imbalanced-learn` Python toolbox [46], which performs a random and strategic selection of samples from the majority class, reducing their representativeness and, thus, levelling the data set for subsequent analyses, lowering the disparity between the interest categories (signal and background). As a result of this balancing procedure, a final data set consists of 33.234 events, 16.617 of which are signal and 16.617 background.

In the subsequent stage, after balancing, the data set was divided into two sets to perform the classification task: training data and validation data. The training data set (70% of the events) is intended for selecting the most relevant characteristics using feature selection techniques. In comparison, the validation data set (30% of the events) is reserved for the evaluation and validation of the results obtained.

The described approach is followed for obtaining solid and reliable models capable of dealing with unbalanced data, selecting the most informative features, and validating

their effectiveness in independent test environments, thus contributing to more accurate analyses and informed decisions.

Table A2 details the parameters used in the employed feature selection techniques, which include Kbest, Random Forest, linear regression, and correlation. We chose to use the default values of the respective software libraries to ensure reproducible results.

Figure 4 presents a visual comparison between the selected methods for the identified classes (Charge, Track Position, Beta, Hit number, Photoelectrons, and PMT number). Each method is represented in an individual graph: the values on the radial axes represent the percentage of selected features for each class in relation to the number of features originally contained in the same class, allowing a quick and comprehensive comparative analysis of the selected features of each method in the different classes. Furthermore, Table 1 complements this visual representation by listing the number (and percentages) of the features selected for each class by each technique.

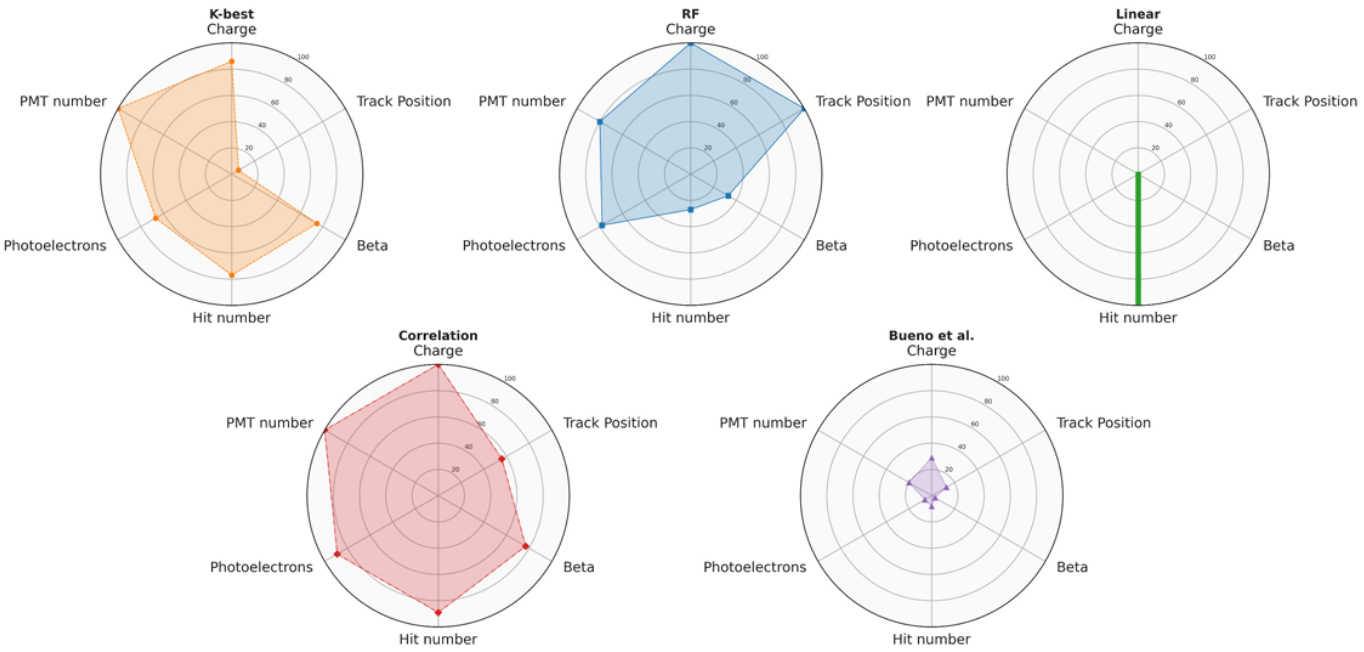


**Figure 4.** Fraction of features selected by each method out of the six classes discussed in this work. See text for discussion.

**Table 1.** Number (and percentages) of the feature selected for each class by the different methods

| | Charge | Track Position | Beta | Hit Number | Photoelectrons | PMT Number | Total |
|---|---|---|---|---|---|---|---|
| Kbest | 6 (86%) | 1 (6%) | 30 (75%) | 20 (77%) | 24 (67%) | 5 (100%) | 86 |
| RF | 7 (100%) | 16 (100%) | 13 (33%) | 7 (27%) | 28 (78%) | 4 (80%) | 75 |
| Linear | 0 (0%) | 0 (0%) | 0 (0%) | 1 (100%) | 0 (0%) | 0 (0%) | 1 |
| Correlation | 7 (100%) | 9 (56%) | 31 (77%) | 23 (89%) | 32 (89%) | 5 (100%) | 107 |
| Bueno et al. | 2 (29%) | 2 (13%) | 1 (3%) | 2 (8%) | 2 (6%) | 1 (20%) | 9 |

The percentage of variables selected by the ML methods for each class strongly depends on the algorithm itself. Nonetheless, there are some classes with a high percentage of features selected for all the methods used: this hints towards a connection between the discrimination power of the features and the physics underlying the detection mechanism, as will be discussed more in detail in Section 4.5.

### *4.2. Analysis of Feature Selection Techniques*

In this study, a boosted decision tree algorithm (AdaBoostClassifier) was used, with the parameter number of estimators (*n_estimators*) set at 100. In addition, *k*-fold cross-validation was used, for training and evaluating the models, with $k = 10$, with $k - 1$ for training and the rest for testing [47,48], thus obtaining the average accuracy rate for each algorithm. The four panels of Figure 5 show the performance of the classification algorithm trained with the five sets of features obtained with the methods discussed in Section 3.2. For comparison, the performance of the classifier trained with all the variables in the database is also computed and labelled as "All". Table 2 shows the evaluation metrics (accuracy, precision, F-1 score, and recall) together with the *p*-value (Shapiro–Wilks test) for the feature selection techniques used in this work. The number of significant digits has been chosen based on the uncertainties reported in Table A3.

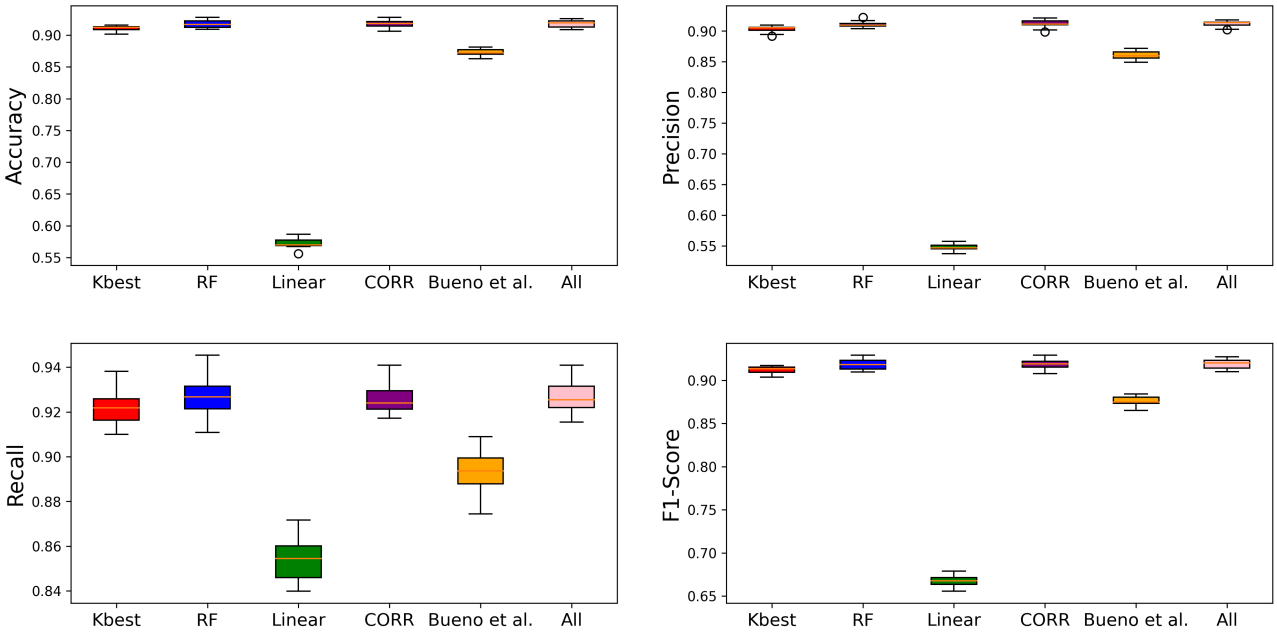


**Figure 5.** Boxplots showing the accuracy, precision, F1-score, and recall for the different selection algorithms used.

**Table 2.** Mean value of assessment metrics and *p*-values for the sets of selected features.

| | Accuracy | Precision | F-1 Score | Recall | *p*-Values (*Shapiro–Wilk*) |
|---|---|---|---|---|---|
| Kbest | 0.911 | 0.903 | 0.912 | 0.923 | 0.860 |
| RF | 0.918 | 0.911 | 0.912 | 0.927 | 0.871 |
| Linear | 0.572 | 0.548 | 0.668 | 0.854 | 0.792 |
| Correlation | 0.918 | 0.911 | 0.919 | 0.927 | 0.063 |
| Bueno et al. | 0.873 | 0.860 | 0.876 | 0.893 | 0.977 |
| All | 0.918 | 0.911 | 0.920 | 0.927 | 0.395 |

The Shapiro–Wilk test was applied to our data set to verify the normality hypothesis and, therefore, determine suitability for parametric or non-parametric tests.The Shapiro–Wilk test is a general test designed to detect all deviations from normality. The test rejects the hypothesis of normality when the *p*-value is less than or equal to 0.05. Failing the normality test allows to assert with 95% confidence that the data do not fit the normal distribution. All techniques yield *p*-values above the 0.05 threshold (see Table 2), corroborating the hypothesis of normal distribution of our data. However, a closer inspection reveals differences between their *p*-values, highlighting distinct distribution profiles for each feature selection technique,

and underscoring the importance of considering both statistical normality and performance metrics in the context of the AMS-02 experiment.

For the parametric test, the *t*-test was used. The *t*-test is a hypothesis test that uses statistical concepts to reject or not reject a null hypothesis. This assumption is usually accepted when the test statistic follows a normal distribution. This parametric method allows excellent reliability in data analysis with conformity to the normal distribution. Paired comparisons with the *t*-test are shown in Table 3, and values less than 0.05 indicate a statistically significant difference between group results.

**Table 3.** *p*-values of the pairwise comparison performed with the *t*-test for the feature selection techniques.

| | **Kbest** | **RF** | **Linear** | **Correlation** | **Bueno et al.** |
|---|---|---|---|---|---|
| RF | 0.001 | - | - | - | - |
| Linear | 0.000 | 0.000 | - | - | - |
| Correlation | 0.027 | **0.704** | 0.000 | - | - |
| Bueno et al. | 0.000 | 0.000 | 0.000 | 0.000 | - |
| All | 0.002 | **1.000** | 0.000 | **0.658** | 0.000 |

The results demonstrate that the features selected by the RF (75 features) and correlation (107 features) techniques do not present statistical differences if all 130 features are used since comparisons between pairs of results are more significant than 0.05. Therefore, only the RF and correlation approaches were considered for the following tests and compared to the approach including all features.

### *4.3. Quantitative Analysis of Computational Complexity*

As shown in the previous section, the RF and correlation techniques do not show statistically significant differences compared to using the set including all variables. Hence, it is crucial to understand how these methods differ in processing time, which can be a limiting factor in practical applications. In this section, we present a quantitative analysis of computational complexity, specifically focused on the CPU time required to train a BDT. We compare the performance of the RF, of the correlation algorithm, and of the approach that employs all 130 features (All). This analysis is carried out on a computer equipped with an 11th Gen Intel(R) Core(TM) i9-11900KF @ 3.50GHz, 3.40 GHz processor to provide a solid basis for evaluation.

Table 4 presents the average CPU times used for each technique and their standard deviation, based on 100 runs, offering a detailed perspective on the variability and consistency between runs. We observe that the time required to train the BDT using the RF, correlation, and all techniques are, respectively, 659, 644, and 718 s.

**Table 4.** Average processing time for the RF, and correlation techniques, compared with the performance for all features.

| | **Average Time [s]** | **Standard Deviation [s]** |
|---|---|---|
| RF | 659 | 8.3 |
| Correlation | 644 | 12.6 |
| All | 718 | 9.4 |

The similarity in processing times among the three techniques indicates that, despite minor differences, these variations do not result in a notable advantage or disadvantage in terms of time performance. In essence, selecting a technique based solely on processing time may not be a decisive factor, as all methods are relatively efficient. Therefore, the choice between the RF, correlation, and all techniques should be based on a comprehensive assessment that considers performance both in terms of evaluation metrics and computational efficiency.

### *4.4. Assessment of Predictive Models*

To evaluate the implemented models, the validation database previously separated was used to evaluate the implemented models. It is worth mentioning that the validation database contains data that were not used in training and, therefore, are unknown to the models. Figure 6 presents the ROC-AUC metric of the RF and correlation techniques compared to the "All" approach, using all features. Additionally, Table 5 provides the values of all metrics for better cross-technique analysis.

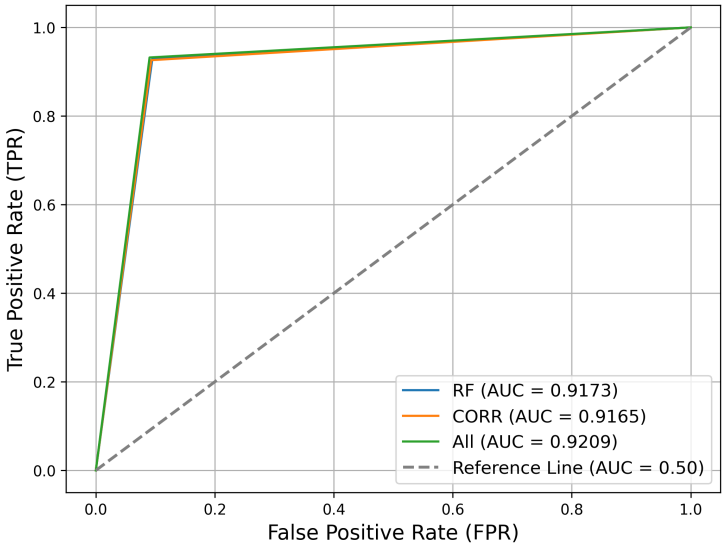


**Figure 6.** ROC-AUC metric for the Random Forest and correlation techniques compared to the one obtained using all features.

**Table 5.** Mean value of assessment metrics of result sets.

| | Accuracy | Precision | F-1 Score | Recall | ROC AUC |
|---|---|---|---|---|---|
| RF | 0.917 | 0.917 | 0.917 | 0.917 | 0.917 |
| Correlation | 0.916 | 0.916 | 0.916 | 0.916 | 0.917 |
| All | 0.921 | 0.921 | 0.921 | 0.921 | 0.921 |

Finally, Figure 7 presents the confusion matrix—a confusion matrix is a tabular representation typically used in supervised learning to visualise algorithm performance—where the rows represent actual labels, and the columns represent predicted labels generated by the analyzed techniques. Classes 0 and 1 correspond to background and signal, respectively.

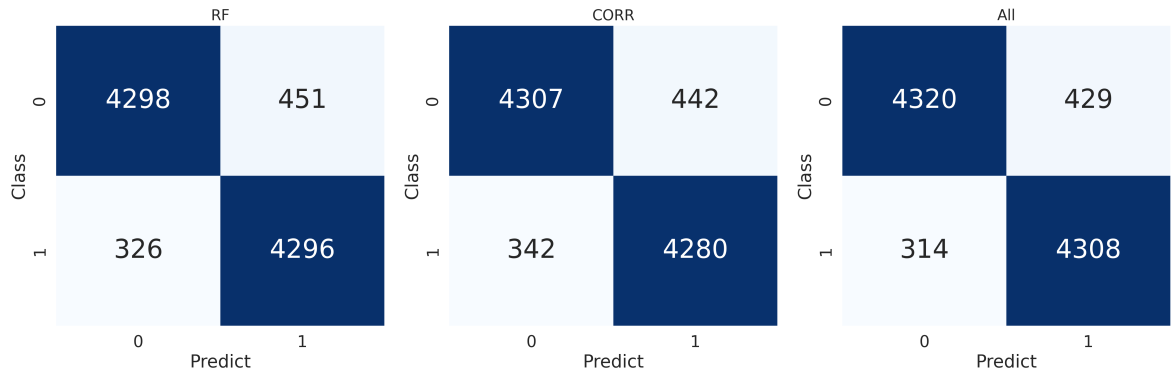


**Figure 7.** Confusion matrices for the Random Forest and correlation techniques compared to the one obtained using all features.

The comparison of different techniques reveals that the performance metrics obtained for the set of features including all the features available are marginally higher than the ones obtained for the other two methods. However, detailed analysis in Section 4.2 and Section 4.3 indicates that the set of variables selected by the Random Forest (RF) and correlation techniques do not exhibit statistical differences compared to the set composed by all features while resulting in lower CPU time. Therefore, employing the set of 75 variables selected by the RF technique can achieve statistically similar outcomes to using all 130 characteristics while enhancing model generalisation, reducing overfitting risks, slightly decreasing training time, and boosting computational efficiency.

### *4.5. Discussion on Performance of Random Forest Algorithm*

All the ML feature selection algorithms apart from the linear regression perform better than the physics-driven selection presented in Bueno et al. [15] both in terms of evaluation metrics (see Table 2) and in terms of statistical compatibility with the complete set of variables (see Table 3). In particular, the values of recall, precision, and F1-score in Table 2 demonstrate improvements in the classification of AMS-02 data compared to the physics-driven selection approach proposed in Bueno et al. [15]. The higher values scored for recall for the Kbest, Random Forest, and correlation methods with respect to the approach of Bueno et al. [15] should be interpreted as a more accurate classification of signal events (true positives), while the higher precision values mark a reduction in the false positives. The efficiency in the classification of signal and background events is confirmed in Figure 6, illustrating that a true positive rate (i.e., signal efficiency) of 92% is achieved with a false positive rate (i.e., background efficiency, denoted as $\epsilon_{bkg}$) of 0.1%. These results imply a background rejection, defined as $1 - \epsilon_{bkg}$, of 90%, showcasing the impact of the implemented methods on the data. This performance translates into an efficient separation of the signal and background events, allowing for a cleaner data sample and, hence, a more accurate mass reconstruction and identification of deuteron isotopes.

Since the Random Forest method is the best-performing selection technique among the ML algorithms compared in this work, the set of features selected by this method will be described in more detail. As shown in Figure 4 and in Table 1, all features in the "Charge" and "Track Position" classes were selected by the Random Forest algorithm. A fraction of 80% of the features was chosen for the "PMT number" class, and a similar percentage of 77.8% was selected from the "Photoelectrons" class. The classes "Beta" and "Hit number" had the lowest feature selection percentages, with 32.5% and 26.9%, respectively.

It is worth noting that the classes "Charge", "Photoelectrons", and "PMT number" present minor variability when the results of the different selection methods are compared, confirming that the power of separation of the variables belonging to these classes is independent of the underlying characteristics of a particular selection method and is closely linked to the physical phenomena underlying the RICH detection mechanism.

The features in the classes "Photoelectrons" and "PMT number" are related to the number of photons detected and used to reconstruct the ring. Because the signal detected in the PMT plane originates from the Cherenkov photons created by the cosmic ray particle that hits the radiator plane, the higher the number of photons, the less likely the signal will be disrupted by noise (Section 3.1). Therefore, variables from these two classes are expected to behave differently for background and signal events, resulting in a good classification performance. In particular, events with lower detected "Photoelectrons" or "PMT number" are more prone to be poorly reconstructed and are expected to be part of the background.

Variables belonging to the "Charge" class are indirectly influenced by the same processes. For example, the charge of the reconstructed particle is often higher for background events than for signal events (see Figure 3) due to the inclusion of additional hits in the reconstructed Cherenkov ring and the feature's distribution for background events is more shifted towards high charge values. Likewise, the Kolmogorov probability, which also belongs to the class "Charge", shows values for background events on average lower than for signal events since the charge distribution along the ring is not uniform (see Section 3.1).

Finally, the percentage of features selected for the "Track Position" class has higher variability, but it is still possible to trace its discrimination power back to the RICH structure. In particular, the impact point of the extrapolated tracker track on the radiator plane is a sensitive observable for this study, as also discussed in Bueno et al. [15]. Particles impacting some radiator areas (e.g., tile borders) tend to produce fewer detectable Cherenkov photons. They are, therefore, more likely to produce a weaker signal subject to incorrect reconstruction, making it possible to separate background and signal events.

## 5. Conclusions

An efficient reduction in the background consisting of events whose velocities are misreconstructed in the RICH detector is needed to identify positive singly charged cosmic ray isotopes with the AMS-02 detector. ML methods that can be used for this purpose, such as BDTs, are often trained on a set of features selected on the basis of the knowledge of the detector and of the classification task. However, ML algorithms can be used to perform automated feature selection improving the efficiency and accuracy of the analysis. In this paper, we applied automated feature selection methods to the background reduction analysis for the identification of cosmic ray deuterons with six months of data collected by the AMS-02 detector. The aim of the study was twofold: to choose among 130 variables associated with the RICH detector the best set of features to reject background events due to misrecontruction of the velocity measured by the RICH detector, and to assess the potential of these ML techniques in improving the background reduction efficiency in RICH compared to a traditional physics-driven approach. We used five feature selection algorithms widely used in the literature, namely Kbest, Random Forest, linear regression, and correlation, together with the method described in [15] which has the same scope but uses a set of physically motivated features. We used a boosted decision tree to perform the classification task and a K-fold cross-validation to validate our results.

We assessed the performance of the six methods with different evaluation metrics and found that the methods Kbest, Random Forest, and correlation outperform the approach described in Bueno et al. [15] in terms of accuracy, precision, F-1 score, and recall. Moreover, it is worth noting that only the set of variables selected by the Random Forest and correlation methods do not present statistical differences when compared to the complete set of 130 variables. Hence, the Random Forest method stands out as the best-performing algorithm, since it shows similar performance metrics compared to the complete set of 130 variables while reducing the risk of overfitting and training time and increasing the computational efficiency. The results obtained from the ML algorithms demonstrate that there is still some potential for improvement, which is crucial for deuteron identification due to the critical need to reduce the background composed by events with poorly reconstructed RICH velocity, as explained in Section 1. We conclude by investigating whether this ML method also maintains the connection between selected variables and the underlying physical phenomena related to RICH detection mechanisms, finding that the Random Forest algorithm maintains a meaningful link between selected variables and the physics of isotope identification.

Finally, the study focuses exclusively on feature selection for BDTs within the context of the AMS-02 experiment. This approach provides a direct and relevant comparison with the previous work of Bueno et al. [15]. However, feature selection techniques can also be applied to a broader spectrum of ML models. Therefore, future work will investigate their performance across different ML paradigms, thereby seeking to expand the results obtained.

**Author Contributions:** Conceptualisation, M.B., M.V. and L.M.; methodology, M.B., L.M. and M.V.; software, L.M.; validation, M.B. and M.V.; formal analysis, M.B. and L.M.; investigation, M.B., L.M. and M.V.; resources, M.V.; data curation, M.B. and L.M.; writing—original draft preparation, M.B. and L.M.; writing—review and editing, M.V. and F.B.; visualisation, M.B. and L.M.; supervision, M.V. and F.B.; project administration, M.V.; funding acquisition, M.V. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is part of the project "Cosmic ray antideuterons as a probe for new physics" with project number OCENW.KLEIN.387 (Budget Number 11680) of the research programme Grant Open Competition Domain Science, which is financed by the Dutch Research Council (NWO). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001 (Processo 300904/2023-1).

**Data Availability Statement:** The datasets presented in this article are not readily available because property of the AMS Collaboration. Requests to access the datasets should be directed to the ASM Collaboration.

**Acknowledgments:** We would like to thank Eduardo Bueno and Alberto Oliva for insightful discussion.

**Conflicts of Interest:** The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## Appendix A.

**Table A1.** List of the features used in the analysis.

| Class | Feature Description |
|---|---|
| Photoelectrons | Maximum number of p.e. in a PMT including or excluding crossed PMTs<br>Number of expected p.e. in the absence of a reconstructed Cherenkov ring<br>Number of p.e. collected inside the Cherenkov ring<br>Photoelectrons associated to the ring for different windows sizes<br>Expected number of p.e. for a Z = 1 Cherenkov ring with reconstruction and input parameters of the current event<br>Expected number of p.e. for a Z = 1 Cherenkov ring with reconstruction and input parameters of the current event and $\beta = 1$<br>Number of photons inside the Cherenkov ring<br>Number of photons outside the Cherenkov ring<br>Number of p.e. detected in the first 5 PMTs by number of p.e.<br>Total number of p.e. for $\beta = 1$ hypothesis<br>Total number of p.e. out of the ring for a particle with $\beta = 1$ |
| Charge | Reconstructed charge (CIEMAT reconstruction) [29]<br>Kolmogorov test of the distribution of charge along the ring<br>Statistical test to check if the hit-by-hit charge is consistent PMT-by-PMT<br>Expected charge resolution<br>Expected charge resolution RMS<br>Reconstructed charge (LIP reconstruction) [28] |
| PMT number | Number of crossed PMTs<br>Number of PMTs inside the Cherenkov ring<br>Expected number of PMTs for a Z = 1 Cherenkov ring with reconstruction and input parameters of the current event |
| Hit Number | Total number of hits<br>Number of hits in the first 5 PMTs by number of p.e.<br>Total number of hits compatible with $\beta = 1$ hypothesis (direct and reflected)<br>Number of hits out of the Cherenkov ring compatible with $\beta = 1$ hypothesis (direct and reflected)<br>Number of hits inside the Cherenkov ring<br>Number of hits which are consistent with reflected photons<br>Distribution for unused hits which do not belong to the PMTs crossed by a charged particle<br>Number of hits outside the Cherenkov ring |
| Track Position | Distance from the tracker track [cm] for the first 5 PMTs by number of p.e.<br>PMT number for the first 5 PMTs by number of p.e.<br>Tile ID for the tile crossed by the particle<br>Impact point of the tracker track to the radiator entrance [cm]<br>Theta of the tracker track to the radiator entrance [rad]<br>Phi of the tracker track to the radiator entrance [rad]<br>Distance of the tracker track impact point on the radiator to the border of the radiator tile |

**Table A1.** *Cont.*

| Class | Feature Description |
|---|---|
| Beta | Reconstructed particle velocity $\beta$ (CIEMAT reconstruction) [29]<br>Expected velocity resolution<br>Expected resolution RMS<br>Raw $\beta$<br>Refit $\beta$<br>$\beta$ corrected for impact point and direction<br>Clusters of signal used for the reconstruction of the (particle) velocity $\beta$.<br>Size of first 10 clusters<br>Average beta of first 10 clusters<br>RMS of first 10 clusters<br>Reconstructed particle velocity $\beta$ (LIP reconstruction) [28]<br>$\beta$ estimated from rigidity |

**Table A2.** Parameters used in each feature selection technique used in this work, together with the corresponding Python library/method used to implement them.

| Techniques | Parameters | Library/Method |
|---|---|---|
| Kbest | score_fun=f_classif, k=10, X=sparse matrix of shape (n_samples, n_features), y=array-like of shape (n_samples,) | sklearn.feature_selection.SelectKBest |
| RF | n_estimators=100, criterion=gini, max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weigh_fraction_leaf=0.0, max_features=sqrt, max_leaf_nodes=None, min_impurity_decrease=0.0, bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None, monotonic_cst=None | sklearn.ensemble.RandomForestClassifier |
| Linear | estimator=LinearRegression(), fit_intercept=True, copy_X=True, positive=False, step=1, min_features_to_select=1, cv=10, scoring=None, verbose=0, n_jobs=None, importance_getter=auto | sklearn.feature_selection.RFECV |
| Correlation | method=pearson, min_periods=None, numeric_only=False, threshold=0.005 | pandas.DataFrame.corr |

**Table A3.** Uncertainties on assessment metrics for the selected features.

| | Accuracy | Precision | F-1 Score | Recall | ROC AUC |
|---|---|---|---|---|---|
| Kbest | 0.004137 | 0.005627 | 0.004245 | 0.008718 | 0.003883 |
| RF | 0.006519 | 0.005536 | 0.006671 | 0.010665 | 0.003592 |
| Linear | 0.008393 | 0.005494 | 0.006611 | 0.010761 | 0.008763 |
| Correlation | 0.006291 | 0.007005 | 0.006229 | 0.008382 | 0.003002 |
| Bueno et al. | 0.005628 | 0.007487 | 0.005734 | 0.010649 | 0.005049 |
| All | 0.006056 | 0.005169 | 0.006127 | 0.008921 | 0.003326 |

## References

1. Gaisser, T.K.; Engel, R.; Resconi, E. *Cosmic Rays and Particle Physics*, 2nd ed.; Cambridge University Press: Cambridge, MA, USA, 2016. [CrossRef]
2. Coste, B.; Derome, L.; Maurin, D.; Putze, A. Constraining Galactic cosmic-ray parameters with $Z \leq 2$ nuclei. *Astron. Astrophys.* **2012**, *539*, A88. [CrossRef]

3. Adelberger, E.G.; García, A.; Robertson, R.G.H.; Snover, K.A.; Balantekin, A.B.; Heeger, K.; Ramsey-Musolf, M.J.; Bemmerer, D.; Junghans, A.; Bertulani, C.A.; et al. Solar fusion cross sections. II. The *pp* chain and CNO cycles. *Rev. Mod. Phys.* **2011**, *83*, 195–245. [CrossRef]
4. Adriani, O.; Barbarino, G.C.; Bazilevskaya, G.A.; Bellotti, R.; Boezio, M.; Bogomolov, E.A.; Bongi, M.; Bonvicini, V.; Bottai, S.; Bruno, A.; et al. Measurements of Cosmic-Ray Hydrogen and Helium Isotopes with the PAMELA experiment. *Astrophys. J.* **2016**, *818*, 68. [CrossRef]
5. de Nolfo, G.A.; Barbier, L.M.; Christian, E.R.; Davis, A.J.; Golden, R.L.; Hof, M.; Krombel, K.E.; Labrador, A.W.; Menn, W.; Mewaldt, R.A.; et al. A measurement of cosmic ray deuterium from 0.5–2.9 GeV/nucleon. *AIP Conf. Proc.* **2000**, *528*, 425–428. [CrossRef]
6. Papini, P.; Piccardi, S.; Spillantini, P.; Vannuccini, E.; Ambriola, M.; Bellotti, R.; Cafagna, F.; Ciacio, F.; Circella, M.; De Marzo, C.N.; et al. High-energy deuteron measurement with the CAPRICE98 experiment. *Astrophys. J.* **2004**, *615*, 259–274. [CrossRef]
7. Aguilar, M.; Cavasonza, L.A.; Ambrosi, G.; Arruda, L.; Attig, N.; Barao, F.; Barrin, L.; Bartoloni, A.; Başeğmez-du Pree, S.; Bates, J.; et al. The Alpha Magnetic Spectrometer (AMS) on the international space station: Part II—Results from the first seven years. *Phys. Rep.* **2021**, *894*, 1–116. [CrossRef]
8. Delgado, C. Precision Measurement of Cosmic Ray Deuterons with Alpha Magnetic Spectrometer. In Proceedings of the 38th International Cosmic Ray Conference—PoS(ICRC2023), Nagoya, Japan, 26 July–3 August 2023; Volume 444, p. 079. [CrossRef]
9. Yang, H.J.; Roe, B.P.; Zhu, J. Studies of boosted decision trees for MiniBooNE particle identification. *Nucl. Instrum. Methods Phys. Res. Sect. A Accel. Spectrometers Detect. Assoc. Equip.* **2005**, *555*, 370–385. [CrossRef]
10. Graziani, G.; Anderlini, L.; Mariani, S.; Franzoso, E.; Pappalardo, L.L.; di Nezza, P. A Neural-Network-defined Gaussian Mixture Model for particle identification applied to the LHCb fixed-target programme. *J. Instrum.* **2022**, *17*, P02018. [CrossRef]
11. Aguilar, M.; Aisa, D.; Alpat, B.; Alvino, A.; Ambrosi, G.; Andeen, K.; Arruda, L.; Attig, N.; Azzarello, P.; Bachlechner, A.; et al. Precision Measurement of the $(e^+ + e^-)$ Flux in Primary Cosmic Rays from 0.5 GeV to 1 TeV with the Alpha Magnetic Spectrometer on the International Space Station. *Phys. Rev. Lett.* **2014**, *113*, 221102. [CrossRef]
12. Accardo, L.; Aguilar, M.; Aisa, D.; Alpat, B.; Alvino, A.; Ambrosi, G.; Andeen, K.; Arruda, L.; Attig, N.; Azzarello, P.; et al. 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.* **2014**, *113*, 121101. [CrossRef]
13. Aguilar, M.; Alberti, G.; Alpat, B.; Alvino, A.; Ambrosi, G.; Andeen, K.; Anderhub, H.; Anderhub, L.; Azzarello, P.; Bachlechner, A.; et al. First Result from the Alpha Magnetic Spectrometer on the International Space Station: Precision Measurement of the Positron Fraction in Primary Cosmic Rays of 0.5–350 GeV. *Phys. Rev. Lett.* **2013**, *110*, 141102. [CrossRef] [PubMed]
14. Graziani, M. Electron/proton separation and analysis techniques used in the AMS-02 $(e^+ + e^-)$ flux measurement. In Proceedings of the 37th International Conference on High Energy Physics (ICHEP), San Francisco, CA, USA, 10–14 October 2016; pp. 2351–2353. [CrossRef]
15. Bueno, E.F.; Barão, F.; Vecchi, M. Machine learning approach to the background reduction in singly charged cosmic-ray isotope measurements with AMS-02. *Nucl. Instrum. Meth. A* **2023**, *1056*, 168644. [CrossRef]
16. Di Luca, A.; Cristoforetti, M.; Follega, F.M.; Iuppa, R.; Mascione, D. Automated feature selection procedure for particle jet classification. *Nucl. Phys. B* **2023**, *990*, 116182. [CrossRef]
17. Finke, T.; Krämer, M.; Manconi, S. Classification of Fermi-LAT sources with deep learning using energy and time spectra. *Mon. Not. R. Astron. Soc.* **2021**, *507*, 4061–4073. [CrossRef]
18. Gavrikov, A.; Ratnikov, F. The use of Boosted Decision Trees for Energy Reconstruction in JUNO experiment. *EPJ Web Conf.* **2021**, *251*, 03014. [CrossRef]
19. Luo, S.; Leung, A.P.; Hui, C.Y.; Li, K.L. An investigation on the factors affecting machine learning classifications in gamma-ray astronomy. *Mon. Not. R. Astron. Soc.* **2020**, *492*, 5377–5390. [CrossRef]
20. Herrera, L.J.; Peixoto, C.J.T.; Baños, O.; Carceller, J.M.; Carrillo, F.; Guillén, A. Composition Classification of Ultra-High Energy Cosmic Rays. *Entropy* **2020**, *22*, 998. [CrossRef] [PubMed]
21. Giovacchini, F.; Casaus, J.; Oliva, A. The AMS-02 RICH detector: Status and physics results. *Nucl. Instrum. Methods Phys. Res. Sect. A Accel. Spectrometers Detect. Assoc. Equip.* **2020**, *952*, 161797. [CrossRef]
22. Bueno, E.; Barão, F.; Vecchi, M. A parametric approach for the identification of single-charged isotopes with AMS-02. *Nucl. Instrum. Methods Phys. Res. Sect. A Accel. Spectrometers Detect. Assoc. Equip.* **2022**, *1031*, 166564. [CrossRef]
23. Jackson, J.D. *Classical Electrodynamics*; Wiley: Hoboken, NJ, USA, 1998.
24. Arruda, L.; Barao, F.; Goncalves, P.; Pereira, R. The Ring Imaging Cherenkov detector of the AMS experiment: Test beam results with a prototype. *Nucl. Phys. B Proc. Suppl.* **2007**, *172*, 32–35. [CrossRef]
25. Aguilar-Benitez, M.; Arruda, L.; Barao, F.; Barreira, G.; Barrau, A.; Baret, B.; Berdugo, J.; Buénerd, M.; Casaus, J.; Cernuda, I.; et al. In-beam aerogel light yield characterization for the AMS RICH detector. *Nucl. Instrum. Meth. A* **2010**, *614*, 237–249. [CrossRef]
26. Giovacchini, F. Space application: The AMS RICH. *Nucl. Instrum. Meth. A* **2020**, *970*, 163657. [CrossRef]
27. Pereira, R.; AMS RICH Collaboration. The AMS-02 RICH detector: Performance during ground-based data taking at CERN. *Nucl. Instrum. Methods Phys. Res. A* **2011**, *639*, 37–41. [CrossRef]
28. Barao, F.; Aguilar-Benitez, M.; Arruda, L.; Baret, B.; Barrau, A.; Barreira, G.; Belmont, E.; Berdugo, J.; Borges, J.; Buenerd, M.; et al. The AMS-RICH velocity and charge reconstruction. In Proceedings of the 30th International Cosmic Ray Conference, Yucatan, Mexico, 3–7 July 2007; Volume 2, pp. 457–460.

29. Delgado Mendez, C.J. Medida de la velocidad de muones y nucleos ligeros con un prototipo del contador RICH del experimento AMS. Ph.D. Thesis, Universidad Autonoma de Madrid, Madrid, Spain, 2003.
30. Eadie, W.T.; Drijard, D.; James, F.E. *Statistical Methods in Experimental Physics*; World Scientific Publishing Company: Hong Kong, China, 1971.
31. Bommert, A.; Sun, X.; Bischl, B.; Rahnenführer, J.; Lang, M. Benchmark for filter methods for feature selection in high-dimensional classification data. *Comput. Stat. Data Anal.* **2020**, *143*, 106839. [CrossRef]
32. Dvornik, N.; Schmid, C.; Mairal, J. Selecting relevant features from a multi-domain representation for few-shot classification. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Part X 16, Glasgow, UK, 23–28 August 2020; pp. 769–786.
33. Han, T.; Li, Y.F.; Qian, M. A hybrid generalization network for intelligent fault diagnosis of rotating machinery under unseen working conditions. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–11. [CrossRef]
34. Jia, W.; Sun, M.; Lian, J.; Hou, S. Feature dimensionality reduction: A review. *Complex Intell. Syst.* **2022**, *8*, 2663–2693. [CrossRef]
35. Qian, W.; Huang, J.; Xu, F.; Shu, W.; Ding, W. A survey on multi-label feature selection from perspectives of label fusion. *Inf. Fusion* **2023**, *100*, 101948. [CrossRef]
36. Chen, C.W.; Tsai, Y.H.; Chang, F.R.; Lin, W.C. Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. *Expert Syst.* **2020**, *37*, e12553. [CrossRef]
37. Effrosynidis, D.; Arampatzis, A. An evaluation of feature selection methods for environmental data. *Ecol. Inform.* **2021**, *61*, 101224. [CrossRef]
38. Liu, C.H.; Tsai, C.F.; Sue, K.L.; Huang, M.W. The feature selection effect on missing value imputation of medical datasets. *Appl. Sci.* **2020**, *10*, 2344. [CrossRef]
39. Alves, V.; Cury, A. An automated vibration-based structural damage localization strategy using filter-type feature selection. *Mech. Syst. Signal Process.* **2023**, *190*, 110145. [CrossRef]
40. Genuer, R.; Poggi, J.M.; Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognit. Lett.* **2010**, *31*, 2225–2236. [CrossRef]
41. Polat, H.; Polat, O.; Cetin, A. Detecting DDoS attacks in software-defined networks through feature selection methods and machine learning models. *Sustainability* **2020**, *12*, 1035. [CrossRef]
42. Seeram, E. An overview of correlational research. *Radiol. Technol.* **2019**, *91*, 176–179.
43. Liu, Y.; Mu, Y.; Chen, K.; Li, Y.; Guo, J. Daily activity feature selection in smart homes based on pearson correlation coefficient. *Neural Process. Lett.* **2020**, *51*, 1771–1787. [CrossRef]
44. Hossin, M.; Sulaiman, M.N. A review on evaluation metrics for data classification evaluations. *Int. J. Data Min. Knowl. Manag. Process* **2015**, *5*, 1.
45. Yacouby, R.; Axman, D. Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, Online, 20 November 2020; pp. 79–91.
46. Lemaître, G.; Nogueira, F.; Aridas, C.K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J. Mach. Learn. Res.* **2017**, *18*, 1–5.
47. Mano, L.Y. Emotional condition in the Health Smart Homes environment: Emotion recognition using ensemble of classifiers. In Proceedings of the 2018 Innovations in Intelligent Systems and Applications (INISTA), Thessaloniki, Greece, 3–5 July 2018.
48. Mano, L.Y.; Faiçal, B.S.; Gonçalves, V.P.; Pessin, G.; Gomes, P.H.; de Carvalho, A.C.; Ueyama, J. An intelligent and generic approach for detecting human emotions: A case study with facial expressions. *Soft Comput.* **2020**, *24*, 8467–8479. [CrossRef]