# PID in the NICA experiment using machine learning techniques

MSc. Julio César Maldonado González
Co-authors: Ph.D. Isabel Domínguez Jiménez, Ph.D. Pedro Luis Manuel Podesta Lerma, Ph.D. Ines Fernando Vega López

Universidad Autónoma de Sinaloa
MexNICA Collaboration

**The XXVI International Scientific Conference of Young Scientists and Specialists (AYSS-2022)**

October 25, 2022

# Overview

Signal $dE/dx$ vs $p$ in the MPD-TPC detector for Bi-Bi collisions of $\sqrt{s} = 11$ GeV,

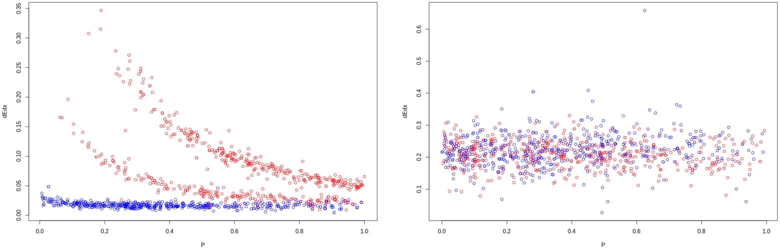



Figure: Protons, Kaons (red), and Pions (blue), for two normalized datasets of 1000 tracks, for low and high momentum.

# Methodology

- Preparing data for three balanced classes for training and testing.
- Ranking and selecting variables (features).
- Analyzing data using MLP (RSNNS), DT (C5.0) and SVM (e1071) classifiers.
- Comparing methods with Bayesian method using confusion matrix

# Simulation and reconstruction data

- Detector: MPD (TPC, TOF)
- Input file: rectestf140.root file (DST)
- Event generator: UrQMD
- Bi-Bi a 11 GeV (MB)
- of events: 10k
- Macro: CompareSpectra.C, anaDST.C

# Data

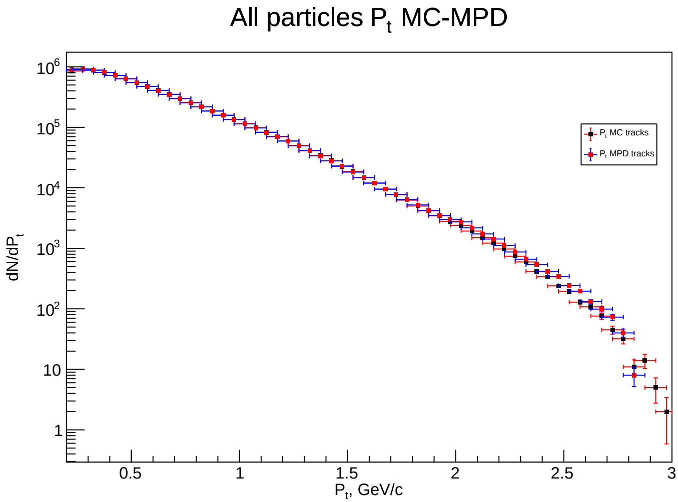



Figure: $P_T$ distribution for MC and reconstruction tracks.

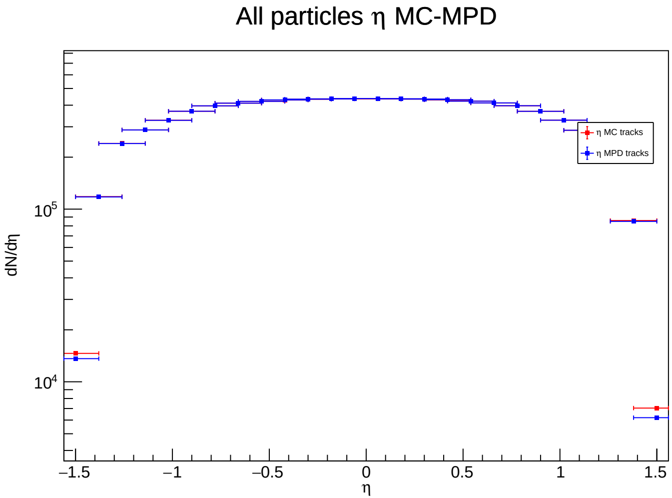



Figure: $\eta$ distribution for MC and reconstruction tracks.

- Training dataset MPDTRACKSTR of 30k elements (tracks) selecting three classes (particles): Protons, pions, kaons.
- Testing dataset MPDTRACKSTE of 15k elements (tracks).

# Feature selection

From dataset we have a feature list which describes the reconstruction tracks.

| | |
|---|---|
| ID | Integer |
| TofBeta | Float |
| TofMass2 | Float |
| dEdx | Float |
| TofHitIndex | Entero |
| TofFlag | Float |
| Chi2 | Float |
| Pt | Float |
| Theta | Float |
| Phi | Float |
| Eta | Float |
| Charge | Float |
| P | Float |
| PDGID | Integer |

Using TMVA libraries, we ranked each feature computing the separation, $\langle S^2 \rangle$,

$$\langle S^2 \rangle = \frac{1}{2} \int \frac{(y_S(y) - y_B(y))^2}{y_S(y) + y_B(y)} dy$$

$y_S$ and $y_B$ are the PDFs of $y$ for signal and background.

| Rank | Feature | Separation |
|---|---|---|
| 1 | Charge | 3.379e-01 |
| 2 | dEdx | 1.480e-02 |
| 3 | Phi | 1.398e-02 |
| 4 | FirstPointX | 1.334e-02 |
| 5 | FirstPointY | 1.201e-02 |
| 6 | P | 1.184e-02 |
| 7 | DCAGlobalX | 1.146e-02 |
| 8 | DCAX | 1.130e-02 |
| 9 | DCAGlobalY | 1.087e-02 |
| 10 | DCAY | 1.060e-02 |
| 11 | PhiError | 1.045e-02 |
| 12 | PtError | 9.873e-03 |
| 13 | TofBeta | 8.765e-03 |
| 14 | ThetaError | 8.711e-03 |
| 15 | Pt | 8.297e-03 |

# Bayesian method

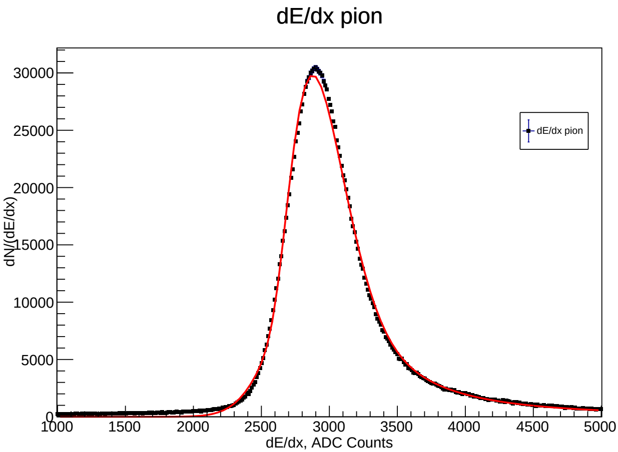
The probability of a particle $i$, if $s$ signal is observed,

$$P(i|s) = \frac{r(s|i)C_i}{\sum_k r(s|k)C_k}$$

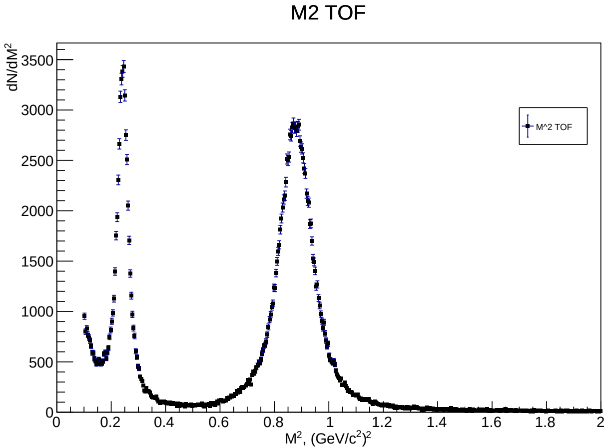
$r(s|i)$ probability density function of the observed signal $s$ of the detector, if a particle $i(e, \mu, \pi, K, p, ...)$ is detected.

$C_i$ frecuency of the observed particle.

# $dE/dx$ histogram and fit



# Frequency of the particle



Output format for a Neural Network of three classes:

| 0 | 1 | 2 | Partícula |
|---|---|---|---|
| 0 | 0 | 1 | $K^{\pm}$ |
| 0 | 1 | 0 | $\pi^{\pm}$ |
| 1 | 0 | 0 | $p$ |

Using the function `multiclass`,

- ROC probability for a proton = 0.9678
- ROC probability for a pion = 0.9581
- ROC probability for a kaon = 0.8866

# Confusion matrix for the three classes model

Selecting a threshold value fixed for the compare probability, we compute the True Positive, False Negative, True Negative and False Positive percentage

| | TP % | FN % | TN % | FP % |
|---|---|---|---|---|
| $p$ | 96.22 | 3.78 | 98.18 | 1.82 |
| $\pi^{\pm}$ | 90.38 | 9.62 | 94.89 | 5.11 |
| $K^{\pm}$ | 85.72 | 14.28 | 95.15 | 4.85 |

Table: MLP model prediction

# Bayesian Method, Multi-Layer Perceptron, Decision Tree and Support Vector Machine comparison

Selecting the dataset for $0.2 \leq P < 1.0$

| | Bayes | MLP | DT | SVM |
|---|---|---|---|---|
| TP | 83.84 % | 99.18 % | 99.14 % | 98.38 % |
| FN | 16.16 % | 0.82 % | 0.86 % | 1.62 % |
| TN | 95.08 % | 99.52 % | 99.46 % | 99.44 % |
| FP | 4.92 % | 0.48 % | 0.54 % | 0.56 % |

Selecting the dataset for $1.8 \leq P < 2.6$

| | Bayes | MLP | DT | SVM |
|---|---|---|---|---|
| TP | 0.8 % | 90.38 % | 89.54 % | 93.18 % |
| FN | 99.2 % | 9.62 % | 10.46 % | 6.82 % |
| TN | 99.37 % | 94.89 % | 94.63 % | 93.24 % |
| FP | 0.63 % | 5.11 % | 5.37 % | 6.76 % |

Table: Results for $\pi$ classification.

# Conclusions

- Bayesian method is only useful in the region of low-momentum ($1.8 \leq P < 2.6$).
- Confusion matrix and ROC computation show machine learning techniques have acceptable results in both regions, of low-momentum and high-momentum of the Bethe-Bloch distribution.
- Machine learning techinques allow us to implement a data classification for more than binomial classification (signal and background) and for more than two features ($P$ and $dEdx$).

Thanks!