Machine learning based TOF charged particle identification at BM@N detector of NICA collider

V A Roudnev¹, SP Merts², S A Nemnyugin¹ and M M Stepanova¹

¹ St-Petersburg State University
² Joint Institute for Nuclear Research

E-mail: v.rudnev@spbu.ru

Abstract. In the article results of charged particles identification for BM@N experiment being performed at NICA acceleration complex of Joint Institute for Nuclear Research are presented. A standard neural network-based technique of constructing a classifier is applied to the data sets obtained both from modelling of a realistic experimental setup and three synthetic data sets. The carried-out analysis demonstrates that the estimated data accuracy is insufficient to make a clear distinction between electrons, muons and pions, and also between α-particles and deuterons. The problem could be solved by using an extra data from the detector or by improving the accuracy of the experimental data by two orders of magnitude.

1. Introduction
NICA (Nuclotron-based Ion Collider fAcility) is a collider complex being built at Joint Institute for Nuclear Research in Dubna (figure 1). The main goal of NICA research program is to study dense baryonic matter properties. The accelerator complex includes three detectors which are at different stages of development: MPD (Multi-Purpose Detector), SPD (Spin Physics Detector) and BM@N (Baryonic Matter at Nuclotron).

The beginning of NICA exploitation is scheduled for 2021 and the whole NICA acceleartor complex is expected to start functioning on a regular basis in 2023 [1].

BM@N experiment is aimed at studying heavy ion collisions with fixed targets. The experiment assumes precise measurements of particle trajectories and TOF (Time Of Flight) measurements for particle identification. The BM@N scheme is presented in figure 2. The detailed description of the BM@N detector can be found in the technical report [2]. The first technical runs of the BM@N that have been aimed at checking and tuning the equipment were performed in 2015–2018. The central BM@N tracker is based on silicon micro-strip detectors and three-layer gas electron multipliers (GEMs) [3] placed after the target inside an analysing dipole magnet, and drift chambers (MWPC) are placed next to the magnet. These detectors provide data for reconstruction of the particles trajectories. Two TOF detectors for identification of light nuclei, protons, π- and K-meson of momentum up to a few GeV/c are placed down the beam. Identification of these particles is the problem we discuss in this work.

2. Data models and the evaluation procedure
Particle identification based on trajectory geometry and TOF data analysis [4] has been performed using a standard neural network-based machine learning technique [6]. The features
used for identification are the particle momentum to charge ratio (rigidity) and the mass to charge ratio (figure 3) which we extract from the particle pass through the detector data modeled with BmnRoot software package [5]. The identification is performed by a single hidden layer feed-forward fully connected neural network. The input layer is fed with the rigidity and mass to charge ratios, the output layer consists of nine softmax elements that correspond to the probabilities of identifying the particle as one of the nine types: $e^\pm$, $\mu^\pm$, $\pi^\pm$, $K^\pm$, $p$, $d$, $t$, $^3\text{He}^{++}$, $^4\text{He}^{++}$. Categorical cross-entropy is taken for the loss function.

Neural networks have been trained using several data sets. The first data set has been generated using a realistic modeling of the experimental setup with BmnRoot package. The rest of the data sets has been synthesized to mimic some statistical properties of the first data set while varying the dispersion.
In order to evaluate the quality of particle identification we have used the identification efficiency

$$E = \frac{N_{id}^+}{N_{data}},$$

where $N_{id}^+$ is the number of correctly identified particles of the given class, $N_{data}$ is the number of particles of the given class in the data set. The second characteristic that we use is the degree of contamination

$$C = \frac{N_{id}^-}{N_{predict}},$$

where $N_{id}^-$ is the number of the given class particles that were identified incorrectly, and $N_{predict}$ is the number of particles that were identified as belonging to the given class.

2.1. Modeled data

The first data set has been generated with the BmnRoot package so that only the particles of a single given type remained in each of the subsets. This allowed us to use the data for neural network training and evaluation.

The data set of 103214 particles has been filtered to exclude tracking errors. For instance, the particles with mass to charge ratio squared exceeding 12 and the particles of superluminal velocity have been excluded. After filtering the set contained 88140 particles. One half of the data has been used as a training set, the rest of the data has been used for testing. Only 77.9 % of the particles in the training set and 77.4 % of the testing set have been classified correctly. This low value is dealt with insufficient accuracy of light particle trajectory parameters extraction ($e^\pm$, \ldots).
$\mu^\pm, \pi^\pm$) in the whole region of rigidity from 0.5 to 3.5 GeV/c/e as well as practical coincidence of the mass to charge ratio for deuteron and $\alpha$-particles.

In figure 4 we show identification quality estimates for electrons, muons and pions. The pion identification efficiency reaches its maximum about 80% at low momentum and falls off rapidly as the particle momentum grows, the degree of contamination also rises with momentum. Does this low quality of the particle identification reflect a deficiency of the algorithm or the property of the input data? The light particle trajectories in the given momentum range are close to straight lines, what makes the evaluation of the momentum on the base of the trajectory curvature difficult and leads to high dispersion for the mass to charge ratio. This can be seen in figure 3, where the light particle data practically merge into a single cluster. So we can infer that reliable identification of light particles solely on the base of TOF analysis is not possible.

![Figure 4](image-url)

**Figure 4.** Identification efficiency and contamination for a) electrons; b) muons; c) pions.
Figure 5 illustrates the quality of identification for deuterons and α-particles. For other light nuclei – such as protons, $^3\text{He}^{++}$ and $^3\text{H}^+$ – that are separated well (see figure 3) the identification efficiency remain close to 100 % with rather low contamination for the whole range of momenta. Distinction of deuterons from α-particles, however, appears a difficult problem. Unlike the light particles, the differentiation between these light nuclei is complicated not by big data inaccuracies only, but also by a very small difference of the mass to charge ratio. This leads to over-contamination (up to 4000 %) of the deuteron channel from α-particles: all the particles with rigidity 1.5 GeV/c/e and lower are identified as α-particles.

![Figure 5](image-url)

**Figure 5.** Identification efficiency and contamination for a) α-particles and b) deuterons.

In order to evaluate the resolution required for reliable identification of the charged particles we have repeated the procedure with synthetic data sets.

### 2.2. Synthetic data

The synthetic data set of 4096 particles of each type has been produced with normal distribution. The mean values of the mass to charge ratio distribution for each type of the particles has been taken form the modelled data. The smooth nearly linear dependence of the dispersion on the particle momentum has been neglected. The dispersion of the mass to charge ratio for each type of the particle has been chosen to achieve a clear separation from data corresponding to the nearest type of the particle in the data set.

In order to characterize the quality of particle identification which can be expected for a given data dispersion we have generated three synthetic sets of data. In the first data set the dispersion for each particle type has been taken as $1/5$ of the distance between the mean mass/charge ratio for the given particle type and the mean value for the closest particle type. This data set has been used for training the neural network. The other two data sets were generated similarly to
the first one, but the dispersions have been chosen as 1/2 and 1 of the corresponding distances. The corresponding parameters of the data sets and mean values of identification efficiency and degree of contamination are reported in table 1.

Table 1. Distribution properties for all the data sets and particle identification quality characteristics.

|                  | Synthetic data set 1 | Synthetic data set 2 | Synthetic data set 3 | Modeled data set |
|------------------|-----------------------|-----------------------|-----------------------|-----------------|
|                  | $e^\pm$ | $\mu^\pm$ | $\pi^\pm$ | $K^\pm$ | $p$ | $d$ | $t$ | $^{3}\text{He}^{++}$ | $^{4}\text{He}^{++}$ |
| $\langle (M/z)^2 \rangle$, ($GeV/e^2/c^2$) | 0.004 | 0.016 | 0.025 | 0.249 | 0.891 | 3.537 | 7.946 | 2.002 | 3.548 |
| $\sigma([M/z]^2)$, ($GeV/e^2/c^2$) | 0.003 | 0.002 | 0.002 | 0.045 | 0.128 | 0.002 | 0.880 | 0.222 | 0.002 |
| Efficiency | 0.993 | 0.988 | 0.988 | 0.993 | 0.993 | 0.988 | 0.994 | 0.993 | 0.987 |
| Contamination | 0.001 | 0.006 | 0.006 | 0.000 | 0.001 | 0.006 | 0.000 | 0.001 | 0.006 |
|                  | $e^\pm$ | $\mu^\pm$ | $\pi^\pm$ | $K^\pm$ | $p$ | $d$ | $t$ | $^{3}\text{He}^{++}$ | $^{4}\text{He}^{++}$ |
| $\langle (M/z)^2 \rangle$, ($GeV/e^2/c^2$) | 0.004 | 0.016 | 0.025 | 0.249 | 0.891 | 3.537 | 7.946 | 2.002 | 3.548 |
| $\sigma([M/z]^2)$, ($GeV/e^2/c^2$) | 0.004 | 0.003 | 0.003 | 0.075 | 0.214 | 0.003 | 1.466 | 0.371 | 0.003 |
| Efficiency | 0.952 | 0.907 | 0.916 | 0.955 | 0.926 | 0.987 | 0.959 | 0.934 |
| Contamination | 0.042 | 0.085 | 0.079 | 0.012 | 0.039 | 0.069 | 0.006 | 0.035 | 0.060 |
|                  | $e^\pm$ | $\mu^\pm$ | $\pi^\pm$ | $K^\pm$ | $p$ | $d$ | $t$ | $^{3}\text{He}^{++}$ | $^{4}\text{He}^{++}$ |
| $\langle (M/z)^2 \rangle$, ($GeV/e^2/c^2$) | 0.006 | 0.004 | 0.004 | 0.112 | 0.321 | 0.005 | 2.199 | 0.556 | 0.005 |
| $\sigma([M/z]^2)$, ($GeV/e^2/c^2$) | 0.006 | 0.004 | 0.004 | 0.112 | 0.321 | 0.005 | 2.199 | 0.556 | 0.005 |
| Efficiency | 0.866 | 0.734 | 0.785 | 0.906 | 0.832 | 0.816 | 0.949 | 0.878 | 0.839 |
| Contamination | 0.128 | 0.254 | 0.221 | 0.086 | 0.160 | 0.184 | 0.047 | 0.114 | 0.151 |
|                  | $e^\pm$ | $\mu^\pm$ | $\pi^\pm$ | $K^\pm$ | $p$ | $d$ | $t$ | $^{3}\text{He}^{++}$ | $^{4}\text{He}^{++}$ |
| $\langle (M/z)^2 \rangle$, ($GeV/e^2/c^2$) | 0.005 | 0.018 | 0.017 | 0.093 | 0.197 | 0.006 | 0.991 | 0.951 | 0.451 |
| $\sigma([M/z]^2)$, ($GeV/e^2/c^2$) | 0.059 | 0.060 | 0.092 | 0.078 | 0.123 | 0.166 | 0.249 | 0.064 | 0.136 |
| Efficiency | 0.830 | 0.198 | 0.197 | 0.975 | 0.993 | 0.067 | 0.991 | 0.951 | 0.451 |
| Contamination | 1.335 | 2.452 | 2.239 | 0.004 | 0.001 | 12.50 | 0.000 | 0.000 | 0.546 |

3. Discussion and conclusion

Comparing dispersion for the modeled and synthetic data sets (table 1) we observe that in order to separate the signals for light particles reliably, the accuracy for the observables should be improved by two orders of magnitude. The same two orders of magnitude resolution improvement is required to distinguish $\alpha$-particles from deuterons. Another problem which we should mention concerns identification of kaons. In figure 6 we see, that even though the efficiency of kaon identification is reasonably high for the particles of lower energy, the higher energy particles are not identified that efficiently due to the growing contamination from the lighter particles. This suggests that the light particle identification is also crucial for correct identification of kaons, which is crucial for the success of the experiment.

As is mentioned above, our analysis suggests that improvements of the tracking procedure, and, possibly, of the detector resolution are needed. Further error analysis is needed to identify the most sensitive parts of the data processing procedure. The need for modifications of the detector hardware could not be entirely excluded, and two orders of magnitude improvements in the hardware resolution might be achieved by using pixel detectors [7].

Another way to improve the quality of particle identification without potentially expensive improvements of the data accuracy can be sought for in adding independent experimental
parameters into the data set. For instance, energy loss in GEM planes could contribute to better separation between deuterons and α-particles.

4. Acknowledgements
This work is supported by Russian Foundation for Basic Research grant 18-02-40104 mega.

References
[1] Official site of NICA accelerator complex (http://nica.jinr.ru)
[2] Afanasiev S et al. 2016 BM@N Project report (http://bmnshift.jinr.ru/wiki/lib/exe/fetch.php?media=bmnproject_2016.pdf)
[3] Baranov D et. al 2017 J. Instrum. 12 06041
[4] Stroynowski R 1989 Time-of-flight technique Proc. Study of tau, charm and J/psi physics. (Stanford) 956
[5] Baranov D, Kapishin M, Mamontova T, Pokatashkin G, Rusanov I, Vasendina V and Zinchenko A 2018 KnE Energ. Phys. 3 291
[6] TensorFlow documentation. (www.tensorflow.org/tutorials/keras/classification)
[7] Feofilov G 2016 Journal of Physics: Conference Series 668 012012

Figure 6. Identification efficiency and contamination for kaons.