JOINT INSTITUTE FOR NUCLEAR RESEARCH
Veksler and Baldin laboratory of High Energy Physics

FINAL REPORT ON THE START PROGRAMME

Machine Learning Approach for Particle Identification in the MPD Experiment

Supervisor:
Dr. Alexey Aparin

Student:
Volobueva Diana, Russia
MIREA — Russian Technological University

Participation period:
July 07 – September 06,
Summer Session 2025

Dubna, 2022

Contents

Abstract **3**

1 Introduction **3**

1.1 Time Projection Chamber 4

1.2 Time of Flight System 5

2 Current particle identification methods for MPD **6**

2.1 Physical principles of PID 7

2.2 The $n\sigma$ method 7

3 Machine learning approach for particle identification **8**

3.1 Exploratory data analysis and data preprocessing 8

3.1.1 Dataset description and initial cleaning 8

3.1.2 Feature analysis 9

3.1.3 Class imbalance and evaluation metric 10

3.2 LightGBM 12

3.2.1 Gradient Boosting Method 12

3.2.2 Features of the LightGBM 14

3.3 Hyperparameters optimization 15

3.4 Results evaluation 16

4 Conclusion **19**

Abstract

This work presents the application of a machine learning-based approach for particle identification in the Multi-Purpose Detector (MPD) experiment at the NICA collider. Particle identification (PID) is a central challenge in high-energy nuclear physics.

We developed a robust PID method using a LightGBM classifier. The model was trained on simulated data using track features. To address class imbalance, the macro F_1-score was used as the primary metric for hyperparameter optimization, which was conducted using the Optuna framework.

The optimized model achieved a macro F_1-score of 0.972 and an overall accuracy of 99.3% on a held-out test set. A comparative analysis demonstrates that the LightGBM model noticeably outperforms the traditional $n\sigma$ method.

1 Introduction

The study of a strongly interacting matter under extreme temperatures and baryon density is an important topic in modern high-energy physics. Quantum Chromodynamics (QCD), the theory of the strong interaction, predicts a phase transition from hadronic matter to a quark-gluon plasma. A primary goal of experimental heavy-ion physics is to explore the QCD phase diagram and to locate the hypothesized critical point of that transition.

The Nuclotron-based Ion Collider fAcility (NICA) at the Joint Institute for Nuclear Research (JINR) is designed to uniquely probe the high-baryon-density region of the phase diagram, which remains less explored. This is achieved by colliding heavy ions, such as Au or Bi, at center-of-mass energies ranging from $\sqrt{s_{NN}} = 4$ to 11 GeV.

The Multi-Purpose Detector (MPD) is one of the two main experiments at NICA, its goal is to investigate the properties of nuclear matter at high baryonic

densities and search for signs of phase transitions. To achieve this, the MPD will collect data for track reconstruction, calorimetry and charged-particle identification.

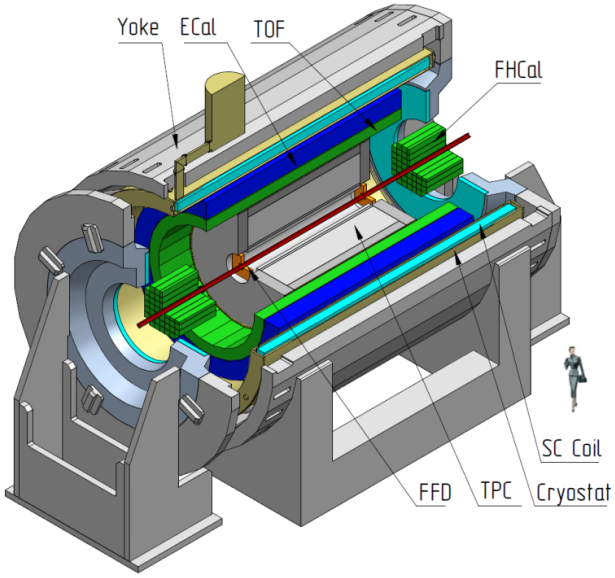


Figure 1.1. Set-up of the MPD and the spatial arrangement of detector subsystems.

The MPD detector is a complex system with a cylindrical geometry, its design incorporates several subsystems, with the Time Projection Chamber (TPC) and the Time-of-Flight (ToF) system playing important roles in the task of particle identification, which will be the focus of this work.

1.1 Time Projection Chamber

The Time Projection Chamber is the main tracking detector of the MPD. It allows precise reconstruction of charged particle trajectories in a magnetic field. As a charged particle traverses the TPC's gas volume, it ionizes the gas atoms. Then electrons drift towards the segmented anode plane of the chamber, where their signal is recorded. These measurements enable reconstruction of the three-dimensional trajectory.

TPC also measures the energy loss (dE/dx), which depends on the particle's

speed and mass according to the Bethe-Bloch law:

$$-\frac{dE}{dx} = K\frac{z^2 Z}{A\beta}\left[\frac{1}{2}\ln\left(\frac{2m_e c^2\beta^2\gamma^2 T_{\max}}{I^2}\right) - \beta^2 - \frac{\delta}{2}\right], \tag{1}$$

where z is the charge of the incident particle, m_e is the electron mass, $K = 4\pi N_A r_e^2 m_e c^2$ is a constant, $\beta = v/c$, γ is a Lorentz factor, T_{max} is the maximum transferable energy to an electron, I is the mean ionization potential, δ is the density-effect correction, charge of the substance Z and its atomic weight A.

This formula predicts a characteristic dependence of (dE/dx), which allows particle identification. For a given momentum, different particle species (e.g., pions, kaons, protons) have different masses and thus different β, leading to distinct values of specific energy loss. This feature is shown in the Figure 1.2 below, as we can see these dependencies for the pions, kaons and protons.

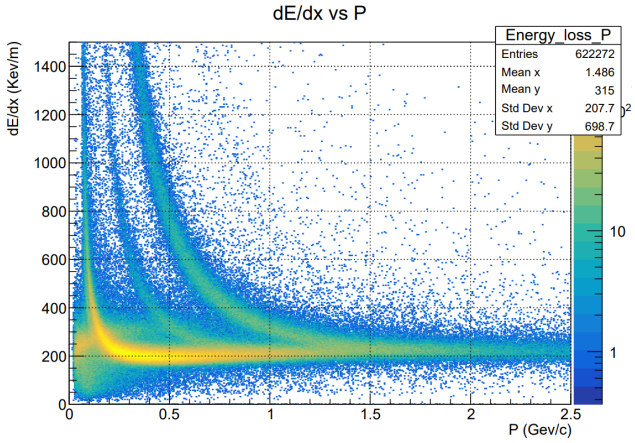


Figure 1.2. Distribution of energy losses vs. total momentum

1.2 Time of Flight System

The Time of Flight system complements the PID capabilities of the TPC, particularly in the intermediate momentum range where the Bethe-Bloch bands for different species begin to overlap. The ToF detector measures the time t taken for a particle to travel a known path length L from the collision vertex. This provides

a direct measurement of the particle's speed β:

$$\beta = \frac{L}{ct}. \tag{2}$$

Combining the speed measurement with the momentum measurement p from the TPC, the particle's mass can be calculated:

$$m^2 = p^2(\frac{1}{\beta^2} - 1). \tag{3}$$

This mass hypothesis is a powerful discriminator between different hadron species (e.g., $\pi^{\pm}$, $K^{\pm}$, p).

The combination of the TPC's (dE/dx) measurement and the ToF's β measurement provides a robust foundation for particle identification in the MPD experiment. However, as will be discussed in the following sections, traditional analysis methods face significant challenges in regions where the detector responses for different particles overlap, necessitating the use of advanced techniques such as machine learning.

2 Current particle identification methods for MPD

After reconstructing the trajectories and kinematic properties of charged particles in the MPD using the TPC and TOF, one obtains a primary experimental dataset. However, information about track coordinates, flight times, and energy losses does not directly show the particle type. In high-multiplicity heavy-ion events, with substantial overlap of proton, kaon, and pion distributions, automated PID becomes a non-trivial task.

2.1 Physical principles of PID

PID in the MPD is based on the relations between a particle's velocity, mass, and charge. The movement of a charged hadron in the TPC's magnetic field allows its momentum p to be determined. The specific energy loss (dE/dx) in the TPC gas follows the Bethe-Bloch law. Combined with the flight time t measured by the TOF system, one obtains an independent measurement of β. The differences in β and ionization losses for protons, kaons, and pions at the same momentum enable their separation within certain momentum boundaries.

2.2 The $n\sigma$ method

In practice, the primary PID method employed in MPD is the $n\sigma$ technique. For each particle species hypothesis h (where $h = \pi^{\pm}, K^{\pm}, p, \bar{p}$), a parameterization of the expected detector response (e.g., (dE/dx) vs. p or $1/\beta$ vs. p) is constructed from calibration data and theoretical curves.

The deviation of the measured value from the expected value for a given hypothesis is normalized to the experimental resolution $\sigma_h(p)$ at that momentum. This defines the $n\sigma$ variable. For pions the ionization loss, for example, it is calculated as:

$$n_\pi = \frac{(dE/dx)_{\text{meas}} - (dE/dx)_\pi (p)}{\sigma_\pi(p)}, \tag{4}$$

with analogous formulas for kaons, protons, and for the $1/\beta, m^2$ values from the TOF detector.

A particle is identified as a given species if its $|n\sigma|$ value for that hypothesis falls below a predefined cut value (typically 2 or 3), effectively selecting particles within a certain number of standard deviations from the expected response.

The $n\sigma$ method is simple, transparent, and provides excellent separation for species where their respective curves are significantly separated, typically at low to intermediate momenta.

However, its limitations become severe at higher momenta ($p > 1.5$ GeV/c), as illustrated in Fig. 1.2. The Bethe-Bloch curves for different species converge, and the mass-squared resolution from TOF degrades, leading to significant overlap in the values. Consequently, the misidentification rate increases substantially in these critical regions.

These limitations of the traditional $n\sigma$ method motivate the search for more robust and universal analysis techniques. Machine learning algorithms, capable of learning complex, multi-dimensional decision boundaries, present a powerful alternative, which is the focus of this work.

3 Machine learning approach for particle identification

3.1 Exploratory data analysis and data preprocessing

The foundation of any robust machine learning model is a well-understood and carefully preprocessed dataset. This section outlines the Exploratory Data Analysis (EDA) and the preprocessing steps undertaken to prepare the data for training the LightGBM model.

3.1.1 Dataset description and initial cleaning

The simulated data were obtained by the Monte Carlo method using the generators UrQMD (Ultra relativistic Quantum Molecular Dynamics model) and the entire chain of reconstructions, simulating the condition of real Bi-Bi collisions of the MPD experiment with $\sqrt{s_{NN}} = 9.2$ GeV. Each entry corresponds to a single reconstructed charged particle track and contains the following features: momentum (p), transverse momentum (pt), azimuthal angle (phi), polar angle (theta), pseudorapidity (eta), energy loss in the TPC (dEdx), squared mass (m2), time-of-flight

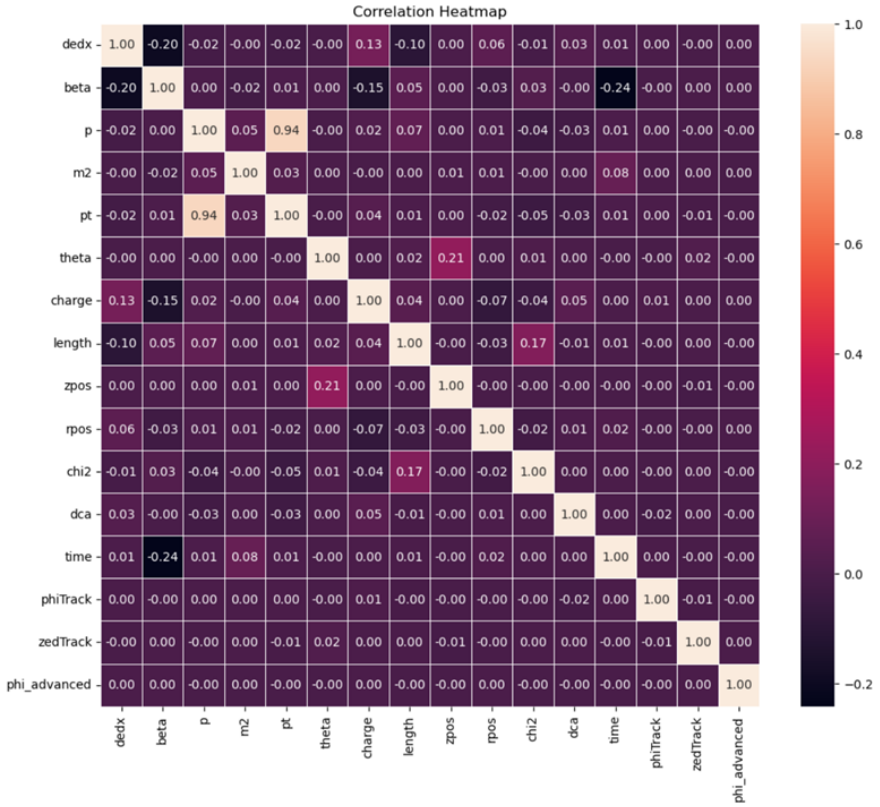


Figure 3.1. Correlation matrix for all features in the dataset.

inverse velocity (beta), number of associated TPC hits (nHits), distance of closest approach to the primary vertex (dca), vertex coordinates (Vx, Vy, Vz) and the target feature, the particle type (type), an integer label where 0: π^+, 1: π^-, 2: K^+, 3: K^-, 4: p, 5: $\bar{p}$.

The first preprocessing step is a quality control. Tracks with a low number of associated hits in the TPC (nHits < 15) were excluded from the dataset, as a insufficient number of hits can lead to significant errors in measurements.

3.1.2 Feature analysis

The correlation analysis presented at the Figure 3.1 revealed a linear relationship between the total momentum p and the transverse momentum pt, as expected from their definition. To avoid redundant information and potential multicollinearity issues, which can be detrimental for some models, the pt feature was removed from the dataset.

Moreover, the choice of choosing p was further supported by the observa-

tions in Figures 3.3, where the Beta-block curves for different particles appear more distinct when plotted as a function of p rather than pt. Plots functions of p confirm the theoretical expectations: clear bands are visible for pions, kaons, and protons at lower momenta, which progressively converge and overlap as momentum increases.

3.1.3 Class imbalance and evaluation metric

A critical aspect of the dataset is a significant class imbalance. The relative abundances of pions, kaons, and protons produced in heavy-ion collisions are not equal; pions are the most abundant, followed by kaons and then protons. This imbalance is clearly visible in the provided pie chart (Figure 3.2). Among the observed particles, 37.5% are π^+, 33.7% are π^-, 22.8% are p, 3.8% are K^+, 2.0% are K^-, 0.2% are $\bar{p}$.

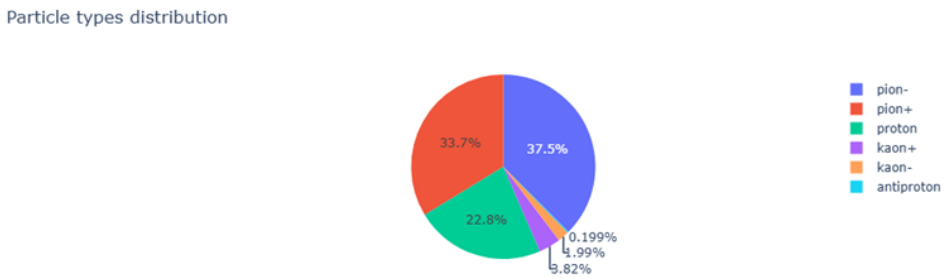
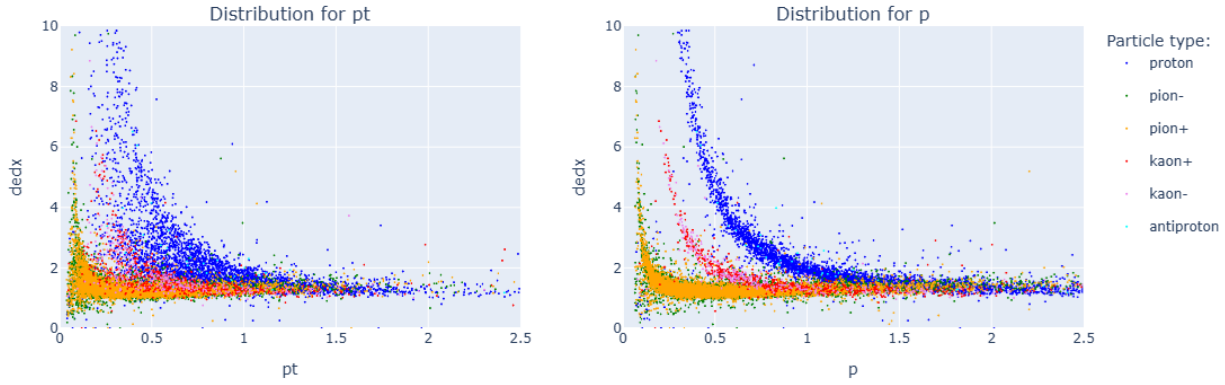
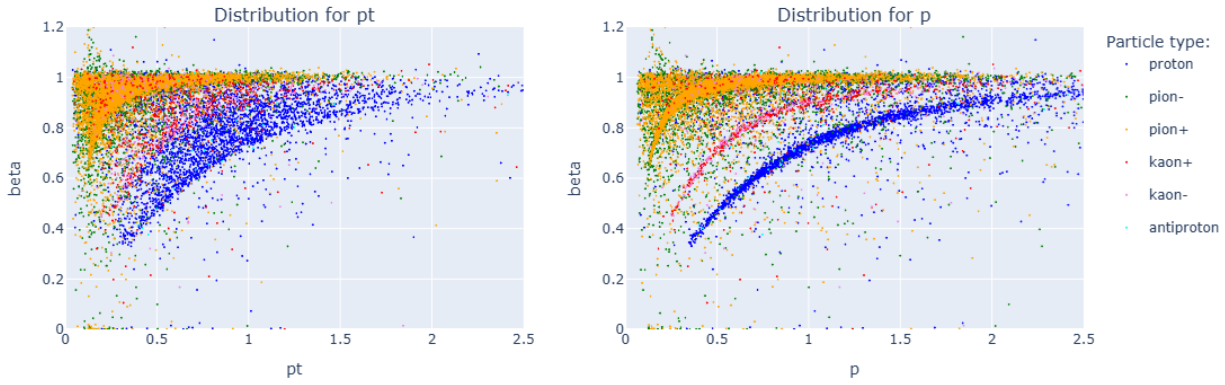


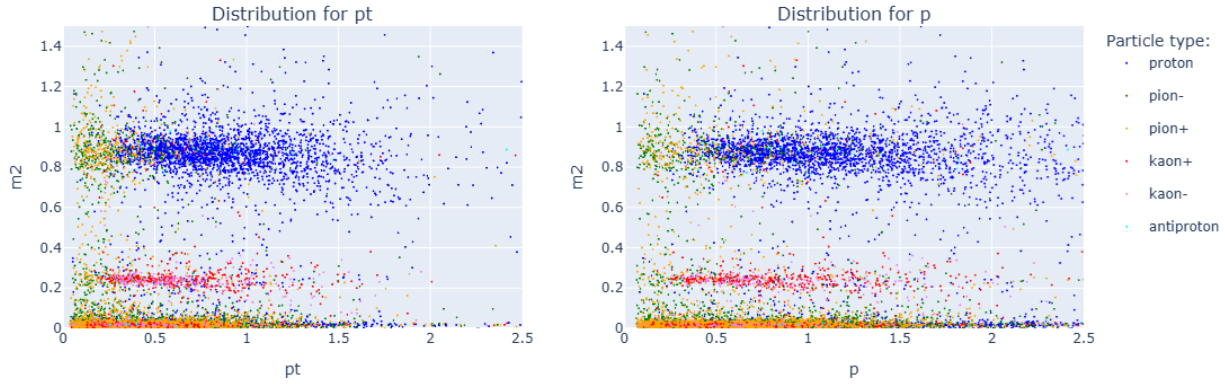
Figure 3.2. Distribution of particle types in the dataset.



(a) Comparison of the plots of dE/dx versus p and dE/dx versus p_t.



(b) Comparison of the plots of β versus p and β versus p_t.



(c) Comparison of the plots of m^2 versus p and m^2 versus p_t.

Figure 3.3. Comparisons of dE/dx, β and m^2 plots as functions of the total momentum p and the transverse momentum p_t.

Training a model on an imbalanced dataset without adjustment can lead to a bias towards the majority classes (pions and protons), resulting in poor identification performance for the minority classes (kaons, antiprotons). To avoid this, the F_1-score was chosen as the primary evaluation metric for model selection and hyperparameter optimization. The F_1-score is the harmonic mean of precision and

recall, providing a balanced measure of a model's accuracy on each individual class. It is defined as:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

where TP are true positives, FP are false positives, and FN are false negatives for a given class. We use the macro-averaged F_1-score, which computes the metric independently for each class and then takes the average, treating all classes equally. This ensures that good performance on the minor classes is valued appropriately during the model's training.

3.2 LightGBM

The task of particle identification (PID) is fundamentally a classification problem, where the goal is to assign each detected track to a specific particle class (e.g., $\pi^{\pm}$, $K^{\pm}$, p, $\bar{p}$) based on its feature vector. Among ML algorithms, gradient-boosting frameworks have consistently demonstrated superior performance in tabular data tasks, such as the one presented by PID. LightGBM (Light Gradient Boosting Machine), a highly efficient and high-performance implementation of gradient boosting, was selected for this study.

3.2.1 Gradient Boosting Method

Gradient boosting is a powerful ensemble learning technique that builds a strong predictive model by combining the outputs of multiple weak learners, typically decision trees, in a sequential, additive manner. Unlike bagging algorithms like Random Forest, which build trees independently, boosting focuses on correcting the errors of previous models.

The core principle can be formalized as follows. Given a dataset with n samples $\{(x_i, y_i)\}_{i=1}^{n}$, the goal is to find an approximation $F(x)$ that minimizes the expected value of a specified differentiable loss function $L(y, F(x))$.

The model is built in an iterative way. Let $F_0(x)$ be an initial model, often chosen as a constant value that minimizes the loss (e.g., the mean of the target values for regression):

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma).$$

For each subsequent boosting round $m = 1$ to M, the algorithm executes the following steps:

1. Compute negative gradients: For each instance i, calculate the negative gradient of the loss function with respect to the current model prediction $F_{m-1}(x_i)$:

$$r_{im} = -\left[\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}\right] \quad \text{for } i = 1, \ldots, n.$$

   These gradients represent the direction and magnitude of the errors made by the current ensemble.

2. Fit a weak learner: train a decision tree $h_m(x)$ to predict the pseudo-residuals r_{im}, effectively modeling the error of the current model. The tree is built on a random subset of the data and/or features to improve robustness.

3. Update the additive model by adding the new weak learner, scaled by a learning rate ν (also called shrinkage, $0 < \nu \leq 1$), which controls the contribution of each tree and helps prevent overfitting:

$$F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x).$$

The final model $F_M(x)$ is the sum of the initial prediction and all subsequent tree models.

3.2.2 Features of the LightGBM

While the previously mentioned gradient boosting principle is general, LightGBM introduces several features that optimize the training process for both speed and memory efficiency, making it exceptionally suitable for large datasets like the one generated by the MPD experiment. In this context, the key features of the LightGBM are:

- Gradient-based One-Side Sampling (GOSS). Traditional boosting uses all data instances to compute gradients. GOSS retains all instances with large gradients (which are under-trained and contribute more to the information gain) and randomly samples instances with small gradients. This focuses computational resources on the more difficult examples, significantly speeding up training without substantially sacrificing accuracy.

- Histogram-Based Algorithm. Instead of examining every possible split point for a feature, LightGBM bins the feature values into discrete histograms. This drastically reduces the computational cost of finding the best split in a tree node, as the algorithm only needs to evaluate the boundaries between bins.

- Leaf-wise (Best-first) Tree Growth Strategy. Many boosting algorithms grow trees level-wise, splitting all leaves at a given depth. In contrast, LightGBM grows trees *leaf-wise*, choosing the leaf with the maximum delta loss to split at each step. This strategy can lead to significantly lower loss and higher accuracy for a fixed number of leaves, though it may also increase the risk of overfitting on small datasets if not properly regularized.

For the PID task in MPD, the combination of high efficiency, handling of large feature spaces, and improved accuracy is crucial. LightGBM ability to train quickly on large datasets allowed for extensive hyperparameter tuning and cross-

validation, which was necessary for developing a robust and generalizable particle identification model.

3.3 Hyperparameters optimization

The performance of a gradient boosting model is highly dependent on its hyperparameters. To achieve optimal performance for the PID task, a systematic hyperparameter optimization study was conducted. The key parameters tuned for the LightGBM model were:

- `learning_rate`: controls the step size shrinkage during gradient descent. A lower value often yields better generalization but requires more iterations;
- `num_leaves`: the maximum number of leaves in one tree. This is the main parameter to control the complexity of a tree model;
- `max_depth`: limits the maximum depth of a tree. This parameter is used to prevent overfitting by restricting the model's complexity;
- `n_estimators`: the number of boosting rounds (trees) to build.

The optimization was performed using the Optuna framework, which employs optimization strategy to efficiently navigate the hyperparameter space. The search was conducted within the following bounds:

- `learning_rate` $\in [0.01, 0.2]$;
- `num_leaves` $\in [5, 255]$;
- `max_depth` $\in [4, 12]$;
- `n_estimators` $\in [5, 4000]$.

After numerous trials, the hyperparameter set that maximized the validation F_1-score was identified as:

- `learning_rate` = 0.012;
- `num_leaves` = 197;
- `max_depth` = 7;
- `n_estimators` = 2615.

This configuration represents a model with a low learning rate and a high number of estimators, a combination that typically leads to high performance at the cost of longer training time. The chosen values for `num_leaves` and `max_depth` indicate a preference for moderately complex trees, balancing the capacity to learn patterns in the data with the need to prevent overfitting.

3.4 Results evaluation

The performance of the optimized LightGBM model was evaluated on a test dataset, which was not used during training or hyperparameter optimization. The overall macro-averaged F_1-score, chosen to account for class imbalance, was equal to 0.972 (changes of this metric during training is presented in the Fig. 3.4). In comparison, F_1-score for the $n\sigma$ method was equal to 0.756. The overall classification accuracy was equal to 99.3% (comparing to 92.3% accuracy score of the $n\sigma$ method). These high scores indicate that the model is both precise and reliable in its predictions. A more detailed picture of the performance per particle types is provided in the confusion matrix shown in Figure 3.5. The most frequent misclassifications occur between species with similar masses, such as pions and kaons, primarily in the high-momentum region where their detector responses overlap significantly,

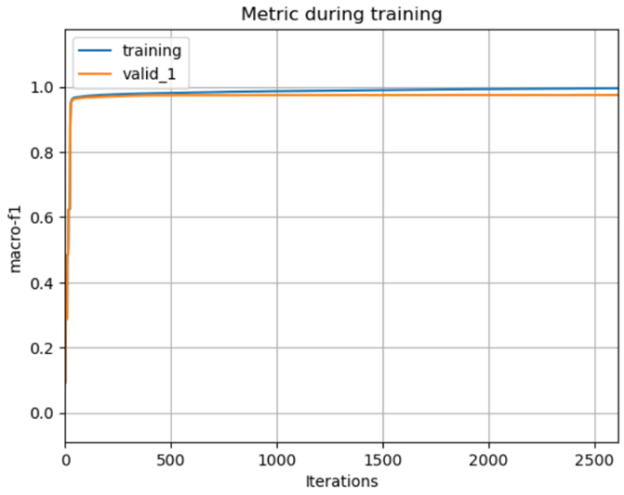


Figure 3.4. F1-score at the different iterations of training of the LightGBM model

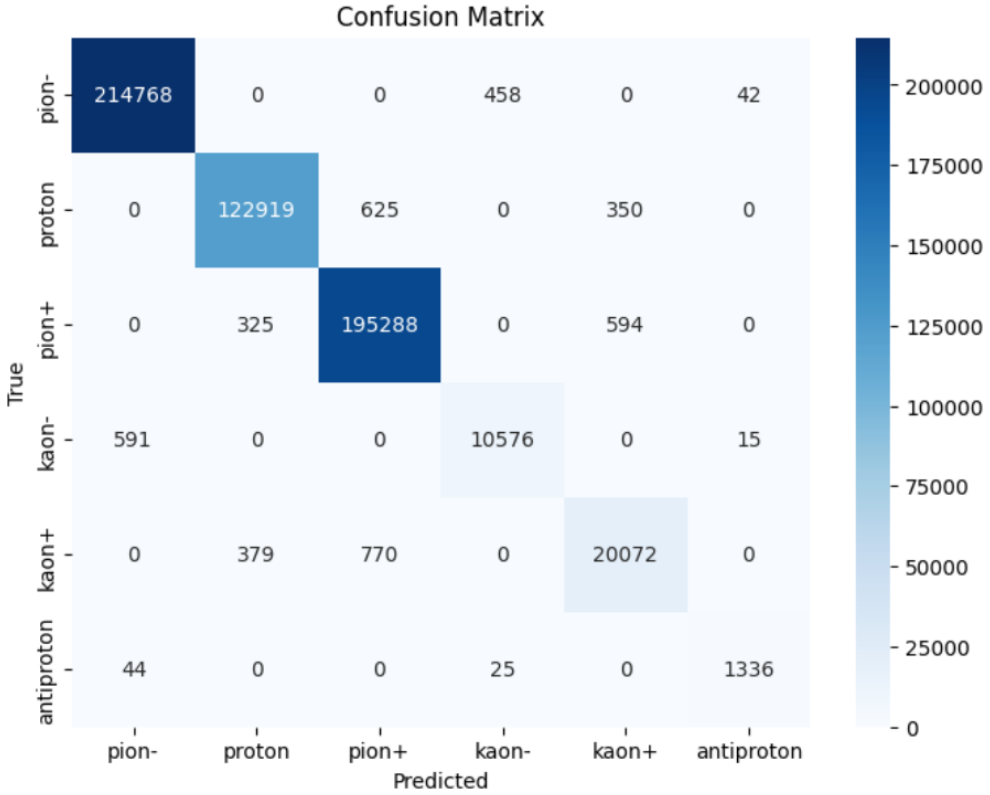
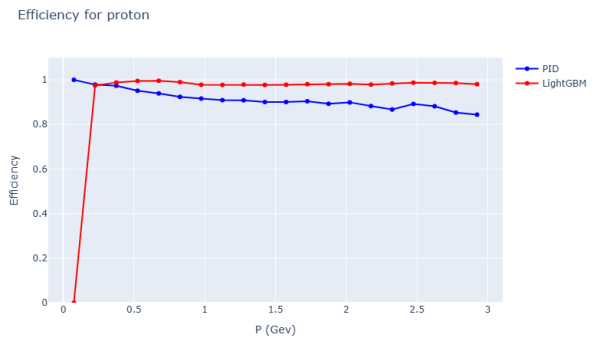
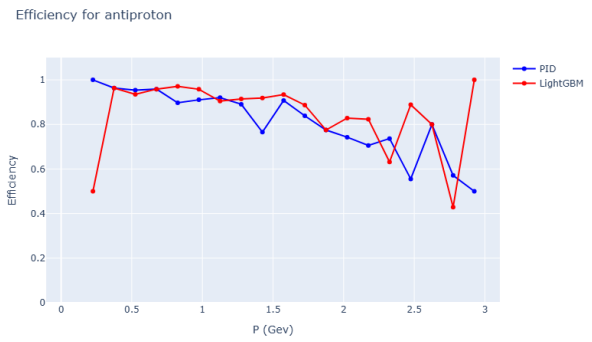


Figure 3.5. Confusion matrix for the LightGBM model on the test dataset

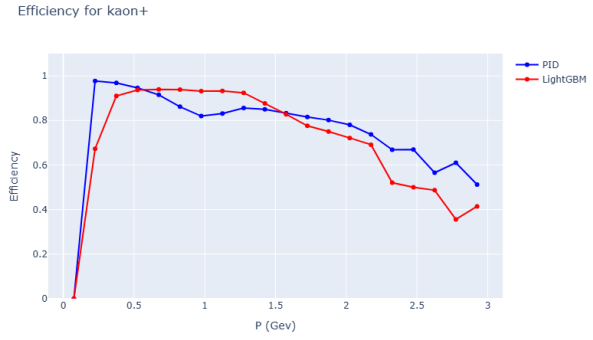
Figure 3.6 presents a direct comparison of the identification efficiency for pions, kaons, and protons as a function of momentum for both the LightGBM model and the $n\sigma$ method. The efficiency is defined as the ratio of correctly identified particles of a given true species to the total number of that true species.



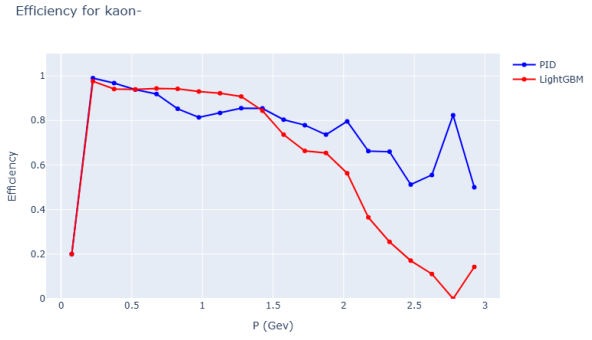
(a) Efficiency plot for p



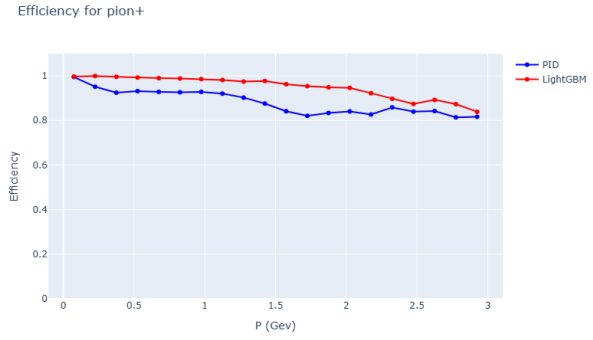
(b) Efficiency plot for $\bar{p}$



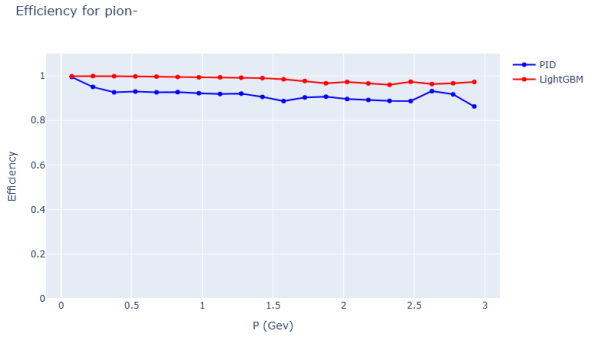
(c) Efficiency plot for K^+



(d) Efficiency plot for K^-



(e) Efficiency plot for π^+



(f) Efficiency plot for π^-

Figure 3.6. Comparison of particle identification efficiency between the LightGBM model and the traditional $n\sigma$ method (labeled as PID).

At the Figure 3.6 presented comparison of the effectiveness of trained LightGBM model and standart $n\sigma$ method. The results are conclusive: the LightGBM model consistently outperforms the $n\sigma$ method across proton and $\pi^{\pm}$ particles and throughout the entire momentum spectrum. For the less represented particle types ($K^{\pm}$, antiproton) model shows poorer perfomance in the area of the higher momentum, but still outperform the $n\sigma$ method at the lower momentum. This char-

acteristic of the trained model may be attributed to the class imbalance present in the dataset. However, the use of the F_1-score ensured that these classes were still taken into account, allowing the model to achieve satisfactory performance in the low-momentum region. The LightGBM model maintains a high efficiency, prooving its ability to learn complex, multi-dimensional decision boundaries from the combined feature set.

This performance difference shows a limitation of the $n\sigma$ method: its reliance on independent, one-dimensional cuts on each detector response. The LightGBM model, by simultaneously considering all relevant features, can achieve a more robust separation, effectively reducing misidentification rates and increasing the purity of particle samples selected for physics analysis. The successful application of LightGBM provides a powerful and reliable tool for PID in the MPD experiment.

4 Conclusion

This work successfully implemented and evaluated a machine learning-based approach for particle identification in the MPD experiment. The developed model, based on the LightGBM framework, demonstrates a significant advancement in charged-hadron identification capabilities.

Through data preprocessing and a hyperparameter optimization process using Optuna, the model was tuned to achieve a macro F_1-score of 0.972, effectively handling the inherent class imbalance in the data. The LightGBM model outperforms the traditional $n\sigma$ technique across p and $\pi^{\pm}$ particles over the entire momentum spectrum. For the less represented particle types ($K^{\pm}, \bar{p}$) the model shows reduced performance in the high-momentum region; however, it still surpasses the $n\sigma$ method at lower momentum. This behavior can be attributed to the class imbalance in the training data, yet the use of the F_1-score ensured that these classes were not neglected, enabling satisfactory performance in the low-momentum region.

In conclusion, this study establishes machine learning, and specifically gradient boosting with LightGBM, as a reliable tool for particle identification in the MPD experiment. The implementation of this model will enable conducting precise physics analyses in the high-baryon-density region of the QCD phase diagram that NICA is designed to explore.

References

[1] 1. Kekelidze, V. ”Project NICA at JINR: status and prospects.” Physics of Particles and Nuclei, 48(5), 727-741.

[2] 2. Golosov, O. MPD Collaboration. ”Multi-Purpose Detector to study heavy-ion collisions at NICA.” Nuclear Instruments and Methods in Physics Research Section A, 1014, 165735.

[3] Abgaryan, V., Acevedo Kado, R., Afanasyev, S. V., Agakishiev, G. N., Alpatov, E., Altsybeev, G., et al. ”Status and initial physics performance studies of the MPD experiment at NICA”. The European Physical Journal A, 58(7), 140. (2022)

[4] MPD Collaboration. ”MPD physics performance studies in Bi+ Bi collisions at $\sqrt{s_{NN}} = 9.2$ GeV.” arXiv preprint arXiv:2503.21117 (2025).

[5] Babkin, V. A., et al. ”Bayesian approach to particles identification in the MPD experiment.” Journal of Instrumentation 19(08) (2024): p. 08007.

[6] Yu Shi, Guolin Ke, Zhuoming Chen, Shuxin Zheng, Tie-Yan Liu. ”Quantized Training of Gradient Boosting Decision Trees”. Advances in Neural Information Processing Systems 35 (NeurIPS 2022), pp. 18822-18833.

[7] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu. ”LightGBM: A Highly Efficient Gradient Boosting Decision Tree”. Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 3149-3157.

[8] Akiba, Takuya and Sano, Shotaro and Yanase, Toshihiko and Ohta, Takeru and Koyama, Masanori ”Optuna: A Next-generation Hyperparameter Optimization Framework”. Proceedings of the 25th International Conference on Knowledge Discovery and Data Mining. (2019)