DeepScience
Multidisciplinary Research: Innovations in Science, Arts and Commerce
Published in October 2025 by Deep Science Publishing (USA | UK | India | Turkey)
ISBN: 978-93-7185-157-2 | E-ISBN: 978-93-7185-737-6 | DOI: https://doi.org/10.70593/978-93-7185-737-6

Leveraging Machine Learning for Accurate Determination of N_{part} in Heavy-Ion Collision Events

Dipankar Basak

Assistant Professor, Department of Physics, Kokrajhar University, Kokrajhar

Email: dipankar0001@gmail.com

Abstract

This work explores the prospects of using of Machine Learning techniques in determining the value of number of participant nucleons (N_{part}) in ultra-relativistic heavy-ion collisions. Several machine learning models were trained on AMPT simulated Au+Au collisions at $\sqrt{s_{NN}}$ = 200 GeV to reconstruct the N_{part} values from raw experimental observables like charged-particle multiplicity at mid-rapidity and the average transverse momentum of charged particles. The study demonstrates that the ML approach significantly achieves higher prediction accuracy in N_{part} estimation on an event-by-event basis compared to traditional methods.

Keywords: Heavy-ion collisions, N_{part}, Machine Learning

1 Introduction

The primary objective of ultra-relativistic heavy-ion collision experiments is to study the phase structure of strongly interacting matter governed by the theory of quantum chromodynamics (QCD), particularly the transition to and the properties of the Quark-Gluon Plasma (QGP) (*Bass et al., 1999*). The QGP is a state of the deconfined color partons, believed to have existed shortly after the Big Bang. The nature and the subsequent space-time evolution of the created matter in these collision experiments depend strongly on the initial collision geometry. One of the key parameters characterizing the geometry in such collisions is the number of participant nucleons (N_{part}), defined as the number of nucleons that undergo at least one inelastic interaction during the collision. Along with the impact parameter (b) and the number of binary nucleon–nucleon collisions (N_{coll}), it forms the basis for determining the centrality of an event which quantifies the nuclear overlap and, hence, the initial geometry of heavy-ion collisions. However, one cannot directly measure these quantities in an experiment. Instead, the Glauber model (*Miller et al., 2007*) is used to calculate centrality theoretically using the final state observables.

Published in October 2025 by Deep Science Publishing (USA | UK | India | Turkey)
ISBN: 978-93-7185-157-2 | E-ISBN: 978-93-7185-737-6 | DOI: https://doi.org/10.70593/978-93-7185-737-6

In recent years, artificial intelligence (AI) has emerged as a powerful tool in high-energy physics for complex tasks, especially for centrality determination (*Li et al., 2020; Mallick et al., 2021; Kuttan et al., 2021; Xiang et al., 2022*). While the majority of works determine centrality in terms of the impact parameter, our previous work (*Basak and Dey, 2023*) showed that centrality can be estimated in terms of N_{part} using Deep Learning (DL). In this work, ML techniques have been used to determine the centrality of Au+Au collisions at $\sqrt{s_{NN}}$ = = 200 GeV in terms of N_{part}.

2 Machine Learning Models

ML, a subset of artificial intelligence, learns complex correlations between the input and target variable from training data. In this study, we employed four different supervised ML models for the regression task.

2.1 Polynomial Regression (PR)

PR (*Peckov, 2012*) is a higher-order extension of linear regression model. Here the original input features are converted into a higher-dimensional space and then training is done on the transformed features using linear model. Regularization using Ridge regression was applied to prevent overfitting.

2.2 K-Nearest Neighbor (KNN)

KNN (*Taunk et al., 2019*) is a supervised, non-parametric, instance-based machine learning technique. It is a simple algorithm that predicts the value of a target variable by averaging the values of the k of its closest (nearest) neighbors in the feature space.

2.3 Decision Trees (DT)

Decision Trees (*Saltykov, 2020*) are tree-based models that recursively partition the feature space through a series of binary splits based on threshold values of individual features to reduce the variance in the target variable within each subset. This process continues until a stopping criterion is met. The model predicts a continuous value at the leaf node, which corresponds to the average of the target values in that node.

Multidisciplinary Research: Innovations in Science, Arts and Commerce
Published in October 2025 by Deep Science Publishing (USA | UK | India | Turkey)
ISBN: 978-93-7185-157-2 | E-ISBN: 978-93-7185-737-6 | DOI: https://doi.org/10.70593/978-93-7185-737-6

Table 1: Results of hyperparameter optimization

ML-Model	Hyperparameters	Values and Ranges	Optimal Hyperparameters
PR	poly_degree	[2, 3, 4, 5, 6, 7, 8]	5
	ridge_alpha	[0.01, 0.1, 1, 10, 100]	1
	ridge_solver	['auto', 'svd', 'cholesky', 'Isqr']	'auto'
KNN	algorithm	['auto', 'ball_tree', 'kd_tree', 'brute']	'auto'
	metric	['euclidean', 'manhattan', 'minkoski']	'euclidean'
	n_neighbors	[1 - 100]	56
	weights	['uniform', 'distance']	'distance'
DT	max_depth	[None, 5, 10, 15, 20]	5
	max_features	[None, 'sqrt', 'log2']	None
	min_samples_leaf	[1, 2, 4]	1
	min_samples_split	[2, 5, 10]	2
LightGBM	learning_rate	[0.1, 0.05, 0.01]	0.05
	max_depth	[3, 5, 7, 10]	5
	n_estimators	[100, 200, 500,600]	200
	num_leaves	[31, 50, 70]	31

2.4 Light Gradient Boosting Machine (LightGBM)

LightGBM (*Ke et al., 2017*) creates an ensemble of weak decision trees through gradient boosting, where each tree corrects errors from previous iterations. LightGBM employs leaf-wise tree growth and advanced techniques including Gradient-based One-Side Sampling and histogram-based splitting for enhanced efficiency and accuracy.

3 Event Generator

We employed A Multi-Phase Transport Model (AMPT) (*Lin et al., 2005*) to simulate minimum bias Au+Au collision events. AMPT is a widely used hybrid transport model for simulating heavy-ion collisions at relativistic energies. The AMPT model consists of four main components that sequentially describe different phases of the collision: (i) Initial Conditions, (ii) Partonic Interactions, (iii) Hadronization, and (iv) Hadronic Rescattering. We used the string melting version of AMPT to generate training and testing datasets for our machine learning models.

Published in October 2025 by Deep Science Publishing (USA | UK | India | Turkey)
ISBN: 978-93-7185-157-2 | E-ISBN: 978-93-7185-737-6 | DOI: https://doi.org/10.70593/978-93-7185-737-6

4 Methodology

4.1 Data generation and feature selection

The dataset used in this study was generated using AMPT for Au+Au collisions at $\sqrt{s_{NN}} =$ 200 GeV. A total of 50K events were generated, with an 80:20 split between training and test sets. Each generated event contains both the target variable (N_{part}) and various final-state observables, enabling supervised learning. Two experimentally measurable final-state observables namely charged particle multiplicity at mid-rapidity $\langle dN\mathrm{ch}/d\eta \rangle$ and the average transverse momentum of charged particles $\langle p_{\mathrm{T}} \rangle$ were used as inputs or features for training the ML models. Only the charged hadrons within mid-rapidity ($|\eta| < 1$) and a transverse momentum cut ($p_{\mathrm{T}} > 0.2$ GeV/c) were taken into consideration during input preparation. Prior to training, all input features were standardized using z-score normalization to ensure they were on the same scale.

4.2 Hyperparameter Optimization

To achieve optimal performance and prevent overfitting, each machine learning model was fine-tuned through GridSearchCV with 5-fold cross validation on the training set. Table **1** shows the optimized hyperparameters for each ML algorithm. All the ML-models were implemented using the Scikit-Learn libraries (*Pedregosa et al., 2011*) in python.

Table 2: *Performance of the ML models.*

ML-Model	MAE	RMSE	R^2
PR	8.4557	11.9622	0.9879
KNN	8.9093	11.8987	0.9881
DT	8.9809	12.1730	0.9875
LightGBM	8.5203	11.6212	0.9886

DeepScience
Multidisciplinary Research: Innovations in Science, Arts and Commerce
Published in October 2025 by Deep Science Publishing (USA | UK | India | Turkey)
ISBN: 978-93-7185-157-2 | E-ISBN: 978-93-7185-737-6 | DOI: https://doi.org/10.70593/978-93-7185-737-6

5 Result and Discussion

All the ML models, after hyperparameter optimization, were trained with 40K AMPT-generated events and the performance of the models was evaluated using the rest 10K events. The primary metrics were used to quantify the performance and precision of the N_{part} prediction for all models. They are−Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Coefficient of Determination (R^2) value. A summary of the quantitative results is provided in Table **2** which shows that all four models were successfully trained to predict N_{part} from the simulated observables. The LightGBM model achieved the best overall performance, with the smallest errors and highest R^2 value, demonstrating its ability to capture complex nonlinear dependencies among observables. Figure **1** plots the correlations between the true values of N_{part} against the ML predicted values. It is evident that they are in agreement as shown by the overall diagonal distribution. Figure **2** shows the ratio of N_{part} predicted with ML models to the true values as a function of true values of N_{part}. Except for central and peripheral collisions, the predicted values are in good agreement with the true values.

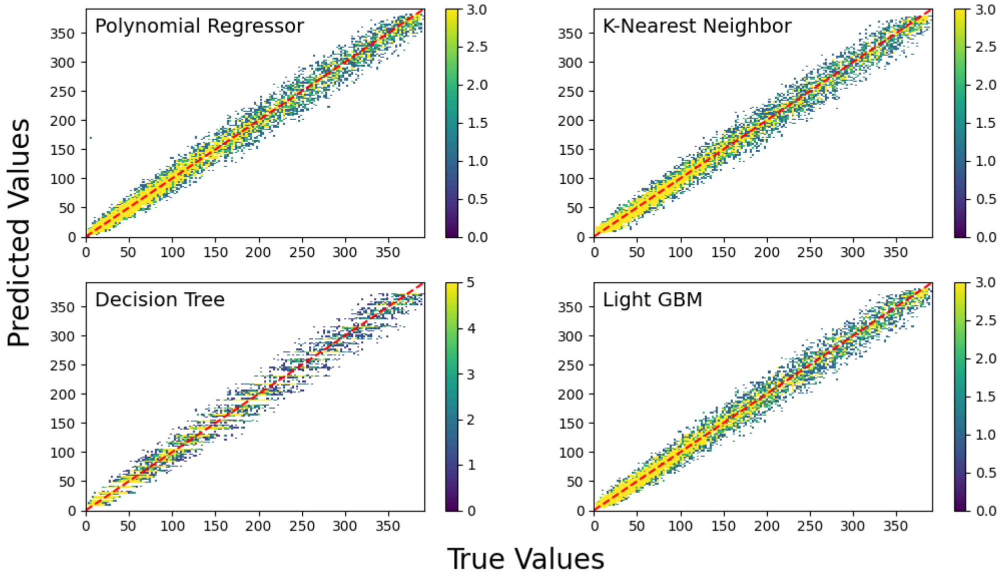

Figure 1: Correlation plot between the true values of Npart and the ML predicted values.

6 Summary

In this work, we have demonstrated a data-driven framework for estimating the number of participant nucleons (N_{part}) in high-energy heavy-ion collisions using machine learning regression techniques. Raw observables from the AMPT-generated data were employed as input features, and several regression algorithms-Polynomial Regression, K-Nearest

ISBN: 978-93-7185-157-2 | E-ISBN: 978-93-7185-737-6 | DOI: https://doi.org/10.70593/978-93-7185-737-6

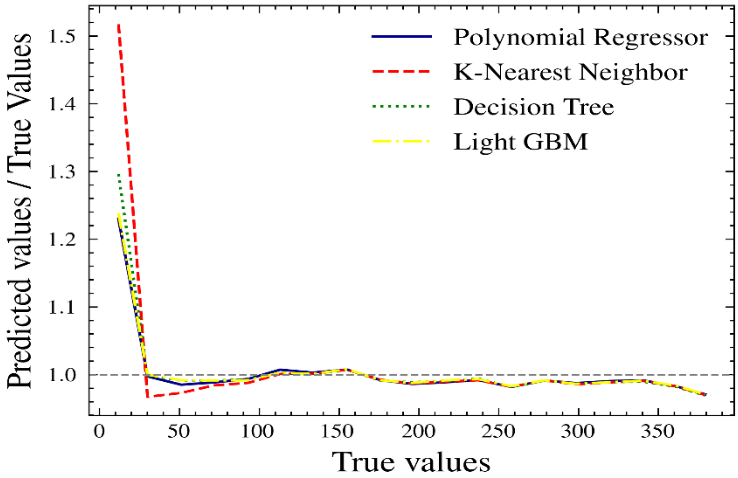

Figure 2: Ratio of ML predicted values of Npart to the true values.

Neighbors, Decision Trees, and Light Gradient Boosting Machine were systematically trained. The results show that machine learning methods can accurately reproduce N_{part} with minimal bias across the entire centrality range except for very central and peripheral collisions. Among the models studied, LightGBM achieved the highest predictive performance (R^2 = 0.9886). The strong correlation between predicted and true values, together with low residual errors, confirms that ML-based models effectively capture the nonlinear relationships between final-state observables and the underlying collision geometry.

References

Basak, D. and Dey, K. (2023). Estimation of collision centrality in terms of the number of participating nucleons in heavy-ion collisions using deep learning. The European Physical Journal A, 59(7):174.

Bass, S. A. et al. (1999). Signatures of quark-gluon plasma formation in high energy heavy-ion collisions: a critical review. Journal of Physics G: Nuclear and Particle Physics, 25(3):R1.

Ke, G. et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In Guyon, I. et al., editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

Kuttan, M. O. et al. (2021). Deep learning based impact parameter determination for the cbm experiment. Particles, 4(1):47–52.

Li, F. et al. (2020). Application of artificial intelligence in the determination of impact parameter in heavy-ion collisions at intermediate energies. Journal of Physics G: Nuclear and Particle Physics, 47(11):115104.

Published in October 2025 by Deep Science Publishing (USA | UK | India | Turkey)
ISBN: 978-93-7185-157-2 | E-ISBN: 978-93-7185-737-6 | DOI: https://doi.org/10.70593/978-93-7185-737-6

Lin, Z.-W., Ko, C. M., Li, B.-A., Zhang, B., and Pal, S. (2005). Multiphase transport model for relativistic heavy ion collisions. Phys. Rev. C, 72:064901.

Mallick, N., Tripathy, S., Mishra, A. N., Deb, S., and Sahoo, R. (2021). Estimation of impact parameter and transverse spherocity in heavy-ion collisions at the lhc energies using machine learning. Phys. Rev. D, 103:094031.

Miller, M. L. et al. (2007). Glauber modeling in high-energy nuclear collisions. Annual Review of Nuclear and Particle Science, 57(1):205–243.

Peckov, A. (2012). A machine learning approach to polynomial regression. Ljubljana, Slovenia.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., et al. (2011). Scikit-learn: Machine learning in python. Journal of Machine Learning Research, 12(85):2825–2830.

Saltykov, S. (2020). Algorithm of building regression decision tree using complementary features. In 2020 13th International Conference"Management of large-scale system development" (MLSD), pages 1–5.

Taunk, K., De, S., Verma, S., and Swetapadma, A. (2019). A brief review of nearest neighbor algorithm for learning and classification. In 2019 International Conference on Intelligent Computing and Control Systems (ICCS), pages 1255–1260.

Xiang, P., Zhao, Y.-S., and Huang, X.-G. (2022). Determination of the impact parameter in high-energy heavy-ion collisions via deep learning. Chinese Physics C, 46(7):074110.

*****