Estimation of impact parameter and transverse spherocity in heavy-ion collisions at the LHC energies using machine learning

Neelkamal Mallick,[1] Sushanta Tripathy,[3] Aditya Nath Mishra,[4] Suman Deb,[1] and Raghunath Sahoo[1,2,*]

[1]*Department of Physics, Indian Institute of Technology Indore, Simrol, Indore 453552, India*
[2]*CERN, CH 1211, Geneva 23, Switzerland*
[3]*INFN—sezione di Bologna, via Irnerio 46, 40126 Bologna BO, Italy*
[4]*Wigner Research Center for Physics, H-1121 Budapest, Hungary*

(Received 2 March 2021; accepted 22 April 2021; published 25 May 2021)

Recently, machine learning (ML) techniques have led to a range of numerous developments in the field of nuclear and high-energy physics. In heavy-ion collisions, the impact parameter of a collision is one of the crucial observables that has a significant impact on the final state particle production. However, the calculation of such a quantity is nearly impossible in experiments as the length scale ranges in the level of a few fermi. In this work, we implement the ML-based regression technique via boosted decision trees to obtain a prediction of the impact parameter in Pb-Pb collisions at $\sqrt{s_{\rm NN}} = 5.02$ TeV using a multiphase transport model. In addition, we predict an event shape observable, transverse spherocity in Pb-Pb collisions at $\sqrt{s_{\rm NN}} = 2.76$ and 5.02 TeV using a multiphase transport and PYTHIA8 based on Angantyr model. After a successful implementation in small collision systems, the use of transverse spherocity in heavy-ion collisions has potential to reveal new results from heavy-ion collisions where the production of a quark gluon plasma medium is already established. We predict the centrality dependent spherocity distributions from the training of minimum bias simulated data and find that the predictions from the boosted decision trees based ML technique match with true simulated data. In the absence of experimental measurements, we propose to implement a machine learning based regression technique to obtain transverse spherocity from the known final state observables in heavy-ion collisions.

DOI: 10.1103/PhysRevD.103.094031

I. INTRODUCTION

A deconfined state of quarks and gluons, also known as quark gluon plasma (QGP) is believed to be produced in ultrarelativistic heavy-ion collisions at the Large Hadron Collider (LHC). However, due to its very short lifetime we do not have any direct evidence of possible QGP formation; instead, several indirect signatures such as strangeness enhancement, quarkonia suppression, direct photon measurements, elliptic flow etc. suggest that formation of QGP is highly probable in such collisions [1]. Such observables are usually studied as a function of centrality classes of the collisions which are determined by the impact parameter (b). However, obtaining the impact parameter values from experiments is still challenging as its value ranges in few femtometers (fm). Thus, in experiments the centrality classes are inferred from final state charged-particle multiplicities and sometimes from the transverse energy distribution. In hindsight, it would benefit the experiments if one can successfully implement machine learning (ML) based techniques to obtain the impact parameter in a precise way.

Historically, the results from proton-proton (pp) collisions are considered as a baseline for understanding the results obtained for heavy-ion collisions. To understand the recent measurements of heavy-ion-like behaviors [2,3] in pp collisions at the LHC, a new event classifier known as transverse spherocity, an event shape observable, has been introduced. [4–12]. After its successful implementation in small collision systems, the use of transverse spherocity in heavy-ion collisions has a potential to reveal new physics where the production of a QGP medium is already established. In our recent publication [13], we have explicitly used transverse spherocity in heavy-ion collisions for the first time to study the final state particle correlations and azimuthal anisotropy as a function of transverse spherocity in a multiphase transport (AMPT) model. A strong anticorrelation of transverse spherocity with the ellipticity of the events in heavy-ion collisions was observed. It was found that low transverse spherocity

*Corresponding author.
Raghunath.Sahoo@cern.ch

events contribute significantly to the elliptic flow while high transverse spherocity events have nearly zero elliptic flow. This indicates that transverse spherocity can be used as a new event classifier in heavy-ion collisions. However, so far no measurement has been performed in heavy-ion collisions as a function transverse spherocity in any of the LHC experiments due to the fact that such a measurement becomes computationally challenging in heavy-ion collisions. Thus, the application of ML-based regression techniques to obtain transverse spherocity from the known final state experimental observables would be very useful in the current scenario.

Recently, machine learning techniques have led to a range of numerous developments in the field of high-energy physics (HEP) along with in different fields of physics [14–20]. For several years, different machine learning algorithms have been used to determine the impact parameter [21–26]. Thus, it is timely to implement ML-based techniques to obtain the impact parameter and transverse spherocity distributions at the LHC energies. Machine learning methods are designed to exploit large datasets in order to reduce complexity. They also help to find new features in data. Currently, the most frequently used machine learning algorithms in high-energy physics are boosted decision trees (BDTs) [27] and neural networks (NNs). Usually, machine learning model is trained for variables relevant to a particular physics problem, which can be classified into either a classification or regression problem. In both cases, training the model is the most time consuming step for both humans and computer CPUs, while the inference stage is relatively inexpensive. Thus, machine learning models are gaining lots of popularity in different fields of basic sciences. BDTs and NNs are typically used to classify particles and events. However, they are also used for regression, where a continuous function is learned and gives a prediction of an observable which is usually cumbersome to obtain from real experiments. In this work, we implement ML-based regression technique via BDTs to obtain predictions for impact parameter and spherocity distributions in Pb-Pb collisions at $\sqrt{s_{\rm NN}} = 2.76$ and 5.02 TeV using AMPT [28] and PYTHIA8 (Angantyr) models [29]. For machine learning, we have used a PYTHON based machine learning package, called the SCIKIT-LEARN software package [30]. We have specifically used the *GradientBoostingRegressor* module inside the *sklearn.ensemble* framework. For our study, we use final state charged-particle multiplicity and mean transverse momentum as the input variables for the predictions of the impact parameter and transverse spherocity.

This paper is organized as follows. We start with a brief introduction on event generation and target observables for ML in Sec. II. Then, in Sec. III we provide a detailed procedure of ML-based regression techniques along with a few quality assurance plots to obtain the impact parameter and transverse spherocity from heavy-ion collisions at the LHC energies using event generators such as the AMPT and PYTHIA8 (Angantyr) models. In Sec. IV, we provide a detailed discussion on the results and we summarize our findings in Sec. V.

II. EVENT GENERATION AND TARGET OBSERVABLES

In this section, we begin with a brief introduction on event generators. Then, we proceed to define impact parameter and transverse spherocity.

A. A multiphase transport model

AMPT model contains four main components, namely, initialization, followed by the parton transport, hadronization mechanism, and hadron transport [28]. The initialization of the model is similar to the HIJING model [31], where the produced partons calculated in pp collisions are converted into heavy-ion collisions. They are incorporated via the nuclear overlap and shadowing function using an inbuilt Glauber model. The initial low-momentum partons are produced from parametrized colored string fragmentation mechanisms and they are separated from high momentum partons by a momentum cutoff. The produced partons are then initiated into the parton transport part, ZPC [32]. In the string melting version of AMPT, at the start of the ZPC melting of the colored strings into low-momentum partons takes place, which are calculated using the Lund FRITIOF model. Then the partons undergo multiple scatterings which take place when any two partons are within a distance of minimum approach and the transported partons are finally hadronized using the spatial coalescence mechanism [33,34]. The produced hadrons further undergo a final evolution in the ART mechanism [35,36] via hadron interactions. The particle flow and spectra at the mid-$p_{\rm T}$ regions are well explained by the quark coalescence mechanism for hadronization [37–39], which is embedded in the string melting mode in AMPT. Thus, we have used the AMPT string melting mode (AMPT version 2.26t7) for all of our calculations. The AMPT settings in the current work are the same as those reported in Refs. [13,40]. For the input of impact parameter values for different centrality classes in Pb-Pb collisions at $\sqrt{s_{\rm NN}} = 2.76$ and 5.02 TeV, we have used Ref. [41].

B. PYTHIA8 (Angantyr)

PYTHIA8 [42], which was initially developed for small collision systems such as e^+e^-, pp, and $p\bar{p}$ collisions, now includes the Angantyr model to describe heavy-ion collisions. The main idea of the Angantyr model in PYTHIA8 is to extrapolate dynamics from pp collisions to heavy-ion collisions, retaining as much as possible from pp collisions [29]. In order to make predictions for heavy-ion collisions, different parts of a standard PYTHIA8 simulation was modified and it was tuned with the results from e^+e^-,

pp, and ep collisions. So far, the model does not use any heavy-ion data to tune it. Thus, the current model retains the production mechanisms from small collision systems. However, it is successful in reproducing several features of pA and AA collisions [29]. In this work, we have used the predictions from the PYTHIA8 (Angantyr) model to show the model dependence of the proposed ML technique in obtaining the transverse spherocity distributions.

C. Impact parameter

The interpretation of several results measured in heavy-ion collisions largely depends on the overlap region of two colliding nuclei in a given impact parameter (b). Obtaining the impact parameter values from experiments are still challenging as its value ranges in a few femtometers (fm). However, theoretical techniques, using the so-called Glauber formalism [43–46], have been developed to allow the estimation of impact parameters and number of participants from experimental data, which consider multiple scattering of nucleons in nuclear targets. AMPT and PYTHIA8 (Angantyr) model internally depend on the Glauber picture to model the early stage of heavy-ion collisions with a proper computation of the number of inelastic subcollisions for a particular centrality class [47]. Here, we briefly describe how the total inelastic cross section, number of binary collisions, and number of participants are related to the impact parameter.

For a collision of two heavy nuclei, A and B at relativistic speeds with impact parameter $\mathbf{b}$, the inelastic cross section can be defined as

$$\sigma_{AB}^{\text{inel}}(\mathbf{b}) = \int d\mathbf{b}[1-(1-T_{AB}(\mathbf{b})\sigma_{NN}^{\text{inel}})^{AB}] \tag{1}$$

$$\simeq \int d\mathbf{b}[1-\exp[-ABT_{AB}(\mathbf{b})\sigma_{NN}^{\text{inel}}]], \tag{2}$$

where $T_{AB}(\mathbf{b})$ is known as the nuclear overlap function and $\sigma_{NN}^{\text{inel}}$ is the nucleon-nucleon inelastic cross section. For such nucleus-nucleus collisions, the total number of binary collisions is

$$N_{\text{coll}}^{AB}(\mathbf{b}) = \sum_{n=0}^{A} nP(n,\mathbf{b}) = ABT_{AB}(\mathbf{b})\sigma_{NN}^{\text{inel}}, \tag{3}$$

and the number of participants (or wounded nucleons) of nucleus A for a given impact parameter $\mathbf{b}$ is given by

$$N_{\text{part}}^{A}(\mathbf{b}) = B\int T_B(\mathbf{s}-\mathbf{b})\{1-[1-T_A(\mathbf{s})\sigma_{\text{inel}}^{\text{NN}}]^A\}d^2s. \tag{4}$$

The number of participants in nucleus A is proportional to the nuclear profile function at transverse positions $\mathbf{s}$, $T_{AB}(\mathbf{s})$, weighted by the sum over the probability for a nucleon-nucleon collision at transverse position $(\mathbf{s}-\mathbf{b})$ in nucleus B. Thus, at a given $\mathbf{b}$, the number of participants is given by

$$N_{\text{part}}(\mathbf{b}) = N_{\text{part}}^{A}(\mathbf{b}) + N_{\text{part}}^{B}(\mathbf{b}). \tag{5}$$

Theoretical calculations in heavy-ion physics use $\mathbf{b}$ as an input to compare theoretical results to the experimental ones. $N_{\text{part}}(\mathbf{b})$ or $N_{\text{coll}}(\mathbf{b})$ are calculated using the Glauber model at a given $\mathbf{b}$, which are subsequently related to multiplicities [48]. In this article, we use machine learning techniques to predict the impact parameter distribution using the observables measured after the collision.

D. Transverse spherocity

Transverse spherocity is defined for a unit vector $\hat{n}$ that minimizes the following ratio in the transverse plane:

$$S_0 = \frac{\pi^2}{4}\left(\frac{\Sigma_i|\vec{p}_{T_i}\times\hat{n}|}{\Sigma_i p_{T_i}}\right)^2. \tag{6}$$

By definition, transverse spherocity is infrared and collinear safe [6] and the extreme limits of transverse spherocity are related to specific configurations of events in the transverse plane. The value of transverse spherocity ranges from 0 to 1. Transverse spherocity becoming 0 means the events have back-to-back structure and are called jetty events while 1 would mean the events are isotropic in nature. The isotropic events are the results of soft processes while the jetty events are usually the hard events. The spherocity distributions are obtained for the events with at least 5 charged particles in the pseudorapidity range of $|\eta| < 0.8$ with $p_{\text{T}} > 0.15$ GeV/c to recreate the similar conditions as in ALICE at the LHC. In recent years, there have been several applications of transverse spherocity at the LHC energies, which can be found in Refs. [4–12].

III. MACHINE LEARNING BASED REGRESSION

ML techniques could be applied to solve numerous real-life problems. First, the ML model is built by training the model with a training dataset. The performance is tested with a new independent set of data and further tuning of the model parameters are made if necessary. Once the predictions or the estimations are sufficiently satisfying, the model is saved and is ready to be applied to actual data to solve the problem. Machine learning addresses mainly classification, regression, and clustering kinds of problems. The problem, that we have in hand, is of the supervised regression kind; i.e., for each set of the input variables, we have a finite numerical value as the target variable. Each set of the data refers to one event of the heavy-ion collisions. We have used charged-particle multiplicity ($\langle dN_{\text{ch}}/d\eta\rangle$), charged-particle multiplicity in the transverse region ($\langle N_{\text{ch}}^{\text{TS}}\rangle$), and mean transverse momentum ($\langle p_{\text{T}}\rangle$) as the input variables and the target variables as the impact

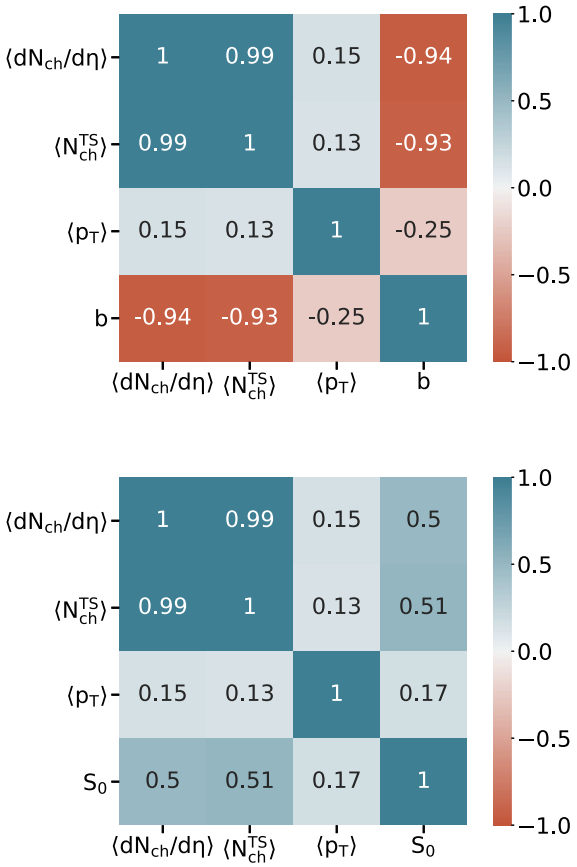

FIG. 1. Correlation matrix of the input variables and target observables in Pb-Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV in AMPT model. The numbers show the correlation coefficients. The top panel shows the correlation matrix for impact parameter while the bottom panel shows the correlation matrix for transverse spherocity.

parameter (b) and transverse spherocity (S_0). For our problem, the gradient boosting decision trees (GBDT) algorithm has been chosen. Figure 1 represents the correlation matrix for the input variables and the target variables for Pb-Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV minimum bias events. The numbers in the boxes represent the correlation coefficient which ranges from -1 to 1 and give the correlation strength between the intersecting variables in the matrix. The correlation coefficient (ρ) for two variables x and y is given by

$$\rho = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y} \tag{7}$$

where $\mathrm{cov}(x, y)$ is the covariance and σ_x and σ_y are the standard deviations of x and y, respectively.

A. Decision trees regression

Machine learning comprises several statistical predictive models. The algorithm learns from the data and builds the model, which then makes predictions or decisions based on its learning experience. Out of many such algorithms, decision trees are the most popular machine learning algorithms known for their simplicity yet intelligent and powerful predictions. Decision trees regression [49] makes predictions for a target variable having continuous finite values (such as real numbers). In this study, these are the impact parameter (b) and transverse spherocity (S_0). Decision trees are built in a top-down approach. Trees can be understood as continuous piecewise structures that take decisions based on certain rules giving rise to binary splitting of the nodes. The tree begins from the root node and then keeps on splitting recursively into further nodes. The splitting process continues until the preset maximum depth of the tree is reached. Each such split points are termed as internal nodes and the criteria of splitting is different for the type of the problem, i.e., classification or regression. Splitting is often governed by minimizing the node impurity. Impurity criteria is a mathematical measure of selecting the best features for splitting and growing the tree. In decision tree regression, there are two common impurity measures, i.e., least squares and least-absolute deviation. In least squares, splits are chosen to minimize the sum of squared error between the observation and the mean in each node. In least-absolute deviation, splits are chosen to minimize the mean-absolute deviation from the median in each node. Mean-absolute deviation is more robust to outliers as compared to mean-squared error; however, it fits slower.

B. Gradient boosting decision trees

GBDTs [50] use decision trees of a fixed size as the base estimators. These base estimators are called as the weak learners in the context of gradient boosting. Gradient boosting is an iterative process and it builds an additive model in a forward stage-wise fashion where addition of a new weak learner compensates the shortcomings of the existing weak learners. These shortcomings are identified as the gradients [Eq. (11)]. In any regression problem, we have a set of target variables y and a set of input (observed) variables $\mathbf{x} = \{x_1, \ldots, x_n\}$. The training sample $\{y_i, \mathbf{x_i}\}_1^N$ has all the known values of $(y, \mathbf{x})$ for N events. The goal is to train the ML model to obtain the functional value $F(\mathbf{x})$ which satisfies $y_i = F(\mathbf{x_i})$. In the gradient boosting method, this estimation can be achieved by adding the outcomes of several weak learners h_m as

$$y_i = F_M(\mathbf{x_i}) = \sum_{m=1}^{M} h_m(\mathbf{x_i}). \tag{8}$$

The parameter M corresponds to the number of trees in each decision tree estimator. Now at each stage, the additive process can be written as

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu h_m(\mathbf{x}). \quad (9)$$

The parameter ν is called as the learning rate. There is a direct tradeoff between the learning rate and number of trees (the number of weak learners), specified by the parameter M. Smaller values of the learning rate require larger numbers of weak learners (more number of trees) to maintain a constant training error. Usually a model is built with a small value of learning rate as it performs better and achieves a minimal testing error. The newly added tree $h_m(\mathbf{x})$ is fitted in order to minimize the sum of a loss function $l(y_i, F_{m-1}(\mathbf{x_i}) + h_m(\mathbf{x_i}))$. We can use the Taylor's first order expansion and approximate the value of l as

$$l(y_i, F_{m-1}(\mathbf{x_i}) + h_m(\mathbf{x_i})) \approx l(y_i, F_{m-1}(\mathbf{x_i})) + h_m(\mathbf{x_i}) \left[\frac{\partial l(y_i, F(\mathbf{x_i}))}{\partial F(\mathbf{x_i})}\right]_{F=F_{m-1}}. \quad (10)$$

For the squared loss function, which is of the form, $l(y_i, F(\mathbf{x_i})) = \frac{1}{2}(y_i - F(\mathbf{x_i}))^2$,

$$-g_i = -\left[\frac{\partial l(y_i, F(\mathbf{x_i}))}{\partial F(\mathbf{x_i})}\right]_{F=F_{m-1}} = y_i - F(\mathbf{x_i}). \quad (11)$$

Here, g_i is the gradient and $(y_i - F(\mathbf{x_i}))$ is called as the residual. We can interpret residuals as negative gradients. Now, to improve the model predictions and build a more robust model, a suitable loss function is chosen, which is then minimized using the gradient descent algorithm. In the GBDT algorithm, we have three types of loss functions, i.e., least squares, least-absolute deviation, and the Huber loss functions.

Least squares:

$$l(y_i, F(\mathbf{x_i})) = \frac{1}{2}(y_i - F(\mathbf{x_i}))^2$$

Least-absolute deviation:

$$l(y_i, F(\mathbf{x_i})) = |y_i - F(\mathbf{x_i})|$$

Huber:

$$l(y_i, F(\mathbf{x_i})) = \begin{cases} \frac{1}{2}(y_i - F(\mathbf{x_i}))^2, |y_i - F(\mathbf{x_i})| \le \delta \\ \delta(|y_i - F(\mathbf{x_i})| - \delta/2), |y_i - F(\mathbf{x_i})| > \delta \end{cases}$$

Here, δ is known as the transition point that defines those residual values that are considered to be outliers, subject to absolute rather than squared-error loss. For residual less than or equal to δ, the Huber loss function becomes the least-squares loss.

C. Quality assurance

In the GBDT algorithm, there are essential parameters such as the number of decision trees, and the maximum depth and learning rate, which play crucial role in the fitting process. The task is to obtain the best fit of the model to the training data by optimizing these parameters. These parameters require manual tuning by observing the performance of the model. For this study, we have taken the number of trees to be 100, maximum depth to be 40, learning rate is fixed at 0.1, and all other parameters are set to their default values. The accuracy of the trained model could be evaluated by calculating the mean-absolute error of the target variables for the training dataset. The mean-absolute error for the impact parameter is given by

$$\Delta b = \frac{1}{N_{\text{events}}} \sum_{n=1}^{N_{\text{events}}} |b_n^{\text{true}} - b_n^{\text{pred}}|. \quad (12)$$

Here, b_n^{true} is the true value of the impact parameter from the simulated data and b_n^{pred} is the predicted value from the GBDT-ML model. Mean-absolute error for transverse spherocity (ΔS_0) could be estimated in the similar fashion. The learning process of the ML model is greatly influenced by the size of the training data. We can see this by evaluating the values of Δb and ΔS_0 for 10K events of independent testing data with the ML model trained with different sets of events. The results are mentioned in Table I. As we can see, with a greater number of events in the training data, the model learns better; hence, the mean-absolute error decreases with an increase in training data size. This behavior is expected as the model should gather more information with more training data, and thus its prediction gets improved. As we increase training data size from 2 to 60 K events, Δb decreases from 0.71 to 0.52 fm and ΔS_0 decreases from 0.079 to 0.055. However, with training size greater than 50 K events, Δb saturates to a constant value, and the decrease in ΔS_0 is too small to make any difference. Therefore, for this study, we have taken 60 K events for the training of the model for both the target variables.

Loss function is another important parameter in the GBDT algorithm, which needs to be chosen carefully. We have obtained the Δb and ΔS_0 values against boosting iterations (number of trees) for three kinds of loss

TABLE I. Size dependence of the GBDT-ML model for the training data in Pb-Pb collisions at $\sqrt{s_{\text{NN}}} = 5.02$ TeV. Δb and ΔS_0 are the mean-absolute error on the independent testing data having 10K events for impact parameter and transverse spherocity, respectively.

Size of training data	2K	10K	20K	40K	50K	60K
Δb [fm]	0.71	0.62	0.58	0.53	0.52	0.52
ΔS_0	0.079	0.068	0.062	0.058	0.056	0.055

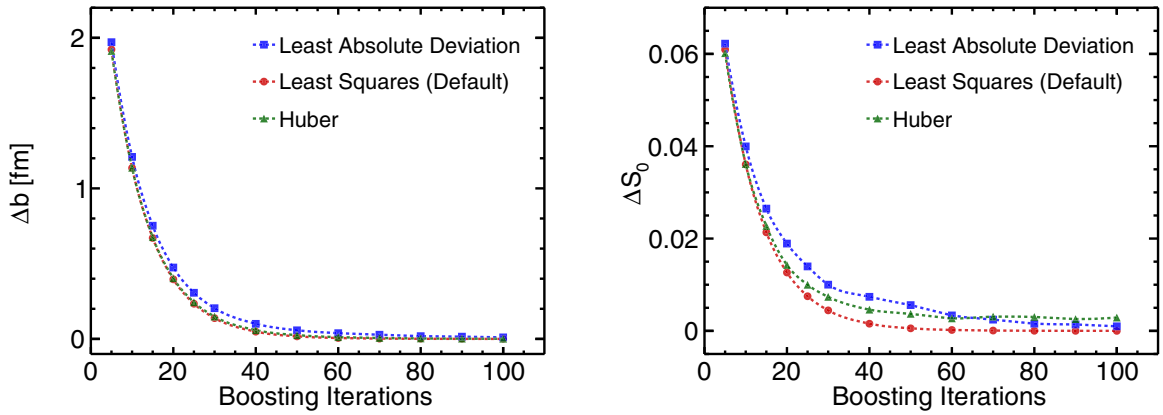

FIG. 2. Performance of different loss functions in the GBDT-ML model in the training dataset with 60K events each. The x axis denotes boosting iterations (number of trees) and the y axis denotes the corresponding mean absolute error.

functions, i.e., least squares, least-absolute deviation, and the Huber loss functions. Figure 2 shows the performance of these loss functions in the training dataset with 60K events for both the target variables. The x axis denotes boosting iterations and the y axis denotes the corresponding mean-absolute error of the training data. As we can see, Δb and ΔS_0 decrease by growing more trees in the model. The values of Δb and ΔS_0 fall exponentially, moving from 10 to 60 number of trees and then the descent is very small. For boosting iterations greater than 80, the mean-absolute error seems to saturate and remains fairly constant. Small values of Δb and ΔS_0 in the training data indicate that the model is learning better. To be fair, we stop at 100 trees. Among these three loss functions, the least-square loss performs better and its training is faster than the mean-absolute deviation and the Huber loss. So, we have chosen the least-square loss function as a default method for this study.

By fixing the training data sample size, optimizing the model parameters and minimizing the mean-absolute errors, i.e., Δb and ΔS_0 in the training data, now it is time to look into the performance of the trained model. We can predict the values of the impact parameter and transverse spherocity using this ML model. Figure 3 shows the predicted values of the variables using ML models versus the true values of the variables from AMPT model simulation for 16K events of minimum bias Pb-Pb collisions at $\sqrt{s_{\rm NN}} = 5.02$ TeV in an independent testing dataset. For the most accurate predictions, the points on the plot should populate a straight line inclined at an angle 45 degrees to the x axis. Though we see a little spread in the plots, the straight lines are distinctly visible, suggesting a good agreement between the predictions from the ML model and true values from the simulation. We have also computed the testing accuracy and found that, for the impact parameter, $\Delta b = 0.52$ fm while for transverse spherocity, $\Delta S_0 = 0.055$ for the testing data.

For the subsequent plots, the total number of accepted events for minimum bias Pb-Pb collisions at $\sqrt{s_{\rm NN}} =$ 5.02 TeV are 76.5K, out of which 60K events are used for the ML training (to minimize the mean-absolute errors)

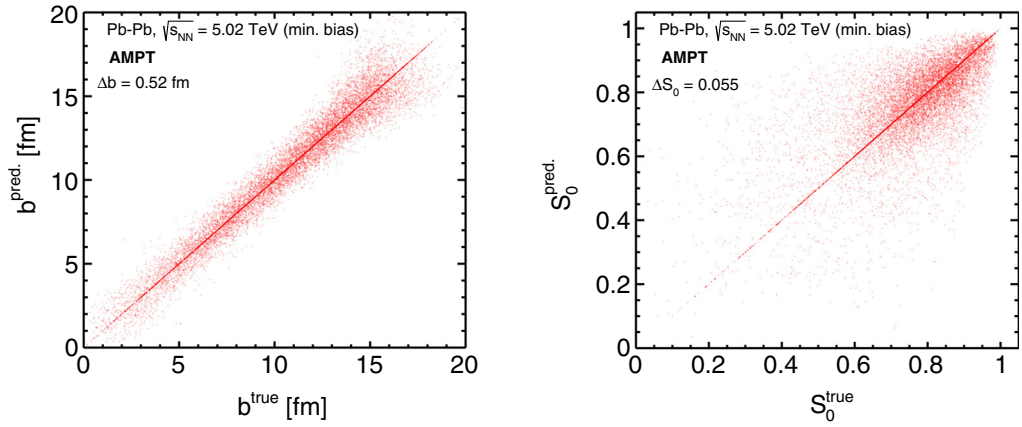

FIG. 3. Predicted values of impact parameter (left) and transverse spherocity (right) in the GBDT-ML model versus their true values in the testing data with 16K events of Pb-Pb collisions at $\sqrt{s_{\rm NN}} = 5.02$ TeV (min. bias) in AMPT model.

and the rest of the events are used for the testing purpose. The maximum deviation among ML predictions from different loss functions with respect to the least-square loss function method (default method) is used as systematic uncertainties in the predicted values. They are summed in quadrature with the statistical uncertainties and shown as red-colored bands in the plots. The statistical uncertainties in the true values are shown as bars. In predicted to true ratio plots, the black-colored band denotes the statistical uncertainties in the true values while the red-colored band denotes the quadratic sum of statistical and systematic uncertainties.

IV. RESULTS AND DISCUSSIONS

The top panel of Fig. 1 shows the correlation matrix of the input variables and impact parameter in Pb-Pb collisions at $\sqrt{s_{\rm NN}} = 5.02$ TeV. The numbers show the correlation coefficients (ρ) which are obtained from Eq. (7). We see a significant anticorrelation of the impact parameter and the final state charged-particle multiplicities, which is evident from the values of ρ. Also, the impact parameter was found to be anticorrelated with the mean transverse momentum of an event. This behavior is evident in Fig. 4, where the correlation with each input variable with the impact parameter is shown. Figure 6 shows the predictions for impact parameter distribution using ML for Pb-Pb collisions at $\sqrt{s_{\rm NN}} = 5.02$ TeV in AMPT model. The lower panel shows the ratio of predicted distribution to the true distribution. One can clearly see that the proposed ML framework with $\langle dN_{\rm ch}/d\eta\rangle$, $\langle N_{\rm ch}^{\rm TS}\rangle$, and $\langle p_{\rm T}\rangle$ as the input variables does a nice job of predicting the impact parameter distribution in Pb-Pb collisions at $\sqrt{s_{\rm NN}} = 5.02$ TeV.

The bottom panel of Fig. 1 shows the correlation matrix of the input variables and transverse spherocity in Pb-Pb collisions at $\sqrt{s_{\rm NN}} = 5.02$ TeV. The numbers show the correlation coefficients (ρ) which are obtained from Eq. (7). The values of ρ for intersecting variables in the matrix suggest there is a good dependency of the chosen input variables and the transverse spherocity. Based on Eq. (6), one would naively expect that spherocity would be highly correlated to the charged-particle multiplicity and mean transverse momentum of an event. Thus, we have chosen total charged-particle multiplicity, charged-particle multiplicity in the transverse region, and the mean transverse momentum as the input variables for the ML prediction. From Fig. 1, it is found that the transverse spherocity has very high correlation with the total charged-particle multiplicity and charged-particle multiplicity in the transverse region of an event. Although, the correlation with the mean transverse momentum is small it is still significant for a proper prediction of transverse spherocity through ML. To understand the correlation between the input variables and the transverse spherocity we have shown the correlation

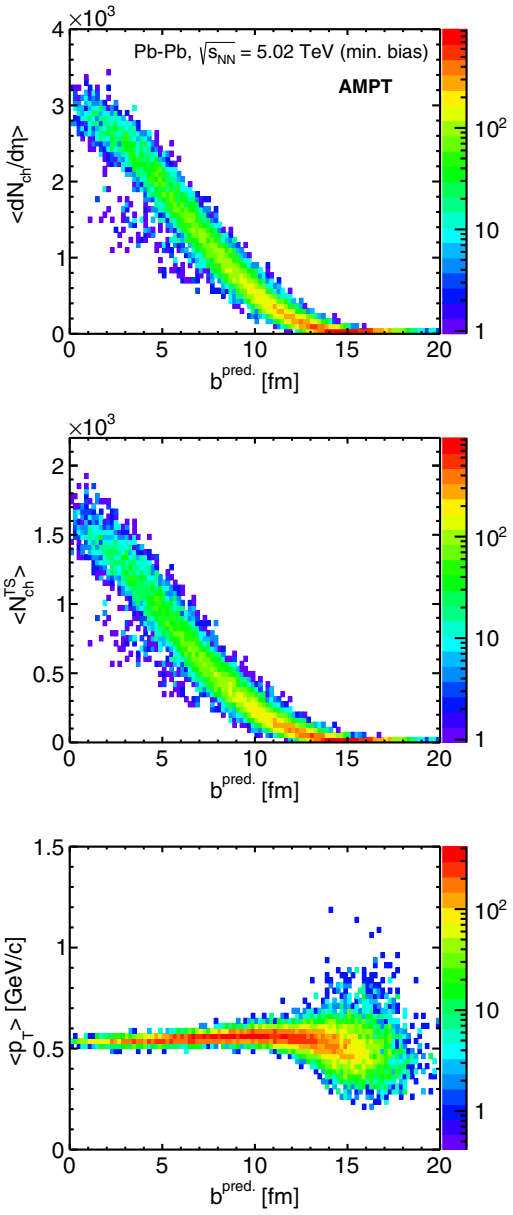

FIG. 4. Correlation plots between each input variable and the predicted value of the impact parameter in Pb-Pb collisions at $\sqrt{s_{\rm NN}} = 5.02$ TeV in the AMPT model.

between each input variable and the predicted value of transverse spherocity in Pb-Pb collisions at $\sqrt{s_{\rm NN}} = 5.02$ TeV in Fig. 5. One could observe that in the top and middle plots, the high-spherocity region is highly correlated with $\langle dN_{\rm ch}/d\eta\rangle$ and $\langle N_{\rm ch}^{\rm TS}\rangle$ of an event. We observe that the events with high spherocity consist of a large number of final state charged particles. However, the low-spherocity region tends to a back-to-back structure and consequently the correlation between transverse spherocity and charged-particle multiplicity decreases. However, in this region, the $\langle p_{\rm T}\rangle$ plays a bigger role as the transverse

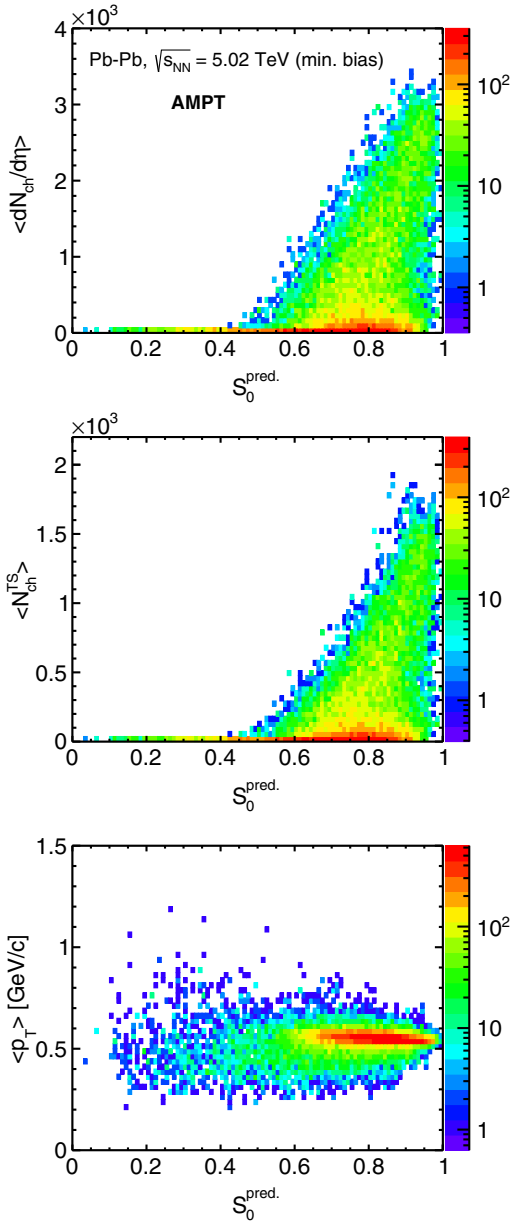

FIG. 5. Correlation plots between each input variable and the predicted value of transverse spherocity in Pb-Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV in AMPT model.

momentum of the produced particles is expected to be high. We have also studied the correlation of transverse spherocity with the leading-transverse momentum of an event and charged-particle multiplicity in the towards and away region. However, their effects are found to be quite negligible in the ML prediction. Thus, we have only considered the shown input variables in Fig. 5 for our present study. Let us now move to the predictions of the transverse spherocity and see how they compare with their true values.

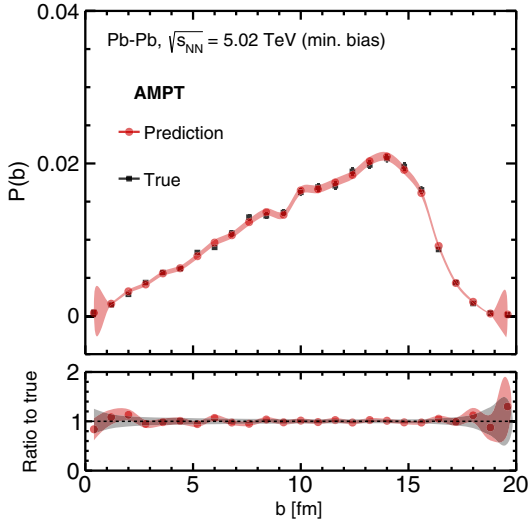

FIG. 6. Predictions for the impact parameter distribution using the gradient boosting decision trees algorithm for Pb-Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV in AMPT model. The quadratic sum of the statistical and systematic uncertainties are shown as a red-colored band for the predicted values. The statistical uncertainties in the true values are shown as bars. In the ratio,the black-colored band denotes the statistical uncertainties in the true values while the red-colored band denotes the quadratic sum of statistical and systematic uncertainties.

Figure 7 shows the predictions for transverse spherocity distribution in Pb-Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV. Here we have also compared the predicted values with the true spherocity distribution obtained from AMPT. One can clearly see that the proposed ML framework with $\langle dN_{\mathrm{ch}}/d\eta\rangle$, $\langle N_{\mathrm{ch}}^{\mathrm{TS}}\rangle$, and $\langle p_{\mathrm{T}}\rangle$ as the input variables predicts the spherocity distribution accurately in Pb-Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV. However, at low-spherocity regions, we see deviation from the true distribution and this could be due to the fact that in heavy-ion collisions the statistics of having events with back-to-back structure are expected to be quite less compared to events that are isotropic in nature. Thus, we believe that this deviation could be due to limited statistics in the low-spherocity region, which can also be seen by the black-colored band in the lower panel. In the bottom plot, we have obtained the estimation of spherocity distribution from the input variables in Pb-Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 2.76$ TeV with the ML training obtained from Pb-Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV. We observe that ML could successfully predict the spherocity distribution at $\sqrt{s_{\mathrm{NN}}} = 2.76$ TeV in wide spherocity ranges. This suggests that the correlation of spherocity distributions with the input variables are quite similar across the LHC energies.

To understand if the proposed algorithm is affected by a particular Monte-Carlo (MC) model, we have used the similar ML algorithm in PYTHIA8 (Angantyr) model in

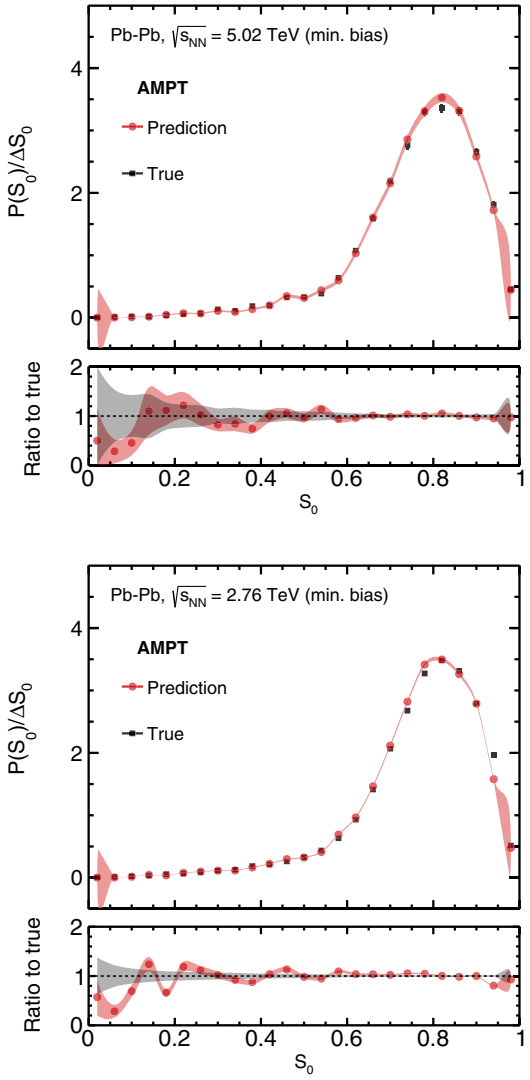

FIG. 7. Predictions for transverse spherocity distribution using ML and their comparison with true values in Pb-Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV (top) and $\sqrt{s_{\mathrm{NN}}} = 2.76$ TeV (bottom) in AMPT model. The lower panels show the ratio of the predicted values to the true values. The quadratic sum of the statistical and systematic uncertainties are shown as a red-colored band for the predicted values. The statistical uncertainties in the true values are shown as bars. In the ratio, the black-colored band denotes the statistical uncertainties in the true values while the red-colored band denotes the quadratic sum of statistical and systematic uncertainties.

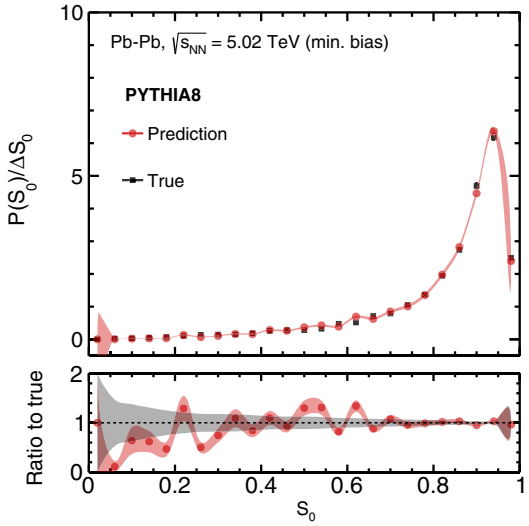

FIG. 8. Predictions for transverse spherocity distribution using gradient boosting decision trees algorithm for Pb-Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV in PYTHIA8 model. The lower panel shows the ratio of the predicted values to the true values. The quadratic sum of the statistical and systematic uncertainties are shown as a red-colored band for the predicted values. The statistical uncertainties in the true values are shown as bars. In the ratio, the black-colored band denotes the statistical uncertainties in the true values while the red-colored band denotes the quadratic sum of statistical and systematic uncertainties.

Fig. 8. As evident in Sec. II, the physics mechanisms in AMPT model and PYTHIA8 (Angantyr) are quite different. However, in Fig. 8, we observe that the predictions for transverse spherocity distribution for Pb-Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV in the PYTHIA8 model is quite accurate compared to the true distribution. After we confirm that the proposed ML algorithm does not have any significant bias due to a particular event generation model, we now move to the predictions of spherocity distribution for different centrality classes in Pb-Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV with the ML training with minimum bias simulated data. Figure 9 shows the predictions of transverse spherocity distributions for (0–10)%, (20–30)%, (40–50)% and (60–70)% centrality classes in Pb-Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV. Here the used input variables are for specific centrality classes but the ML training is from minimum bias Pb-Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV. Also, the predicted results are compared with true spherocity distribution and it is found that for high-spherocity regions, the prediction is quite consistent with the true distribution (evident in the lower panels).

The obtained results from AMPT are quite interesting and encouraging. In the absence of experimental data, the proposed ML algorithm gives an important tool to obtain the impact parameter and spherocity distributions using the available observables from experiments such as final state charged-particle multiplicity and mean transverse momentum. It would be very interesting to see how our results compare with the same from experiments. So, it is quite evident that the current study will act as a baseline for future experimental explorations in this direction.

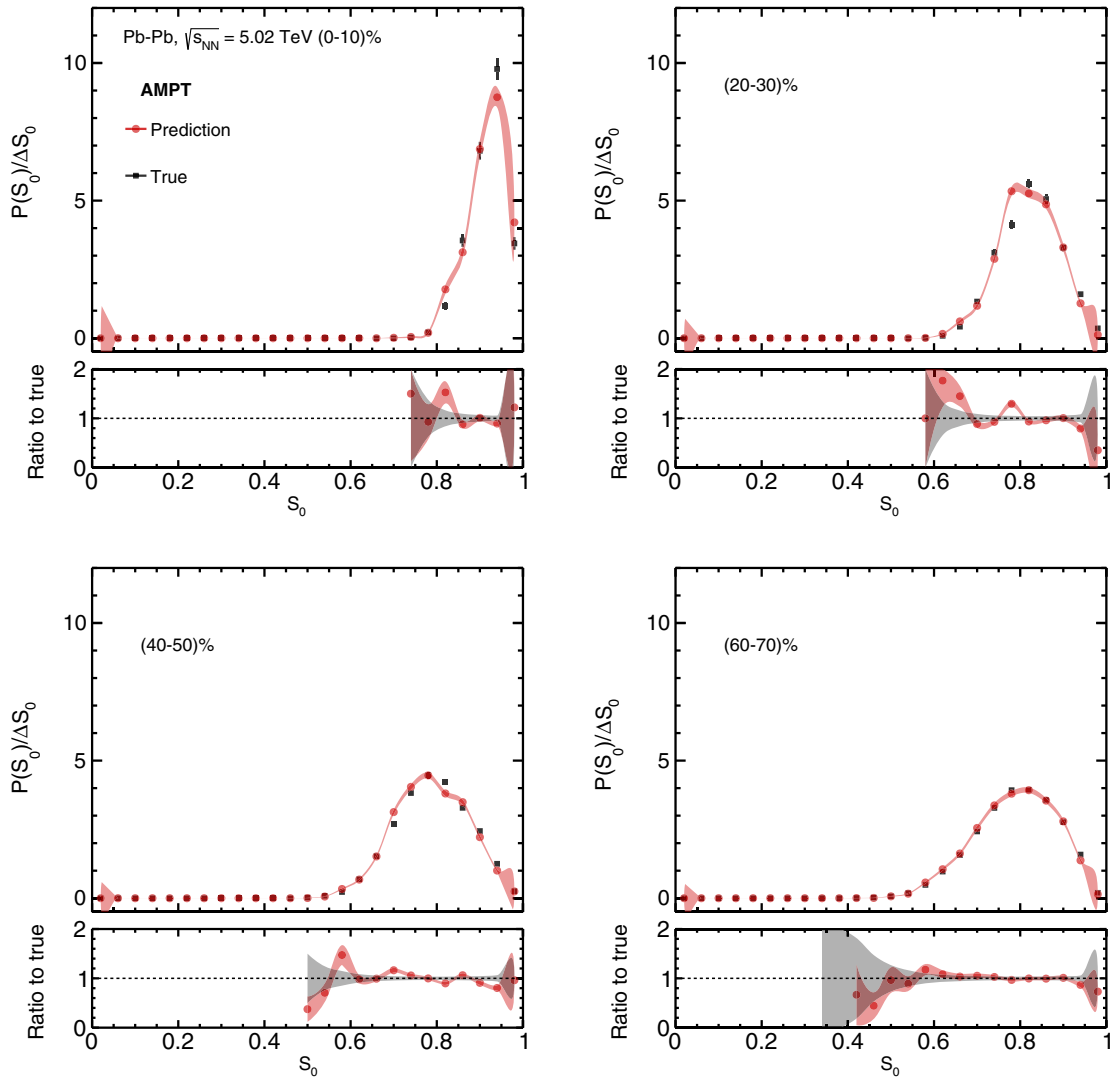

FIG. 9. Predictions of transverse spherocity distributions for different centrality classes in Pb-Pb collisions using the ML (GBDT) model from minimum bias Pb-Pb collisions at $\sqrt{s_{NN}} = 5.02$ TeV in AMPT model. The lower panels show the ratio of the predicted values to the true values. The quadratic sum of the statistical and systematic uncertainties are shown as a red-colored band for the predicted values. The statistical uncertainties in the true values are shown as bars. In the ratio, the black-colored band denotes the statistical uncertainties in the true values while the red-colored band denotes the quadratic sum of statistical and systematic uncertainties.

V. SUMMARY

In summary, we implement the ML-based regression technique via BDTs to obtain a prediction of impact parameter and transverse spherocity in Pb-Pb collisions at the LHC energies using AMPT model. We obtain the predictions for centrality dependent spherocity distributions from the training of minimum bias simulated data and find that the predictions from BDTs based ML technique matches with true simulated data. In the absence of experimental measurements, we propose to implement the machine learning based regression technique to obtain transverse spherocity from the known final state quantities in heavy-ion collisions.

We would like to mention here that the ML-based training with the correlations of input observables using a MC model is quite useful, when the MC model describes the input observables as close as possible to the experimental data. This method will be useful to handle the physics associated with unmeasured quantities in the experiment. In addition, to handle heavy computational problems of central heavy-ion collisions of high-energy experimental data, such a ML-based training using minimum bias data could be used to deal with centrality dependent behavior of observables for a given collision energy and colliding species.

ACKNOWLEDGMENTS

R. S. acknowledges the financial support under the CERN Scientific Associateship and the financial grants

under DAE-BRNS Project No. 58/14/29/2019-BRNS. The authors would like to acknowledge the use of resources of the LHC grid computing facility at VECC, Kolkata. S. T. acknowledges the support from a INFN postdoctoral fellowship in experimental physics. S. T. also acknowledges the discussions related to machine learning tools with Dr. Antonio Ortiz. A. N. M. thanks the Hungarian National Research, Development and Innovation Office (NKFIH) under Contracts No. OTKA K135515, No. K123815, and No. NKFIH 2019-2.1.11-T ET-2019-00078, No. 2019-2.1.11-T ET-2019-00050, and the Wigner GPU Laboratory.

[1] S. A. Bass, M. Gyulassy, H. Stoecker, and W. Greiner, J. Phys. G **25**, R1 (1999).
[2] V. Khachatryan *et al.* (CMS Collaboration), Phys. Lett. B **765**, 193 (2017).
[3] J. Adam *et al.* (ALICE Collaboration), Nat. Phys. **13**, 535 (2017).
[4] E. Cuautle, R. Jimenez, I. Maldonado, A. Ortiz, G. Paic, and E. Perez, arXiv:1404.2372.
[5] A. Ortiz, G. Paic, and E. Cuautle, Nucl. Phys. **A941**, 78 (2015).
[6] G. P. Salam, Eur. Phys. J. C **67**, 637 (2010).
[7] G. Bencdi (ALICE Collaboration), Nucl. Phys. **A982**, 507 (2019).
[8] A. Banfi, G. P. Salam, and G. Zanderighi, J. High Energy Phys. 06 (2010) 038.
[9] A. Khuntia, S. Tripathy, A. Bisht, and R. Sahoo, J. Phys. G **48**, 035102 (2021).
[10] S. Tripathy, A. Bisht, R. Sahoo, A. Khuntia, and P. S. Malavika, Adv. High Energy Phys. **2021**, 8822524 (2021).
[11] S. Tripathy (ALICE Collaboration), J. Phys. Conf. Ser. **1690**, 012126 (2020).
[12] S. Tripathy (ALICE Collaboration), *Proc. Sci.*, ICHEP2020 (2020) 512.
[13] N. Mallick, R. Sahoo, S. Tripathy, and A. Ortiz, J. Phys. G **48**, 045104 (2021).
[14] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Rev. Mod. Phys. **91**, 045002 (2019).
[15] T. J. Armitage, S. T. Kay, and D. J. Barnes, Mon. Not. R. Astron. Soc. **484**, 1526 (2019).
[16] A. Ortiz, A. Paz, J. D. Romo, S. Tripathy, E. A. Zepeda, and I. Bautista, Phys. Rev. D **102**, 076014 (2020).
[17] A. Ortiz and E. Zepeda, arXiv:2101.10274.
[18] A. Radovic, M. Williams, D. Rousseau, M. Kagan, D. Bonacorsi, A. Himmel, A. Aurisano, K. Terao, and T. Wongjirad, Nature (London) **560**, 41 (2018).
[19] K. Albertsson, P. Altoe, D. Anderson, J. Anderson, M. Andrews, J. P. Araque Espinosa, A. Aurisano, L. Basara, A. Bevan, W. Bhimji *et al.*, J. Phys. Conf. Ser. **1085**, 022008 (2018).
[20] HEP ML Community: A Living Review of Machine Learning for Particle Physics, https://iml-wg.github.io/HEPML-LivingReview/.
[21] S. A. Bass, A. Bischoff, C. Hartnack, J. A. Maruhn, J. Reinhardt, H. Stoecker, and W. Greiner, J. Phys. G **20**, L21 (1994).
[22] F. Li, Y. Wang, H. Lü, P. Li, Q. Li, and F. Liu, J. Phys. G **47**, 115104 (2020).
[23] C. David, M. Freslier, and J. Aichelin, Phys. Rev. C **51**, 1453 (1995).
[24] S. A. Bass, A. Bischoff, J. A. Maruhn, H. Stoecker, and W. Greiner, Phys. Rev. C **53**, 2358 (1996).
[25] F. Haddad, K. Hagel, J. Li, N. Mdeiwayeh, J. B. Natowitz, R. Wada, B. Xiao, C. David, M. Freslier, and J. Aichelin, Phys. Rev. C **55**, 1371 (1997).
[26] J. De Sanctis, M. Masotti, M. Bruno, M. D'Agostino, E. Geraci, G. Vannini, and A. Bonasera, J. Phys. G **36**, 015101 (2009).
[27] B. P. Roe, H. J. Yang, J. Zhu, Y. Liu, I. Stancu, and G. McGregor, Nucl. Instrum. Methods Phys. Res., Sect. A **543**, 577 (2005).
[28] Z. W. Lin, C. M. Ko, B. A. Li, B. Zhang, and S. Pal, Phys. Rev. C **72**, 064901 (2005).
[29] C. Bierlich, G. Gustafson, L. Lnnblad, and H. Shah, J. High Energy Phys. 10 (2018) 134.
[30] F. Pedregosa *et al.*, J. Mach. Learn. Res. **12**, 2825 (2011), https://dl.acm.org/doi/10.5555/1953048.2078195.
[31] X. N. Wang and M. Gyulassy, Phys. Rev. D **44**, 3501 (1991).
[32] B. Zhang, Comput. Phys. Commun. **109**, 93 (1998).
[33] Z. w. Lin and C. M. Ko, Phys. Rev. C **65**, 034904 (2002).
[34] Y. He and Z. W. Lin, Phys. Rev. C **96**, 014910 (2017).
[35] B. Li, A. T. Sustich, B. Zhang, and C. M. Ko, Int. J. Mod. Phys. E **10**, 267 (2001).
[36] B. A. Li and C. M. Ko, Phys. Rev. C **52**, 2037 (1995).
[37] V. Greco, C. M. Ko, and P. Levai, Phys. Rev. C **68**, 034904 (2003).
[38] R. J. Fries, B. Muller, C. Nonaka, and S. A. Bass, Phys. Rev. Lett. **90**, 202303 (2003).
[39] R. J. Fries, B. Muller, C. Nonaka, and S. A. Bass, Phys. Rev. C **68**, 044902 (2003).
[40] S. Tripathy, S. De, M. Younus, and R. Sahoo, Phys. Rev. C **98**, 064904 (2018).
[41] C. Loizides, J. Kamin, and D. d'Enterria, Phys. Rev. C **97**, 054910 (2018); **99**, 019901(E) (2019).
[42] T. Sjstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, Comput. Phys. Commun. **191**, 159 (2015).
[43] M. L. Miller, K. Reygers, S. J. Sanders, and P. Steinberg, Annu. Rev. Nucl. Part. Sci. **57**, 205 (2007).
[44] C. Loizides, Phys. Rev. C **94**, 024914 (2016).
[45] R. J. Glauber and G. Matthiae, Nucl. Phys. **B21**, 135 (1970).

[46] C. Y. Wong, *Introduction to High-Energy Heavy-Ion Collisions* (World Scientific, Singapore, 1994).
[47] D. d'Enterria and C. Loizides, arXiv:2011.14909.
[48] P. F. Kolb, U. Heinz, P. Huovinen, K. J. Eskola, and K. Tuominen, Nucl. Phys. **A696**, 197 (2001).
[49] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees* (Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA, 1984), p. 358, https://doi.org/10.1002/cyto.990080516.
[50] J. H. Friedman, Ann. Stat. **29**, 1189 (2001).