CBM performance study of K_S^0 yield using ML techniques for Au+Au collision at p_{beam}=12 $AGeV$

Shantanu Sudhir Bhalerao

Supervisor - Prof.Dr. Hans Rudolf Schmidt
Co-supervisors - Dr. Ilya Selyuzhenkov,
Dr. Andrea Dubla

Thesis presented for the degree of
Master of Science Astro and Particle Physics

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Science
Department of Physics
10^{th} November 2022

Abstract

There is a large scientific interest to explore the phase diagram of strongly interacting matter. The Compressed Baryonic Matter (CBM) experiment at the FAIR facility is devoted to study the phase diagram of strongly interacting matter at high net baryon density and low temperature where critical point and first-order phase transition from hadronic matter to deconfined matter is expected. For the precise measurement of the properties of this deconfined QCD matter, the multi-differential yield measurement of the strange hadrons is required. The enhanced production of the strange hadrons such as K_S^0 is an important probe of the new deconfined state of the QCD matter. Here in this study, the CBM performance for the reconstruction of the K_S^0 particle by its decay to π^+ and π^- are presented. Decay topology reconstruction is furnished by Particle-Finder Simple(PFSimple) package with Machine Learning algorithm for efficient decay reconstruction and to obtain a high signal-to-background ratio. The implemented yield extraction procedure is used for the yield extraction of K_S^0 double differentially(p_T - y_{LAB}).

Declaration of originality

Hereby I confirm, Shantanu Sudhir Bhalerao, Matr. No. 5397178, that this assignment is my own work and that I have only sought and used tools mentioned. I have clearly referenced in the text and the bibliography all sources used in the work (printed sources, internet or any other source), including verbatim citations or paraphrases. I am aware of the fact that plagiarism is an attempt to deceit which, in case of recurrence, can result in a loss of test authorization. Furthermore, I confirm that neither this work nor parts of it have been previously, or concurrently, used as an exam work – neither for other courses nor within other exam processes.

Place and date .... Signature ....

Dedication

To my mother Asha Sudhir Bhalerao and my father Sudhir Shivram Bhalerao

Acknowledgement

Thanks to Prof.Dr.H.R.Schmidt for offering me such an interesting topic.

I would like to thank Dr. Ilya Selyuzhenkov and Dr. Andrea Dubla for their co-supervision, ML meetup at GSI which was like a boost week, suggestions, comments, and decent criticism for my thesis work.

Many many thanks to Shahid Khan who was there always to answer my questions, for inspiring me in many situations, helping in various academical things, for fruitful talks at mensa, and for his support!

I would like to thank Oleksii lubynets for his help in PFSimple-related things, comments and suggestions for my thesis work.

Contents

Chapter 1

Introduction/motivation

1.1 Matter under extreme conditions

Humans are always curious about the questions like, what is the fundamental constituent of the matter? What is the interaction between them?, etc. After numerous discoveries by various scientists today we know that the fundamental particles are quarks, leptons, and bosons. Everyday matter which we see in nature is mostly made up of up and down quarks and electrons. The bosons are the force carrier particles. The four known forces of nature i.e., strong force, weak force, electromagnetic force, and gravitational force describe the interaction of the know particles of the universe with each other. In everyday matter, the strong force is seen in the confinement of the quarks and gluons inside the nucleon and binding of the nucleons by the residual strong force to form a nucleus. The strong force is explained with the help of quantum chromodynamics (QCD) theory. QCD tells about the interaction between the quarks and gluons, which are the constituent part of the hadron. There are three main features of QCD [1][2] 1) Asymptotic freedom - The strength of the coupling constant α_S becomes large for large distances and small for small distances. 2) Colour confinement - No free quark and/or gluon can be found in nature [3]. 3) Chiral symmetry breaking - It is the spontaneous symmetry breaking that gives the mass to a hadron.

In normal nuclear matter, quarks are confined in the hadrons and cannot be seen as free in nature. Asymptotic freedom suggests that quarks behave freely when they come very close to each other. The quarks can be brought close by different methods - 1) By increasing the temperature, the rate of random excitation of the hadrons in the QCD vacuum increases. After some critical temperature T_c they start to overlap and the system is formed where quarks and gluons are free. 2) If the baryons are kept in the box and they are pressed adiabatically they will start to overlap after some critical pressure ρ_c. After pressing further the quarks will come so close that they will move freely. The system formed in both of the cases under these extreme conditions is called the deconfined state of the strongly interacting matter. If this deconfined state is in thermal equilibrium it is called the quark-gluon plasma (QGP).

1.2 Phase diagram of the strongly interacting matter

The different phases of the strongly interacting matter can be seen from the figure 1.1.

The normal nuclear matter exists at low temperatures and moderate baryon chemical potential (≈ 938 MeV), where baryon chemical potential (μ_B) means the difference between the baryon and the anti-baryon in MeV. When the temperature and/or density is increased the hadronic matter transforms into the deconfined state of the QCD matter. According to Lattice QCD (LQCD), the phase transition is expected around 160 MeV temperature [4] from hadronic matter to deconfined matter. These LQCD calculations were done at low μ_B. The color super-conductor phase of the strongly interacting matter is expected at very high μ_B and low temperature. Another phase of strongly interacting matter is predicted which has the properties of both the dense baryonic matter and the quark matter [5][6]. This phase is called quarkyonic matter.

In nature, the QGP state of strongly interacting matter can be accessed using different ways. It is expected that during the early stages of the universe, it was in a QGP state at a very high temperature and zero μ_B. In the current universe, neutron stars are the candidate where high μ_B and low temperature is expected. In the interior of the neutron star, the deconfined state of the QCD matter might exist. Another way to produce a QGP is through the heavy ion collision experiments. The LHC and RHIC experiments trace the QCD phase diagram at low μ_B and very high temperature. At these conditions, a second-order phase transition is expected. In the year 2000, the QGP was first detected at the CERN [8].

The QGP is produced in the laboratory by colliding heavy ion nuclei at very high speed. The different stages of the strongly interacting matter in the heavy ion collision are depicted in the figure 1.2

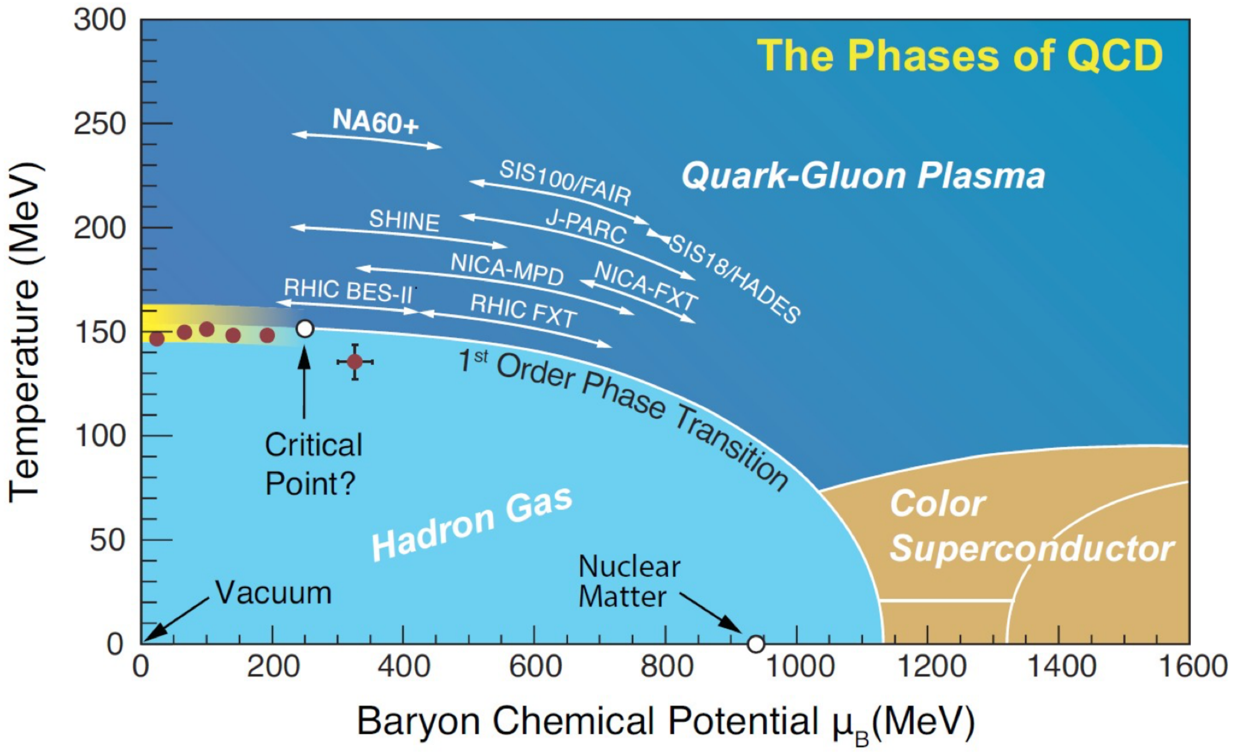


Figure 1.1: QCD phase diagram [7]

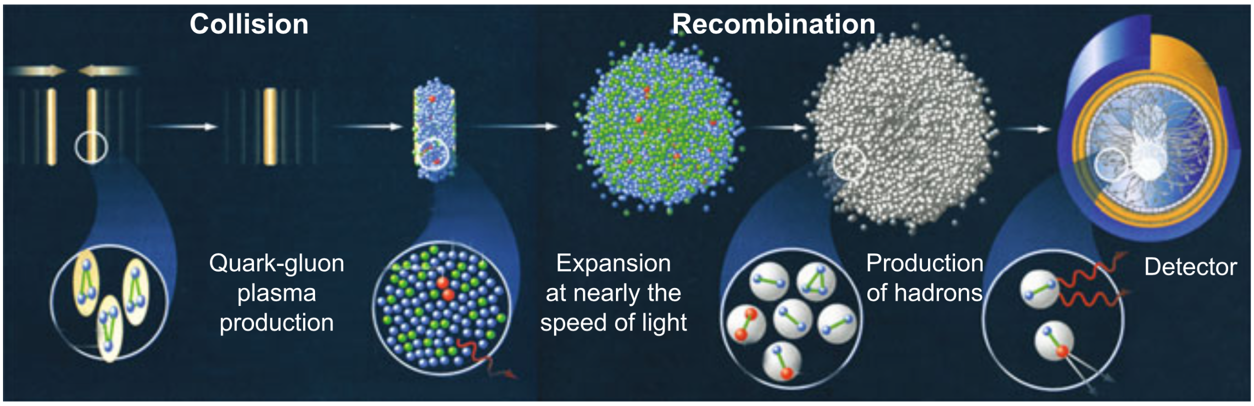

Collision
Recombination
Quark-gluon plasma production
Expansion at nearly the speed of light
Production of hadrons
Detector

Figure 1.2: Evolution of strongly interacting matter in heavy ion collision [9]

- Initial phase: Two Lorentz-contracted heavy ion nuclei collide with each other at relativistic speed.
- QGP: The deconfined state of matter is formed at the central region where the energy density is highest. If the local thermal equilibrium is reached it is called QGP. In the QGP the quarks and gluons interact with each other.
- Hydrodynamic evolution: The created fireball expands almost at the speed of light and cools down thereafter quarks combine into hadrons.
- Detection: After some critical temperature the interaction between hadrons freezes out and they fly toward the detector.

The formation time of the QGP is predicted as 1 fm/c. At the end of the heavy-ion collision, hadrons and other particles are registered in the detector. By using the information of these registered particles the properties of the fireball created in the heavy-ion collision can be investigated. Some of the probes of the dense strongly interacting matter produced in the collision reaction are strangeness enhancement, J/ψ suppression, jet quenching, etc.

The compressed baryonic matter (CBM) experiment at the faculty for anti-proton and ion research (FAIR) is devoted to find the properties of the strongly interacting matter at high μ_B and low temperature.

At such extreme conditions, the 1^{st} order phase transition is expected from the hadronic matter to the deconfined state of the QCD matter. The future CBM experiment will be able to find the answers to the following questions [10] -

1. What is the equation of state of the strongly interacting matter at high μ_B and what are the relevant degree of freedom for these densities?
2. Does the quarkyonic state of QCD matter exist?
3. What is the property of the highly dense baryonic matter? Are the indications of the chiral symmetry restoration accessible?
4. What is the limit of the nuclei chart towards the third dimension (strangeness) by producing single and double strange hypernuclei?

1.3 Strangeness a probe of the new deconfined state of QCD matter

The new deconfined state of strongly interacting matter is produced in the collision of two heavy ion nuclei. The strangeness enhancement is one of the important probe of this new state of matter [11]. In normal nuclear matter, strange quarks are not present. A hadron containing strange quarks will decay weakly into other stable particles. Strange quarks/anti-quarks found in the experiments are made during the heavy ion collision reaction. The yield of the strange particles depends on the density of the deconfined matter formed in the collision.

There are different mechanisms for the production of strangeness [12]. Hadron-hadron collision can produce some strange particles at high collision energies ($\sqrt{s_{NN}}$=700 MeV), ex.,

$$p + p \to p + \Lambda + K^+ \tag{1.1}$$

$$\pi + \pi \to K + \bar{K} \tag{1.2}$$

While in the case of the deconfined state of matter strange particles can be produced at $\sqrt{s_{NN}}$= 300 MeV by the following mechanism

$$g + g \to s + \bar{s} \tag{1.3}$$

$$q + q \to s + \bar{s} \tag{1.4}$$

The equation 1.3 is the dominant production mechanism of the strange quark in quark matter. The strangeness density in the new state of matter is 0.3 fm^{-3} while in the case of hadron gas it is 0.1 fm^{-3} [12]. These predictions suggest that strangeness enhancement is an important probe of the new state of strongly interacting matter.

1.4 K^0_S particle

The figure 1.3 shows the yield the hadrons for Ar+KCl collison at $\sqrt{s_{NN}} = 2.61 GeV$.

Here K^0_S is the 3^{rd} most abundant strange particle produced in the heavy ion collision reaction. Therefore it is worth studying the K^0_S particle.

The main decay channels of the K^0_S particle are as follows [14] -

$$K^0_S \to \pi^+ + \pi^- (69.20 \pm 0.05\%) \tag{1.5}$$

$$K^0_S \to \pi^0 + \pi^0 (30.69 \pm 0.05\%) \tag{1.6}$$

The eqation 1.5 decay channel will be studied here. The mean lifetime of the K^0_S is 8.594×10^{-11} s that corresponds to ct 2.68 cm. The invariant mass of the K^0_S is 0.4976 GeV. Λ is the most abundant strange particle produced by the fireball. Different from the Λ decay, the K^0_S decays symmetrically into two soft pions, which allows an additional probe performance of the tracking for the secondary reconstruction.

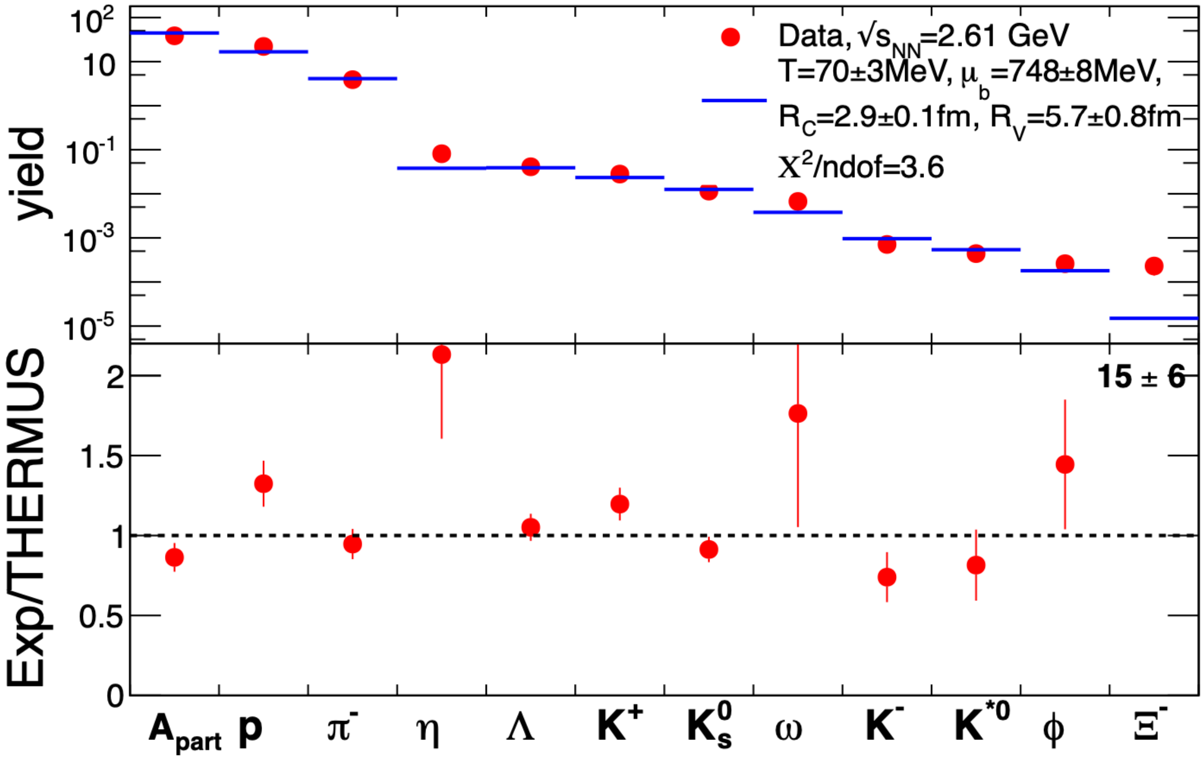


Figure 1.3: Yield of hadrons in Ar+KCl collision at $\sqrt{s_{NN}} = 2.61GeV$ [13]

Chapter 2

The CBM experiment

2.1 CBM experiment

The FAIR is envisaged by the international science community and GSI laboratory. The goals of the FAIR is to do a multifaceted science program, with beams from stable and unstable nuclei as well as antiproton at wide range of intensities and energies [15]. The skecth for the FAIR facility and GSI laboratry is shown in figure 2.1.

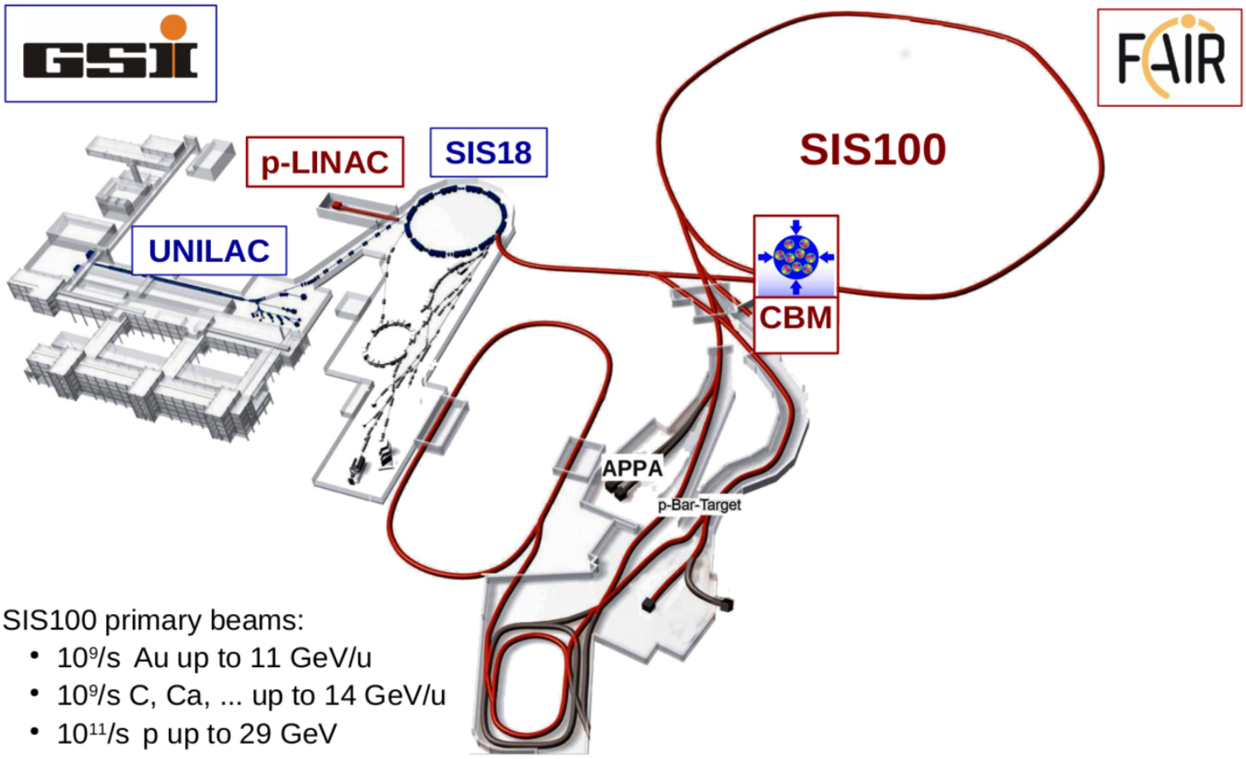

GSI
FAIR
p-LINAC
SIS18
UNILAC
SIS100
CBM
APPA
p-Bar-Target
SIS100 primary beams:
• 10^9/s Au up to 11 GeV/u
• 10^9/s C, Ca, ... up to 14 GeV/u
• 10^{11}/s p up to 29 GeV

Figure 2.1: FAIR together with existing GSI facilities [16]

FAIR comprises of a synchrotron (SIS 100), different rings for collecting and storage of the beams, different experiments for nuclear physics, Astrophysics, Plasma physics, and Atomic physics purpose [17].

The CBM experiment is designed to investigate the strongly interacting matter at very high baryonic densities. Such a densities are expected to exist in the core of the neutron stars [10]. The goal of the CBM experiment is to measure the yield, spectra, collective flow, event-by-event fluctuation, and correlation of the different hadrons, electrons, and muons produced in the heavy-ion collision at FAIR beam energy range with the unprecedented precision and statistics for different collision scenarios (i.e., $A + A$, $A + p$, $p + p$ collisions) [18].

2.2 CBM detectors

For the diagnosis of the rare probes of the deconfined matter, the CBM experiment will operate with the event rates up to 10^7 Au+Au collisions per second to gather enough statistics. Therefore for the measurements fast and radiation hard detectors are required [17]. The figure 2.2 shows the different detectors of the CBM experiment.

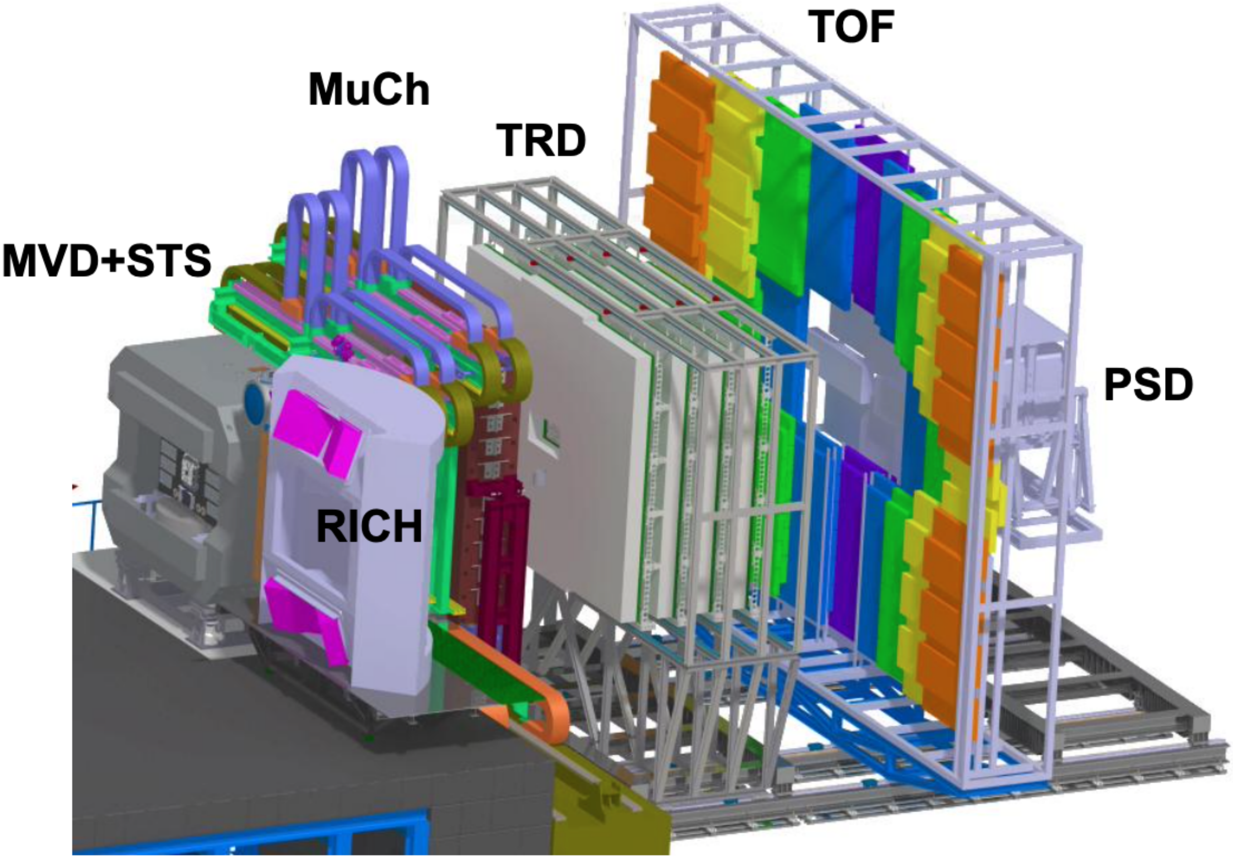


Figure 2.2: The detector setup of the CBM experiment [16]

- Dipole magnet - The tracking detectors will be placed inside a superconductive 1 Tm dipole magnet.
- Micro-Vertex Detector (MVD) - It will be used for the track reconstruction of the different particles. One of the possible application of MVD is the identification of the D mesons with the high vertex resolution (50 - 100 μm) and to reduce the combinatorial background in the case of electron measurements. The MVD will be placed inside the dipole magnet. The MVD uses monolithic active pixel sensors (MAPS). 3-4 layers of the MAPS will be placed downstream of the target up to 20 cm.
- Silicon Tracking System (STS) - It will be used for the track reconstruction and momentum determination of the particles. For this purpose, it will be placed inside the dipole magnet. In the current version, STS consists of 8 different silicon detector layers. They will be located downstream of the target from 30 cm to 100 cm.
- Ring Imaging CHerenkov detector (RICH) - It detects the electron using the Cherenkov radiation technique. It will also be used for the suppression of the pions of momentum less than 8 GeV/c^2.
- Muon Chamber system (MuCh) - MuCh is used for the detection of muons.
- Transition Radiation Detector (TRD) - TRD will be used for the tracking and identification of the electron and positron. It consists of 3 detector layers which will be situated at 5 m, 7.2 m, and 9.5 m downstream of the target.
- Time of flight (TOF) detector - It is used for the identification of the hadrons using TOF information. It is located 10 m downstream of the target.

- Electromagnetic CALorimeter (ECAL) - It is a "shashlik" type of calorimeter that will be used for the measurement of the direct and indirect photons produced in the heavy ion collison reaction.
- Projectile Spectator Detector(PSD) - It will be used for the detection of the centrality and the orientation of the reaction plane.

The TOF, STS and MVD are the important detectors for the short-lived particles. The tracks of the daughter particle is reconstructed using the MVD and STS detectors. The particle momentum is calculated using the STS information and identified using the TOF detector.

Chapter 3

Reconstruction of K_S^0 using PFSimple package

3.1 Data simulation

Transport theory describes the microscopic dynamics description of heavy ion collisions. UrQMD [19] is a microscopic transport model which is based on phase space description of the heavy ion collision reaction. DCM-QGSM-SMM [20] is another transport model which can simulate the product of heavy ion collision reaction from the hundred MeV to hundred GeV energy range.

For the analysis in this study, the data is simulated from Au+Au collision containing $\approx$ 2M events for UrQMD and $\approx$ 5M events for DCM-QGSM-SMM model at p_{beam}=12 $AGeV$ ($\sqrt{s_{NN}}$= 4.93 GeV), minimum bias. The Au+Au collision produces different particles, these particles are passed through the CBM detector setup using GEANT4 [21] transport engine to register the response of each detector to these particles. The produced charge particles generate hits in the tracking detector along the trajectory of the particle. Cellular Automaton (CA) [22] track finder is used to reconstruct the trajectories (tracks) of the charged particles. The CA method produces tracklets using the hits in the neighboring detectors and links them to build the track candidates.

Here in this study, the DCM-QGSM-SMM model is taken as pure signal therefore simulated data. By excluding the signal in the 5σ range of invariant mass of the K_S^0 particle the UrQMD model data can be treated as background for further study.

3.2 Reconstruction of K_S^0 using PFSimple package

K_S^0 is a short-lived particle, which means it decays before or short within the tracking detector. The decay length of the K_S^0 particle is 2.68 cm [14]. Therefore K_S^0 can be reconstructed only indirectly using its decay products. The main decay channel of the K_S^0 is shown in the equation 3.1

$$K_S^0 \rightarrow \pi^+ + \pi^- \qquad (69.20 \pm 0.05\%) \tag{3.1}$$

The negative and positive pion from K_S^0 decay produces tracks in MVD and STS detectors. All negative and positive charged particle tracks are combined to make the candidate for the K_S^0. Reconstructed and Monte Carlo tracks are matched using the matching algorithm (in GEANT4 [21] simulation). Only those K_S^0 candidates which come from true K_S^0 decay (by applying the cut, MC = 1) are termed as a signal and other candidates are called background (by applying the cut, MC = 0).

Figure 3.1 shows the cartoon of topological variables.

There are mainly two types of tracks - 1) Primary track - These tracks are made by the particles produced during the collision of two heavy ion nuclei at relativistic speed. 2) Secondary tracks - The secondary particles are the decay products of the primary particles. Secondary tracks are made by secondary particles. Primary tracks overlap with the primary vertex (PV) within the errors while the secondary tracks do not. Therefore to distinguish primary tracks and secondary tracks χ^2 criterion is used [23]. Where χ^2 is the distance between two tracks or a track and a vertex normalized on their total errors.

KF Particle package [24] uses the Kalman Filter method for the complete reconstruction of short-lived particles. PFSimple package [25] is the simplified version of the KF Particle package based on KF Particle mathematics. PFSimple package takes track information as input and outputs candidates kinematics and topological variables.

The topological variables given by PFSimple are as follows -

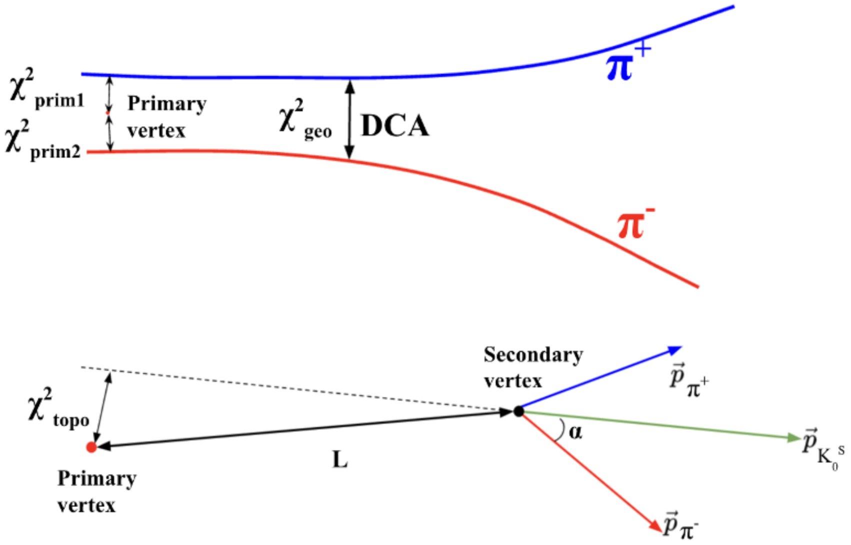


Figure 3.1: Pion tracks in tracking detector a) Variables associated with the decay tracks of K^0_S candidate. b) Variables associated with the momentum vector

- Distance of the Closest Aproach (DCA) - The DCA (cm) between π^+ and π^- track candidates
- χ^2_{geo} - Squared distance between daughter tracks divided by its error
- $\chi^2_{prim\pi+}$ - Squared distance between PV (collison point) and π^+ track divided by its error
- $\chi^2_{prim\pi-}$ - Squared distance between PV and π^- track divided by its error
- L - Distance between PV and secondary vertex (SV)
- L/ΔL - L divided by its error
- χ^2_{topo} - Squared distance between PV and the extrapolated trajectory of K^0_S divided by its error
- cosine second - cosine of the angle between $\overrightarrow{P}_{K^0_S}$ and $\overrightarrow{P}_{\pi^+}$
- cosine first - cosine of the angle between $\overrightarrow{P}_{K^0_S}$ and $\overrightarrow{P}_{\pi^-}$
- cosine topological - cosine of the angle between $\overrightarrow{P}_{K^0_S}$ and the line joining PV to SV

PFSimple uses manual selection criteria for maximizing the signal-to-background ratio which is dependent on the collision energy, decay channel, and detector configuration. While with the help of a machine learning model, the selection can be done in a non-linear fashion for multi-dimensional phase space in an automatized way and different data scenarios i.e., different collision energies, p_T, y, etc.

3.3 Selection of variables for training the model

A set of preselection criteria for topological variables is used to remove the numerical artifacts from K^0_S candidates. The numerical values of the preselection criteria of the topological variables are shown in the table 3.1

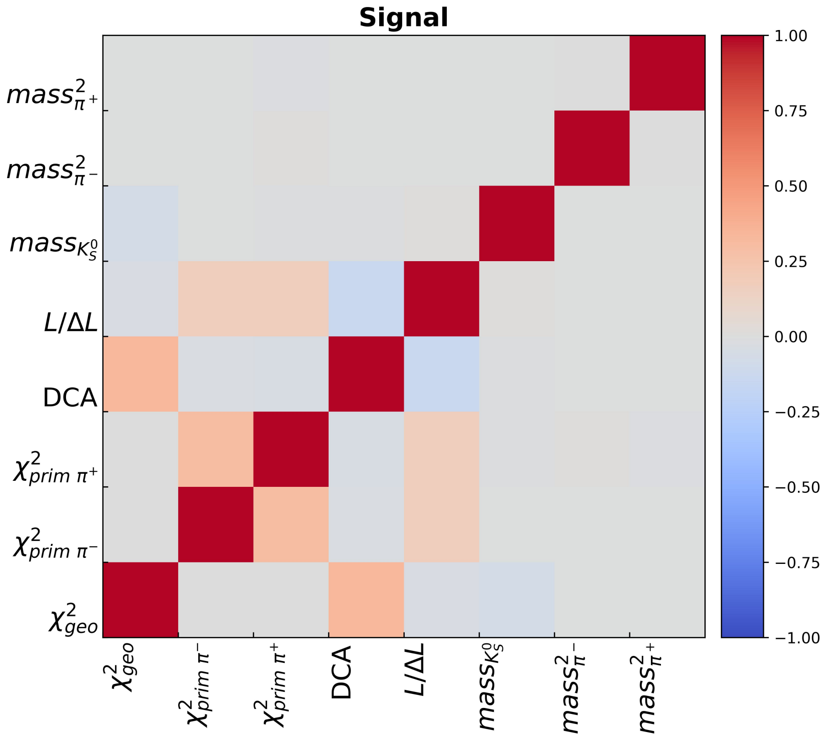
Parameter	χ^2_{geo}	χ^2_{topo}	χ^2_{prim}	L/ΔL	DCA(cm)
Selection criterion	$0 - 10^3$	$0 - 3 \times 10^5$	$0 - 3 \times 10^8$	(-25) - 15000	0 - 100

Table 3.1: PFSimple Quality assurance cuts for K^0_S candidate

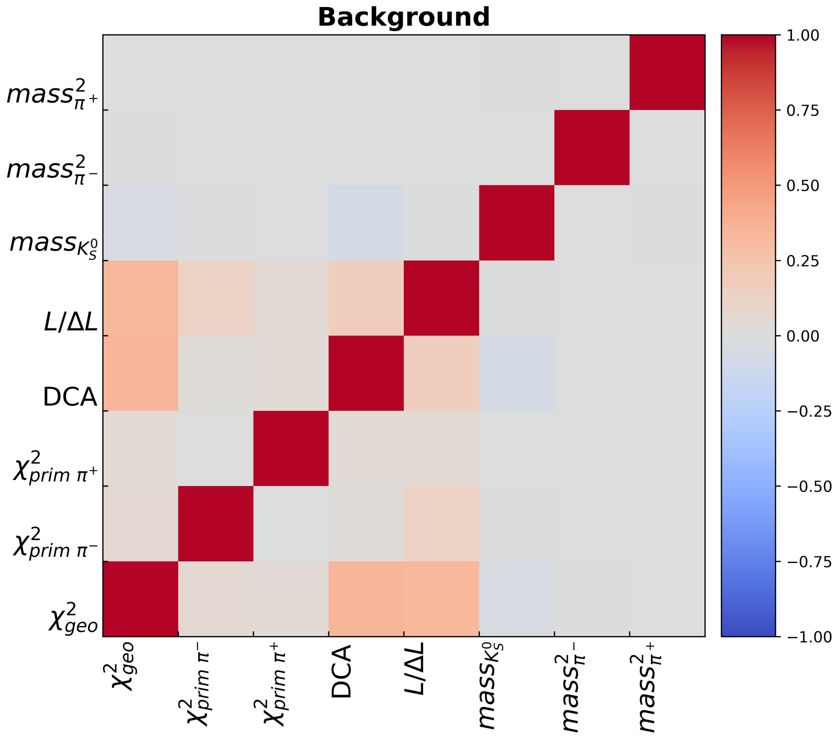
The Pearson correlation coefficient describes the linear correlation between the two variables [26]. It is the ratio of the covariance between two variables and the product of their standard deviations. The Pearson correlation coefficient has a value between -1 to 1, where +1 suggests that there is a linear dependency between the two variables when the value of X increases the value of Y also increases, and vice versa for -1. The formula for the Pearson correlation coefficient is given in the equation 3.2

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{3.2}$$

The Pearson correlation for all the variable of K^0_S is shown in figure 3.2.



(a) DCM-QGSM-SMM



(b) UrQMD

Figure 3.2: Pearson correlation coefficient plot

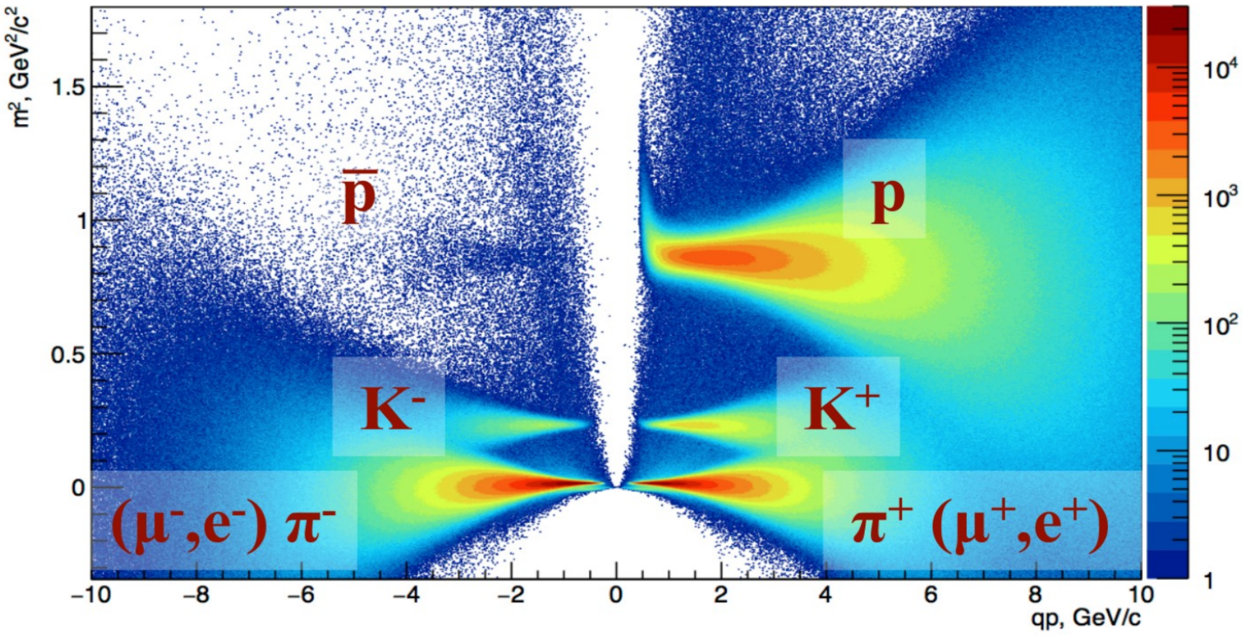


Figure 3.3: Distribution of m^2 versus charge times momentum for Au+Au collison at $p_{beam} = 10\ AGeV/c$ [16]

All negative and positive charge particle tracks are combined to build the K_S^0 candidates. The m^2 variable information is important in differentiating between the candidates as seen from the figure 3.3, the m^2 value for π is less than the other major charge particles. $m^2_{\pi+}$ is the m^2 of the positive charge particle and considered as π^+ particle while $m^2_{\pi-}$ is the m^2 of the negative charged particle and it is considered as π^- particle. The m^2 variable is calculated using the TOF information. Where p is calculated using track of the particle and β is calculated using TOF information. The formula for m^2 is as follows -

$$m^2 = p^2 \left(\frac{1}{\beta^2} - 1 \right) \tag{3.3}$$

Here the goal is to reconstruct the K_S^0 candidates properly and therefore plot the invariant mass spectra of K_S^0 candidates at the end. When the K_S^0 mass variable is added in the training of the machine learning algorithm, it greedily picks this variable and mostly uses it for the prediction of the particular K_S^0 candidates, so the model gets biased and produces the false peaks under the K_S^0 invariant mass peak . Therefore the topological variables which have an insignificant correlation with the mass variable of K_S^0 candidates are chosen to train the model. The variables used for training the model are - DCA, χ^2_{geo}, $\chi^2_{prim\pi+}$, $\chi^2_{prim\pi-}$, L/ΔL, $m^2_{\pi+}$, $m^2_{\pi-}$.

Chapter 4

Optimization of selection criteria for K_S^0 using machine learning

4.1 Introduction to machine learning

Different topological and kinematical variables are associated with the daughter track of the K_S^0 particle. The optimization of the selection criteria is required for the suppression of the combinatorial background. Conventionally the selection is done linearly using the box cuts on the different variables of the daughter track. These box cuts depend on the collision energy, multiplicity, centrality, etc., therefore the conventional optimization of selection criteria is computationally expensive and manually laborious work. Machines are thought to be the best solution to solve this cognitive problem. Machine learning algorithms can learn hidden patterns in the data and therefore optimization of the selection criteria can be done in a non-linear fashion by an automatized way on multi-dimensional phase space. Machine learning (ML) is the subfield of computer science that gives the computer the capability to learn from the data without being explicitly programmed. Therefore different models can be built using ML algorithms for various data scenarios such as multiplicity, centrality, collision energy, etc., to optimize the selection criteria of K_S^0 and the work can be done smartly. ML algorithms are used in various fields such as email filtering, image detection, language translator, etc.

There are three main types of machine learning [27]. 1) Supervised machine learning - In a supervised machine learning algorithm, the sample data is given to the machine that contains various features (variables represented as X) and the correct output value (represented as y). The algorithm then learns the pattern and rules in the data. After learning the patterns and the rules it produces the model that predicts the output value. As the features and the correct output value is known to the machine, the sample dataset in supervised machine learning is called a labeled dataset. The different types of supervised machine learning are regression, decision trees, k-nearest neighbors, etc. Supervised machine learning is then subdivided into two types a) Classification - Where the model predicts the type/class of the target variable. b) Regression - Where the model predicts a continuous value or a number. 2) Unsupervised Machine learning - In the case of unsupervised machine learning, an unlabeled dataset is given to the machine. Here the primary goal is to discover the hidden patterns in the dataset. One example of unsupervised machine learning is k-means clustering. In k-means clustering the data points are clustered to find the new labels (features) in the dataset. 3) Reinforcement machine learning - In supervised and unsupervised machine learning algorithms, an undetermined endpoint is reached after generating the model with training and test data. In the reinforcement machine learning algorithm, the model is fixed to train by continuous learning. Positive or negative feedback improves the model in each iteration of continuous training. Q-learning is an example of reinforcement machine learning.

Various types of ML algorithms can be used both for regression and classification. A decision tree (DT) algorithm is one of them. DT is the sequential model that unites the basic logical tests efficiently and cohesively, where a numerical attribute is compared to the threshold value in each test [28][29]. The advantage of the DT over "black-box" models such as neural networks is its lucidity. The logical rules in the DT are much easier to understand than the numerical weight given to a connection of each node in the neural networks. DT uses greedy search through all possible combinations of the logical rules to find the best split to divide the instances into a separate class, forming a tree. Entropy is the measure of the randomness or the impurity of the dataset. The best split is chosen to get the greatest gain/minimum entropy at the particular node.

4.2 XGBoost

The learning algorithms whose performance we wish to boost are called weak learners [30]. Boosting is the ensemble technique where weak learners are joined together consecutively to produce a strong learner. The single DT can perform well than the random guess but its performance can be enhanced using the boosting technique. In gradient boosting machine (GBM) algorithm the strong learner is built by back-fitting and non-parametric regeression [31]. The functions which are used to minimize the difference between the predicted value (by model) and the actual value (data) are called loss functions. GBM exploits the gradient descent method to minimize the loss function of the initial model and thus a more precise model is built. The eXtreme Gradient Boosting (XGBoost) [32] is the advanced version of gradient boosting where it not only uses the gradient of the loss function but its second-order derivative is used for the better estimation of the steps towards the minimum of the loss function.

XGBoost is a highly scalable end-to-end tree-boosting system, which means it can be used for smaller as well as larger datasets in distributed processing frameworks. It uses a gradient tree-boosting machine-learning algorithm. XGBoost was used by many machine learning competition-winning teams. XGBoost uses decision trees in a sequential manner. The decison trees in XGBoost are called as boosted decision tree (BDT). Various regularizing parameters (i.e., alpha, gamma, lambda, etc.) are used in XGBoost to avoid overfitting. Regularizing parameters are applied to each BDT, therefore each BDT learns a small portion of the pattern in the data and gives the prediction. The wrong prediction of BDT is corrected by successive BDTs.

There are various advantages of using XGBoost such as: 1) It uses the sparsity-aware algorithm. This algorithm learns the preferred direction of sparse data (missing value entries) during BDT learning. 2) It is scalable, and portable and also works on distributed processing frameworks 3) It has package implementation for Python, R, C++, Java, Scala, Julia, and Perl 4) It uses exact greedy algorithm, which tries to find the best split on all features by enumerating over all the possible splits.

The XGBoost model is chosen for the reconstruction of K^0_S candidates in this study. The K^0_S candidate is either signal or background therefore the classification is binary.

4.3 Hipe4ml

Hipe4ml [33] is the minimal Heavy ion physics environment for machine learning. It has different classes. ModelHandler class is used to handle the ML classifier, build the model, dump the model, etc. TreeHandler class is used for storing and managing the data i.e., .root file, Pandas Dataframe, etc. Other classes are used for analysis and plotting purposes. Here in this study Hipe4ml package is used.

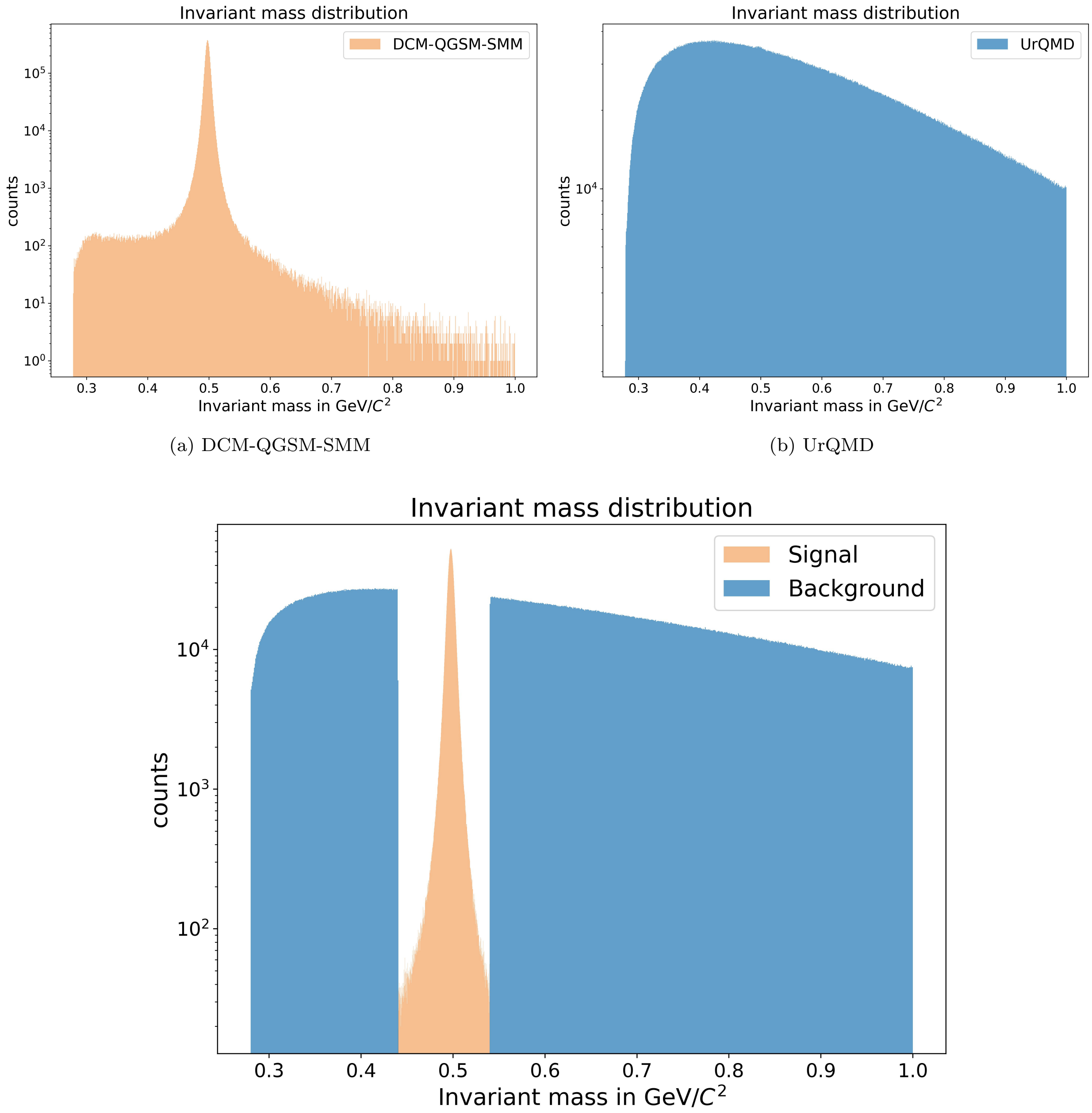
The Optuna is the next generation optimization softeware [34]. Optuna package is used for the optimization of the hyperparameters of the XGBoost model in the hipe4ml environment.

4.4 Selection of the data for training and testing the model

The UrQMD model for 1M events of Au+Au collision, produces some ten thousand of the K^0_S candidates and only $\approx$ 6,000 of them were detected by the CBM detector setup. The K^0_S particles produced in the real data are very less compared to the combinatorial background and therefore the signal is under-represented in this data sample. Therefore to train and test the XGBoost model the data is oversampled by increasing the signal-to-background ratio. Signal is taken from DCM-QGSM-SMM model data in the 5σ region of the invariant mass of K^0_S i.e., $0.44 <$ invariant mass < 0.54 (in Gev/c^2). UrQMD model data can be treated as background by applying MC $= 0$ cut and excluding the candidates in the 5σ region of the invariant mass of K^0_S.

The decay channels of the K^0_S particle is very well known therefore microscopic transport models can describe the K^0_S signal to a very good accuracy but the combinatorial background is a random process and therefore hard to simulate using the transport models. When the real data will come from the CBM experiment, the background will be taken from it, but as currently no real data is available the UrQMD model is treated as real data (signal+background). The background is taken from the UrQMD model to train the XGBoost model. As the combinatorial background is taken from the real data the ML optimization procedure is partially dependent on the simulated data rather than fully dependent.

The invariant mass spectra of DCM-QGSM-SMM, UrQMD, and train-test data is shown in figure 4.1 After selecting the data (figure 4.1c) for training the XGBoost model, it is divided into two samples, 50% of the candidates are used for training the XGBoost model and 50% of the candidates are used for testing the XGBoost model.



(a) DCM-QGSM-SMM

(b) UrQMD

(c) train-test data

Figure 4.1: Invariant mass spectra for Au+Au at p_{beam} = 12 $AGeV$ a) Only true K^0_S candidates produced by DCM-QGSM-SMM model b) All K^0_S candidates produced by UrQMD model c) K^0_S candidates chosen from DCM-QGSM-SMM and UrQMD model for training and testing the XGBoost model.

4.5 Evaluation of models performance

The XGBoost model is trained using the train data and the training variables. To check the bias and variance of the XGBoost model different evaluation methods are used, some of them are BDT output distribution, and ROC-AUC. The feature importance of the XGBoost model is shown using the SHAP plots.

4.5.1 Evaluation using BDT output distribution

In the XGBoost model, the final prediction for a particular candidate is given by the vote from all the BDTs. The average of all the votes is called as BDT output score. A candidate is given a score between 0 and 1 dependent on its feature values. Candidates which are more signal-like are placed near 1 and the candidates which are more background-like will be placed near 0. The trained XGBoost model is applied on the train and test data. Figure 4.2 shows the BDT output distribution for the train and test data sample. The red

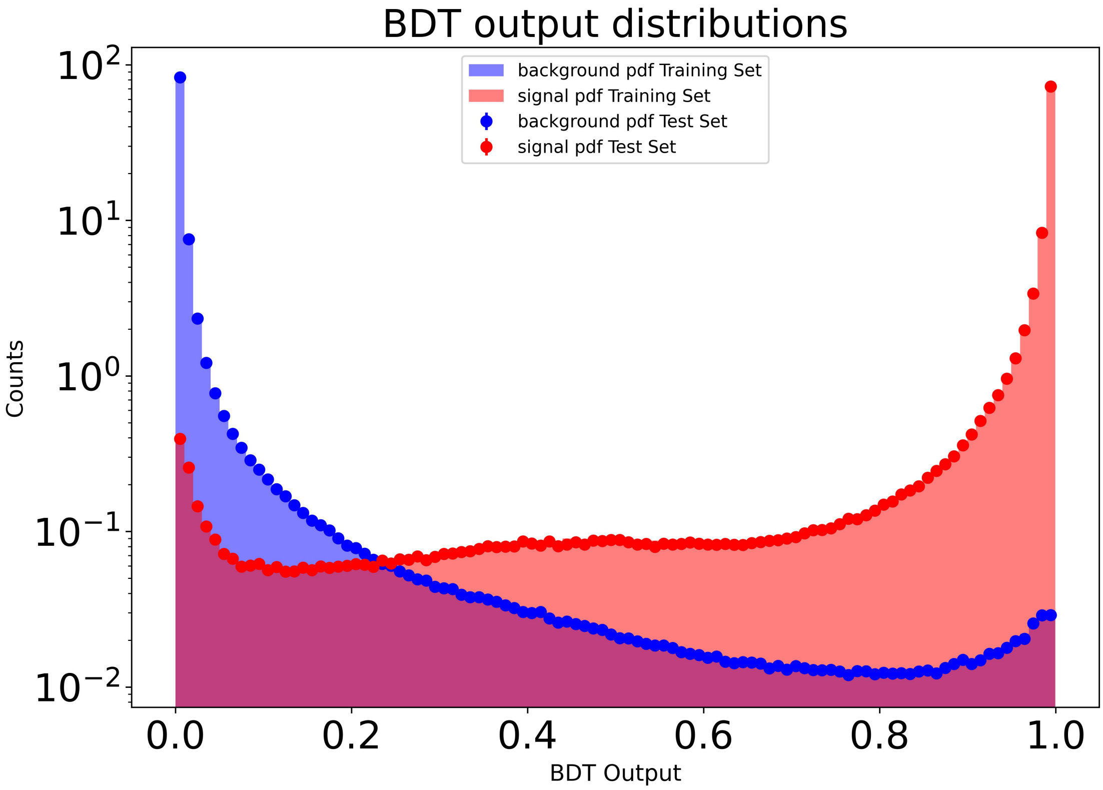
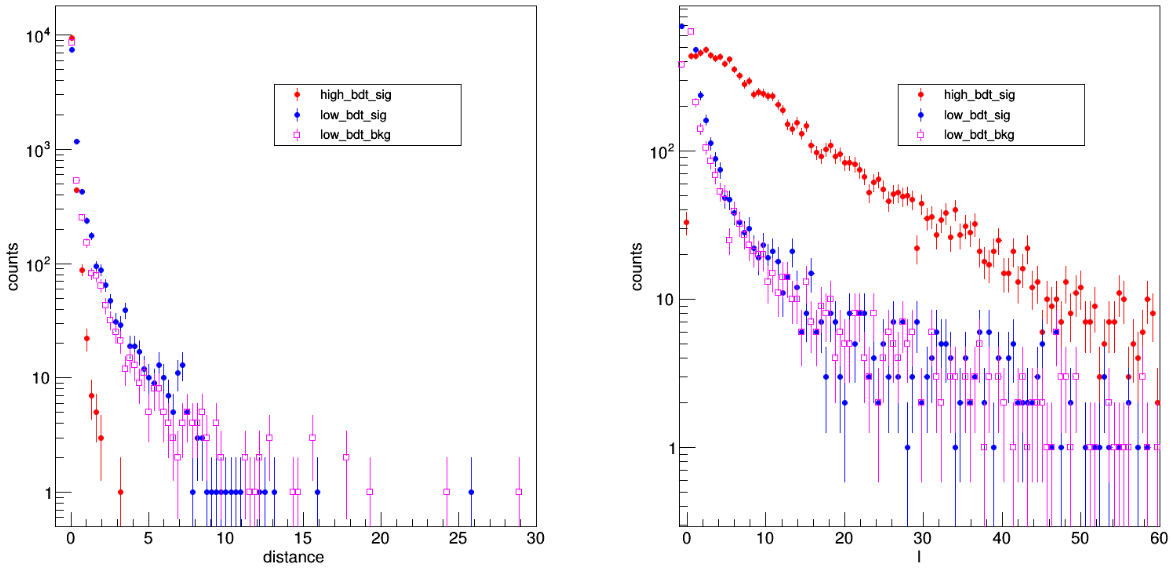


Figure 4.2: Trained XGBoost model applied on train and test data

color in the figure shows the pdf of the true signal candidates while the blue color shows the pdf of the true background candidates. Two peaks are seen at 0 and 1. The K_S^0 candidates which peak at 0 are mostly background candidates and the K_S^0 candidates that peak at 1 are mostly signal candidates. This shows that the XGBoost model can successfully classify most of the signal candidates.



(a) DCA distribution for different bdt samples

(b) L distribution for different bdt samples

Figure 4.3: Topological variable distribution plot for different bdt samples

Some signal candidates peak at 0, which is investigated using the following procedure. The train test data is divided into three data samples according to their BDT output score and MC information - a) Low bdt signal - this data sample contains all signal candidates (MC = 1) in BDT output range 0 - 0.1 b) High bdt signal - this data sample contains all signal candidates (MC = 1) in BDT output range 0.9 - 1 c) Low bdt background - this data sample contains all background candidates (MC = 0) in the BDT output range 0 - 0.1. These three data samples are taken for further investigation. The plots of DCA and L variables for these three data samples are shown in figure 4.3. Figure 4.3b shows that the distance between PV and SV is low for the low bdt signal candidates, which means low bdt signal candidates decay near the PV. As the low bdt signal candidate decay near the PV, its daughters will bend more in the magnetic field and have higher DCA. Therefore low bdt signal candidates have high DCA than the high bdt signal candidates as seen from figure 4.3a. This confirms that some K_S^0 candidates decay near PV and therefore their SV cannot be resolved properly, hence the XGBoost model predict these low bdt signal candidates as background i.e., near to 0 BDT output score. The plots of the other topological variables for these three bdt-based data samples are given in appendix A.

4.5.2 Evaluation using ROC-AUC

Receiver Operating Characteristics Area Under the Curve(ROC-AUC) [35] is used for the performance assessment of the ML model. The ROC plot is a two-dimensional plot of the true positive rate (TPR) versus the false positive rate (FPR). The TPR and FPR rate can be defined as:

$$\text{True Positive Rate} = \frac{\text{S classified as S}}{\text{S classified as S + S classified as B}} \tag{4.1}$$

$$\text{False Positive Rate} \ = \frac{\text{B classified as S}}{\text{B classified as S + B classified as B}} \tag{4.2}$$

Here, S: signal and B: background. The TPR tells how many true signal candidates are classified a signal candidate (based on a selection on a particular BDT ouput score) by the XGBoost model among all the true signal candidates. And FPR tells about how many true background candidates are classified as a signal (based on a selection on a particular BDT ouput score) by the XGBoost model among all the true background candidates. The ROC plot is plotted for all thresholds (i.e., 0 to 1 BDT output score). The ROC-AUC is the

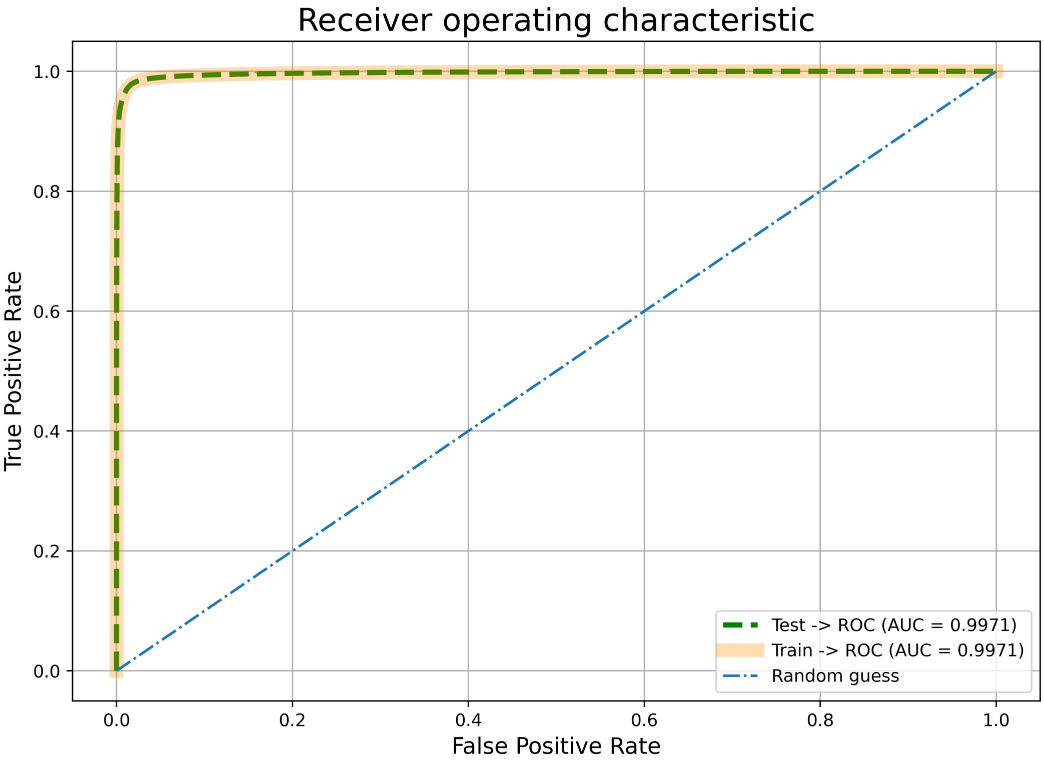


Figure 4.4: ROC plot for train and test data

aggregate measure of the performance of the model for all thresholds therefore it is threshold independent. Different models/algorithms performance can be evaluated using ROC-AUC. An ideal model will have TPR = 1 and FPR = 0.

The ROC plot of the trained XGBoost model applied on the train and test data samples is shown in the figure 4.4. The ROC-AUC is 0.9971 for both the train and test data samples, which shows that the XGBoost model is not over-trained on the train data as it performs the same on the test data. The random guess line is seen in the figure, which tells about the 50% - 50% probability of a given candidate being signal or background.

4.5.3 Evaluation using SHAP plot

XGBoost model is a complex model containing hundreds of BDTs. The prediction for a particular candidate given by BDTs cannot be interpreted easily by looking at the structure of the XGBoost model. SHAP (SHapley Additive exPlanations) [36] plots give insight into how the model was built, and which variables were most used in building the XGBoost model. In the SHAP technique, an explanatory model is built, which is the simplification of the original complex model. The prediction of the explanatory model should match the prediction of the original model. The SHAP value is the contribution of the individual variable in the prediction by the XGBoost model. SHAP value is a unified measure of the feature (variable) importance.

The figure 4.5 shows the feature importance plot for the trained XGBoost model. The variables are ranked according to their contribution to the prediction of the result by the XGBoost model. The highest ranked variable is placed at the top in the SHAP plot, which is $\chi^2_{prim\pi+}$. The blue dots are the low value and the red dots are the high values in the respective variable. The χ^2_{prim} criterion tells if a particular candidate overlaps with the PV within the errors or not, therefore the tracks of the secondary particles will have a high χ^2_{prim} value. Hence in figure 4.5, the red points of $\chi^2_{prim\pi+}$ are given the high positive SHAP value which contributes to the signal. Sometimes color can be misleading as some outlier candidates can have very high or very small feature values relative to the normal distribution of the feature values as seen from appendix D. The two bolbs for the $\chi^2_{prim\pi+}$ variable int the SHAP plots are explained in the appendix E.

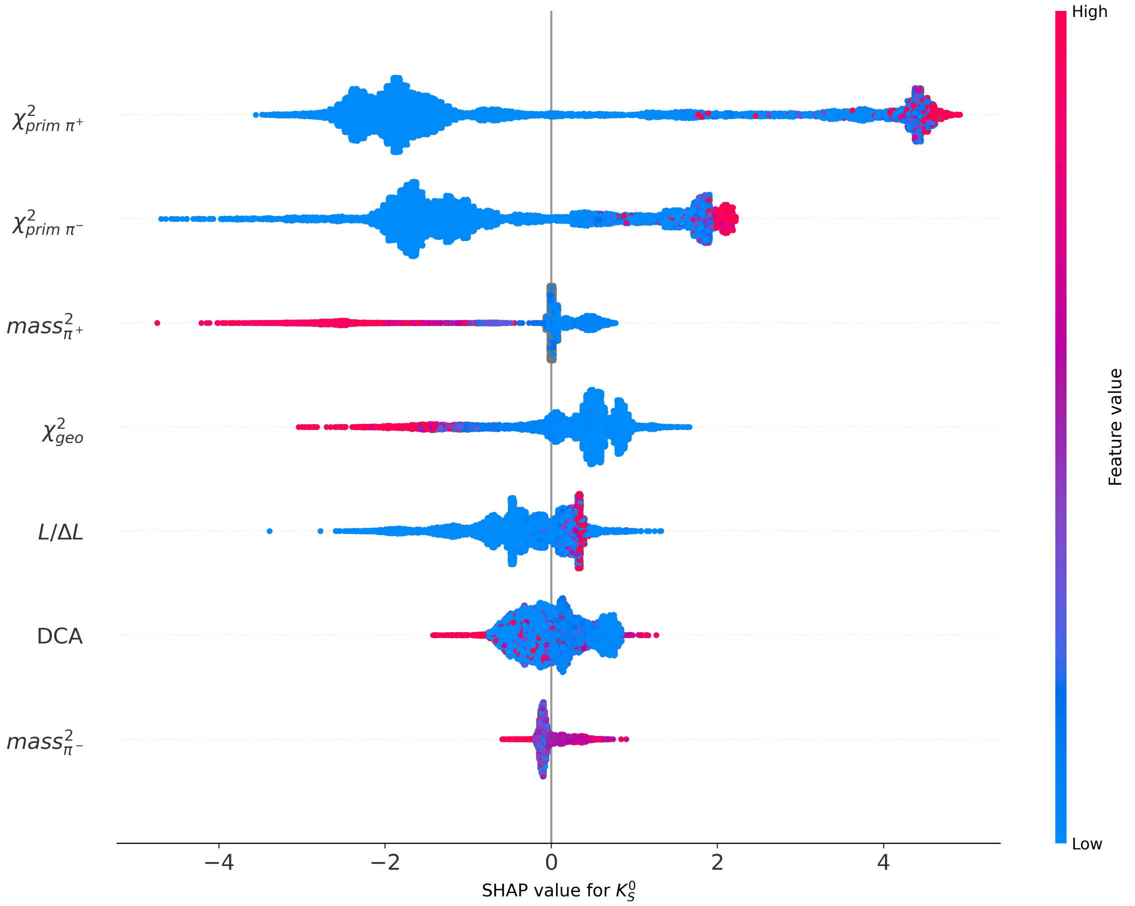


Figure 4.5: SHAP plot of the XGBoost model

For the reconstruction of the K^0_S all positive and negative tracks are combined in the mass range 0.28

GeV to 1 GeV. The positive charge particles in the given mass range are π^+, K^+, and proton, which can be seen from figure A.1b. The π^+ have the lowest mass among all of these positive charge particle particles. Therefore in figure 4.5 the low values of the $mass^2_{\pi+}$ are given positive SHAP value which contributes to the signal and vice versa for the high values of $mass^2_{\pi+}$ variable. The gray dots(and violet due to convolution of colour) for m^2 variable in SHAP plot 4.5 are the NaN value candidates.

Some soft pions cannot reach the TOF detector and therefore the m^2 variable information is not available for them. When the m^2 information is not available NaN value is assigned to the m^2 variable of those particular K^0_S candidates. For the purpose of statistics and the ability of the XGBoost to handle the missing values, these m^2 NaN value candidates (along with the non-NaN values candidates) are used for training and testing the XGBoost model. For investigation, a data sample containing only NaN values of the m^2 variables is chosen and SHAP plot is produced. The effect of these m^2 NaN values on BDTs prediction is investigated using the following SHAP plot figure 4.6. From figure 4.6 it is seen that the least importance is given to the m^2 variable when it has the NaN value. The NaN values of the $m^2_{\pi+}$ variable are given zero SHAP value, while a very small negative SHAP value is given to the NaN values of the $m^2_{\pi-}$ variable. After further investigation, it was found that most of the candidates having NaN value for $m^2_{\pi-}$ variable are the true background candidates. Therefore the BDTs learn this information and a very small negative SHAP value is given to such candidates.

The other plots for the NaN and/or non-NaN data samples are shown in appendix B.

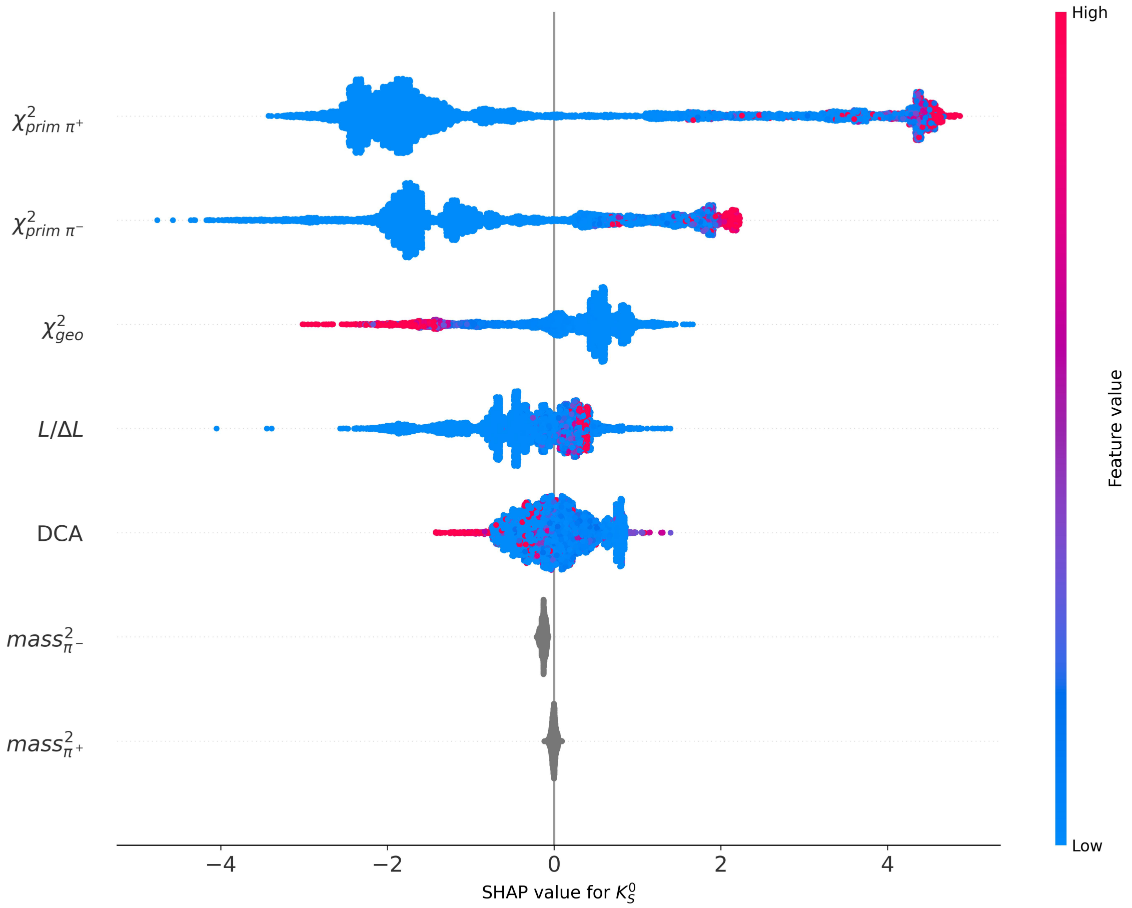


Figure 4.6: SHAP plot of the XGBoost model for m^2 NaN value data sample

4.6 Reconstruciton of K^0_S using XGBoost model

After training, testing, and evaluating the performance of the XGBoost model is applied to the real event-like case, the UrQMD model which contains a signal as an under-represented class in the data. Figure 4.7 shows the invariant mass distribution of the UrQMD data (signal + background).

The >0.99 BDT output score cut is applied to the UrQMD data. The >0.99 BDT output score cut is chosen as it has high signal-to-background ratio which can be seen from the figure 4.2. In figure 4.7, the peak is seen at the mass of K^0_S, which shows that the XGBoost model can reconstruct the K^0_S candidates

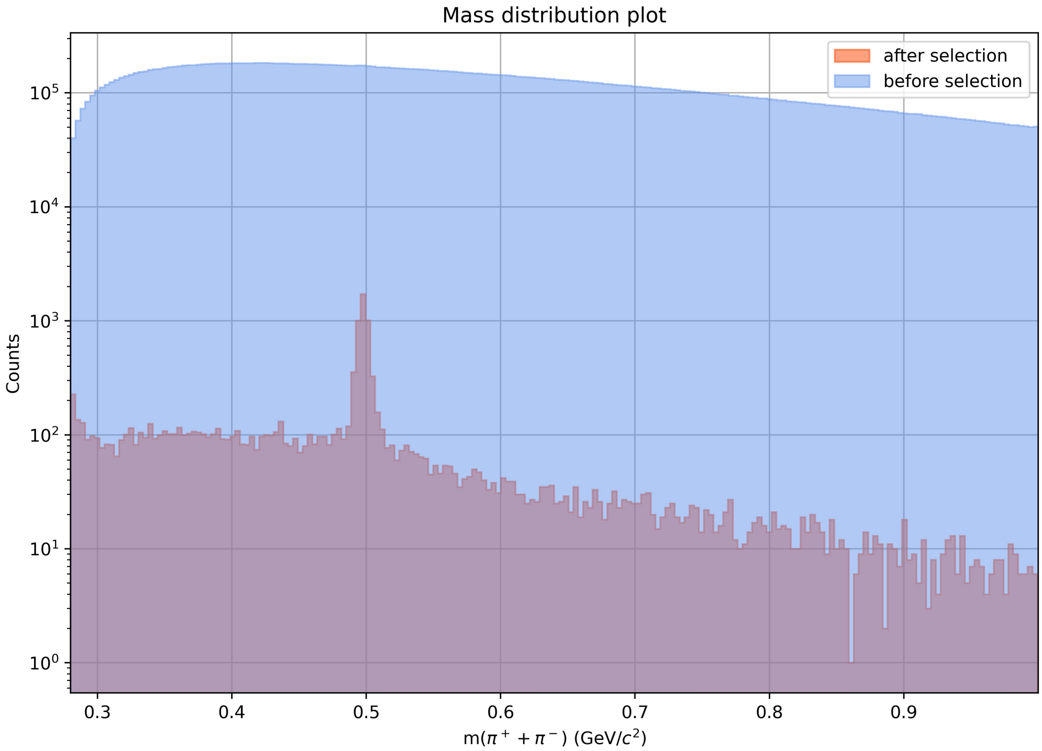


Figure 4.7: XGBoost model applied on UrQMD data at >0.99 BDT output score

successfully. The >0.99 BDT output score selection preserved 76.12% of the true K_S^0 candidates. The lost K_S^0 candidates due to the BDT output score selection are corrected using the yield correction procedure in the next chapter. Using XGBoost method and >0.99 BDT output score critereion, signal to background ratio 2.39 is achieved (within the 3σ region, $0.46 < \text{mass} < 0.52$).

Figure 4.7 without log scale is shown in appendix C.

Chapter 5

Double differential yield extraction of K_S^0

5.1 Procedure used for yield extraction

After reconstruction and selection of the K_S^0 candidates using the XGBoost model, yield can be estimated using a multi-step fitting routine. Figure 5.1 shows the CBM acceptance for K_S^0 signal candidates generated using the DCM-QGSM-SMM model. From figure 5.1 it is seen that most of the K_S^0 candidates are produced near the mid-rapidity region. For the differential study, p_T is divided into 200 MeV/c bins from 0 to 2 and y_{LAB} is divided into 0.3 bins from 0 to 3.

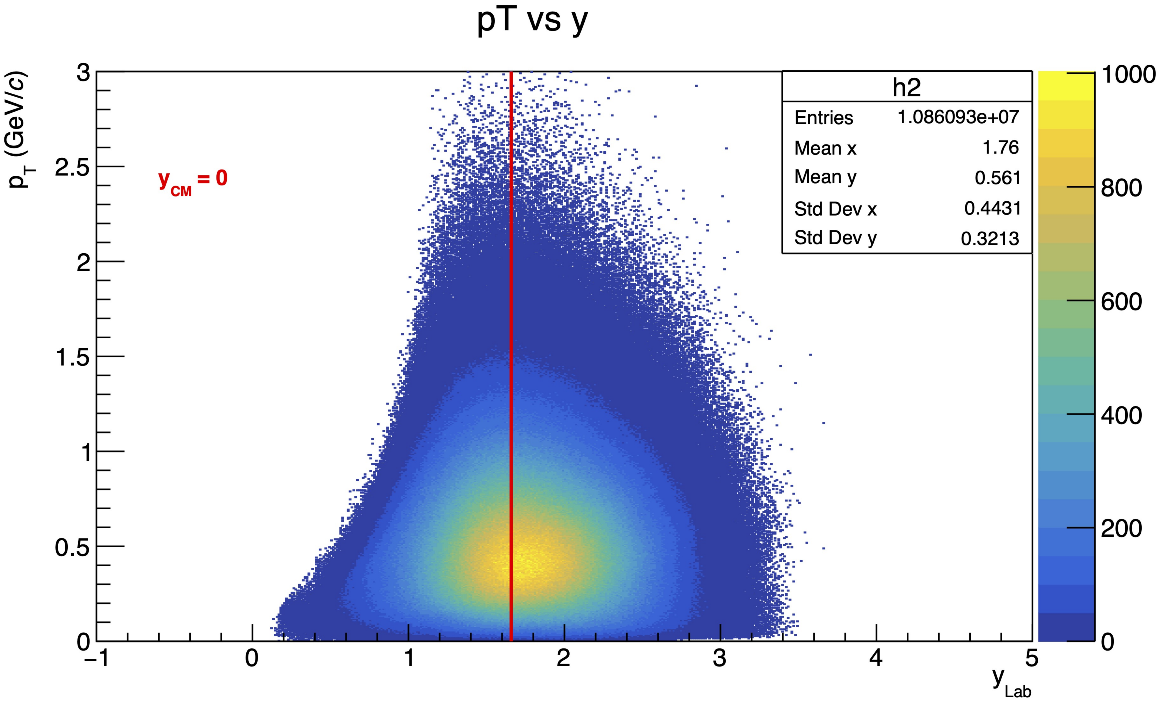
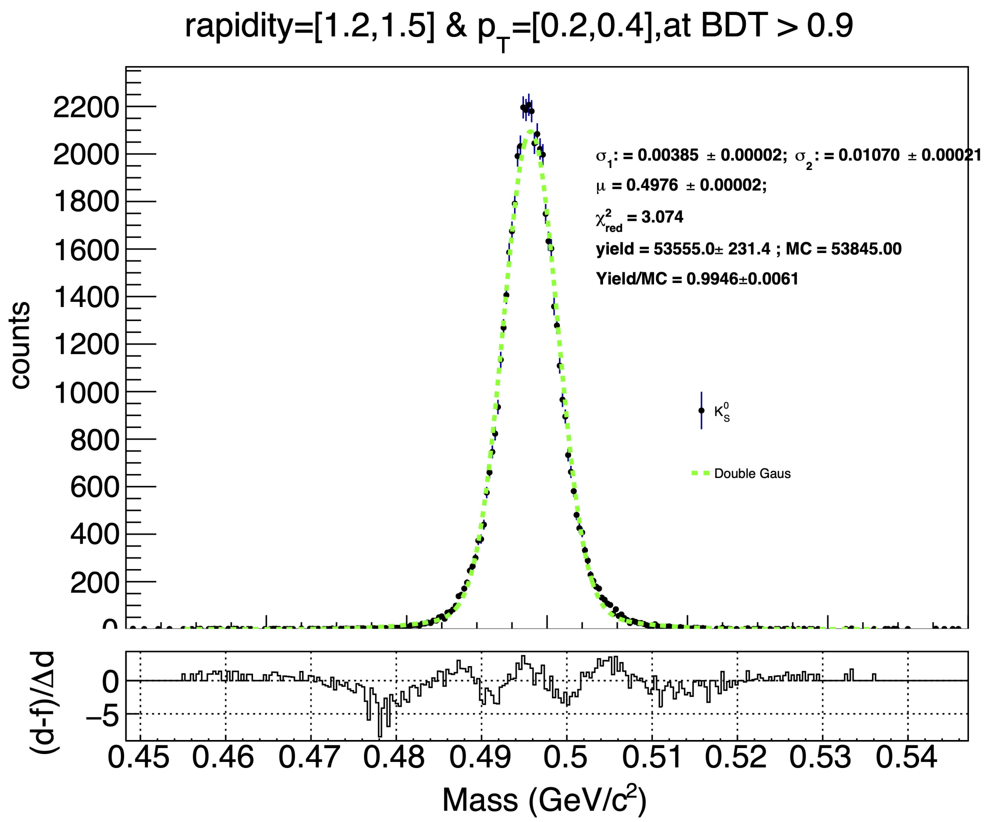


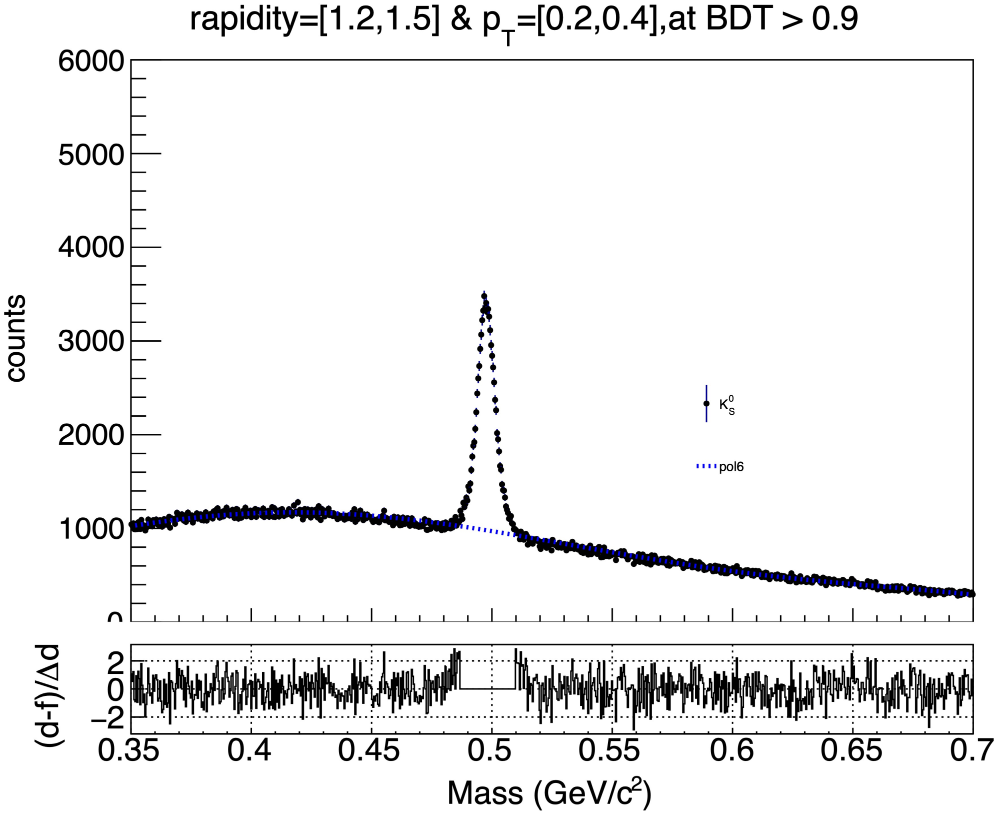
Figure 5.1: CBM acceptance for K_S^0, Au+Au at p_{beam} = 12 $AGeV$

The signal distribution will be approximated by a double gaussian (DG) function and the background distribution by a 6th-order polynomial function. The multi-step fitting procedure is used in the following steps -

1. Step 1 - Fit a double gaussian function on the invariant mass spectra of DCM-QGSM-SMM generated K_S^0 signal only candidates in 4σ region around the peak mass value of K_S^0. The yield is estimated as the integral of the signal function and true signal candidates are calculated using the MC cut in the code. Figure 5.2a shows that yield to MC ratio is around 1 within the uncertainties, which confirms that the signal function and its integral work well.

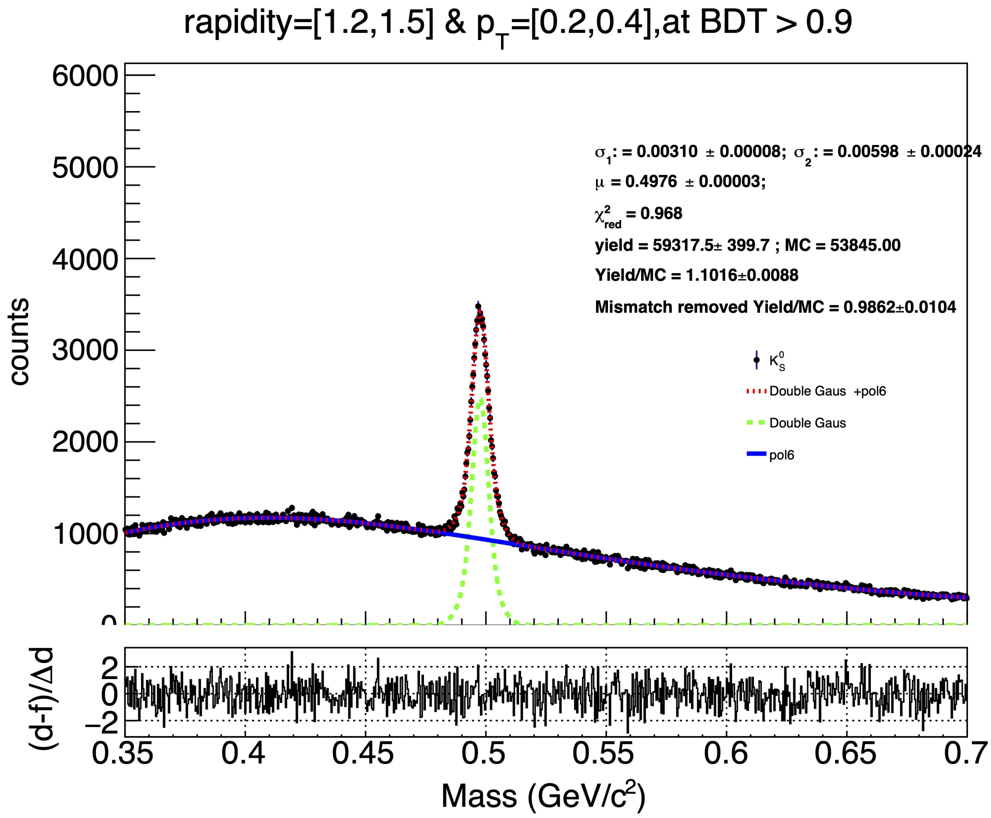


(a) Step 1



(b) Step 2

Figure 5.2: Step involved in double differential yield extraction



(c) step 3

Figure 5.2: Step involved in double differential yield extraction

2. Step 2 - Take the invariant mass spectra of the UrQMD data in the 0.35 - 0.7 (GeV/c^2) range. Exclude the signal in the 4σ region (m > 0.4542 & m < 0.5398) and fit the background with a polynomial of 6th order (Pol6). The pol6 extended in K^0_S peak region is shown in figure 5.2b.

3. Step 3 - A combination of DG and pol6 is taken to fit the full invariant mass distribution of the UrQMD data and the initialization of the parameters is performed with parameters obtained during the first two steps. Step 3 is shown in figure 5.2c.

For the fitting procedure AliHFInvMassFitter [37] class is used. For showing the step 1, 2 and 3, the rapidity = [1.2, 1.5] and p_T = [0.2, 0.4] bin is chosen as it have the maximum yield, with the selection of >0.9 BDT output score.

The DG function is given in equation 5.1 and pol6 function in equation 5.2. In equation 5.1, m is the mass of each K^0_S candidate, A is the integral of the signal function, μ is the mean of the signal function, B is the fraction of the second gaussian, and σ_1, σ_2 are the two variance of the DG function having same μ.

$$DG(m, A, \mu, \sigma_1, B, \sigma_2) = A\left[\frac{(1-B)}{\sqrt{2\pi}/\sigma_1} e^{\frac{-(m-\mu)^2}{2\sigma_1^2}} + \frac{(B)}{\sqrt{2\pi}/\sigma_2} e^{\frac{-(m-\mu)^2}{2\sigma_2^2}}\right] \tag{5.1}$$

$$pol6 = p_0 + \sum_{i=1}^{6} p_i \frac{C^i}{i!} \tag{5.2}$$

In figure 5.2, the yield is calculated as the integral of the signal only fit function and the MC true signal candidates are counted as well. For step 1 in figure 5.2a the yield to MC ratio is around 1 within the error bars. The yield-to-MC ratio for the UrQMD data (figure 5.2c) is greater than 1. The reason for it will be discussed in the following subsection 5.1.1.

5.1.1 Description of the high yield

The yield to MC ratio in figure 5.2c is greater than 1, investigation revealed that by default the MVD detector was not added in the Reco-to-MC matching, therefore if the reconstructed tracks had several hits in the MVD detector but do not have enough hits in the STS detector (at least 2), then this track did not match to any MC track, so the mismatch happened in the matching algorithm. Due to this mismatch, a small fraction of the true K_S^0 decays are labeled as the background. The invariant mass spectra of the background-only candidates of the UrQMD data is shown in figure 5.3. The peak is observed at the mass of K_S^0.

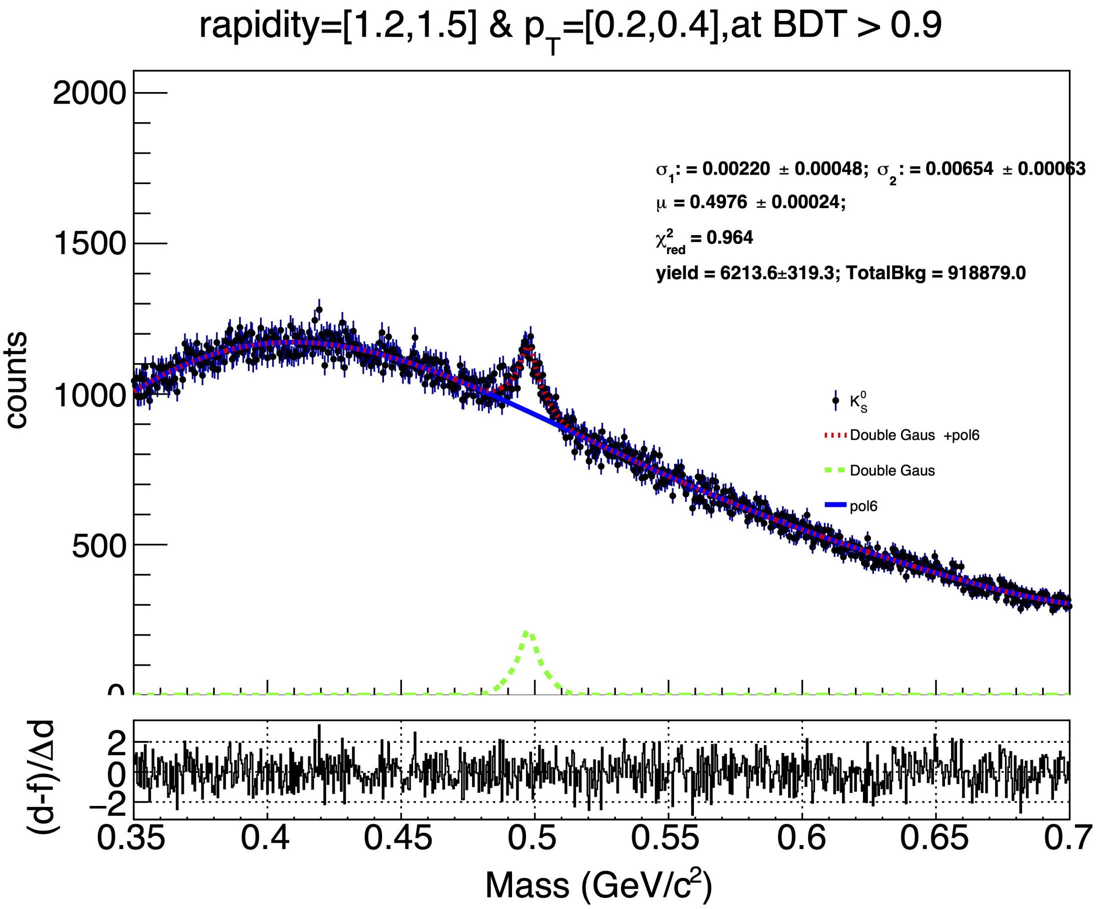


Figure 5.3: UrQMD data background only candidates

The yield from figure 5.3 is then subtracted from the yield in figure 5.2c and called Mismatch removed yield. Now, the ratio of mismatch removed yield to MC in figure 5.2c matches with 1 within the error bars.

The double differential yield extraction is done for all p_T and y_{LAB} bins.

5.2 Efficiency correction

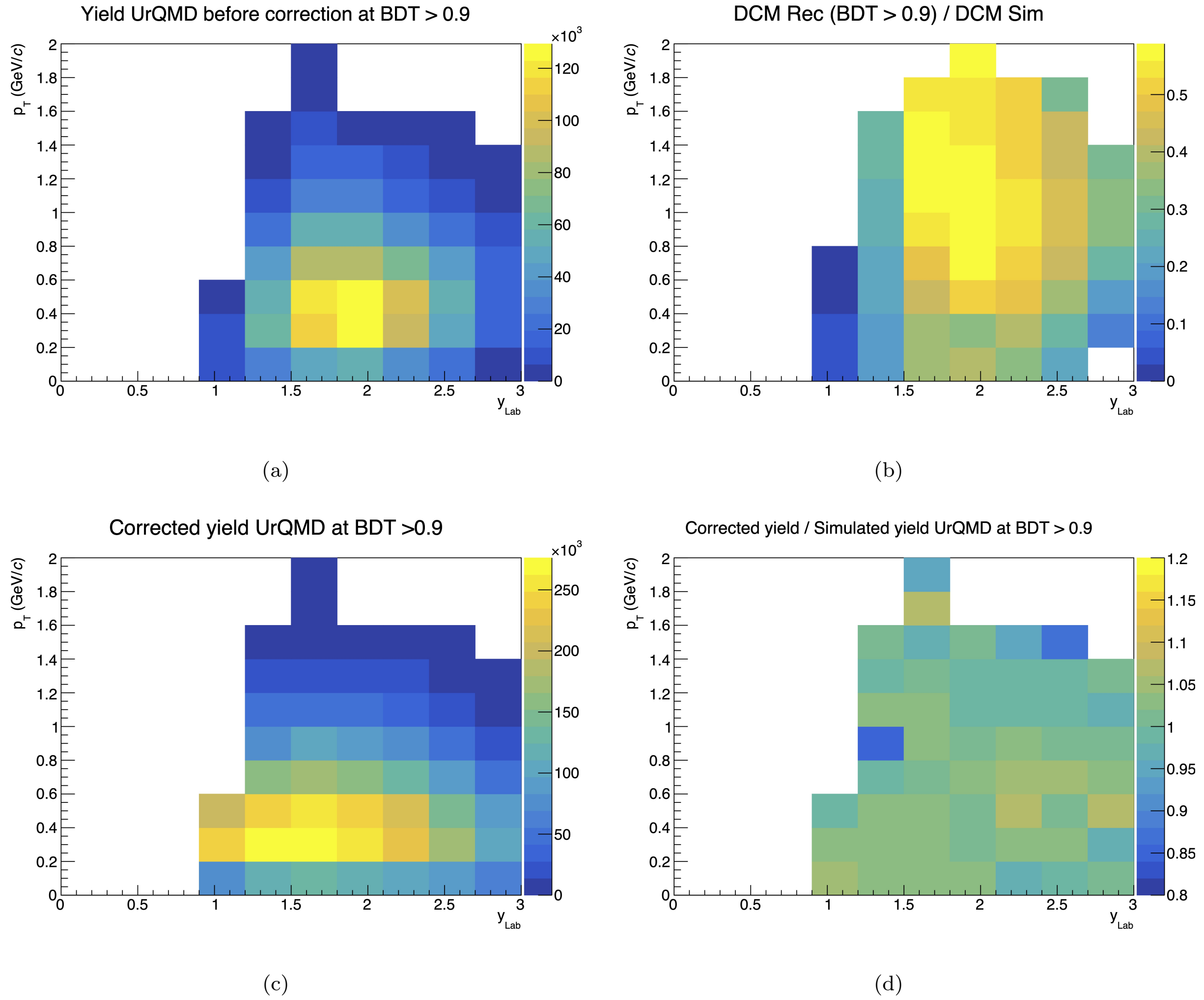
A $Au + Au$ collision at $p_{beam} = 12\ AGeV/c$ simulated by a collision simulator produces x number of K_S^0 particles. These K_S^0 particles then decay due to the weak interaction and produces their daughters $\pi^+ + \pi^-$ with a branching ratio of 69% [14]. So out of the total K_S^0 candidates around 69% percent can be reconstructed by this particular decay channel. Due to the detector's coverage angle $x - y$ some K_S^0 daughters will pass through it while others will be lost and the term acceptance encapsulates this. Similarly, the reconstruction algorithm will also select some candidates and others will be lost and the term efficiency takes this into account. The yield after the reconstruction of the K_S^0 candidates is called a reconstructed yield, if no selection is applied. The selection of K_S^0 candidates through ML selection also has efficiency and the total efficiency will now also contain this factor. The simulated yield can be reproduced from the reconstructed yield using a correction number which is simply the total efficiency × acceptance.

Here in this study, DCM-QGSM-SMM data is taken as simulated data, and UrQMD data is taken as real data. The original simulated yield as well as the reconstructed yield of both the DCM-QGSM-SMM and UrQMD models can be accessed from the data. With the help of the DCM-QGSM-SMM model, the corrected (simulated) yield of the UrQMD model can be estimated. This corrected yield is then compared with the original simulated yield of the UrQMD yield in all p_T - y_{LAB} bins using the following procedure -

1. The correction number is calculated for each (p_T - y_{LAB}) bin of the DCM-QGSM-SMM model by dividing the reconstructed yield of the DCM-QGSM-SMM model by the simulated yield of the DCM-QGSM-SMM model.
2. The corrected yield of the UrQMD model is found by dividing the reconstructed yield of the UrQMD model with the correction number for each p_T - y_{LAB} bin.
3. The ratio of the corrected yield to the original simulated yield of the UrQMD model shows that the original simulated yield of the real data can be obtained using this yield correction procedure.

The figures for the DCM simulated and reconstructed yield, UrQMD simulated yield are included in the appendix F.

Figure 5.4c shows the corrected yield of the UrQMD model at >0.9 BDT output selection. Figure 5.4d shows the ratio of the corrected yield to the original simulated yield of the UrQMD model. The fitting procedure fails in the low statistical bins therefore some bins are empty. As seen from the figure 5.4d most of the K^0_S candidates are retrieved using the fitting and yield extraction procedure. From figure 5.4c it is seen that most of the bins have an efficiency of around 1 within the statistical uncertainties. Therefore the reconstruction of the K^0_S using the XGBoost model and yield extraction using the fitting procedure works very well.



(a)

(b)

(c)

(d)

Figure 5.4: a) UrQMD reconstructed yield b) Correction number (Acceptance × total efficiency) c) Corrected yield of the UrQMD d) Corrected yield / Simulated yield of the UrQMD

Chapter 6

Summary

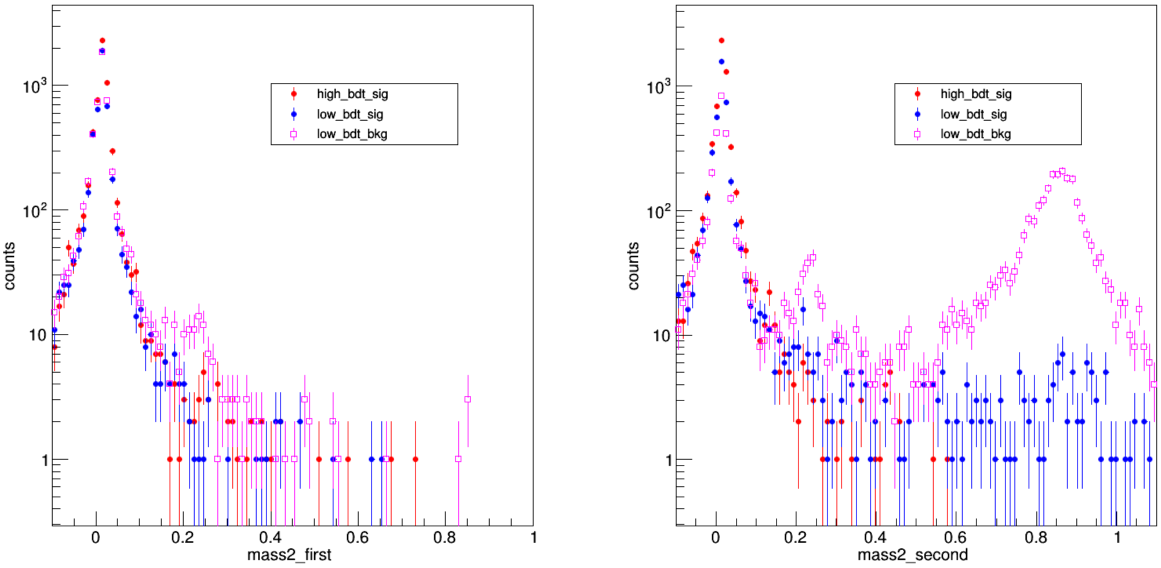
The CBM experiment will investigate the unexplored region of the phase diagram of QCD matter at high μ_B, therefore reconstruction of the (multi-)strange hadrons is important. The strangeness enhancement is an important probe of the deconfined matter. The CBM performance for the K^0_S meson via its decay to $\pi+$ and $\pi-$ is presented. Neutral kaons (K^0_S) selection is implemented in the CBM experiment at the FAIR facility using a machine learning technique. The optimization of selection criteria of K^0_S is done using the XGBoost ensemble method. By using the BDT output score selection high signal purity, background rejection, and high efficiency are achieved. Double differential(p_T - y_{LAB}) yield extraction of K^0_S is performed. The correction of the mismatch is done by extracting and then subtracting the signal (labeled as background due to the matching algorithm) from the background of the UrQMD data. The corrected yield is consistent with the original yield within the statistical uncertainties.

This study can be extended to different p_T, y_{LAB}, multiplicity, and different collision energies. The BDTs training and selection can be done for different p_T, y, and multiplicity bins.

Appendix A

Signal peak at low BDT output score

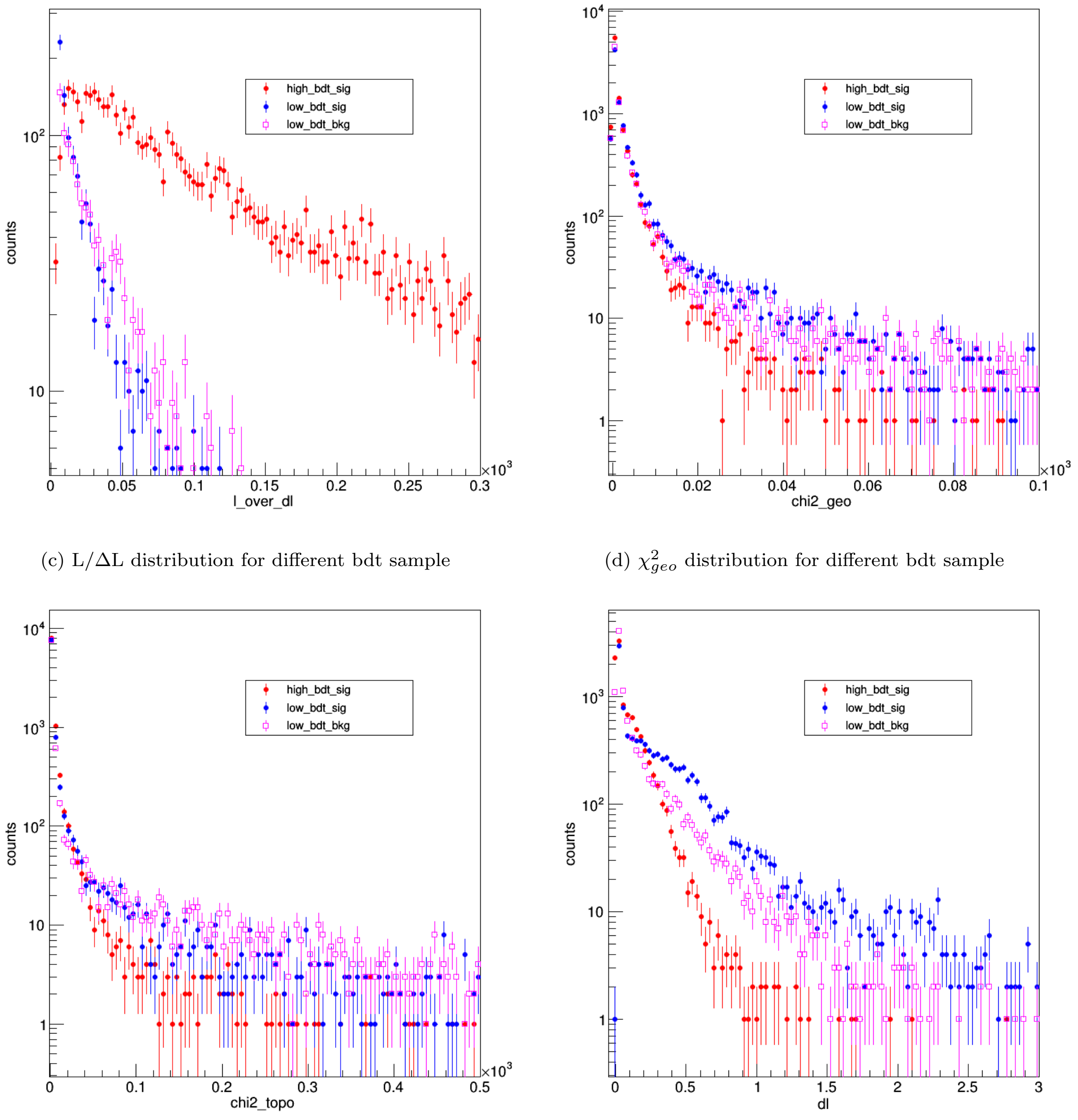
The plots for topological variables for low bdt signal, high bdt signal and low bdt background are shown below.



(a) $m^2_{\pi-}$ distribution for different bdt sample

(b) $m^2_{\pi+}$ distribution for different bdt sample

Figure A.1: Topological variable distribution plot for different bdt samples



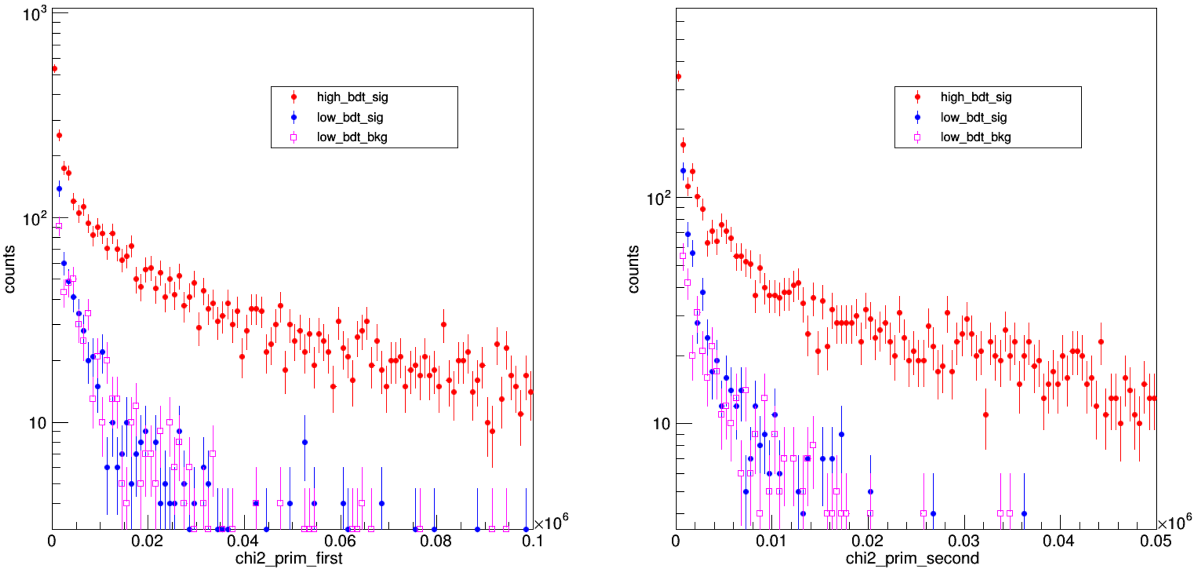
(c) L/ΔL distribution for different bdt sample

(d) χ^2_{geo} distribution for different bdt sample

(e) χ^2_{topo} distribution for different bdt sample

(f) ΔL distribution for different bdt sample

Figure A.1: Topological variable distribution plot for different bdt samples



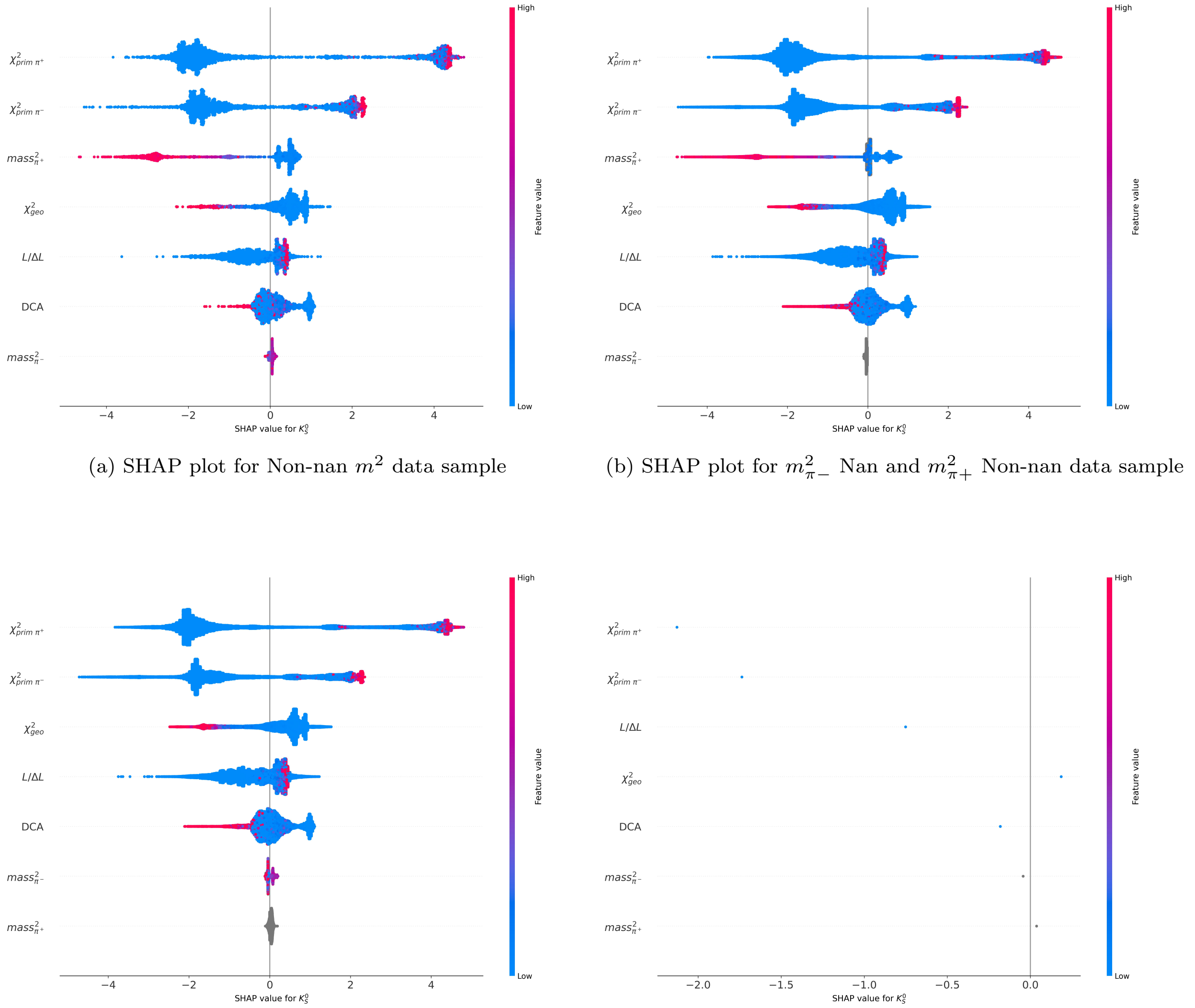
(g) $\chi^2_{prim\pi-}$ distribution for different bdt sample

(h) $\chi^2_{prim\pi+}$ distribution for different bdt sample

Figure A.1: Topological variable distribution plot for different bdt samples

Appendix B

SHAP plot for NaN and/or non-NaN data samples



(a) SHAP plot for Non-nan m^2 data sample

(b) SHAP plot for $m^2_{\pi-}$ Nan and $m^2_{\pi+}$ Non-nan data sample

(c) SHAP plot for $m^2_{\pi-}$ Non-nan $m^2_{\pi+}$ Nan data sample

(d) SHAP plot for a single K^0_S candidate having Nan value for both $m^2_{\pi-}$ variable

Figure B.0: SHAP plots

From figure B.0d it is seen that $m^2_{\pi-}$ variables are at the last rank(as they have nan value) and they do not have a large SHAP value therefore in the total sum of the SHAP value of all the variables for the K^0_S candidate, $m^2_{\pi-}$ variable does not any significant role.

Appendix C

BDT's applied on UrQMD data invariant mass plot

Here the plot for the XGBoost model applied on Real data at >0.99 BDT output score for invariant mass is shown(without log scale on y axis).

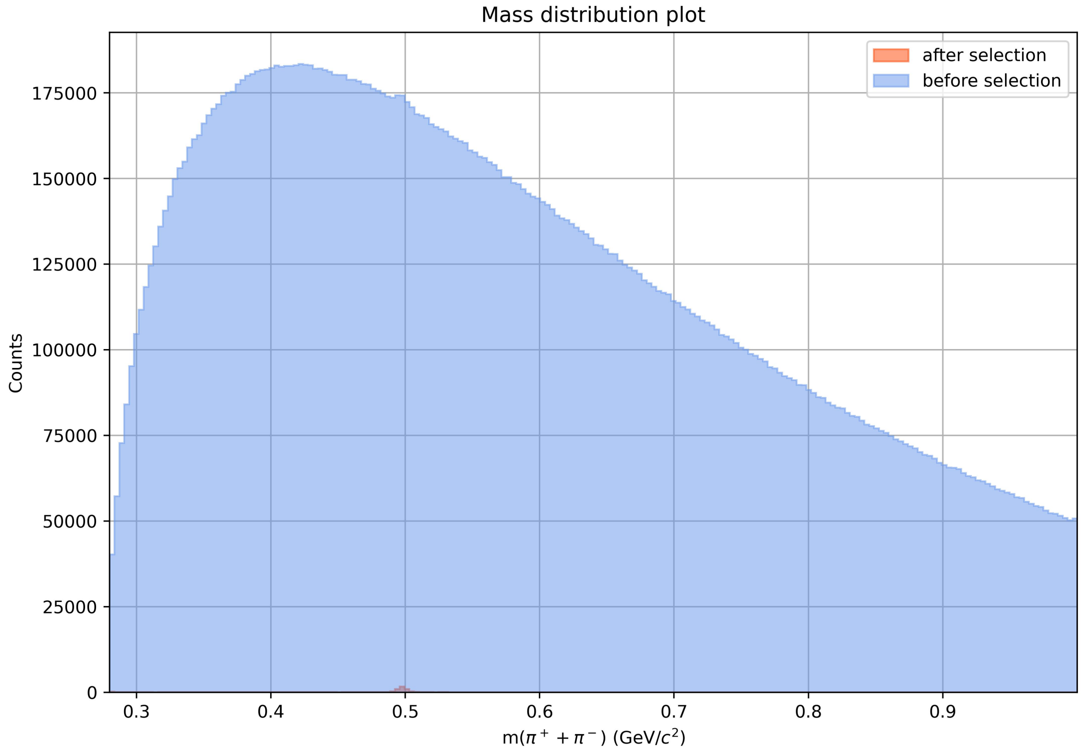


Figure C.1: The XGBoost model applied on UrQMD data at >0.99 BDT output score without log scale

Appendix D

$\chi^2_{prim\pi+}$ distribution plot

This is the plot for the $\chi^2_{prim\pi+}$ variable.

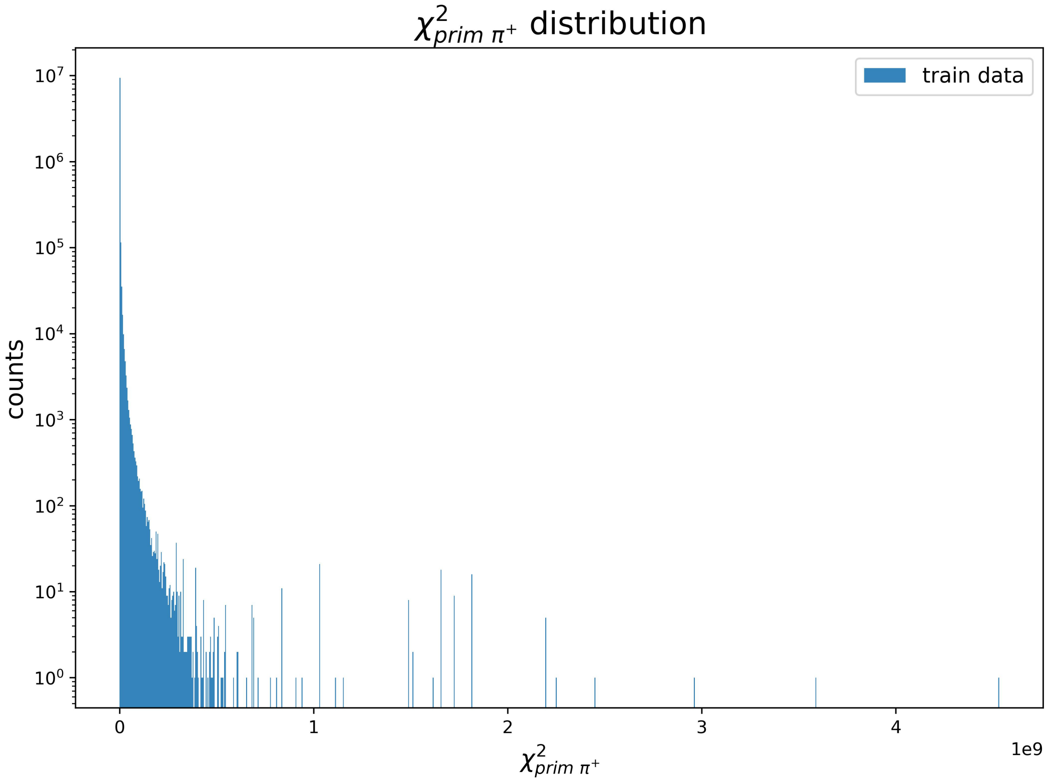


Figure D.1: $\chi^2_{prim\pi+}$ distribution plot for train data

Appendix E

SHAP plot structure of the $\chi^2_{prim\pi+}$ variable

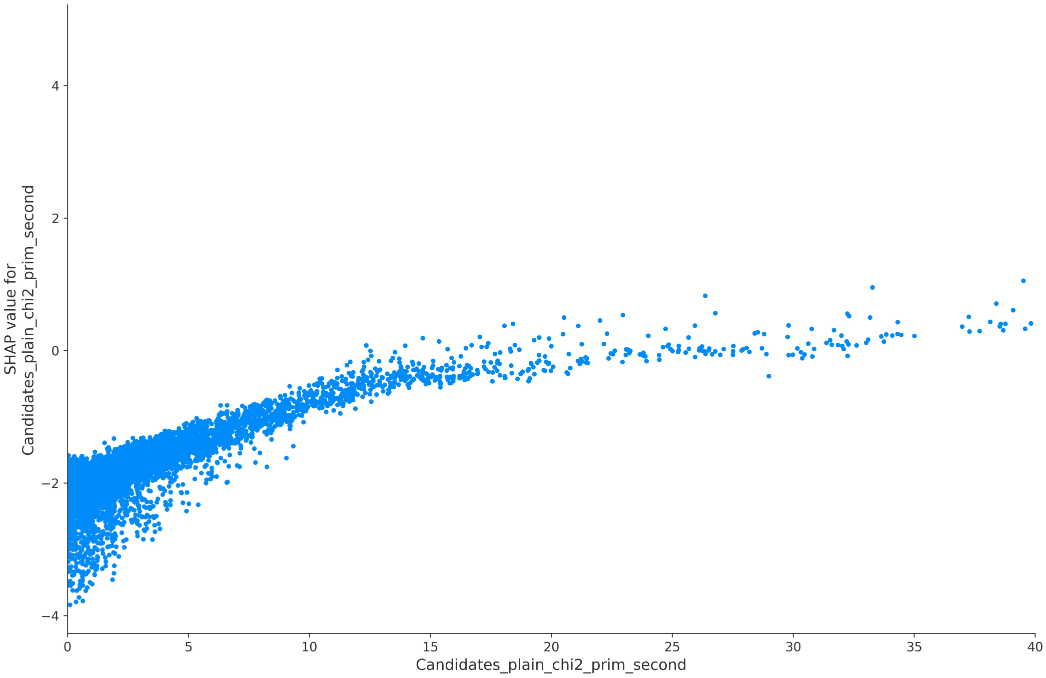
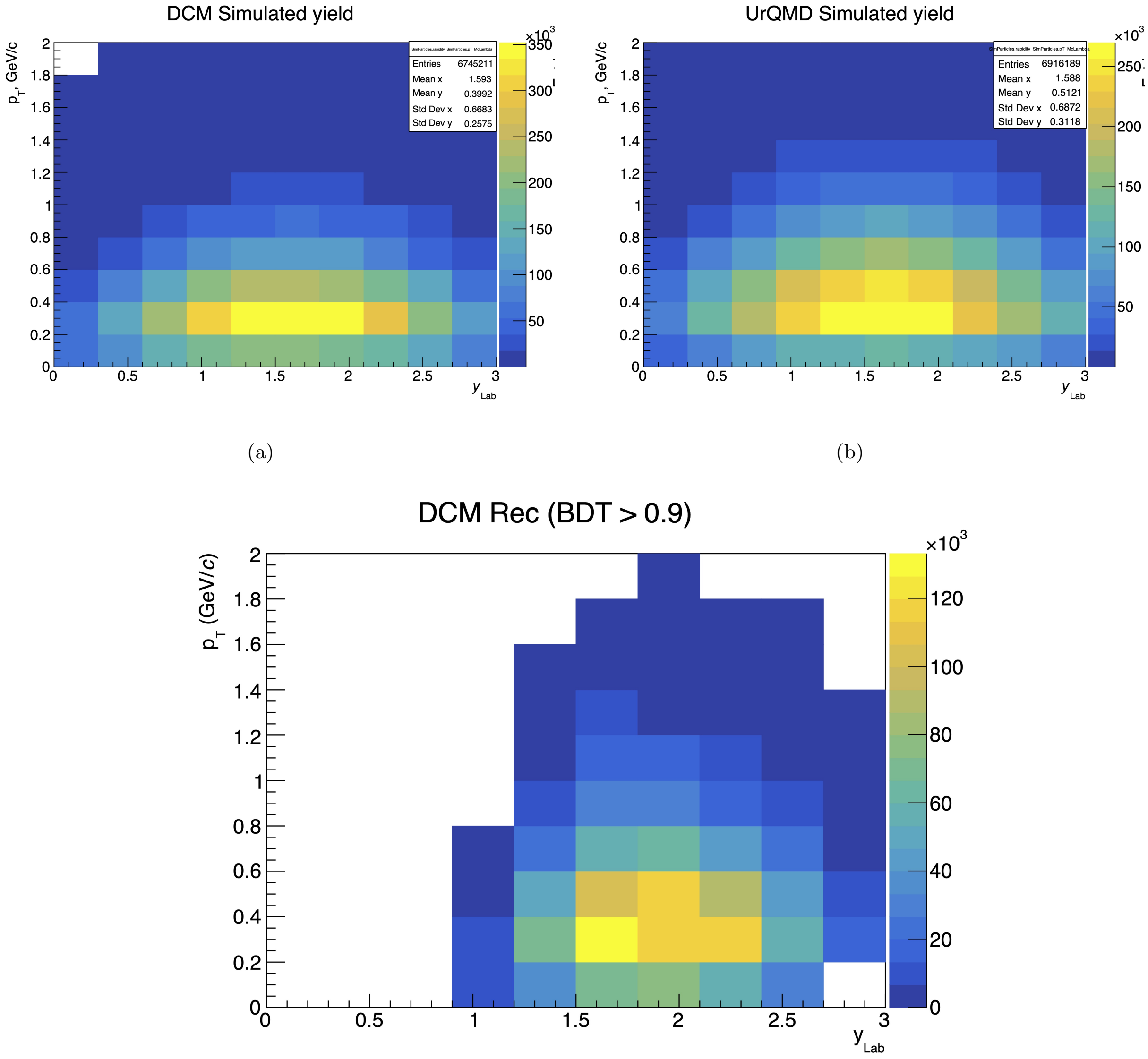


Figure E.1: structure description of the $\chi^2_{prim\pi+}$ variable is SHAP plot 4.5

Here chi2_prim_second is $\chi^2_{prim\pi+}$. As seen from the figure 4.5 there are two blobs for the $\chi^2_{prim\pi+}$ variable. These are the candidates having low $\chi^2_{prim\pi+}$ value and near to the -2 SHAP value as seen from the figure E.1

Appendix F

K_S^0 yield for DCM and UrQMD model



(a)

(b)

(c)

Figure F.1: a) K_S^0 (simulated) yield generated by DCM-QGSM-SMM model before passing through the CBM detector b) K_S^0 (simulated) yield generated by UrQMD model before passing through the CBM detector c) K_S^0 (reconstructed) yield generated by DCM-QGSM-SMM model after passing through the CBM detector

Appendix G

m^2 importance for the XGBoost model

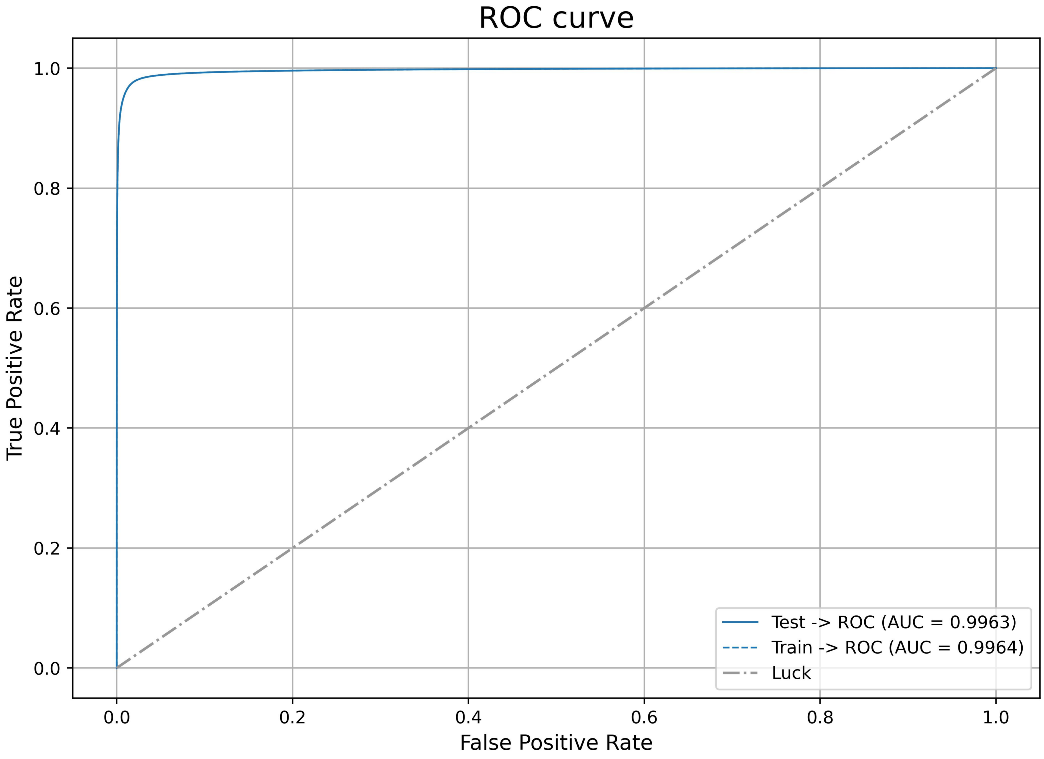


Figure G.1: ROC plot for the XGBoost model train without the m^2 variables

Figure G.1 show the performance of the XGBoost model trained without the m^2 variables. Figure G.1 have the ROC-AUC value less than the ROC-AUC value in figure 4.4, which shows that there is a slight improvement in the performance when the m^2 variable is added for the training of the XGBoost model.

Bibliography

[1] Kohsuke Yagi, Tetsuo Hatsuda, and Yasuo Miake. *Quark-gluon plasma: From big bang to little bang*, volume 23. Cambridge University Press, 2005.

[2] Jerzy Bartke. *Introduction to relativistic heavy ion physics*. World Scientific, 2008.

[3] Kenneth G Wilson. Confinement of quarks. *Physical review D*, 10(8):2445, 1974.

[4] A Bazavov, H-T Ding, P Hegde, Olaf Kaczmarek, Frithjof Karsch, Edwin Laermann, Swagato Mukherjee, P Petreczky, Christian Schmidt, D Smith, et al. Freeze-out conditions in heavy ion collisions from qcd thermodynamics. *Physical review letters*, 109(19):192302, 2012.

[5] Larry McLerran and Robert D Pisarski. Phases of dense quarks at large nc. *Nuclear Physics A*, 796(1-4):83–100, 2007.

[6] Larry McLerran, Krzysztof Redlich, and Chihiro Sasaki. Quarkyonic matter and chiral symmetry breaking. *Nuclear Physics A*, 824(1-4):86–100, 2009.

[7] Alessandro De Falco. Cpod2021 - international conference on critical point and onset of deconfinement. https://indico.cern.ch/event/985460/contributions/4264615/attachments/2211234/3742919/adf_cpod.pdf.

[8] Ulrich Heinz and Maurice Jacob. Evidence for a new state of matter: An assessment of the results from the cern lead beam programme. *arXiv preprint nucl-th/0002042*, 2000.

[9] Michael Riordan and William A Zajc. The first few microseconds. *Scientific American*, 294(5):34A–41, 2006.

[10] T Ablyazimov, A Abuhoza, RP Adak, Marek Adamczyk, K Agarwal, MM Aggarwal, Z Ahammed, F Ahmad, N Ahmad, S Ahmad, et al. Challenges in qcd matter physics–the scientific programme of the compressed baryonic matter experiment at fair. *The European Physical Journal A*, 53(3):1–14, 2017.

[11] Johann Rafelski and Berndt Müller. Strangeness production in the quark-gluon plasma. *Physical Review Letters*, 48(16):1066, 1982.

[12] Johann Rafelski. Discovery of quark-gluon plasma: strangeness diaries. *The European Physical Journal Special Topics*, 229(1):1–140, 2020.

[13] G Agakishiev, O Arnold, A Balanda, D Belver, A Belyaev, JC Berger-Chen, A Blanco, M Böhmer, JL Boyard, P Cabanelas, et al. Statistical model analysis of hadron yields in proton-nucleus and heavy-ion collisions at sis 18 energies. *arXiv preprint arXiv:1512.07070*, 2015.

[14] PDG. k_S^0 pdg website. https://pdg.lbl.gov/2022/listings/rpp2022-list-K-zero-S.pdf.

[15] Hans H Gutbrod, I Augustin, H Eickhoff, KD Groß, WF Henning, D Krämer, and G Walter. Fair-baseline technical report. executive summary. *Darmstadt (September 2006) http://www.gsi.de/fair/reports/btr.html*, 2006.

[16] CBM. Cbm material for conference presentations. https://cbm-wiki.gsi.de/foswiki/bin/view/PWG/Figures.

[17] Bengt Friman, Claudia Höhne, Jörn Knoll, Stefan Leupold, Jorgen Randrup, Ralf Rapp, and Peter Senger. *The CBM physics book: Compressed baryonic matter in laboratory experiments*, volume 814. Springer, 2011.

[18] J Heuser et al. Technical design report for the cbm sts, 2013.

[19] Steffen A Bass, Mohamed Belkacem, Marcus Bleicher, Mathias Brandstetter, L Bravina, Christoph Ernst, Lars Gerland, Max Hofmann, Sigurd Hofmann, Jens Konopka, et al. Microscopic models for ultrarelativistic heavy ion collisions. *Progress in Particle and Nuclear Physics*, 41:255–369, 1998.

[20] Mircea Baznat, Alexander Botvina, Genis Musulmanbekov, Viacheslav Toneev, and Valeriy Zhezher. Monte-carlo generator of heavy ion collisions dcm-smm. *Physics of Particles and Nuclei Letters*, 17(3):303–324, 2020.

[21] Sea Agostinelli, John Allison, K al Amako, John Apostolakis, H Araujo, Pedro Arce, Makoto Asai, D Axen, Swagato Banerjee, GJNI Barrand, et al. Geant4—a simulation toolkit. *Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506(3):250–303, 2003.

[22] Valentina Akishina and Ivan Kisel. Time-based cellular automaton track finder for the cbm experiment. In *Journal of Physics: Conference Series*, volume 599, page 012024. IOP Publishing, 2015.

[23] Maksym Zyzak. *Online selection of short-lived particles on many-core computer architectures in the CBM experiment at FAIR*. doctoralthesis, Universitätsbibliothek Johann Christian Senckenberg, 2016.

[24] S Gorbunov and I Kisel. Reconstruction of decayed particles based on the kalman filter. *CBM-SOFT-note-2007-003*, 7, 2007.

[25] Oleksii Lubynets Viktor Klochkov, Ilya Selyuzhenkov. Cbm-gsi gitlab. https://git.cbm.gsi.de/pwg-c2f/analysis/pf_simple/-/tree/master/.

[26] Wikipedia. Pearson correlation coefficient wikipedia. https://en.wikipedia.org/wiki/Pearson_correlation_coefficient.

[27] Oliver Theobald. *Machine learning for absolute beginners: a plain English introduction*, volume 157. Scatterplot press, 2017.

[28] Sotiris B Kotsiantis. Decision trees: a recent overview. *Artificial Intelligence Review*, 39(4):261–283, 2013.

[29] Bahzad Charbuty and Adnan Abdulazeez. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(01):20–28, 2021.

[30] Yoav Freund. Boosting a weak learning algorithm by majority. *Information and computation*, 121(2):256–285, 1995.

[31] Zhiyuan He, Danchen Lin, Thomas Lau, and Mike Wu. Gradient boosting machine: a survey. *arXiv preprint arXiv:1908.06951*, 2019.

[32] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[33] Luca Barioglio, Fabio Catalano, Matteo Concas, Pietro Fecchio, Fabrizio Grosa, Francesco Mazzaschi, and Maximiliano Puccio. hipe4ml/hipe4ml. https://doi.org/10.5281/zenodo.7014886, April 2022.

[34] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[35] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[36] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[37] C. Bianchin F.Prino, A. Rossi. Alihfinvmassfitter class for the fit of invariant mass distribution. http://alidoc.cern.ch/AliPhysics/master/_ali_h_f_inv_mass_fitter_8h_source.html.