Department of Physics and Astronomy

University of Heidelberg

Bachelor Thesis in Physics

submitted by

Sarah Liedtke

born in Düsseldorf (Germany)

February 2023

Feasibility study of Ξ_c^+ in proton-lead collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02\,\mathrm{TeV}$ with the ALICE detector

This Bachelor Thesis has been carried out
by Sarah Liedtke at the
Physikalisches Institut of the University of Heidelberg
under the supervision of
Prof. Dr. Silvia Masciocchi

Abstract

Heavy-ion collisions at ultra-relativistic energies are used to study the physics of strongly interacting matter under extreme conditions. Extremely high energy densities and temperatures are reached in these collisions and a phase of matter called quark-gluon plasma (QGP) forms. Measurements of heavy-flavour hadron production in high-energy collisions are used to study and characterise the QGP, as heavy quarks (charm and beauty) are created before the QGP can form and therefore interact with the medium and experience the whole evolution of the system. In addition, heavy-flavour meson and baryon production measurements give insight into hadronisation mechanisms in different collision systems. The Ξ_c^+ baryon has not yet been measured in systems larger than proton-proton (pp) collisions, which would be an important step towards further understanding charm production and hadronisation.

This thesis presents a feasibility study for the reconstruction and the observation of the Ξ_c^+ baryon in proton-lead (p-Pb) collisions at the centre-of-mass energy $\sqrt{s_{\mathrm{NN}}} = 5.02\,\mathrm{TeV}$ recorded by the ALICE detector at the LHC. The Ξ_c^+ reconstruction is performed in the transverse momentum range $2 < p_{\mathrm{T}}(\Xi_c^+) < 12\,\mathrm{GeV}/c$, divided into the three intervals: $2 < p_{\mathrm{T}}(\Xi_c^+) < 4\,\mathrm{GeV}/c$, $4 < p_{\mathrm{T}}(\Xi_c^+) < 6\,\mathrm{GeV}/c$ and $6 < p_{\mathrm{T}}(\Xi_c^+) < 12\,\mathrm{GeV}/c$. The short-lived Ξ_c^+ baryon is reconstructed via its hadronic decay into two positively charged pions and a Ξ^- baryon, decaying into a negatively charged pion and a Λ baryon, which further decays into a proton and a negatively charged pion. The reconstruction of the decay is performed with the KFParticle software package. Reconstructed signal and background are classified using the supervised machine learning tool XGBoost. The signal is then extracted with fits to the invariant mass spectrum. In this thesis, several Boosted-Decision-Tree (BDT) models are trained with different training features and preselections and their performances are compared. Throughout all analysed p_{T} intervals a significant signal is found in the Ξ_c^+ invariant mass spectrum, giving a strong indication that a full analysis of Ξ_c^+ baryon production in p-Pb collisions is feasible.

Zusammenfassung

In Schwerionenkollisionen bei ultra-relativistischen Energien kann die Physik der stark wechselwirkenden Materie unter extremen Bedingungen untersucht werden. Dabei werden extrem hohe Energiedichten und Temperaturen erreicht und es bildet sich eine Materiephase, Quark-Gluon-Plasma (QGP) genannt wird. Darüber hinaus eignet sich die Produktion von Heavy-Flavour Hadronen in hochenergetischen Kollisionen zur Untersuchung und Beschreibung des QGP, da die schweren Quarks vor der Entstehung des QGP erzeugt werden und daher mit dem Medium wechselwirken und die Entwicklung des Systems begleiten. Um einen tieferen Einblick in die Hadronisierungsmechanismen in verschiedenen Kollisionssystemen zu erhalten, werden Messungen der Produktion von Baryonen und Mesonen, die Charm-Quarks enthalten, durchgeführt. Das Ξ_c^+ - Baryon wurde bisher nicht in größeren Systemen als Proton-Proton (pp) Kollisionen gemessen, was ein wichtiger Schritt zum detaillierteren Verständnis der Produktion von Baryonen und Mesonen mit Charm und der Hadronisierung von Charm-Quarks wäre.

In dieser Arbeit wird eine Machbarkeitsstudie zur Rekonstruktion und Beobachtung des Ξ_c^+ - Baryons in Proton-Blei (p-Pb) Kollisionen bei einer Schwerpunktsenergie von $\sqrt{s_{\mathrm{NN}}} = 5.02\,\mathrm{TeV}$ vorgestellt, die mit dem ALICE-Detektor am LHC aufgezeichnet werden. Die Rekonstruktion des Ξ_c^+ - Baryons wird im Transversalimpulsbereich $2 < p_{\mathrm{T}}(\Xi_c^+) < 12\,\mathrm{GeV}/c$ durchgeführt, der in drei Intervalle unterteilt ist: $2 < p_{\mathrm{T}}(\Xi_c^+) < 4\,\mathrm{GeV}/c$, $4 < p_{\mathrm{T}}(\Xi_c^+) < 6\,\mathrm{GeV}/c$ und $6 < p_{\mathrm{T}}(\Xi_c^+) < 12\,\mathrm{GeV}/c$. Das kurzlebige Ξ_c^+ - Baryon wird über seinen hadronischen Zerfallskanal in zwei positiv geladene Pionen und ein Ξ^- - Baryon rekonstruiert, das in ein negativ geladenes Pion und ein Λ - Baryon zerfällt, welches wiederum in ein Proton und ein negativ geladenes Pion zerfällt. Zur Rekonstruktion des Zerfalls und der Zerfallstopologie wird das Softwarepaket KFParticle verwendet. Die Klassifizierung von Signal und Hintergrund erfolgt mithilfe des Machine Learning Pakets XGBoost und das Signal wird anschließend durch eine Analyse der invarianten Masse extrahiert. Es werden verschiedene Boosted-Decision-Tree (BDT) Modelle mit unterschiedlichen Trainingsvariablen sowie Vorselektionen trainiert und ihre Leistungen verglichen. In allen untersuchten p_{T}-Intervallen wird ein signifikantes Signal im invarianten Massenspektrum des Ξ_c^+ gefunden, was ein deutlicher Hinweis darauf ist, dass eine vollständige Analyse des Ξ_c^+ - Baryons in p-Pb Kollisionen möglich ist.

Contents

1 Introduction

1.1 Standard model and quantum chromodynamics

The current understanding of matter in the visible universe, which is made from a few fundamental particles governed by four fundamental forces, is best described by the so-called Standard Model of particle physics (SM) [1]. The elementary particles characterised by the SM are 12 fermions, which are *matter particles* of spin $\frac{1}{2}$ and different types of spin 1 *field mediators* called gauge bosons (photons, W and Z bosons, and gluons), which are force carriers that mediate the fundamental interactions between the fermions. The fermions are either quarks or leptons, each grouped in pairs forming three generations with increasing mass. There are charged and electrically neutral leptons, the latter are called neutrinos. All 12 fermions have corresponding antiparticles, which have the same mass but opposite charge. The SM describes the strong, the electromagnetic and the weak interaction. All leptons undergo the weak force mediated by Z- and two charged $W^{\pm}$-bosons, and charged fermions additionally participate in the electromagnetic interaction mediated by photons. Quarks engage in all four fundamental interactions.

The relativistic Quantum Field Theory of the strong interaction between quarks mediated by gluons is called Quantum Chromodynamics (QCD). The equivalent of electric charge associated with strong interactions of QCD is colour charge carried by quarks, which come in 6 different flavours and also have mass and electric charge. The underlying symmetry of QCD is an invariance under SU(3) local phase transformations. The eight generators of this symmetry can be associated to the eight types of massless, coloured gluons coupling to particles that have non-zero colour charge. Since leptons are colour neutral, only quarks participate in the strong interaction. The gluons carry colour and anticolour charge and can therefore also self-interact. The effective strength of the strong interaction between colour charges is determined by the coupling strength $\alpha_s(Q^2)$. The strength between colour charges results from all possible processes, including higher-order corrections to the QCD interaction vertex. The gluon–gluon self-interactions lead to higher order loop diagrams, which in turn leads to the strong coupling strength α_s evolving with the momentum transfer, Q^2. The coupling becomes small for large momentum transfers Q^2 or small distances, and diverges at small Q^2 or large distances. This leads to the concepts of *colour confinement* and *asymptotic freedom*.

At small momentum transfer or large distances quarks are confined together inside colourless bound states, called hadrons. This concept called colour confinement states that only colourless composite particles can propagate as free particles. At large momentum transfer or small distances the coupling becomes small and the quarks and gluons can be treated as quasi-free particles. This concept is known as asymptotic freedom, and perturbative Quantum Chromodynamics (pQCD) calculations become applicable.

1.2 Quark-gluon plasma

QCD predicts different phases of nuclear matter, which can be explained by the evolution of α_s with the energy scale. As shown in Figure 1.1, the different phases of the strongly interacting matter depend on the change of the temperature T and the baryo-chemical potential μ_B quantifying the net-baryon content of the system. Ordinary nuclear matter exists at temperature T≈0 and μ_B=1 GeV. At low tem-

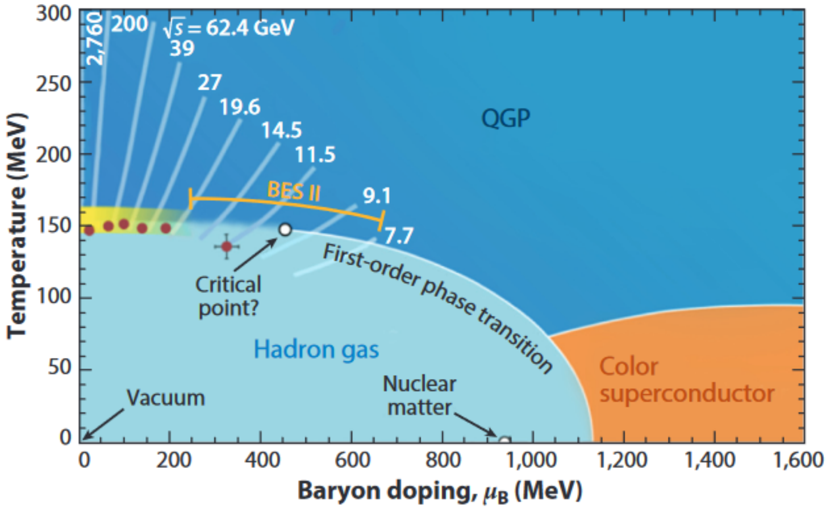


Figure 1.1: QCD phase diagram as function of the temperature T and the baryo-chemical potential μ_B [2].

peratures and low energy densities the quarks and gluons are confined. They only exist in colour neutral hadrons. For sufficient high temperatures and/or densities, quarks and gluons move freely over distances larger than the size of a nucleon and a phase transition between hadronic matter and a medium of so-called quark–gluon plasma (QGP) [3] is expected.

The quark–gluon plasma is a state of matter in which quarks and gluons can be considered free. It is believed that in early stages the Universes existed in the form of QGP. With ultra-relativistic heavy-ion (HI) collisions, it is possible to reproduce the density and temperature of this matter in laboratory conditions. The heavy ions reach nearly the speed of light and are therefore Lorentz contracted when they collide at high energies. The partons inside the nuclei scatter in hard processes. After the collision, the system undergoes a collective expansion and passes through different stages [4], which are shown in a space-time diagram in Figure 1.2. The two nuclei collide at $t = 0$, $z = 0$ and a fireball is created consisting of deconfined quarks and gluons. At this pre-equilibrium stage, the constituents of the fireball interact and the system thermalises, resulting in local equilibrium. After the expected thermalisation time of $\tau_0 \lesssim 1\,\mathrm{fm}/c$, a thermalised QGP forms and the quarks and gluons are no longer confined [6]. During the expansion, the system cools down

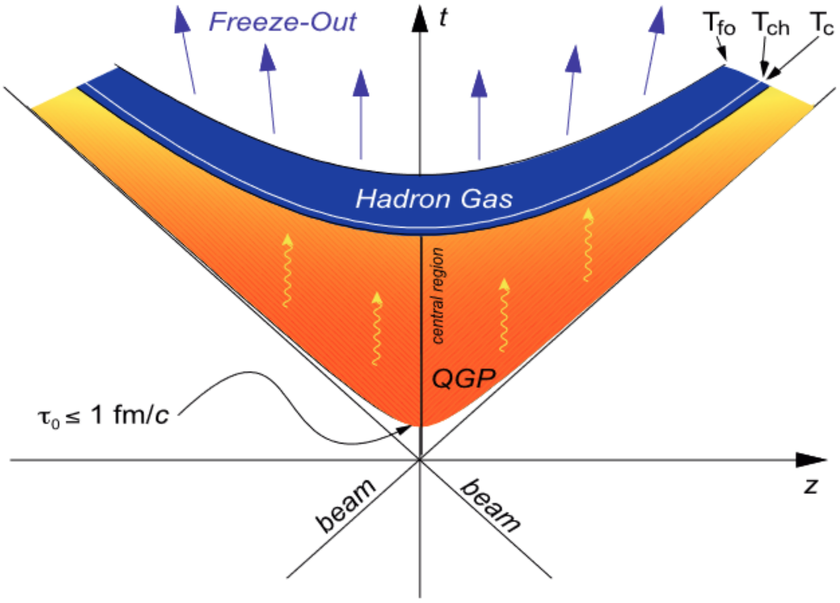


Figure 1.2: Different stages of the evolution of a heavy-ion collision in space-time [5].

as its energy density decreases. When the critical temperature T_C is reached, the quarks and gluons reconfine into hadrons. The hadron gas further expands while inelastic collisions between the constituents still take place. At the chemical freeze-out temperature T_{ch}, the inelastic scattering between the hadrons ceases and the hadron abundances are fixed. The system further expands and cools down until the temperature falls below the kinetic freeze-out limit T_{kin} and the elastic interactions also cease, leading to a fixed momentum distribution of the hadrons.

1.3 Charm production

1.3.1 Factorisation approach

In this thesis, open heavy-flavour hadrons, which are hadrons containing at least one heavy quark (charm or beauty), are studied. Top quarks are not considered because they cannot bind into hadronic states due to their short lifetime [1]. Charm and beauty quarks have large masses [7] and are therefore produced only in the initial hard scattering of high-energy collisions with large momentum transfer $Q^2 > 4m_{b,c}^2$. As stated earlier, the coupling strength α_s is small at large momentum transfer Q^2 and the charm production can be determined by pQCD calculations. The heavy quarks are created before the QGP can form, and therefore can be used to study and describe the QGP because they interact with it and experience the evolution of the system. At the LHC, open heavy flavour hadrons in Pb-Pb, p-Pb and pp collisions are used to study the QGP. Although QGP is not expected in pp collisions, the measurements serve as a reference for the other collisions and are studied to gain more insight into charm baryon production and hadronisation processes. The transverse momentum (p_{T}) - differential production cross section of hadronic collisions producing an open heavy-flavour hadron can be calculated using the QCD factorisation approach [8]:

$$\frac{\mathrm{d}\sigma}{\mathrm{d}p_{\mathrm{T}}}^{pp\to H_cX} = \sum_{i,j=q,\bar{q}g} \underbrace{f_i\left(x_1, Q^2\right) f_j\left(x_2, Q^2\right)}_{\text{PDF}} \cdot \underbrace{\frac{\mathrm{d}\sigma^{ij\to c\bar{c}}}{\mathrm{d}p_{\mathrm{T}}}}_{\text{Partonic Cross Section}} \cdot \underbrace{D_{c\to H_c}(z_c)}_{\text{Fragmentation Function}} , \tag{1.1}$$

where H_c refers to open heavy-flavour hadrons containing a charm quark c, and p_{T} to their transverse momentum. The *parton distribution functions* (PDFs) describe the probability of finding a quark or gluon in the colliding hadrons with a specific fraction

of the total momentum. Because of the large mass of the charm quarks, the *parton hard-scattering cross section* for the production of charm quarks can be computed perturbatively. The *fragmentation functions* characterise the hadronisation of the charm quark c into a particular hadron H_c with the momentum fraction z_c. Unlike the hard-scattering cross section, the PDFs and the FFs describe non-perturbative processes and therefore have to be determined from measurements. The PDFs are parameterised from deep inelastic $\mathrm{e^-p}$ scatterings and the FFs are assumed to be the same for all collision systems and are usually taken from $\mathrm{e^+e^-}$ collisions.
Hadron-to-hadron production ratios are used to study the hadronisation process, since the PDFs and partonic interaction cross sections are independent of the final measured hadron species and almost completely cancel in the ratios. The yield ratio is therefore sensitive only to the hadronisation process described by the fragmentation functions.

1.3.2 Hadronisation

The hadronisation process is not yet fully understood, and various approaches, most notably the fragmentation mechanism and coalescence, are used to describe it and compared with measured data.
The *fragmentation* process is not affected by QGP and takes place in vacuum as it occurs due to the colour confinement of separating quarks and antiquarks. As the distance between two quarks increases, the QCD potential between them grows linearly with the distance and forms a high tension string. The potential becomes sufficient to form new quark-antiquark pairs from the vacuum, and the string fragments into hadrons. Unlike fragmentation, *coalescence* describes the hadronisation of partons by the recombination of quarks and is therefore not possible in vacuum, but in a parton-rich environment. Heavy quarks in a deconfined medium, such as the QGP, recombine with light quarks close in phase space to form hadrons.

1.3.3 Baryon-to-meson ratio

Analyses by the ALICE Collaboration report higher Λ_c^+/D^0 baryon-to-meson ratios in pp and p-Pb collisions [9], and Ξ_c^0/D^0 and Ξ_c^+/D^0 baryon-to-meson ratios in pp collisions compared to previous measurements in $\mathrm{e^+e^-}$ and $\mathrm{e^-p}$ collisions [10][11]. This enhanced production of baryons over mesons in hadronic collisions challenges the assumption that the fragmentation fractions of charm quarks are universal across different collision systems. Measurements of the Λ_c^+/D^0 ratio in pp and p-Pb collisions at $\sqrt{s_{\mathrm{NN}}}\ = 5.02\,\mathrm{TeV}$ measured by the ALICE Collaboration as a function of p_{T} are shown in the left panel of Figure 1.3. For both pp and p–Pb collisions, the

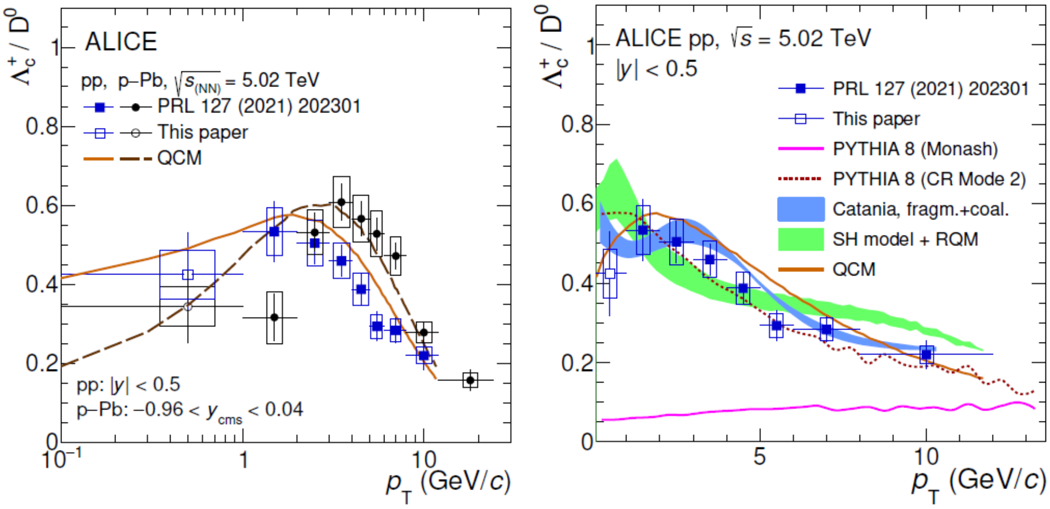


Figure 1.3: Λ_c^+/D^0 ratio measured at $\sqrt{s_{\mathrm{NN}}} = 5.02\,\mathrm{TeV}$ as function of p_{T} [9]. Left: Λ_c^+/D^0 ratio in pp and p–Pb collisions, compared with the QCM model. Right: Λ_c^+/D^0 ratio in pp collisions, compared to different model predictions.

Λ_c^+/D^0 ratio decreases for low p_{T} with large uncertainties and reaches a maximum for intermediate p_{T}, in the range $1 < p_{\mathrm{T}} < 3\,\mathrm{GeV}/c$ for pp and $3 < p_{\mathrm{T}} < 5\,\mathrm{GeV}/c$ for p-Pb collisions. For low p_{T}, the Λ_c^+/D^0 yield ratio for pp collisions exceeds the ratio for p-Pb collisions, while the peak of the ratio in p-Pb collisions is shifted towards higher p_{T}. Computing the average transverse momentum confirms this modified Λ_c^+ production spectrum, as the mean value is significantly higher in p-Pb collisions [9]. In contrast, the average transverse momentum for the D^0 mesons is the same in both collision systems [9]. The modified p_{T} dependence of the Λ_c^+/D^0 ratio in p-Pb collisions compared to pp collisions indicates modifications of the p_{T} shape depending on the multiplicity of the collisions, which could be due to a contribution of collective effect such as radial flow observed in heavy ion collisions. Nevertheless, the p_{T}-integrated Λ_c^+/D^0 ratios of both collisions are consistent with each other, indicating similar overall hadronisation fractions for pp and p-Pb collisions.

The right panel of Figure 1.3 shows the Λ_c^+/D^0 ratio in pp collisions and model calculations based on different hadronisation processes. PYTHIA 8 with the Monash tune used in this analysis is tuned on charm production measurements in e^+e^- collisions and predicts a value of about 0.1 throughout all p_T values. The model underestimates the Λ_c^+/D^0 ratio by a factor of about 6 to 10 for low p_T values and by a factor of about 3 at high p_T values. Models implementing different hadronisation approaches that enhance baryon production, like colour reconnection (CR) beyond the leading-colour (LC) approximation and coalescence or the statistical hadronisation model (SHM) including predictions by the relativistic quark model (RQM), seem to reproduce the Λ_c^+/D^0 yield ratio much better. All models are further explained in Ref. [9].

Figure 1.4 shows the baryon-to-meson ratios Ξ_c^+/D^0 and Ξ_c^0/D^0 measured in pp collisions at $\sqrt{s} = 13\,\text{TeV}$ compared to different model predictions. Both ratios are consistent within their uncertainties and show a similar p_T dependence, as they both decrease for $p_T > 3\,\text{GeV}/c$. Also in this case, the predictions from PYTHIA 8 with the Monash tuned on measurements in e^+e^- collisions significantly underestimate the ratios. Unlike for the Λ_c^+/D^0 ratio, models with CR beyond LC, SH including RQM or coalescence are also unable to describe the Ξ_c^+/D^0 and Ξ_c^0/D^0 ratios well, because they still underestimate the ratios by a factor of about 4 - 6 for $p_T < 4\,\text{GeV}/c$. The Catania model, which implements coalescence along with fragmentation, seems to best describe the trend of the measured results for the whole p_T interval.

The ratio distributions of Λ_c^+/D^0 in pp and p-Pb collisions (Figure 1.3) and of Ξ_c^+/D^0 and Ξ_c^0/D^0 in pp collisions (Figure 1.4) strongly indicate that the hadronisation of charm quarks into baryons and mesons differs across different collision systems and therefore is non-universal. An important step towards further understanding of charm production and hadronisation are measurements of the Ξ_c^+ and Ξ_c^0 baryon in larger systems and at even lower p_T, which have not yet been performed.

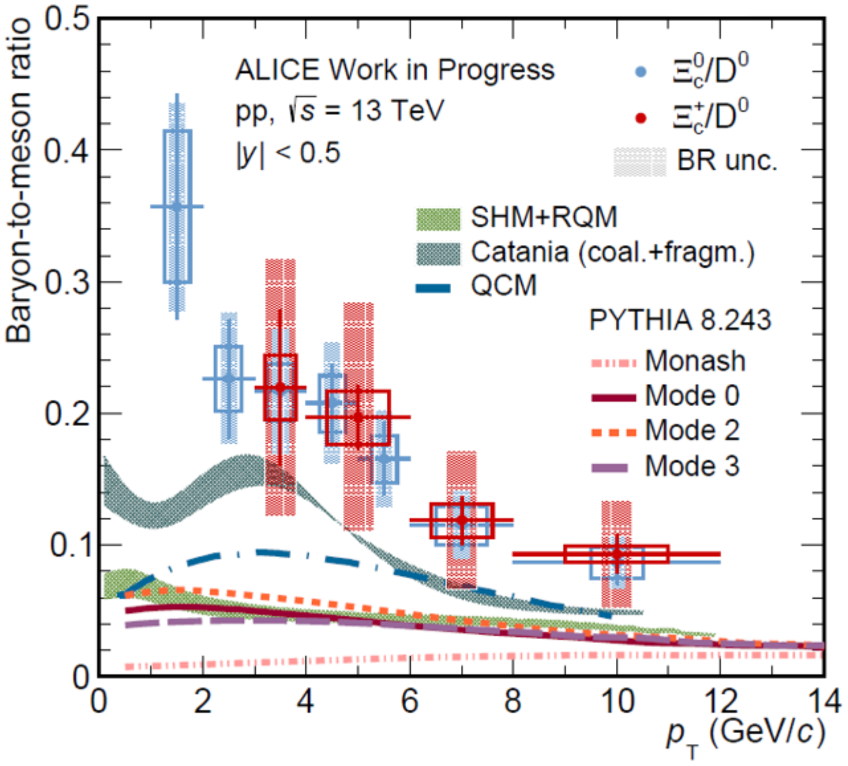


Figure 1.4: Ξ_c^+/D^0 ratio (red) and Ξ_c^0/D^0 ratio (blue) in pp collisions at $\sqrt{s} = 13$ TeV, compared to different model predictions [11].

2 The ALICE experiment

ALICE (A Large Ion Collider Experiment) is one of the main experiments at the Large Hadron Collider (LHC) at the European Organization for Nuclear Research (CERN). The LHC is the world's most powerful and largest particle accelerator with a circumference of 26.7 km [12]. The aim of ALICE is to study the collisions of heavy ions at ultra-relativistic energies. Bunches of particles, like protons and lead ions, are accelerated in the LHC and collide at the centre of the detector, the so-called interaction point (IP). In heavy-ion collisions, extreme energy densities are realised and a phase of matter called quark-gluon plasma (QGP) is created [13]. ALICE is designed to investigate this phase of strongly interacting matter. The maximum centre-of-mass energy achieved per nucleon–nucleon pair is $\sqrt{s} = 13.6$ TeV for pp and $\sqrt{s_{\mathrm{NN}}} = 5.36$ TeV for Pb–Pb in Run 3 [14], and $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV or $\sqrt{s_{\mathrm{NN}}} = 8.16$ TeV for p–Pb collisions in Run 2 [15].

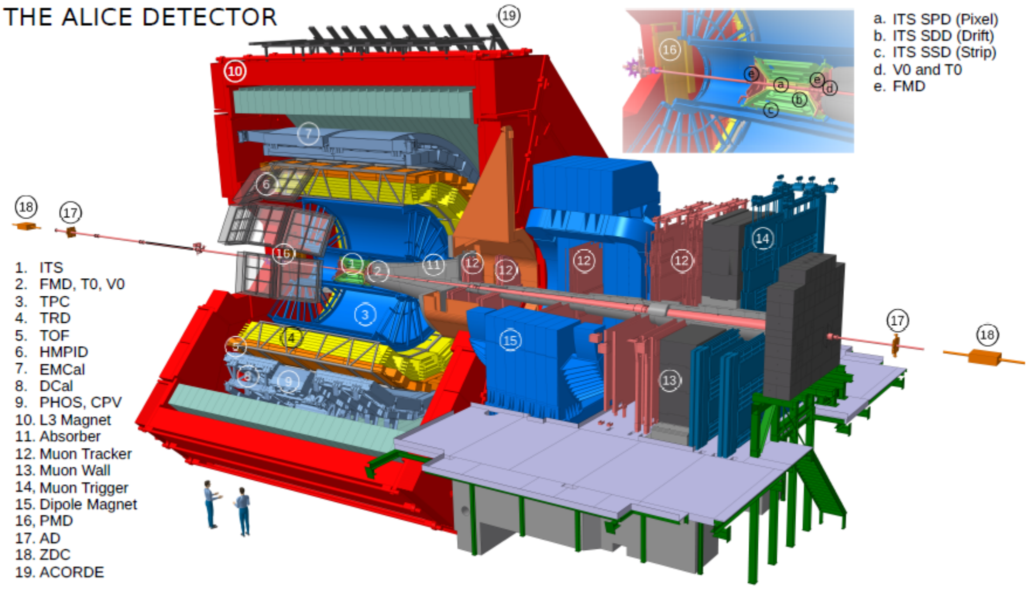


Figure 2.1: Schematic view of the ALICE detector during Run 2 [16].

Figure 2.1 shows an overview of the ALICE detector and its subdetectors as it was installed in Run 2. In total, the detector has a diameter of 16 m and a weight of 10.000 t. There are 18 different subdetector systems, most of which are arranged in the central barrel part of the detector and surrounded by a large solenoid magnet with a magnetic field strength of B = 0.5 T parallel to the beam axis [13]. The coordinate system used for ALICE is a right-handed orthogonal Cartesian system, which is centred in the middle of the central barrel [17]. The z-axis is parallel to the beam direction and the x-axis is perpendicular to the beam pipe pointing towards the centre of the LHC, whereas the y-axis is pointing upwards. The inner singular part of the central barrel consists of two tracking detectors, the Inner Tracking System (ITS) and Time Projection Chamber (TPC), which are surrounded by the Transition Radiation Detector (TRD) and the Time of Flight Detector (TOF). The three outermost detector parts are the High momentum particle identification detector (HMPID) and two electromagnetic calorimeters: the Photon spectrometer (PHOS) and the Electromagnetic calorimeter (EMCal). In Ref.[18] all detectors in ALICE are further described in detail.

In order to describe the kinematics of a particle of known mass inside the detector, the azimuth angle ϕ, the polar angle θ and the particle's transverse momentum p_{T}, which is the momentum in xy-direction can be used. In high energy collisions, the colliding particles and their constituents are Lorentz boosted along the beam axis. Since the polar angle θ is not Lorentz-invariant along the z-axis it is replaced by the rapidity y:

$$y = \frac{1}{2}\ln\frac{E + p_{\mathrm{z}}c}{E - p_{\mathrm{z}}c}, \tag{2.1}$$

where p_{z} is the momentum in the longitudinal direction, E the energy of the particle and c the speed of light.

A commonly used variable is the so-called pseudorapidity η:

$$\eta = \frac{1}{2}\ln\frac{|p| + p_{\mathrm{z}}}{|p| - p_{\mathrm{z}}} = \ln\left[\tan\left(\frac{\theta}{2}\right)\right]. \tag{2.2}$$

For high momentum particles, where $E \approx pc >> mc^2$, the pseudorapidity coincides with the rapidity y. The central barrel covers the pseudorapidity range of $|\eta| < 0.9$ [15].

2.1 Inner Tracking System

The Inner Tracking System (ITS) is the innermost detector located in the central barrel of ALICE. It consists of six layers of three different types of silicon-detectors arranged around the beam pipe. The main functions of the ITS are to reconstruct the primary vertex (PV), which is the measured collision point and to reconstruct the decay points of short-lived particles, the secondary vertices. In addition the ITS can also identify low-momentum particles via measurements of specific energy loss. The two inner layers of the ITS are Silicon Pixel Detectors (SPD) used for a first estimation of the position of the PV. For these layers a track density of up to 50 tracks/cm^2 is expected for the heavy-ion collisions at LHC [13]. The third and fourth layers are Silicon Drift Detectors (SDD), while the next two layers are Silicon Strip Detectors (SSD). These four layers can be used for particle identification (PID) of low momentum particles, as they provide a measurement of the specific ionisation energy loss $\mathrm{d}E/\mathrm{d}x$ [18]. The four outer layers and the high spatial resolution of the two SPD layers can be used to measure the impact parameter of secondary tracks from the weak decay of particles containing heavy-flavour quarks such as charm and beauty, and to reconstruct their secondary vertices. The impact parameter is the distance of closest approach (DCA) between the tracks and the PV.

2.2 Time Projection Chamber

The Time Projection Chamber (TPC) is the primary detector in ALICE for track reconstruction [13]. Together with the other detectors of the central barrel, it provides precise charged-particle momentum measurements with good two-track separation and PID. The TPC encloses the ITS and covers the full azimuth around the beamline and is divided longitudinally by a central high-voltage electrode with a uniform electric field between the high-voltage electrode and the endplate electrodes. The cylindrical field cage is filled with a $\mathrm{NeCO_2N_2}$ or $\mathrm{ArCO_2N_2}$ gas mixture [13], which is ionised by charged particles passing through it. The electrons emerging from the ionisation drift along the electric field towards the end-plates, which consist of Multi-Wire Proportional Chambers (MWPC). The signal gets amplified at the MWPC by further ionisation processes. An avalanche process starts and a mirror signal is induced on the chamber backplanes that is read out by several readout pads. The readout chambers contain 159 pad rows along the radial direction, leading to a maximum of 159 clusters in the TPC for a passing track. The charge induced signal on a pad-row has to exceed a certain threshold and fulfill several quality criteria to be detected as a cluster. Therefore, to determine the number of crossed rows, the number of pad rows without signal but with clusters in both adjacent rows is added to the number of clusters.

The PID information provided by the TPC is based on simultaneously measuring the specific energy loss ($\mathrm{d}E/\mathrm{d}x$), the charge of a particle and its momentum. The TPC allows to reconstruct particles in a transverse momentum range of about 0.1 GeV/c to 100 GeV/c [13] while maintaining a good momentum resolution. To describe the mean energy loss per unit path length of a particle with charge z passing through a material with atomic number Z and mass number A, the Bethe-Bloch formula [19] is used:

$$\left\langle -\frac{\mathrm{d}E}{\mathrm{d}x} \right\rangle = Kz^2 \frac{Z}{A} \frac{1}{\beta^2} \left[\frac{1}{2} \ln \frac{2m_e c^2 \beta^2 \gamma^2 W_{\max}}{I^2} - \beta^2 - \frac{\delta(\beta\gamma)}{2} \right], \tag{2.3}$$

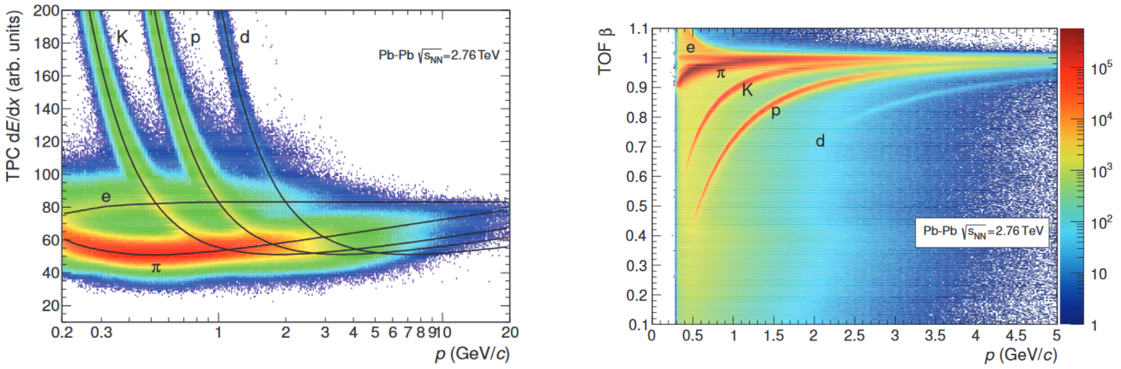
where β is the particle velocity, γ the Lorentz factor, and m_e the electron mass. $W_{\max}$ is the maximum energy transferred to an electron in one collision, I is the mean excitation energy and $\delta(\beta\gamma)$ is a density correction specific of the medium. To get PID information, the measured energy loss in the TPC is compared with the expected $\mathrm{d}E/\mathrm{d}x$ for specific particle species and momenta, since the Bethe-Bloch formula depends on the traversed medium and the particle $\beta\gamma$, where $\beta\gamma = \frac{p}{m}$. A parameterisation of the Bethe-Bloch formula [18] is used to describe the expected mean energy loss of particles traversing the TPC:

$$f(\beta\gamma) = \frac{P_1}{\beta^{P_4}} \left[P_2 - \beta^{P_4} - ln\left(P_3 + \frac{1}{(\beta\gamma)^{P_5}} \right) \right], \tag{2.4}$$

where P_{1-5} are parameters from fits to measured data.

Figure 2.2a shows the distribution of the specific energy loss ($\frac{\mathrm{d}E}{\mathrm{d}x}_{\mathrm{TPC}}$) measured by the TPC as a function of the particle momentum p in Pb-Pb collisions. To be able to select individual particle species i, the deviation n between the measured and the expected energy loss in terms of the measurement resolution σ is used:

$$n^{\sigma}_{i,\mathrm{TPC}} = \frac{\frac{\mathrm{d}E}{\mathrm{d}x}_{\mathrm{TPC}} - \left\langle \frac{\mathrm{d}E}{\mathrm{d}x} \right\rangle_i}{\sigma_{\mathrm{d}E/\mathrm{d}x}}. \tag{2.5}$$



(a) (b)

Figure 2.2: a) Specific energy loss for different particle species measured by the TPC in Pb-Pb collisions as function of the particle momentum. The lines show the expected energy loss. b) The particle velocity β measured by the TOF as function of the particle momentum in Pb-Pb collisions. Both figures are from [18].

2.3 Time-of-Flight detector

The Time-Of-Flight (TOF) detector is a large array of Multi-gap Resistive-Plate Chambers (MRPC), covering the full azimuth and providing PID information for the intermediate momentum range. Each MRPC consists of two stacks of glass plates that form a series of small gas gaps. A uniform electric field is generated over the full gaseous volume. Similarly to the TPC, charged particles traversing the detector ionise the gas between the glass plates of each MRPC. The emerging electrons are accelerated by the electric field and induce signals in the pickup electrodes on the outer surface [13]. The total signal is then calculated as the sum of all signals from the gaps, leading to a time resolution down to 40 ps [13]. For PID the TOF measures the flight times t of the particles from the interaction point to the TOF detector. Since the velocity v of a particle depends on its momentum p and mass m, the particle's mass can be calculated from measuring p, t and L, which is the length along the particle trajectory [20]:

$$\beta = \frac{v}{c} = \frac{L}{tc} = \frac{1}{(\frac{mc}{p})^2 + 1} \implies m = \frac{p}{c}\sqrt{\frac{c^2t^2}{L^2} - 1}. \tag{2.6}$$

Figure 2.2b shows the TOF PID performance, where the measured velocity β is plotted as a function of the particle momentum p measured by the TPC for Pb–Pb collisions. By comparing the two Figures in 2.2 it becomes clear that at momenta above 1 GeV/c where the particle bands are crossing each other for the TPC, making a separation impossible, the TOF detector can complement the PID. As for the measurements in the TPC, the $n^{\sigma}_{i,TOF}$ TOF is used as a discriminating variable and is defined as:

$$n^{\sigma}_{i,TOF} = \frac{|t - \langle t \rangle_i|}{\sigma_{TOF}}, \tag{2.7}$$

where the deviation between the measured time of flight t_i and the expected time of flight $\langle t \rangle$ is divided by the TOF resolution σ_{TOF}.

2.4 Event reconstruction

An event is a snapshot of a collision described by a main interaction point, the PV, and the tracks emerging from this collision measured by the detector. Particles and their decay products induce signals measured by the different detectors. For the event reconstruction these signals are combined, making it possible to reconstruct the full trajectory of the particle. In the central barrel the tracking procedure starts with the clusterisation, converting the detector signals into clusters, which are characterized by positions, signal amplitudes and their associated errors [18]. First the clusters in the SPD are used to determine a preliminary interaction vertex. The interaction vertex is defined as the point where a maximum number of tracklets, which are pairs of clusters in the SPD, converge. The actual track reconstruction in ALICE is accomplished by an inward-outward-inward approach and starts with finding tracks in the TPC. Therefore the preliminary interaction vertex is used along with clusters at high radius in the TPC to build track seeds. The track seed are built in two passes, once with and once without constraint on the determined interaction vertex [18]. The seeds are then propagated inwards to the inner TPC radius and the track parameters are repeatedly updated with each new found cluster using the Kalman Filter algorithm [21]. The 159 tangential pad rows in the TPC readout chambers allow a track to theoretically produce 159 clusters. Therefore tracks with less than 20 out of the 159 maximal possible clusters are rejected and not further considered. These reconstructed TPC tracks are then propagated to the outermost ITS layer and are used as a starting point for track finding in the ITS, similarly to the procedure in the TPC. As the reconstruction efficiency in the TPC decreases for low momenta and many low momentum tracks are not reaching the TPC, a standalone ITS reconstruction is performed with the clusters that are not used for the ITS-TPC tracks.

The second step starts with extrapolating the reconstructed tracks to their point of closest approach to the preliminary interaction vertex. Then the tracks are refitted in the outwards direction, until the fitting reaches the TRD. The track is then matched with a TRD tracklet and later with clusters in the TOF and with signals in the Electromagnetic Calorimeter (EMCal), the High Momentum Particle Identification Detector (HMPID) and the Photon Spectrometer (PHOS).

At the final stage of the track reconstruction, all tracks are again refitted inwards to the interaction vertex, starting from the outer radius of the TPC. The final track properties like the position, direction, inverse curvature, together with the covariance matrix are determined. At the end of the event reconstruction procedure the final position of the PV is determined by a precise vertex fit with the global tracks.

3 Analysis methods

3.1 Decay reconstruction with the Kalman Filter Particle package

The Kalman Filter Particle package (KFParticle package) [22] is developed to reconstruct the full decay chain and the decay topologies of short-lived particles based on the Kalman-Filter method [23]. Unlike traditional vertexing packages that focus only on the reconstruction of the production and decay vertices, the KFParticle package additionally generates an estimate of the decayed particle's parameters and associated covariance matrix [23]. The Kalman Filter (KF) algorithm is a mathematical iterative procedure for estimating unknown variables. For this, the algorithm starts with a certain initial approximation of the estimator and than refines it with each additional measurement [21]. The KF algorithm consists of three steps. The first step is the initial approximation of a state vector $\mathbf{r_0}$ and its covariance matrix $\mathbf{C_0}$. For the reconstruction of a decayed mother particle this refers to an initial approximation of the position of its decay point and an estimation of its momentum and energy. The parameters of the particle are stored in the so-called state vector:

$$\mathbf{r} = (x, y, z, p_x, p_y, p_z, E)^T, \tag{3.1}$$

where $(x, y, z)^T$ is the position along the trajectory, $(p_x, p_y, p_z)^T$ is the particle momentum and E its energy [21]. The respective covariance matrix contains the parameter uncertainties. As the process continues, the parameter $s = \frac{l}{p}$ is added to the state vector, where l is the length of the particle trajectory and p is the momentum [23]. In the second step of the algorithm, predictions of the evolution of $\mathbf{r}$ and $\mathbf{C}$ are made. In this process, the state vector of one daughter particle is extrapolated to its point of closest approximation to the initial approximated decay vertex of the mother particle. The final step is the filtering, where the state vector is updated for each measurement. Refitting with all previous measurements gives an optimal estimate of the vector according to these measurements. This means that the properties of the daughter particles are used to update the parameters of the mother

particle. When all daughter particles are included in the procedure, the optimal state vector of the mother particle is obtained by geometrical fitting.
To improve the precision of the measurement, several constraints treated as one-dimensional measurements by the Kalman Filter [23] can be applied on the features of the particle. In the reconstruction of secondary vertices, a so-called *mass constraint* is used, which requires the mother particle to have a certain mass. An additional constraint can be made on the vertex of the reconstructed particle. This *topological constraint* is used to align the particle so that it points to its expected production vertex or any another vertex.

3.1.1 Output variables

Different quantities describing the vertex fit quality and the decay topology can be extracted after the reconstruction and used as criteria for the selection of specific reconstructed particle candidates.
For the reconstruction of a decayed particle by its daughters, a geometrical fitting procedure is performed. The variable χ^2_{geo}/NDF, where NDF is the number of degrees of freedom corresponding to the number of measurements used for the fitting, expressed the quality of this fit. It describes whether trajectories of daughter particles intersect within their uncertainties [24]. Small values of χ^2_{geo}/NDF indicate a high probability that the daughter particle trajectories intersect within their uncertainties and therefore indicate a high likelihood that they emerge from a common vertex. The χ^2_{topo}/NDF estimates the probability that a particle is actually produced at its assigned production vertex in the case where a topological constraint has been used to assign the candidate to that vertex. A small χ^2_{topo}/NDF indicates a high probability of the hypothesis that the particle is produced at its assigned vertex within the uncertainties of the reconstructed trajectory and vertex [24].

3.2 Machine learning techniques

Machine learning (ML) techniques are an important tool for analyzing data generated in high-energy physics experiments, as they can be used to extract information from large data samples. Machine learning algorithms can solve regression and classification tasks. In this analysis supervised ML is used for the binary classification of signal and background of Ξ_c^+ candidates measured in p-Pb collisions at the centre-of-mass energy $\sqrt{s_{\mathrm{NN}}} = 5.02\,\mathrm{TeV}$.

For supervised ML the input data set has to be labelled. To solve a classification task, the learning goal of the model is to estimate a function that maps the input instances to a set of class labels using the labelled input data and its characteristics. During training, the algorithm learns the correlations between the input variables and their label to fit a model that correctly describes unlabelled data. The model is then tested with an independent data sample to ensure that the algorithm makes correct predictions on unknown data and does not fit the training data set perfectly well without being generalisable. The ML algorithms often used in high-energy physics are Boosted-Decision Trees (BDT), which are a robust classification tool because they handle incomplete or imbalanced data sets well [25].

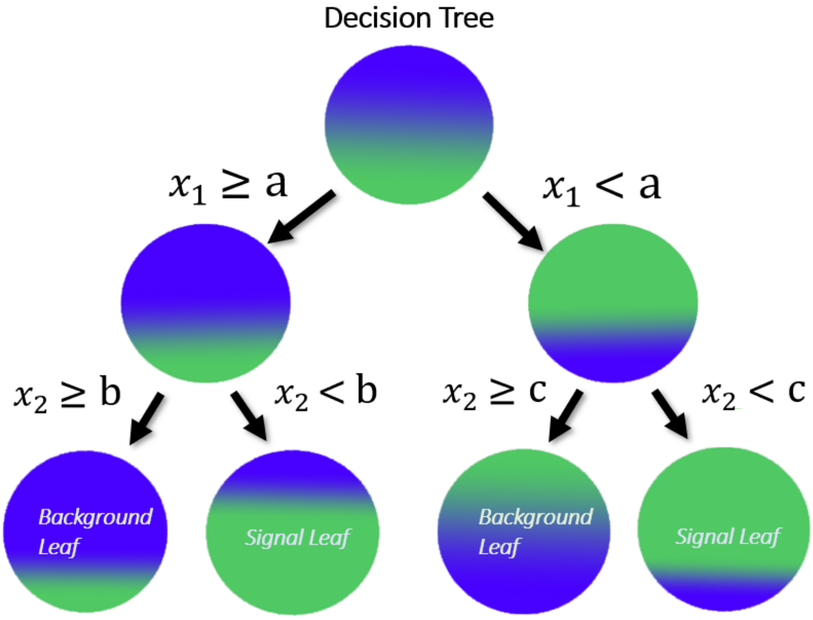


Figure 3.1: Single decision tree for binary classification of signal (green) and background (blue) candidates.

Single decision trees perform the classification by repeatedly splitting the dataset into smaller subsets that have a better class separation. Figure 3.1 shows a single decision tree of a two class problem separating signal and background, with each branch representing a certain subset. The initial node contains all candidates. The separation of background and signal candidates is done by applying selections on different features (x_i) of the candidates so that after each selection the subsets split into two smaller sets. The outcome of the decision tree is several final branches, called leaves, which contain subsets that are classified as signal or background, depending on which candidates are dominant in them. Since single decision trees, also known as weak learners, are unstable, so-called boosting is used to improve the classification by combining multiple weak learners [25].

In this analysis the python boosting algorithm *XGBoost* is used, and booster parameters, which control the structure of the algorithm, are optimized in a bayesian approach as describes in section 4.3.2.

4 Data analysis and results

The Ξ_c^+ baryon is reconstructed via its hadronic decay into two positively charged pions and a Ξ^- baryon. The Ξ^- baryon decays into a negatively charged pion and a Λ baryon, which further decays into a proton and a negatively charged pion. The mass of the Ξ_c^+, given by the Particle Data Group, is $M_{\Xi_c^+} = (2467.71 \pm 0.23)\,\mathrm{MeV}/c^2$ [7]. The Ξ_c^+ baryon is a short-lived particle, with a decay length of $\tau c = 136.6\mu\mathrm{m}$ [7], hence it is not directly detectable in the detector due to its short lifetime.
The goal of this analysis is to perform a feasibility study investigating the possibility of reconstructing and observing the Ξ_c^+ in p-Pb collisions. To reconstruct the decay chain and the decay topologies the KFParticle package (described in sec. 3.1), developed for the reconstruction of short-lived particles, is used [22]. The classification of signal and background is performed using supervised Machine Learning (ML), namely the XGBoost library, working with Boosted Decision Trees (BDTs). The Ξ_c^+ reconstruction is performed in the transverse momentum range $2 < p_\mathrm{T}(\Xi_c^+) < 12\,\mathrm{GeV}/c$, divided into the three intervals: $2 < p_\mathrm{T}(\Xi_c^+) < 4\,\mathrm{GeV}/c$, $4 < p_\mathrm{T}(\Xi_c^+) < 6\,\mathrm{GeV}/c$ and $6 < p_\mathrm{T}(\Xi_c^+) < 12\,\mathrm{GeV}/c$.

4.1 Candidate reconstruction and selection

4.1.1 Event selection

Prior to particle reconstruction, the events relevant to the analysis are selected by applying the selection criteria described in the following. The events are p-Pb collisions at the centre-of-mass energy $\sqrt{s_\mathrm{NN}} = 5.02\,\mathrm{TeV}$ collected by ALICE during LHC Run2 in 2016. To avoid edge effects, only events whose z-coordinate of the reconstructed primary vertex (PV), which is the measured collision point, lies within the range of $\pm$ 10 cm from the nominal interaction point are selected. This ensures uniform detector acceptance. Furthermore, so-called pileup events, recorded events including multiple collisions, are rejected. Since these pileup events have multiple vertices, they can be removed by correlating the information on the reconstructed tracks of the TPC and the ITS [18]. In total about 550 million minimum-bias triggered events are selected, corresponding to an integrated luminosity of $\mathcal{L}_{\mathrm{int}} = (263 \pm 8)\,\mu\mathrm{b}^{-1}$. In addition to the selected data events, signal generated in Monte Carlo (MC) simulations is needed to perform the ML training and the reconstruction efficiency calculation. The MC events are generated using PYTHIA 8 with Monash tune [26], simulating the detector conditions, during the p-Pb data taking. For p–Pb collisions, an underlying p–Pb event generated with the HIJING 1.36 generator [27] was added on top of the PYTHIA 8 event to simulate events with more than one nucleon–nucleon collision. The MC is heavy flavour enhanced, as the events are produced by injecting a $c\bar{c}$ pair. The signal candidates are taken from these events, as each event has to contain at least one Ξ_c^+ baryon, decaying via the hadronic decay of interest.

4.1.2 Decay reconstruction and track preselection

The reconstruction of the Ξ_c^+ is performed with the KFParticle package. The tracks used for reconstruction are preselected to reject poor quality tracks.
First, the Λ baryon is reconstructed. Since neutral particles are not detected in the detector, the Λ baryon is not tracked before decaying. Its neutral decay vertex is displaced from the primary vertex. Neutral particles, decaying into oppositely charged particles, like the Λ baryon, are called V^0 candidates, as they leave a V-shaped signature in the detector. The V^0 reconstruction of the Λ starts with the selection of oppositely charge tracks of protons and pions. The particle identification (PID) of these daughter tracks is realised by the specific energy loss measurement in the TPC and only protons and pions with $|n\sigma_{\mathrm{TPC}}| < 3$ are selected. Then cascade candidates like the Ξ^- are reconstructed. To form a Ξ^- the Λ baryon is combined with a secondary π^- track and the same PID selection as before, $|n\sigma_{\mathrm{TPC}}| < 3$, is applied for the pion. If information from the TOF detector is available for the pion tracks, it is used under the selection $|n\sigma_{\mathrm{TOF}}| < 3$. In order to select good quality candidates and to reject background away from the peak region, the reconstructed Ξ^- mass is required to lie in the range of $\pm\,5\,\mathrm{MeV}/c^2$ around the mass of the Ξ^- ($1321.71 \pm 0.07\,\mathrm{MeV}/c^2$) [7]. Finally, the selected Ξ^- candidates are combined with the tracks corresponding to the two positively charged pions. These pion tracks undergo the same selection criteria for the TOF and the TPC used before. The transverse momenta of the two decay pions coming from the Ξ_c^+ are constrained

to be higher than 0.3. This removes the large combinatorial background for low momenta. Further studies on the selection of these pions' transverse momenta will be discussed in the following chapters.

Besides these selections, topological and kinematic constraints are applied to the reconstructed Ξ_c^+ candidates to reject what is most probably only background. These preselection criteria are shown in Table 4.1. The pointing angle (PA) is defined as the angle between the line connecting the decay vertex of a reconstructed particle with its assigned production vertex and its momentum vector. By definition, correctly assigned production vertices correlate with a small pointing angle. Therefore the selection $\mathrm{PA}(\Lambda \to \Xi^-) < 0.5$ is used. Since the pseudorapidity coverage of the detectors is limited, the pseudorapidity selection $|\eta(\Xi_c^+)| < 0.8$ is made.

The variables χ^2_{topo} and χ^2_{geo} describe the vertex fit quality of the specific reconstructed particle candidates, and are further explained in chapter 3. Only candidates with $\chi^2_{\mathrm{geo}}(\Xi_c^+) < 50$ and $\chi^2_{\mathrm{topo}}(\Xi_c^+ \to \mathrm{PV}) < 50$ are selected.

Table 4.1: Preselection criteria.

Variable	Criterion
$\mathrm{PA}(\Lambda \to \Xi^-)$	< 0.5
$\|\eta(\Xi_c^+)\|$	< 0.8
$\chi^2_{\mathrm{geo}}(\Xi_c^+)$	$>0.$ and $< 50.$
$\chi^2_{\mathrm{topo}}(\Xi_c^+ \to \mathrm{PV})$	$>0.$ and $< 50.$
$\chi^2_{\mathrm{topo}}(\Xi^- \to \mathrm{PV})$	$>0.$
$p_{\mathrm{T}}(\pi^+ \leftarrow \Xi_c^+)$	$> 0.3\,\mathrm{GeV}/c$

4.2 Determination of expected signal

In order to have a first understanding of the feasibility of the analysis, the amount of expected Ξ_c^+ signal, and its significance on top of the large combinatorial background coming from the p-Pb data, is computed starting from pp measurements.

The expected number of signal candidates (s) is determined by rearranging the equation for calculating the p_{T}-differential production cross section $\frac{\mathrm{d}^2\sigma}{\mathrm{d}p_{\mathrm{T}}\mathrm{d}y}|_{\mathrm{p\,p}}$ of the

Ξ_c^+ measured in pp collisions [11]:

$$N_{\mathrm{raw}}^{\Xi_c^+,\Xi^-} = 2 \cdot 208 \left.\frac{\mathrm{d}^2\sigma}{\mathrm{d}p_{\mathrm{T}}\mathrm{d}y}\right|_{\mathrm{p\,p}} \cdot \Delta p_{\mathrm{T}} \Delta y \cdot (\mathrm{Acc} \times \varepsilon) \cdot \mathrm{BR} \cdot \mathcal{L}_{\mathrm{int}}, \tag{4.1}$$

where BR is the total branching ratio of the decay chain, and $\mathcal{L}_{\mathrm{int}}$ is the integrated luminosity. The factor 2 accounts for the presence of both particles and antiparticles in the raw yields, and $\Delta y \Delta p_{\mathrm{T}}$ accounts for the widths of the rapidity and transverse momentum intervals. The rapidity interval is $\Delta y = 1.6$, assuming the Ξ_c^+ rapidity distribution to be uniform in the range $|y| < 0.8$. The factor $(\mathrm{Acc} \times \varepsilon)$ is the product of the geometrical acceptance (Acc) and the reconstruction and selection efficiency (ε) for prompt Ξ_c^+ candidates in the $\Xi_c^+ \rightarrow \Xi^- \pi^+ \pi^+$ channel. The factor 208 is the mass number of lead ion. With some exceptions, it can be assumed that heavy-ion collisions can be considered as a superposition of many binary nucleon-nucleon interactions [28]. Under this assumption the mass number of lead is used to scale the known production cross-section of Ξ_c^+ measured in pp collisions so that it can be used as a proxy for the production cross section in p-Pb collisions.

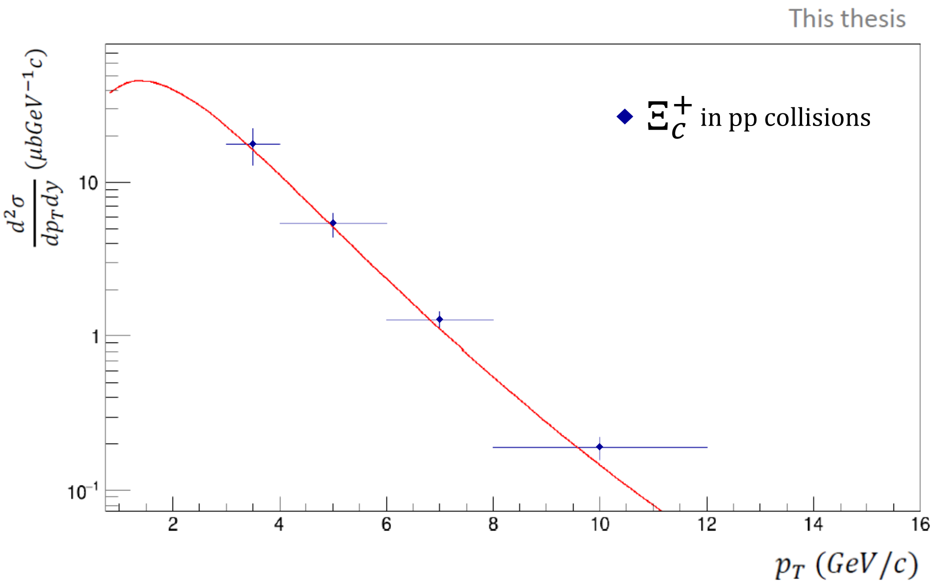


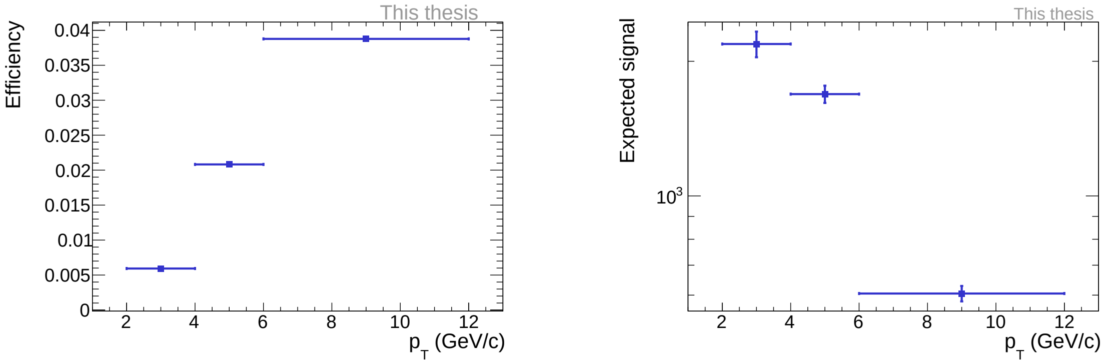
Figure 4.1: The p_{T}-differential production cross section of prompt Ξ_c^+ baryons in pp collisions at $\sqrt{s} = 13$ TeV, fitted with a Tsallis function (red line) [29]. The values of the cross section are taken from [11].

As the analysis is performed in three different p_{T} intervals, the cross sections at

$p_{\mathrm{T}} = 3\,\mathrm{GeV}/c$, $p_{\mathrm{T}} = 5\,\mathrm{GeV}/c$ and $p_{\mathrm{T}} = 9\,\mathrm{GeV}/c$ are required. Therefore the p_{T}-differential production cross section of prompt Ξ_c^+ baryons in pp collisions at $\sqrt{s} = 13\,\mathrm{TeV}$ taken from previous analysis [11] is fitted with a Tsallis-function [29]. The fit is shown in Figure 4.1. The required cross section values are extracted from this fit.

The branching ratio BR is defined as the fraction of a particular decay mode. Therefore, the BR of the analysed decay of the Ξ_c^+ is calculated considering the individual decay components of the Ξ^- and the Λ. The final branching ratio is $\mathrm{BR} = (1.83 \pm 0.08)\%$.

For this analysis, the total efficiency is calculated by dividing the number of reconstructed and preselected candidates by the number of generated prompt Ξ_c^+ candidates in MC. The calculated efficiencies and the expected number of signal candidates (s) for each of the three p_{T} intervals, with the preselections described in section 4.1.2, are reported in Figure 4.2a and 4.2b. The efficiencies for all p_{T} intervals are low in this analysis, which makes signal extraction challenging. The lowest efficiency is 0.005 for the p_{T} interval of 2 - 4 GeV/c, while the highest efficiency is 0.04, obtained for the p_{T} interval of 6 - 12 GeV/c.



(a) Efficiency for the three different p_{T} intervals.

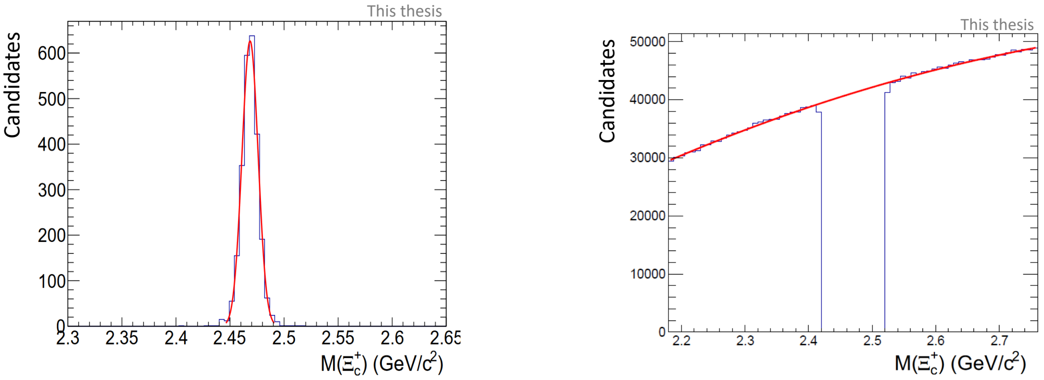
(b) Number of expected signal candidates for the three different p_{T} intervals.

Figure 4.2: Efficiency (left) and number of expected signal candidates (right) for the discussed preselections.

To examine if the expected signal emerges from the large combinatorial background, the so-called pseudo-significance (S) is calculated. The pseudo-significance is defined by the number of expected signal candidates (s) and the background candidates (b):

$$S = \frac{s}{\sqrt{s+b}}. \tag{4.2}$$

A significance higher than 3 indicates a signal peak structure on top of the background spectrum and therefore provides information on whether the extracted signal can be described as a statistical fluctuation or due to real signal. To estimate the number of background candidates, the invariant mass spectrum of the real p-Pb data sample with applied preselections is used. The background candidates of interest lie in the signal region defined by the $3\,\sigma$ range around the mean of the signal peak. The σ is determined by a Gaussian fit to the MC candidates, as shown in Figure 4.3a. To obtain the number of background candidates, the signal region is excluded from the invariant mass spectrum of the data sample and the two side-bands of the spectrum are fitted with a second order polynomial. The polynomial is extrapolated to the signal region and the integral below the fit in this region is used to estimate the number of background candidates. Figure 4.3b shows the background fit for the p_{T} interval of 4 - 6 GeV/c.

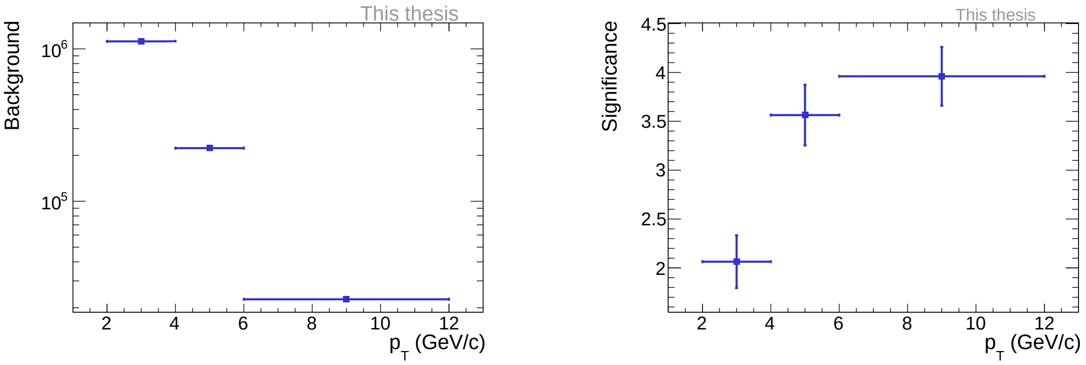


(a) Gaussian fit to the MC candidates.

(b) Background candidates fitted with a second order polynomial.

Figure 4.3: Invariant mass distribution of the Ξ_c^+ for $4 < p_{\mathrm{T}} < 6$ GeV/c.

The number of background candidates for the different p_{T} intervals are reported in Figure 4.4a. Together with the numbers of expected signal candidates the pseudo-significances for the different p_{T} intervals can be calculated. The values of these significances are shown in Figure 4.4b. The significances range from 2 to 4 and increase for higher p_{T}. For low p_{T} the combinatorial background is large and signal has an expected significance of about two. Even though the expected significance for the lowest p_{T} interval is not higher than three, the pseudo-significances can be interpreted as an indication that the signal extraction of the Ξ_c^+ is possible for the analysed data, as the significances might be improved by the use of BDT models.



(a) Number of background candidates for the three different p_{T} intervals.

(b) Pseudo-significance for the discussed preselections as a function of p_{T}.

Figure 4.4: Number of estimated background candidates (left) and Pseudo-significance (right) for the preselections discussed.

4.3 Machine learning

Since the signal of the Ξ_c^+ is rare for p-Pb collisions, the previously described preselections can be improved by using machine learning techniques for the separation of reconstructed signal candidates from combinatorial background. In this analysis, the gradient boosting algorithm XGBoost and thus Boosted Decision Trees (BDT) are used for this binary classification task. Different BDT models are trained for the three p_{T} intervals of the reconstructed Ξ_c^+ separately. Then, the trained models are applied to independent data samples for testing to see whether the model is generalisable and not just a good fit to the training data set. Before training, a set of input features and model hyperparameters have to be selected and optimised. The training of the model based on those variables combines the different selection features into one response variable, the BDT probability.

4.3.1 Input sample and training variables

The analysis is conducted for prompt Ξ_c^+ candidates, which are produced in the primary collision and decay via the decay channel $\Xi_c^+ \to \Xi^- \pi^+ \pi^+$. A small fraction of 10% of this decay channels of interest, decay via a resonance $\Xi_c^+ \rightarrow \Xi(1530)^0\pi^+ \rightarrow \Xi^-\pi^+\pi^+$ [7]. The training and testing of the model is performed with an input sample of background and signal data candidates. While the signal comes from simulated Ξ_c^+ events, the background is selected from real data candidates. To ensure that the background sample does not contain any true signal candidates, the signal region is excluded by selecting only candidates with an invariant mass outside the range 2.411 to 2.525 GeV/c^2. The available MC sample contains resonant and direct decay candidates. However, there are many more background candidates from the data than signal candidates provided by the MC sample. Due to this limitation of simulated events, only 20% of the real data is used to gather the background candidates, while the whole MC sample is exploited for the signal candidates. The proportion of signal to background for the training of the models is 1:2. An exception is the p_T interval between 6 - 12 GeV/c. For this interval, the number of background candidates is low, and all candidates are considered. The total numbers of candidates used for testing and training for the different p_T intervals are shown in Table 4.2. The input sample is randomly divided into two independent parts with 60% of the data for training and 40% for testing.

Table 4.2: Number of signal and background candidates for the BDT training and testing.

p_T (GeV/c)	2-4	4-6	6-12
Prompt, direct candidates	6895	7962	8301
Prompt, resonant candidates	2128	2846	3330
Background	18046	21616	23262

The training variables of the BDT model, used as classification criteria, are selected before the training. As the number of variables increases, the separation between background and signal might improve, but the model also becomes more complex and therefore less generalisable in case of limited statistics in the training sample. To avoid overtraining, only training features with the largest impact on the model performance are used. The variable's impact on the model output is defined by their feature importance, depending on how often the variables are used in the BDT process. Furthermore it has to be checked whether the training variables correlate with the Ξ_c^+ invariant mass. Correlations with the Ξ_c^+ mass for background candidates should be avoided, as they can modify the background shape of the invariant mass spectrum, artificially enhancing or reducing the extracted signal. However, possible correlations between the training variables that occur only in the background or in the signal can further improve the signal and background separation. In this analysis, the variables used for the Ξ_c^+ analysed in pp collisions [11] are used as a starting point for selecting optimal training variables. The signal and background distributions of all the variables used in this analysis are shown in Figure 4.5. Some of the distributions show a significant difference between signal and background, indicating high separation power of the variables. Throughout this analysis the most important feature is the pointing angle (PA) of the Ξ_c^+, which is defined as the angle between the line connecting the decay vertex of the reconstructed Ξ_c^+ with the PV and its momentum vector. For real and correctly reconstructed Ξ_c^+ the value of the PA should be small. Since the Ξ_c^+ has a very short lifetime and therefore decays close to its production vertex, the PA of the Ξ^- to the PV should be small too. Even though Ξ_c^+ is a short-lived particle, the distribution of its PA shown in Figure 4.5 indicates that the reconstruction of the decay vertex is still possible

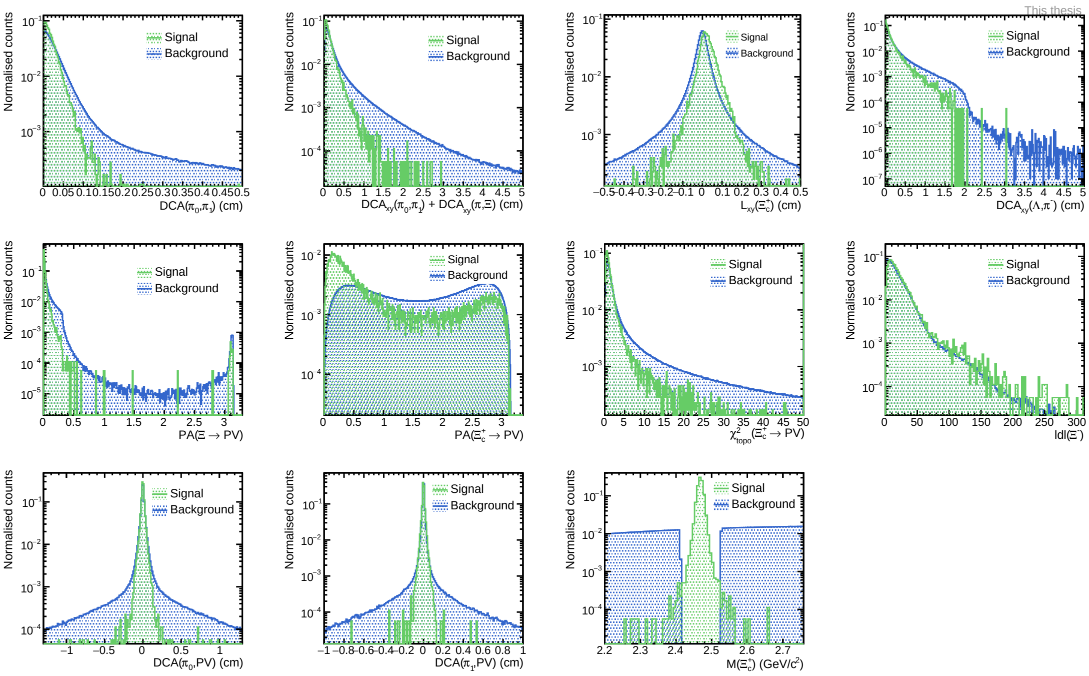
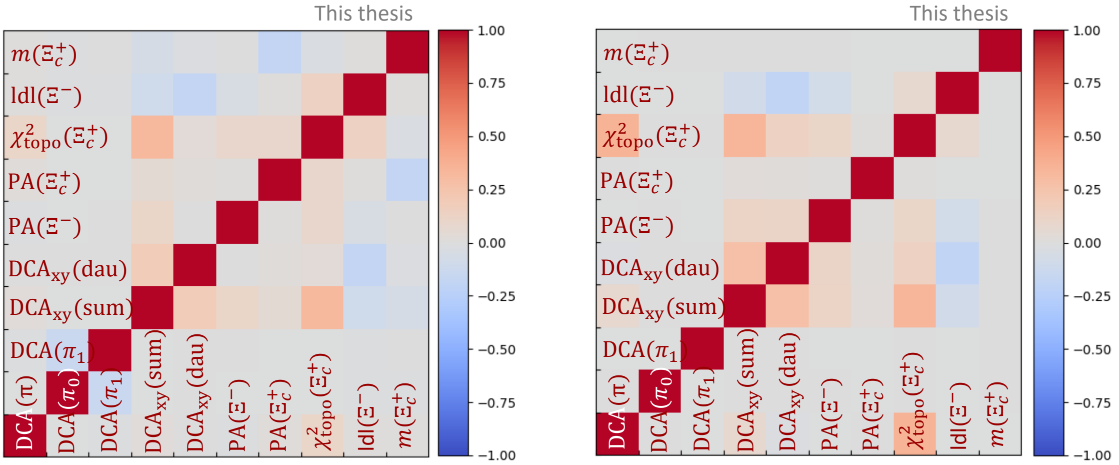


Figure 4.5: Signal (green) and background (blue) distributions of all available decay features in the range $4 < p_{\mathrm{T}} < 6$ GeV/c, normalised to the number of candidates.

since the signal distribution strongly increases for smaller values and differs from the background distribution. The distance of closest approach (DCA) between the two decay pion tracks in three dimensions and the DCA of each of them to the PV, as well as the sum of the DCA between the two pions and the DCA between the pions and the Ξ^- in xy-direction are also used. The DCA between the Ξ^- daughter candidates in xy-direction can also be used as an input feature, becoming less important for higher momenta. The reconstructed primary vertex is used as a constraint to the reconstructed tracks. The $\chi^2_{\text{topo}}/\text{NDF}$ characterises whether the momentum vector of the Ξ_c^+ candidate points back to the reconstructed PV, and is calculated by the KFParticle algorithm [22]. For high momenta, the particles are Lorentz-boosted and have a decay vertex further away from the PV. For these candidates, the decay length of the Ξ_c^+ can be used to separate signal and background, replacing the DCA of the Ξ^- daughters.

Figure 4.6 shows the correlations for the background and signal samples. No strong correlations between the training features and the invariant mass of the Ξ_c^+ are observed.



(a) Signal

(b) Background

Figure 4.6: Correlation matrix of the training features in the range $4 < p_{\text{T}} < 6\,\text{GeV}/c$. Correlations are indicated in red, anticorrelations in blue.

4.3.2 Hyperparameters

In addition to the training variables, a set of hyperparameters has to be chosen and optimised for the BDT model. In machine learning, a hyperparameter is a parameter used to control the learning process. The set of hyperparameters for the training is optimised to exploit the best performance of the BDT process. One possible optimisation procedure is the bayesian optimisation, which is applied in this analysis. Bayesian optimisation is an iterative procedure scanning specific sets of hyperparameters and incorporating information from prior evaluations to choose the next set. To avoid overfitting or selection bias the cross-validation method k-fold is used [30]. In k-fold cross validation, the data sample is randomly split into k equally sized folds. Then k-1 folds are used for training so that each parameter set is evaluated k-1 times while the remaining fold is retained for testing. The hyperparameters are only scanned in predefined spaces, which need to be chosen so that the optimisation process does not always converge to the lower or upper end of the given interval. This process is repeated k times under permutation of the folds, so that each fold is used once for testing. The k results can then be averaged to produce a single estimation. When choosing the parameter ranges different aspects like memory consumption, the risk of overfitting, and conservatism of the resulting model have to be considered. The hyperparameter optimisation in this analysis is conducted for the parameter ranges listed in Table 4.3, where for example the parameter range for the tree depth is restricted to 1 - 3 to avoid overtraining. For each of the models discussed in this thesis, the parameters are re-optimised, but the parameter ranges stay the same.

Table 4.3: Parameter ranges for the hyperparameter optimisation

Parameter	Range
Max. depth	(1,3)
Learning rate	(0.01, 0.1)
Estimators	(100,1000)
Min. child weight	(1,10)
Subsample ratio of rows	(0.8,1)
Subsample ratio of columns	(0.8,1)

Here, only the optimised sets of hyperparameters referring to the final models for the different p_{T} intervals discussed in section 4.4.3.1, 4.4.3.2 and 4.4.3.3 are listed in Table 4.4.

Table 4.4: Optimised Hyperparameters

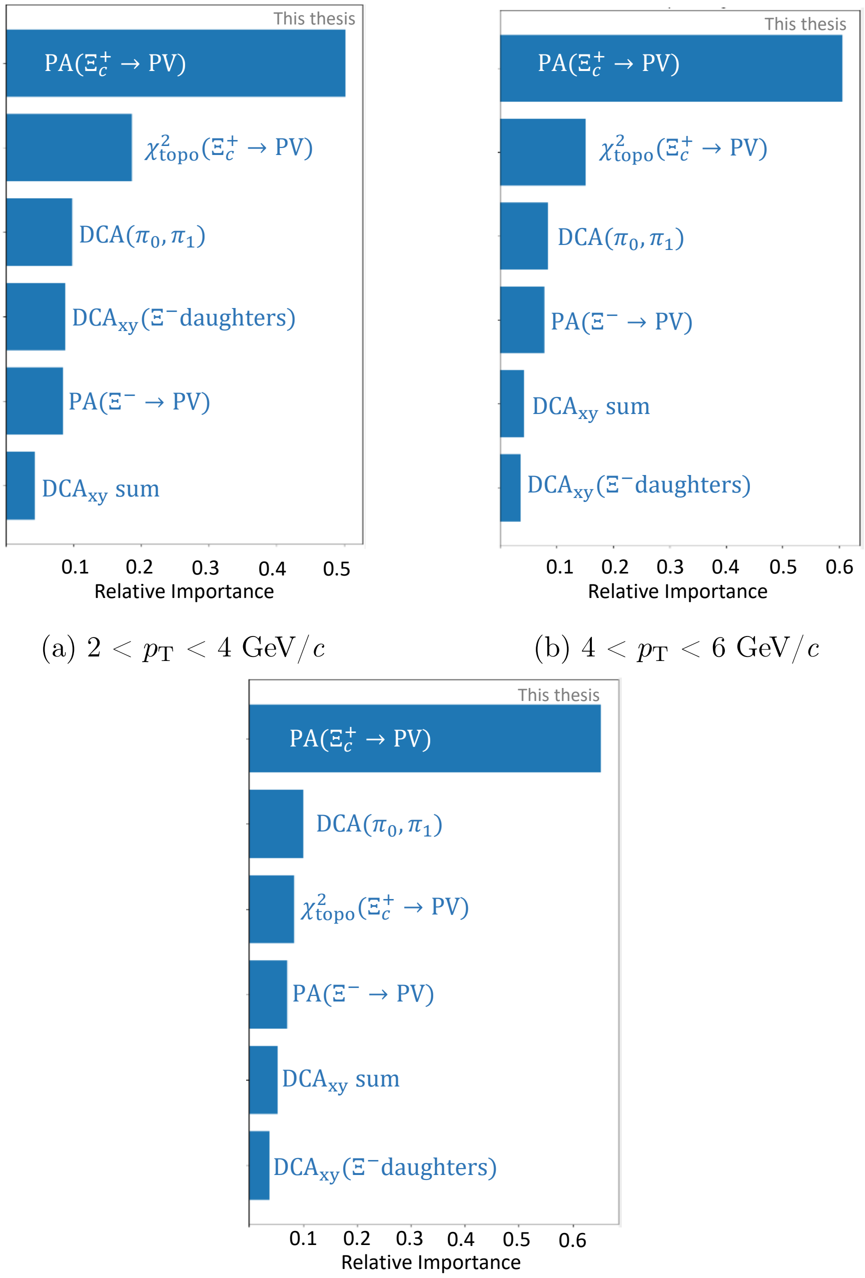
p_{T} (GeV/c)	2-4	4-6	6-12
Max. depth	2	3	3
Learning rate	0.04	0.04	0.06
Estimators	363	240	480
Min. child weight	2.3	7.0	5.1
Subsample ratio of rows	0.87	0.81	0.93
Subsample ratio of columns	0.89	0.83	0.96

4.4 Machine learning models

In the process of this analysis, several models with different training variables and preselections are used for the signal extraction. The models are trained, tested and applied to each p_{T} interval separately.

4.4.1 First models - loose preselection, basic training features

The machine learning part of this analysis starts with a first set of models, trained with loose preselections and six different training features. The preselections are the same as discussed in chapter 4.1.2, except that only decay pions with $p_{\mathrm{T}} > 0.4\,\mathrm{GeV}/c$ are selected, as in the previous analysis of Ξ_c^+ in pp collisions [11]. The six variables selected for the first models are already used in the previous analysis [11] for the Ξ_c^+ signal extraction in pp collisions for low p_{T} ranges and are therefore selected here to check if they are a good choice for extracting the Ξ_c^+ in p-Pb collisions. The feature importance of the training variables for the different p_{T} intervals are shown in Figure 4.7. The PA of the Ξ_c^+ to the PV is the most important feature for all p_{T} intervals, followed by the χ^2_{topo} of the Ξ_c^+, except for the highest p_{T} interval of 6 - 12 GeV/c, where the the DCA between the two decay pion tracks is more important. The DCA between the Ξ^- daughter candidates in xy-direction is the third most important feature for low p_{T} intervals, becoming less important for higher momenta. In the BDT process, the trained models are applied to the test set. To ensure that the model is neither overtrained nor undertrained the so-called learning curves are computed. The learning curves are defined as the root-mean-square error (RMSE) of the training set and the test set. The RMSE is the difference between the



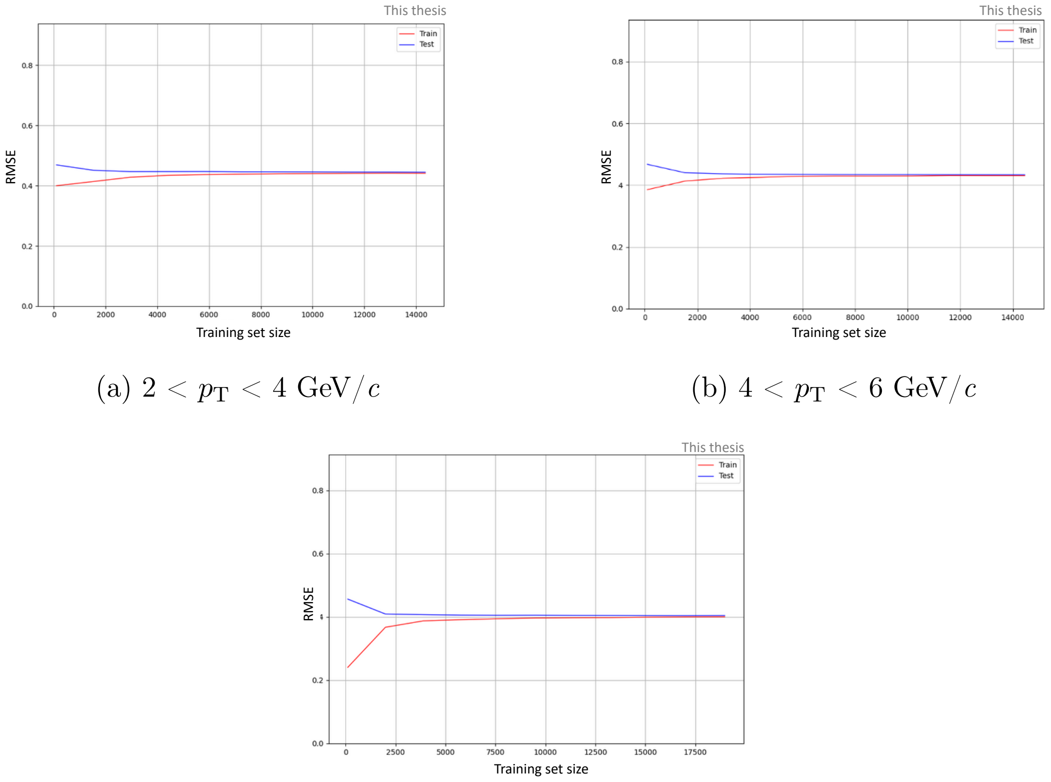
(a) $2 < p_{\rm T} < 4$ GeV/c

(b) $4 < p_{\rm T} < 6$ GeV/c

(c) $6 < p_{\rm T} < 12$ GeV/c

Figure 4.7: Feature importance ranking of the selected training variables for the first set of BDT models in different $p_{\rm T}$ ranges.

values from the model prediction and the observation, as a function of the training set size. The RMSE of the training set is close to zero, when only few instances are used for the training, as the fit is trivial and a nearly perfect description of the data is possible. When more instances are added to the training, the error increases as the fluctuations increase. For a training set with few instances, the model fails to correctly describe the test set, leading to an increased RMSE. For an increasing training set size, the error of the test set decreases, as the model performance and therefore the description of the test data improves. For an optimal model, the deviation between the training set and test set should be negligible. The two curves should stabilise at a certain set size and converge at a common value. The learning curves for the three models trained in the different p_{T} intervals are shown in Figure 4.8. The RMSE seems to stabilize for a training size with about 4000 instances and the two curves converge. The learning curves demonstrate that the number of candidates used for the training and testing provides a stable model that is neither undertrained nor overtrained.



(a) $2 < p_{\mathrm{T}} < 4$ GeV/c

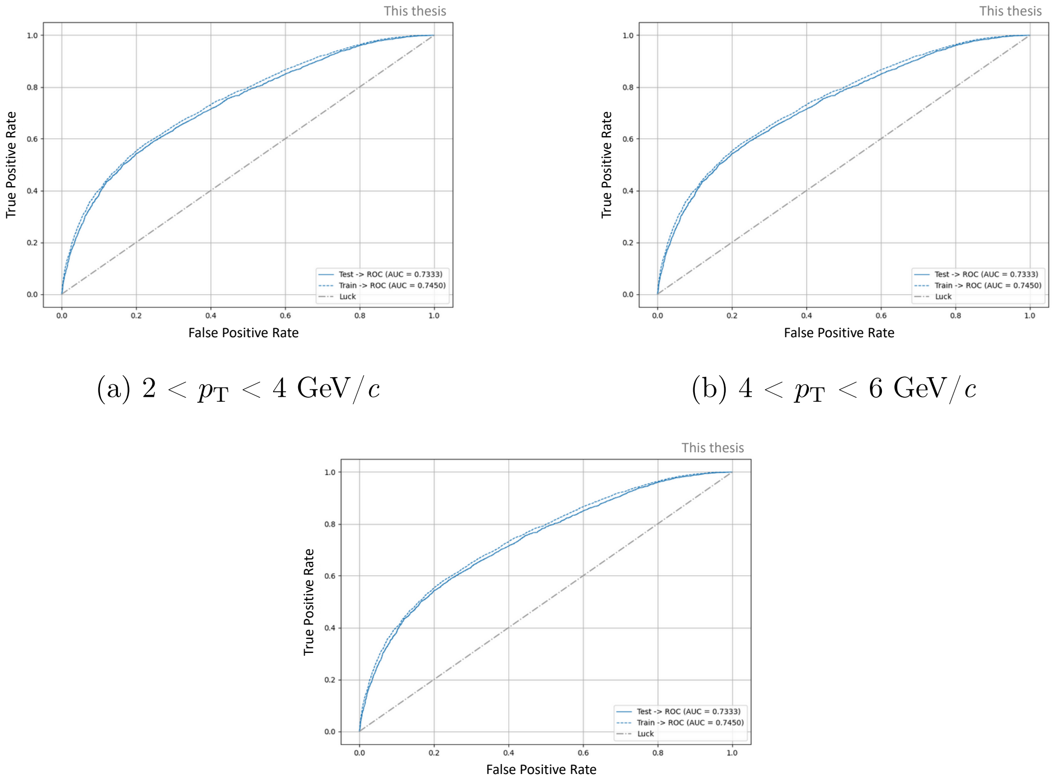
(b) $4 < p_{\mathrm{T}} < 6$ GeV/c

(c) $6 < p_{\mathrm{T}} < 12$ GeV/c

Figure 4.8: Learning curves: Root-mean-square error (RMSE) of training set (red) and test set (blue) as function of the training set size for the first set of BDT models in different p_{T} ranges.

An additional ML output that can be used to check the performance of the BDT model is the so-called Receiver Operating Characteristic (ROC). ROC is a probability curve, while AUC, the Area Under the ROC curve represents the ability of the model to classify the candidates correctly. In binary classification tasks the possible test results are true and false positives and negatives. The true positive rate (TPR) is defined as the fraction of the true positive, the correctly classified instances, and all instances of the signal class, the true positive and the false negative. The false positive rate (FPR) is the fraction of false positives, the wrongly classified instances, and all instances of the background class, the false positives and the true negatives. Therefore the TPR can be referred to as efficiency, while the FPR represents the error rate (1 - purity). The ROC curves for the training and the test set are shown in Figure 4.9. To obtain the curve, the true positive rate (TPR) is plotted as a function of the false positive rate (FPR) for different classification values. The grey line marks describes a random classifier, where the model prediction is correct in half of the cases. The curve of the test set should not deviate significantly from the curve of the training set to ensure a stable model performance without over- or underfitting. Strongly deviating curves show that the model might be too complex and thus not generalisable. Here, the model might also become not generalisable because the statistic of the data sample is too small. Since the AUC is interpreted as the probability that a true positive is correctly classified, the better the model can distinguish between signal and background, the larger it becomes.

The BDT model combines various selection features into a single response variable, the BDT probability. The model classifies each candidate of the training and the test set as background or signal and assigns it a certain BDT probability. In ideal models, the signal distribution would peak at one and decrease at lower probabilities, while the background would behave oppositely, peaking at zero. Additionally, the validation data sample should not deviate significantly from the training set to ensure a good model performance without over- or underfitting. In the process of the analysis trained models are applied to the full data samples for each p_{T} interval and a BDT output probability is chosen, below which all Ξ_c^+ candidates are rejected. The resulting BDT probability outputs of the trained models for the different p_{T} ranges are shown in Figure 4.10. The distribution of the test set follows the training set distribution. Even though the background and signal candidates peak at low and high probabilities, respectively, the signal does not decrease at low probabilities. Since the Ξ_c^+ baryon is a short lived particle with the decay length $\tau c = 136.6 \mu$m,



(a) $2 < p_{\mathrm{T}} < 4$ GeV/c

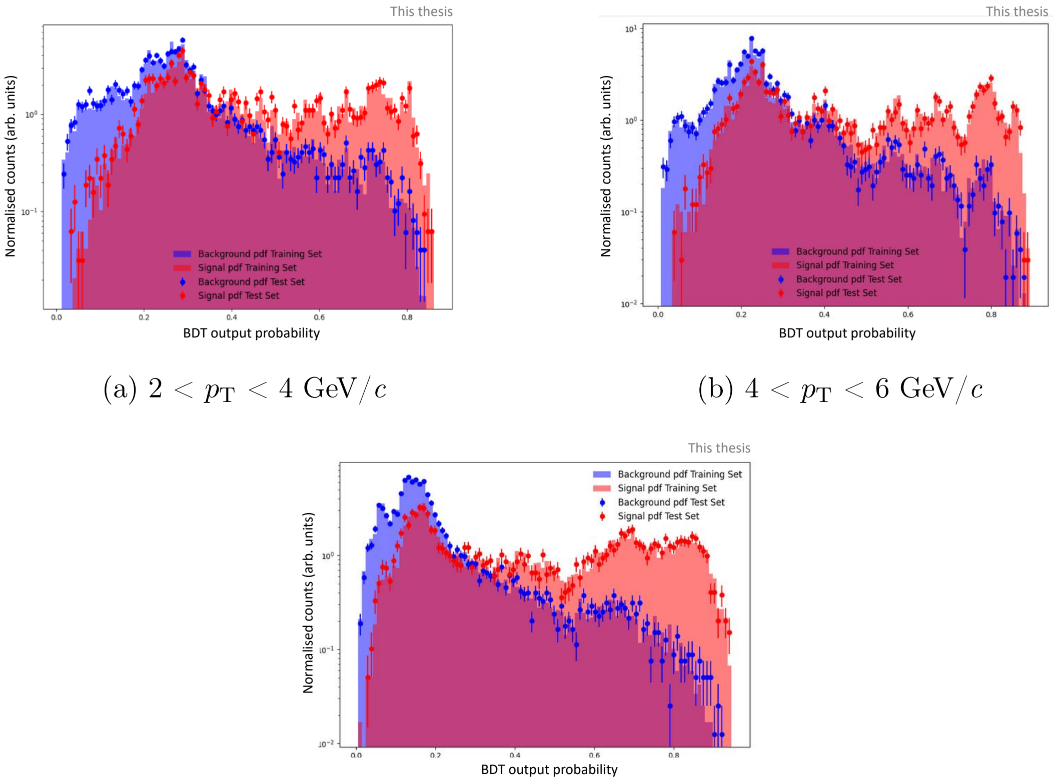
(b) $4 < p_{\mathrm{T}} < 6$ GeV/c

(c) $6 < p_{\mathrm{T}} < 12$ GeV/c

Figure 4.9: Receiver Operating Characteristic (ROC) curves of the training set and the test set for the first set of BDT models in different p_{T} ranges.

vertexing is possible but more difficult, especially for low p_{T}. This can be seen in the separation of the background and signal BDT probability distributions that become more significant for higher p_{T}.

The trained models are applied to the full data sample and after the selection of the candidates through the application of a BDT probability the Ξ_c^+ raw signal candidates are extracted via a fit to the invariant mass distribution. The fit combines a Gaussian fit to the signal region, defined by the 3σ range around the mean, and an exponential function to model the background. The width of the Gaussian function is fixed to the simulated MC to improve the stability of the fit. The overall fit is indicated by the blue, and the exponential fit of the background is described by the red line. For each fit, the number of signal (S) and background (B) candidates in the signal region, the signal-to-background ratio (S/B), the significance (s), the mean and the width are reported in the figures. In Figure 4.11, the fit results of the BDT



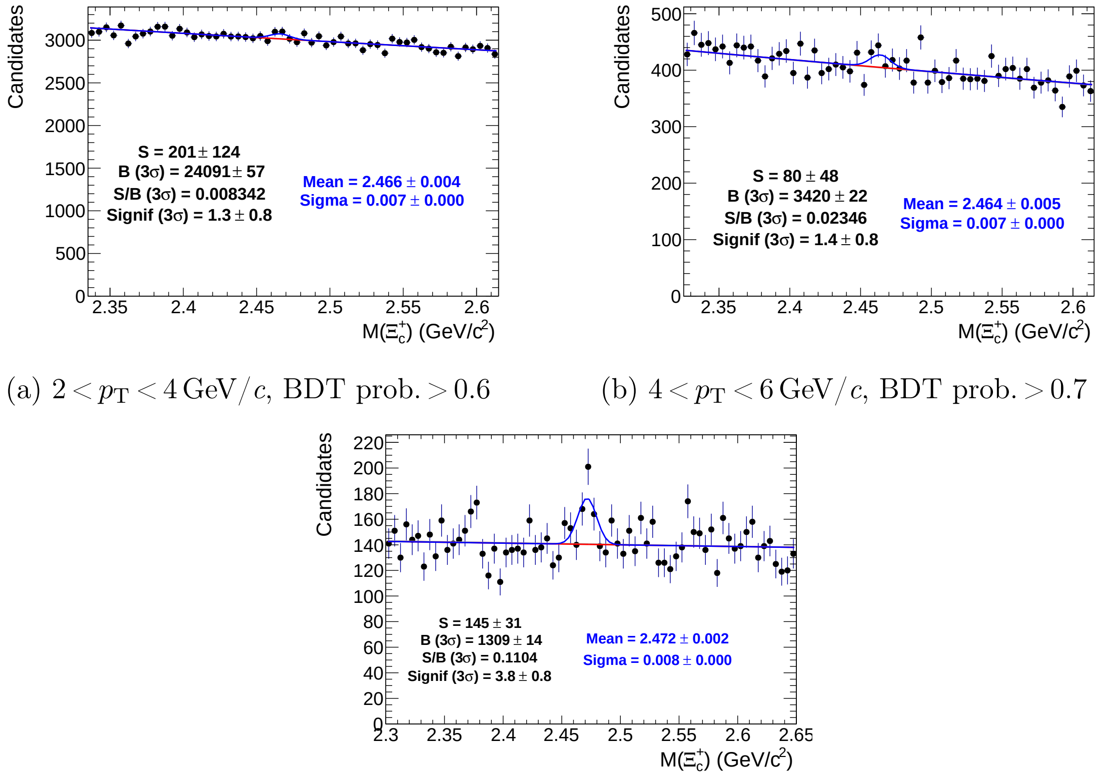
(a) $2 < p_{\mathrm{T}} < 4$ GeV/c

(b) $4 < p_{\mathrm{T}} < 6$ GeV/c

(c) $6 < p_{\mathrm{T}} < 12$ GeV/c

Figure 4.10: Model output probability for signal (red) and background (blue) candidates in the training set (bars) and the test set (full markers) for the first set of BDT models in different p_{T} ranges.

selections with the highest significances for the different p_{T} intervals are shown. In order for the signal not to be considered a fluctuation over the large background, the signal peak must have a significance higher than 3. Only for the p_{T} interval 6 - 12 GeV/c significant signal peaks were found for different BDT selections, with the highest significance for a BDT probability of 0.5. In general, the optimal BDT probability selection should be chosen using a blind approach without looking at the real data to avoid picking up statistical fluctuations. However, in this specific feasibility study the emergence of the signal peak over the background in real data is studied, and therefore no blind selection is performed.



(a) $2 < p_{\mathrm{T}} < 4\,\mathrm{GeV}/c$, BDT prob. > 0.6

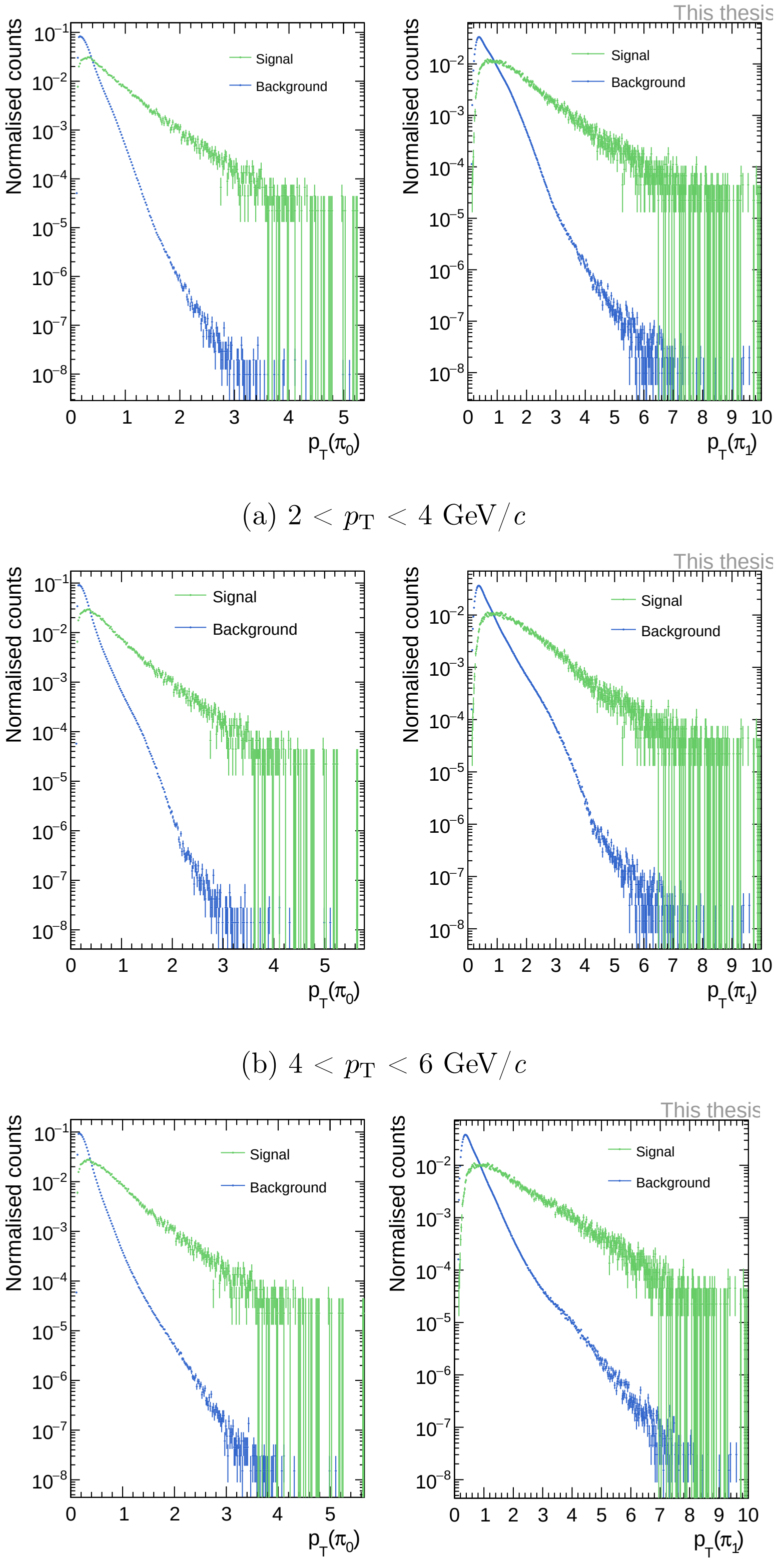
(b) $4 < p_{\mathrm{T}} < 6\,\mathrm{GeV}/c$, BDT prob. > 0.7

(c) $6 < p_{\mathrm{T}} < 12\,\mathrm{GeV}/c$, BDT prob. > 0.5

Figure 4.11: Invariant mass spectrum of Ξ_c^+ candidates for different p_{T} intervals. The number of extracted signal (s) and background candidates (b) are reported with the signal-to-background ratio (s/b) in the signal region and the signal significance (S).

4.4.2 Models trained with preselections on the p_{T} of the two decay pions coming from the Ξ_c^+

To improve the signal and background separation for the p_{T} intervals 2 - 4 GeV/c and 4 - 6 GeV/c, new BDT models with modified preselections are trained. To understand the effect of preselections on the transverse momentum of the two decay pions, the p_{T} distributions of the MC signal and the background from real data for both pions are plotted. To ensure that the real data do not contain true signal candidates, the signal region is excluded by selecting only candidates with an invariant mass outside the range 2.411 - 2.525 GeV/c. Figure 4.12 shows the distributions for all three p_{T} intervals. For both pions, the combinatorial background distribution is more peaked at small p_T values and is more steeply falling than the MC distribution. Thus, the



(a) $2 < p_{\mathrm{T}} < 4$ GeV/c

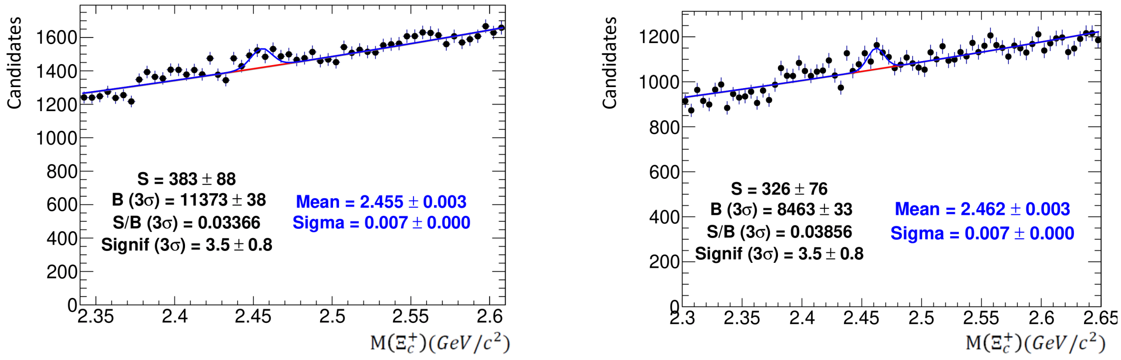
(b) $4 < p_{\mathrm{T}} < 6$ GeV/c

(c) $6 < p_{\mathrm{T}} < 12$ GeV/c

Figure 4.12: p_{T} spectra of the two decay pions for signal (green) and background candidates (blue) in the three different $p_{\mathrm{T}}(\Xi_c^+)$ intervals.

shape of the p_{T} distribution for the real candidates is harder than the one of the background, making it possible to reject background while preserving real signal candidates by applying selections on the p_{T} of the decay pions. Furthermore, the distributions indicate that pions with low and high p_{T} are stored separately, resulting in a hard and a soft spectrum for the pions. In this work π_1 refers to the pions with larger p_{T} in the decay reconstruction, while the pions of the soft p_{T} spectrum are referred to as π_0. The two different spectra indicate that it is reasonable to apply different preselections on the p_{T} of each of the two pions. However, the next set of models is trained with the same selection applied to both pions to explore the different pion p_{T} selections before choosing individual selection criteria later. The models are trained with the same six training variables used before (see Figure 4.7) but with different preselections on the transverse momentum of the two pions. Five selections have been tested ranging from selecting pions with momentum higher than $0.6\,\mathrm{GeV}/c$ to $1\,\mathrm{GeV}/c$, each corresponding to a new model. The range is chosen based on the distributions because the background exceeds the signal at lower transverse momenta. In the $p_{\mathrm{T}}(\pi_1)$ spectrum, the background distribution falls below the MC data at a p_{T} value of $1\,\mathrm{GeV}/c$, while for the π_0 this occurs at about $0.6\,\mathrm{GeV}/c$. The transverse momenta of the two pions are not included in the BDT model because they correlate with the Ξ_c^+ mass for background candidates. Such correlations must be avoided for training features, as ML learning can apply non-liner selection, potentially leading to a modification of the background shape that might enhance or reduce the extracted signal. Therefore, the variables have to be exploited by applying preselections on them. The other preselections remain the same as listed in Table 4.1. The modified preselection on the p_{T} of the two pions hardly changes the feature importance of the variables, and the models' performances, as can be seen in chapter 4.4.1.

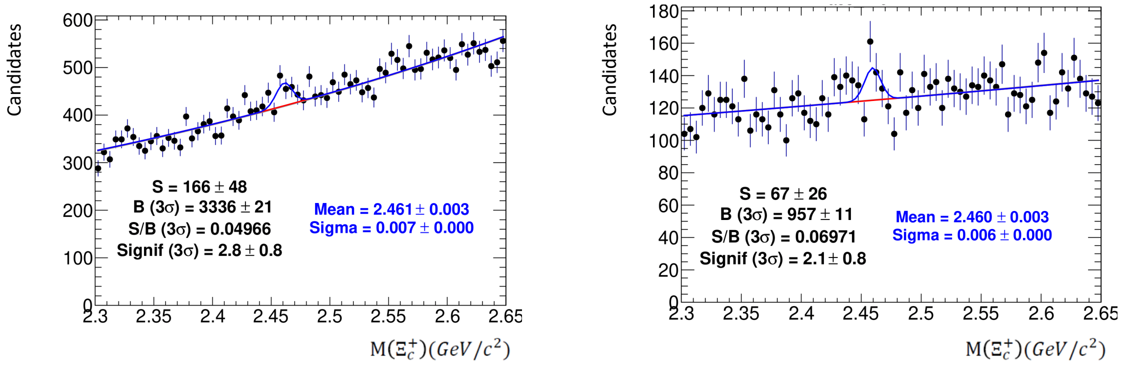
Figure 4.13 and Figure 4.14 show the invariant mass spectrum for the preselections with the highest signal significance for the two p_{T} intervals and different BDT selections. For the p_{T} interval 2 - 4 GeV/c, selecting decay pions with a transverse momentum higher than $0.7\,\mathrm{GeV}/c$ indicates a signal peak, while for all other preselections on the momentum no signal was visible.



(a) BDT probability > 0.4

(b) BDT-probability > 0.5

Figure 4.13: Invariant mass spectrum and fit of Ξ_c^+ candidates in $2 < p_\mathrm{T} < 4\,\mathrm{GeV}/c$ for the BDT model only selecting decay pions with $p_\mathrm{T} > 0.7\,\mathrm{GeV}/c$.



(a) BDT probability > 0.3

(b) BDT-probability > 0.5

Figure 4.14: Invariant mass spectrum and fit of Ξ_c^+ candidates in $4 < p_\mathrm{T} < 6\,\mathrm{GeV}/c$ for the BDT model only selecting decay pions with $p_\mathrm{T} > 1\,\mathrm{GeV}/c$.

Figure 4.13 shows the mass spectrum for the BDT probabilities 0.4 and 0.5 with a signal significances of 3.5. Despite the high significance of the extracted signal, this observation should be interpreted with caution. The mean value of the peak at BDT probability 0.4 is significantly shifted from the known mass $M_{\Xi_c^+} = (2467.71 \pm 0.23)\,\mathrm{MeV}/c^2$ [7], with a deviation of about $4\,\sigma$. Furthermore, the signal peak is not visible for all other BDT probabilities. The uncertainty of the mean includes only statistical uncertainties, since it results from the fitting procedure. Therefore, the σ deviation also includes only statistical uncertainties.

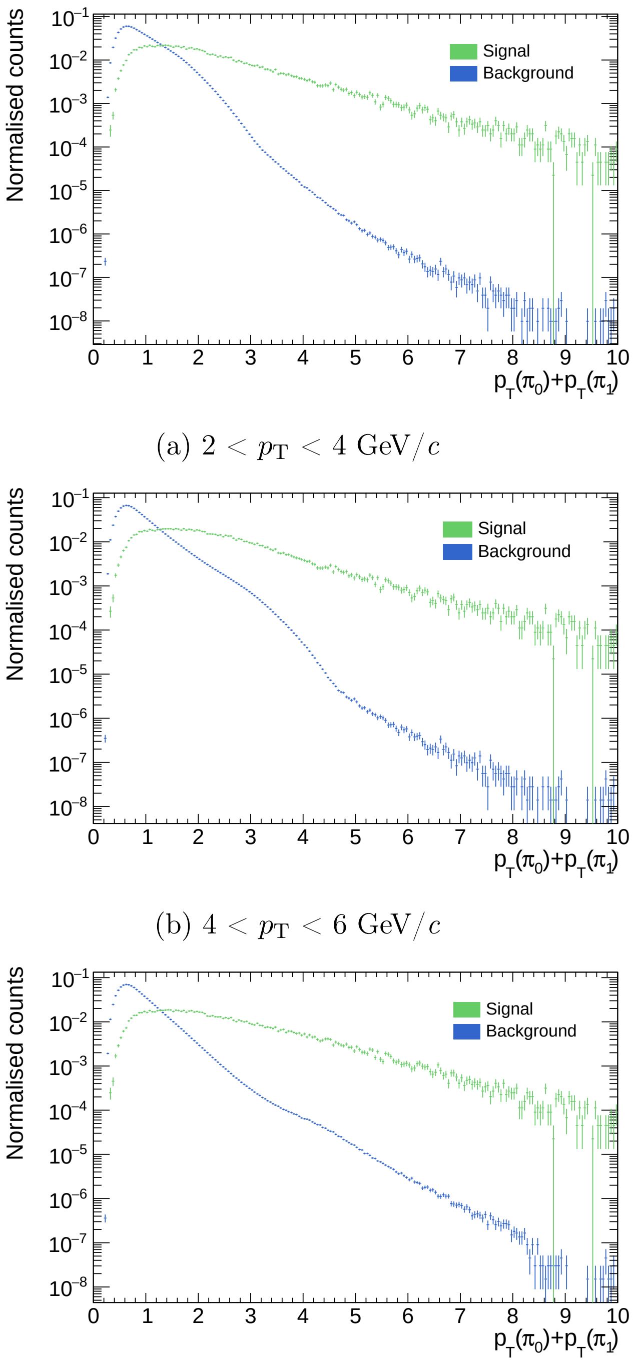
The distributions for the p_T interval 4 - 6 GeV/c do not show any significant signal peak. Figure 4.14 presents the invariant mass spectrum of decay pions with transverse momentum higher than 1 GeV/c for BDT probabilities 0.3 and 0.5. The distributions show strong fluctuations, which make a reliable signal extraction impossible.

4.4.3 Models with additional training features and modified preselections on the p_{T} of the two decay pions

To further improve the separation of signal and background, new training features are chosen, and the preselections on p_{T} of the decay pions are varied. The DCA between each of the two pions and the PV in xy-direction can also be exploited by including it as a training feature. Preselecting the DCA would also be possible but not as effective since the background and MC peak are superimposed, making it difficult to exclude much background without losing too many signal candidates.

As discussed earlier, the p_{T} spectra of both decay pions (Figure 4.12) indicate that different preselections for the p_{T} values of each of the two pions might improve the signal extraction. One possible way to do this is to apply preselections on the sum of the p_{T} values. Therefore, the distribution of the sum of the p_{T} values of both decay pions, $p_{T,sum} = p_{\mathrm{T}}(\pi_0) + p_{\mathrm{T}}(\pi_1)$, is shown in Figure 4.15.

The MC distribution of the sum slowly decreases for high p_{T}, while the background peaks for low p_{T} and then steeply decreases and falls below the signal distribution for a p_{T} sum of 1.5 GeV/c. In particular, for the p_{T} interval of 4 - 6 GeV/c, selections of the sum might work better, since the previous preselections did not significantly improve the signal extraction.



(a) $2 < p_{\mathrm{T}} < 4$ GeV/c

(b) $4 < p_{\mathrm{T}} < 6$ GeV/c

(c) $6 < p_{\mathrm{T}} < 12$ GeV/c

Figure 4.15: $p_{\mathrm{T}}(\pi_0^+) + p_{\mathrm{T}}(\pi_1^+)$ distribution for signal (green) and background candidates (blue) in different p_{T} intervals.

To test the extent to which selecting pion pairs with a p_{T} sum higher than 1.5 GeV/c would reject true signal candidates, the distribution of the MC data before and after selection for the p_{T} interval 4 - 6 GeV/c is shown in Figure 4.16. The pions of the hard spectrum are plotted separately from the pions of the soft spectrum. As expected, more pions of the hard spectrum (π_1) are rejected upon selection, resulting in a peak shifted to higher values. The peak of the distribution of the soft spectrum becomes slightly lower and shifted after selection, but not as much as the π_1 distribution. Instead of using the same model conditions for all p_{T} intervals, the preselections and the training features are tuned separately now. The six training features used in 4.4.1 are chosen according to the training features for low p_{T} in [11]. For higher p_{T} values, the decay length of the Ξ_c^+ becomes more important as a feature variable, since the particles are Lorentz-boosted and decay further away from the PV. Therefore, the decay length is added as a new variable replacing the DCA of the Ξ^- daughters in the xy-direction, which was the least important variable for higher p_{T} ranges. This approach is also used in the previous analysis [11], where the same variables are exchanged. Furthermore, the distance between the Ξ^- production and its decay vertex, normalised by the associated uncertainty of the decay vertex, referred to as ldl(Ξ^-), is added as a training feature. Even though this variable is not used in [11], its high feature importance can be seen there. Additionally, both

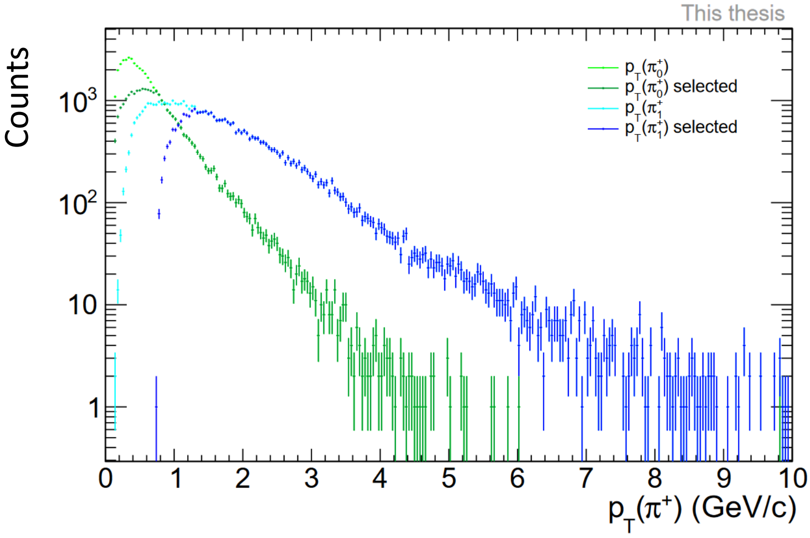


Figure 4.16: p_{T} distribution of the decay pions for MC candidates, before (pink, blue) and after (green, dark blue) the preselection $p_{\mathrm{T}}(\pi_0^+) + p_{\mathrm{T}}(\pi_1^+) > 1.5\,\mathrm{GeV}/c$ in $4 < p_{\mathrm{T}} < 6\,\mathrm{GeV}/c$.

variables, the DCA between each decay pions and the PV are included. For all following models, the preselections described in Table 4.1 apply, except changes in the selection of the p_T of the two decay pions.

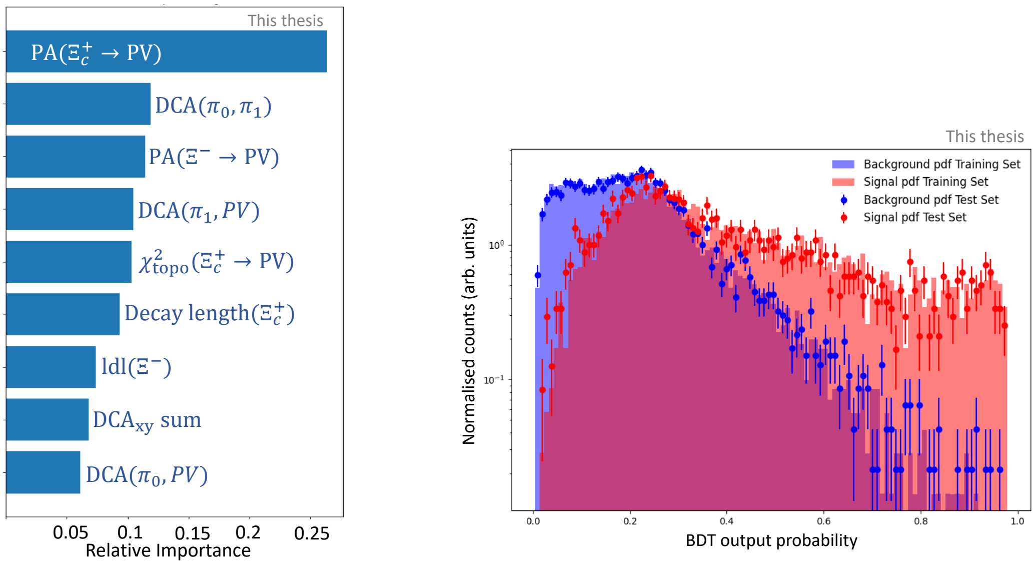
4.4.3.1 Models trained in $2 < p_\mathrm{T}(\Xi_c^+) < 4$ GeV/c

Four different models are trained in the p_T interval 2 - 4 GeV/c. Preselections additional to those listed in Table 4.1 and all training features used for the following models are reported in Table 4.5.

Table 4.5: Training features and additional selection criteria for the models trained in $2 < p_\mathrm{T} < 4$ GeV/c.

model Nr.	**1st**	**2nd**	**3rd**	**4th**
$\mathrm{DCA}(\pi_0, \pi_1)$	×	×	×	×
$\mathrm{DCA_{xy}}(\pi_0, \pi_1) + \mathrm{DCA_{xy}}(\pi^-, \Xi^-)$	×	×	×	×
$\mathrm{PA}(\Xi_c^+ \rightarrow \mathrm{PV})$	×	×	×	×
$\mathrm{PA}(\Xi^- \rightarrow \mathrm{PV})$	×	×	×	×
$\chi^2_\mathrm{topo}(\Xi_c^+)$	×	×	×	×
$\mathrm{DCA_{xy}}(\Lambda,\pi^-)$			×	×
$\mathrm{ldl}(\Xi^-)$	×	×		
$\mathrm{L_{xy}}(\Xi_c^+)$	×	×		
$\mathrm{DCA}(\pi^+,\mathrm{PV})$	×	×	×	×
$p_\mathrm{T}(\pi_0)$			> 0.7	> 0.7
$p_\mathrm{T}(\pi_1)$	> 1		> 0.7	> 0.7
$p_\mathrm{T}(\pi_0) + p_\mathrm{T}(\pi_1)$		> 1.2		
$\mathrm{PA}(\Xi_c^+ \rightarrow \mathrm{PV})$				< 1.5

1st model Figure 4.12 indicates that stronger selections on the $p_{\mathrm{T}}(\pi_1^+)$, the transverse momentum of the hard-spectrum pions, may be useful. Therefore, the next model for the p_{T} interval 2 - 4 GeV/c is trained with the additional variables introduced in the previous section and a selection only on the $p_{\mathrm{T}}(\pi_1^+)$ selecting pions with $p_{\mathrm{T}}(\pi_1^+) > 1\,\mathrm{GeV}/c$. The feature importance of the training variables for the new model are shown in Figure 4.17a. The BDT output probability is depicted in Figure 4.17b. Within their statistical uncertainties, the distribution of the test set follows the training set distribution, which is a sign of a stable model performance without overfitting or underfitting.



(a) Feature importance ranking of the new training variables.

(b) Model output probability for signal and background in the training and test set.

Figure 4.17: Feature importance and BDT probability.

The application of this new model did not allow to extract the signal with high significance. The significances for different BDT probabilities are reported in Figure 4.18. For the BDT probability at 0.3, the significance increases and reaches a value of about 4. However, since it decreases rapidly for lower and higher BDT probabilities, it can be concluded that the high significance is only due to statistical fluctuations.

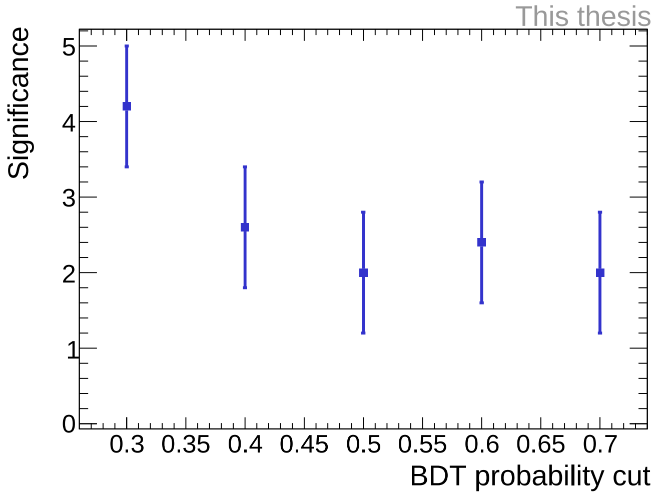
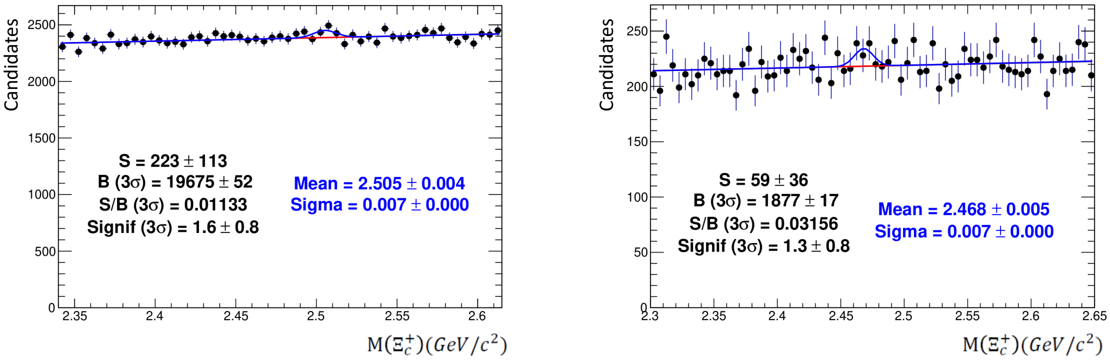


Figure 4.18: Signal significance as function of the BDT probability.

2nd model A second model is trained with the same variables and a preselection applied on the sum of the transverse momenta of both pions ($p_{T,sum} = p_\mathrm{T}(\pi_0) + p_\mathrm{T}(\pi_1)$), rather than to each momentum individually. Based on the distribution depicted in Figure 4.15a, only pion pairs with a p_T sum higher than $1.2\,\mathrm{GeV}/c$ are selected. The invariant mass spectrum for different BDT probabilities is shown in Figure 4.19. Again, the invariant mass spectrum does not shows any significant signal peak.

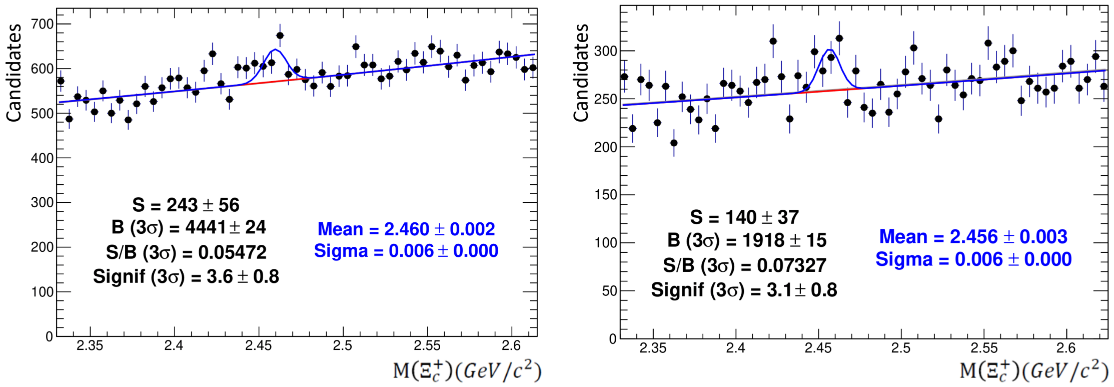


(a) BDT probability > 0.6

(b) BDT probability > 0.8

Figure 4.19: Invariant mass spectrum and fit of Ξ_c^+ candidates.

3rd model The new variables and preselections do not seem to improve the signal extraction for low p_{T}. Therefore, the next model is trained with the DCA of the Ξ^- daughters in xy-direction and without the decay length of the Ξ_c^+ and the ldl of the Ξ^-. The DCA between the two decay pions and the PV are retained because their distributions indicate a clear separation of background and signal, even for low p_{T}. Since a significant signal peak was found in the invariant mass spectrum of the model that selected only decay pions with $p_{\mathrm{T}} > 0.7\,\mathrm{GeV}/c$ (see Figure 4.13b), the same preselection is applied for the next model. The third model is therefore trained with the six variables used in chapter 4.4.1, plus the DCA between the two decay pions and the PV, and with the preselection $p_{\mathrm{T}} > 0.7\,\mathrm{GeV}/c$ applied on both pions (see 4.5). The invariant mass spectrum depicted in Figure 4.20, shows high statistical fluctuations. Although there is evidence of a signal emerging from the background, the mean of the Gaussian for both BDT probabilities again differs by about $4\,\sigma$ from the mass of the Ξ_c^+ reported in [7]. Again, it has to be noted, that the uncertainties of the mean reported in Figure 4.20 are only from the fitting procedure.



(a) BDT probability > 0.6

(b) BDT probability > 0.7

Figure 4.20: Invariant mass spectrum and fit of Ξ_c^+ candidates for the BDT model only selecting decay pions with $p_{\mathrm{T}} > 0.7\,\mathrm{GeV}/c$.

4th model Throughout all models, the pointing angle of the Ξ_c^+ to the PV was the most dominant training feature, with a relative importance much higher than that of all other features. To test whether signal extraction improves by exploiting the separation power of the other features more, a preselection is applied on the PA of the Ξ_c^+ to the PV. Figure 4.21 shows the distribution of the PA for the MC and background candidates. It can be seen that the MC candidates tend to have lower values, while the background is distributed nearly uniformly.

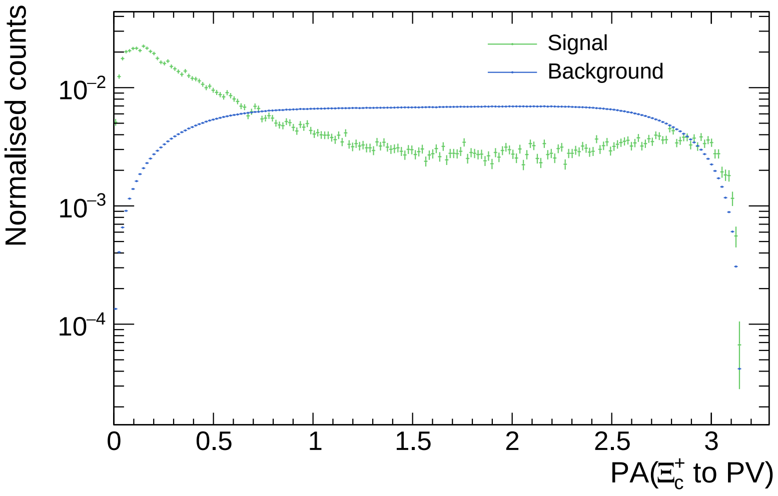
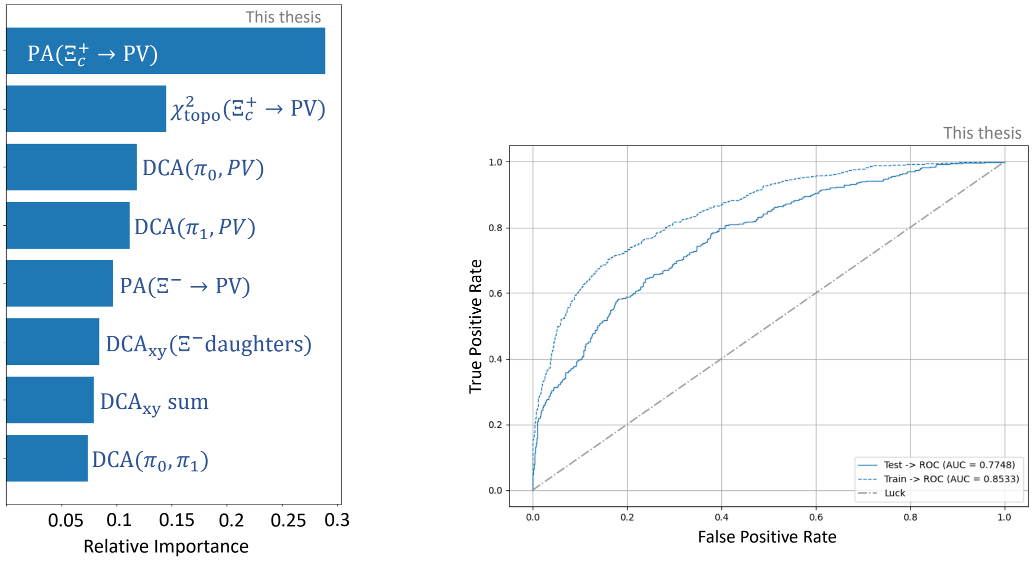


Figure 4.21: PA of Ξ_c^+ to the PV for signal (green) and background candidates (blue) in $2 < p_{\mathrm{T}} < 4\,\mathrm{GeV}/c$.

The next model is trained with the same variables and p_{T} selections for the decay pions as before, but with an additional selection of the PA of the Ξ_c^+, selecting only candidates with a pointing angle smaller than 1.5. The feature importance of the variables is reported in Figure 4.22a, indicating a changed order compared to the previous rankings. The χ^2_{topo} of the Ξ_c^+ and both DCA between the two decay pions and the PV gain importance, while the DCA between the two pions itself becomes the least important feature. The ROC curve in Figure 4.22b shows a deviation between the curve of the test set and the training set, indicating a decreased model performance. This can be explained by the preselection applied on the PA of the Ξ_c^+ to the PV. This preselection significantly reduces the number of signal candidates available for testing and training of the model, making it less generalisable. The same can be seen in the BDT output probability, which is shown in Figure 4.23. The distribution of the test and the training set differ from each other and have large statistical uncertainties.



(a) Feature importance ranking of the training variables.

(b) Receiver Operating Characteristic (ROC) curves of the training set and the test set.

Figure 4.22: Feature importance and ROC of the model only selecting decay pions with $p_{\mathrm{T}} > 0.7\,\mathrm{GeV}/c$ and Ξ_c^+ candidates with PA smaller than 1.5.

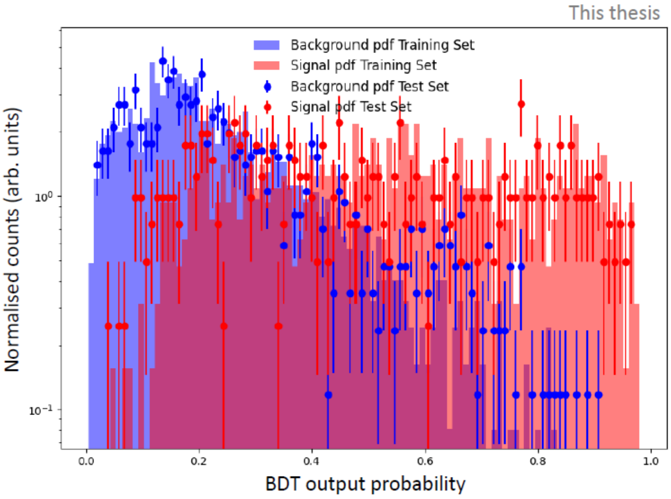
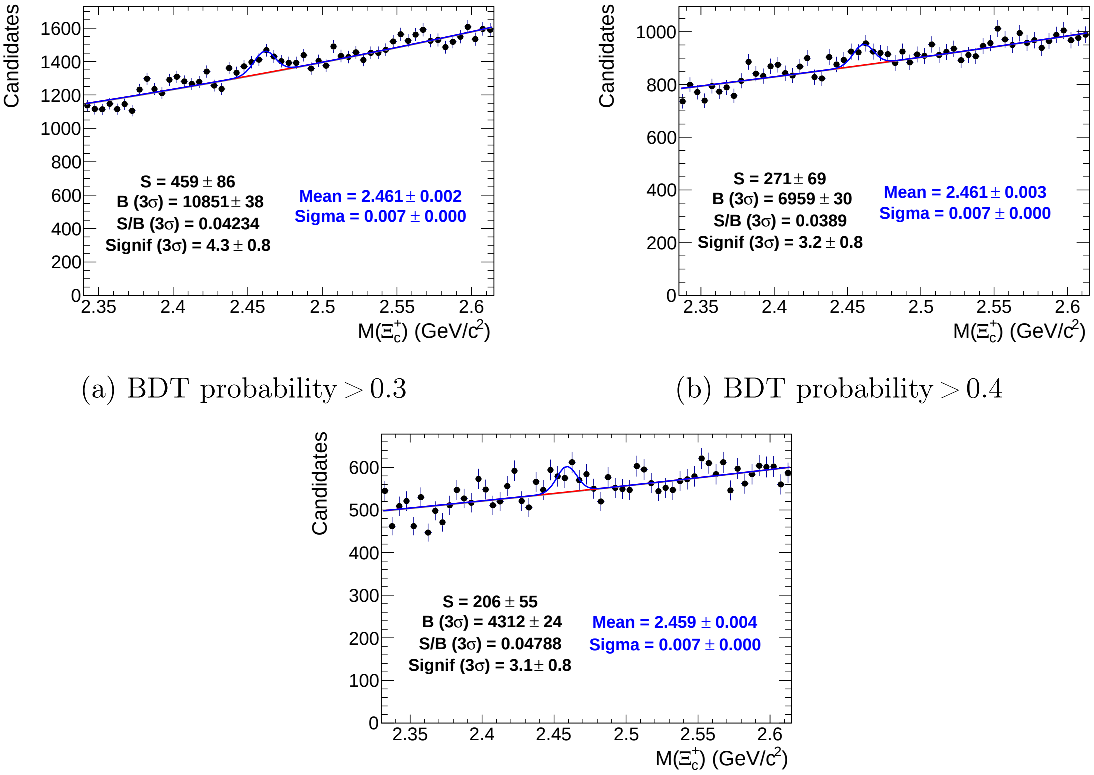


Figure 4.23: Output probability for signal and background in the training and test set.

The invariant mass spectrum and the fit results for the BDT probability selections of 0.3, 0.4 and 0.5 are shown in Figure 4.24, indicating significant signal peaks. The highest significance of about 4 is achieved for the BDT probability of 0.3. However, the mean again deviates from the mass of the Ξ_c^+ [7], as already seen for the 3rd model. For the BDT probability 0.3, the mean deviates by about $3\,\sigma$, while for the other two BDT probabilities the deviation is only around $2\,\sigma$. Again, it has to be noted, that the uncertainties of the means reported in Figure 4.24, are from the fitting procedure and include only statistical uncertainties. Even though these deviations are not significant, they indicate high statistical fluctuations. For further analysis, systematic studies about this deviation of the mean at low p_T should be performed. Despite this small variation, there is good indication that even in this p_T interval the signal extraction might be possible, however probably with large statistical and systematic uncertainties.



(a) BDT probability > 0.3

(b) BDT probability > 0.4

(c) BDT probability > 0.5

Figure 4.24: Invariant mass spectrum and fit of Ξ_c^+ candidates for the BDT model.

4.4.3.2 Models trained in $4 < p_{\mathrm{T}}(\Xi_c^+) < 6$ GeV/c

1st model Since the signal extraction for the p_{T} interval 4 - 6 GeV/c did not work with the BDT models tested so far, new models are trained with the new training features. For the first model, a selection is applied only to the p_{T} of the hard-spectrum pions, selecting only pions with $p_{\mathrm{T}}(\pi_1) > 0.9$ GeV/c. The feature importance of the variables is shown in Figure 4.25. As expected, the decay length of Ξ_c^+ has a greater impact for the higher momentum range. It is the fourth most important feature in the ranking, whereas it was the least important for lower p_{T} (see Figure 4.17a).

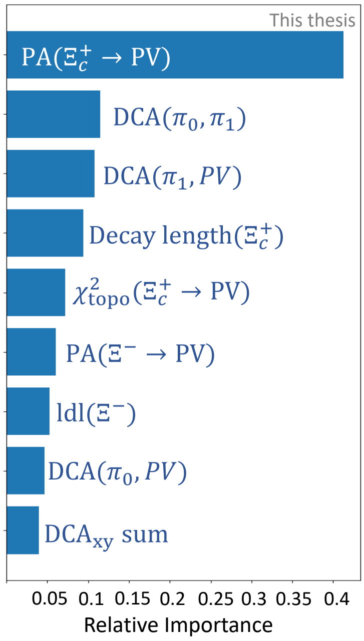


Figure 4.25: Feature importance ranking of the training variables.

The significances for different BDT probability selections from 0.3 to 0.7 are reported in Figure 4.26. A signal significance higher than 3 is only achieved for the probability selection of 0.6. For higher and lower BDT probability selection the significance decreases.

The invariant mass spectrum and the fit results for the probability selection at 0.6 and 0.7 are shown in Figure 4.27. As the significant peak structure disappears for all other BDT probabilities, it could simply be due to statistical fluctuations.

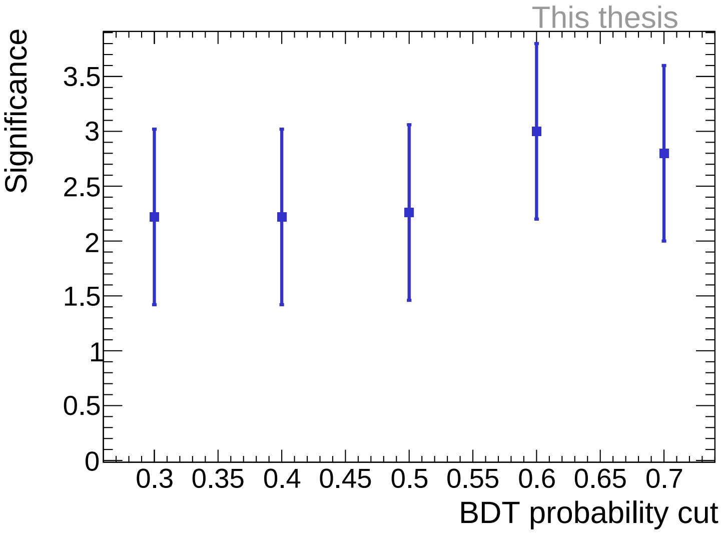
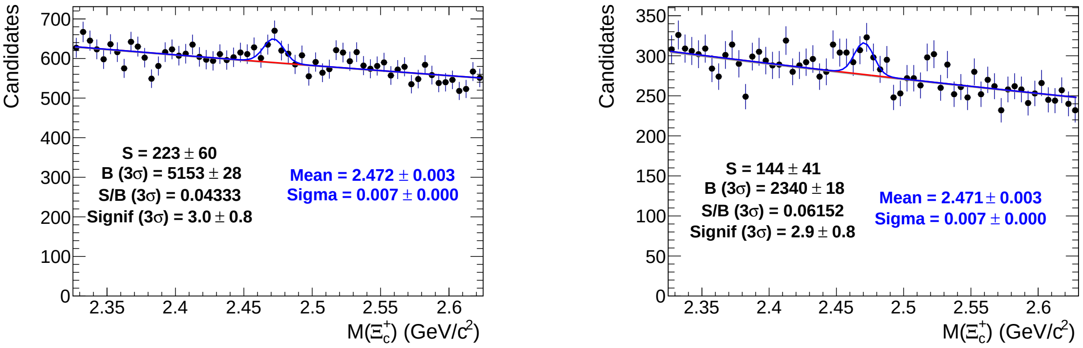


Figure 4.26: Signal significance as function of the BDT output probability.

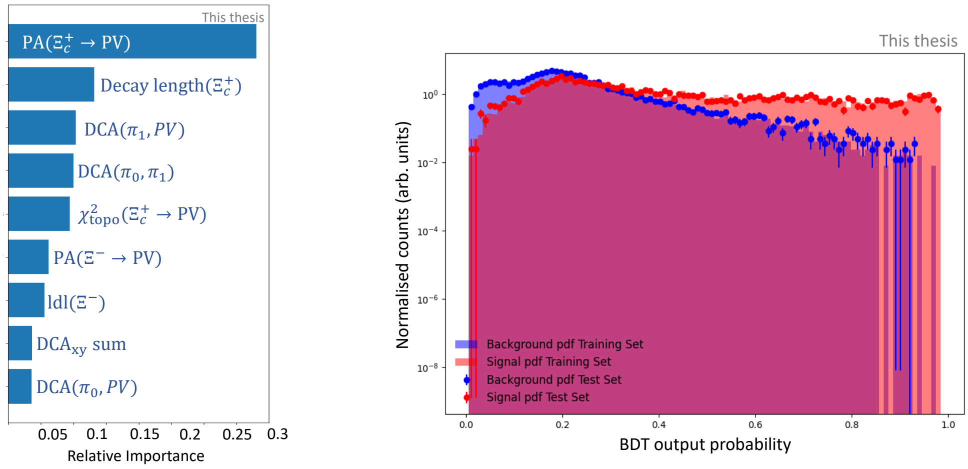


(a) BDT probability > 0.6

(b) BDT probability > 0.7

Figure 4.27: Invariant mass spectrum and fit of Ξ_c^+ candidates.

2nd model Based on the information on the distributions of the sum of the p_{T} of both pions (see 4.15b), the next model is trained only with decay pion pairs whose p_{T} sum is higher than $1.5\,\mathrm{GeV}/c$. The training variables remain the same. The feature importance of the variables is shown in Figure 4.28a, where the decay length of Ξ_c^+ is even more important compared to the previous model.



(a) Feature importance ranking of the training variables.

(b) Model output probability for signal and background in the training and test set.

Figure 4.28: Feature importance and output probability.

As seen in Figure 4.28b the BDT probability distributions of the signal in red and the background in blue. However, the distribution of the test set follows the distribution of the training set, indicating stable model performance.

The signal significances for different BDT probability selections from 0.1 to 0.5 are depicted in Figure 4.29. For probabilities below 0.5, the significance is higher than 3. The best signal significance is achieved for the BDT probabilities 0.2 and 0.3. The invariant mass spectrum and the fit results for this probabilities are shown in Figure 4.30. Both invariant mass fits have a significance of about 3, indicating a possible signal extraction. The significances are the highest reached in this analysis in the p_{T} interval 4 - 6 GeV/c. However, the mean of the fits deviates from the known mass of the Ξ_c^+ [7] by about $2\,\sigma$ and even shows variations among the different BDT probability selections, indicating high statistical uncertainties.

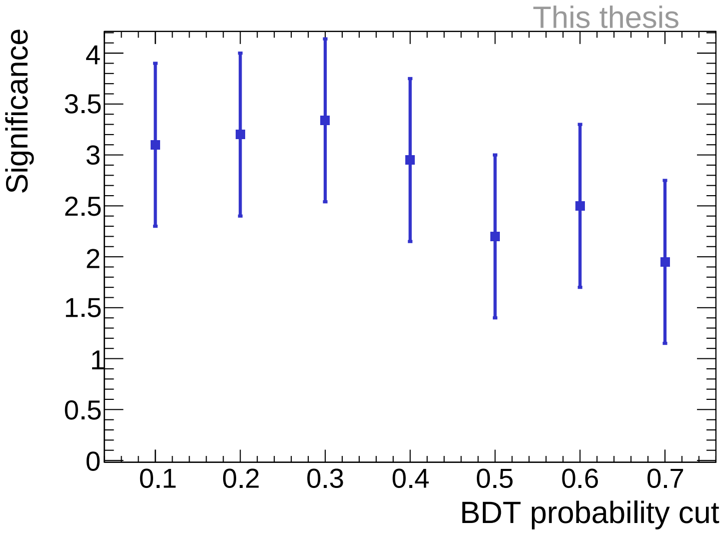
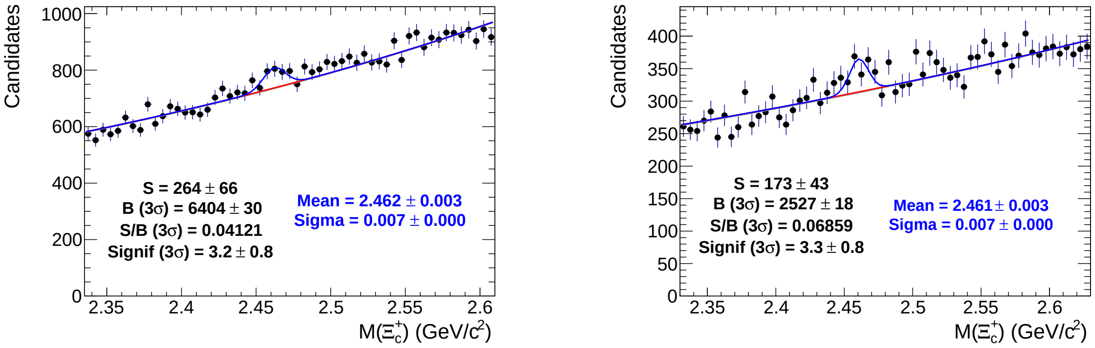


Figure 4.29: Signal significance as function of the BDT output probability.



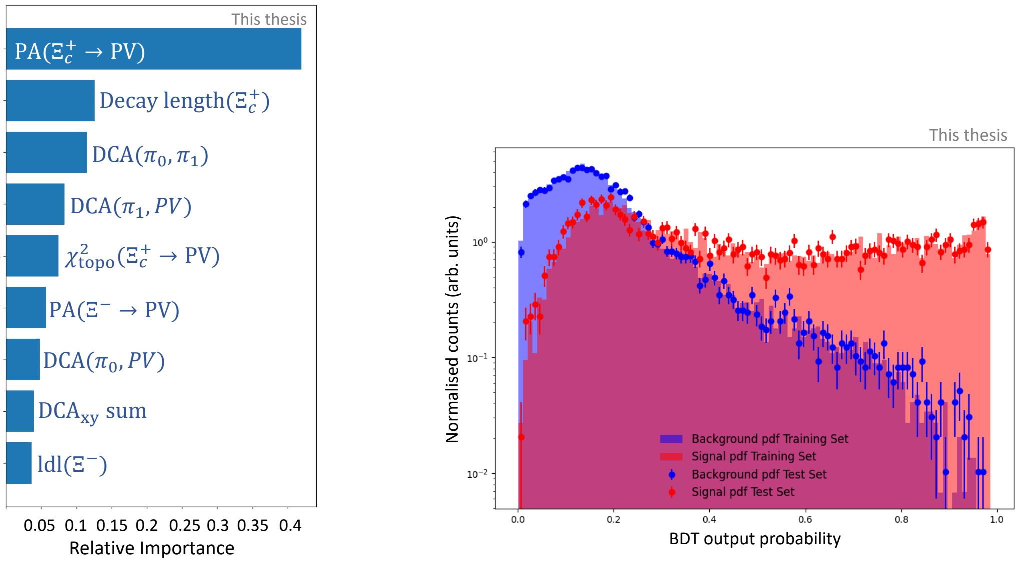
(a) BDT probability > 0.2.

(b) BDT probability > 0.3.

Figure 4.30: Invariant mass spectrum and fit of Ξ_c^+ candidates.

4.4.3.3 Models trained in $6 < p_{\mathrm{T}}(\Xi_c^+) < 12\,\mathrm{GeV}/c$

For high transverse momentum, the signal extraction is easier because the combinatorial background becomes less and the secondary vertex is more displaced from the PV and therefore better reconstructed. As seen in Figure 4.11c, a significant signal peak for the p_{T} interval 6 - 12 GeV/c was already found with the first set of models. Nevertheless, the new variables and a preselection on the sum of the p_{T} of the two pions are also tested in this p_{T} range. Based on the distribution in Figure 4.15c, only pion pairs whose p_{T} sum is higher than 1.3 GeV/c are selected. The feature importance ranking of the variables is depicted in Figure 4.31a, confirming the importance of the decay length of the Ξ_c^+ for high p_{T}. The output probability of the model is shown in Figure 4.31b and indicates a good model performance, as the distribution of the test set follows the training set.



(a) Feature importance ranking.

(b) Model output probability for signal and background in the training and test set.

Figure 4.31: Feature importance and output probability.

The signal significances of the invariant mass spectrum for different BDT probabilities from 0.3 to 0.7 are reported in Figure 4.32. The significance has values between 4 and 5 for BDT probability selections from 0.5 to 0.7, where the model is able to reject a lot of background while preserving enough signal to be extracted. The invariant mass spectrum for the probabilities 0.5 and 0.6 with the fit results are shown in Figure 4.33. A signal significance of up to 5 is achieved. The mean of the fit barely deviates from the Ξ_c^+ mass [7], which improves the signal extraction of the previously tested model.

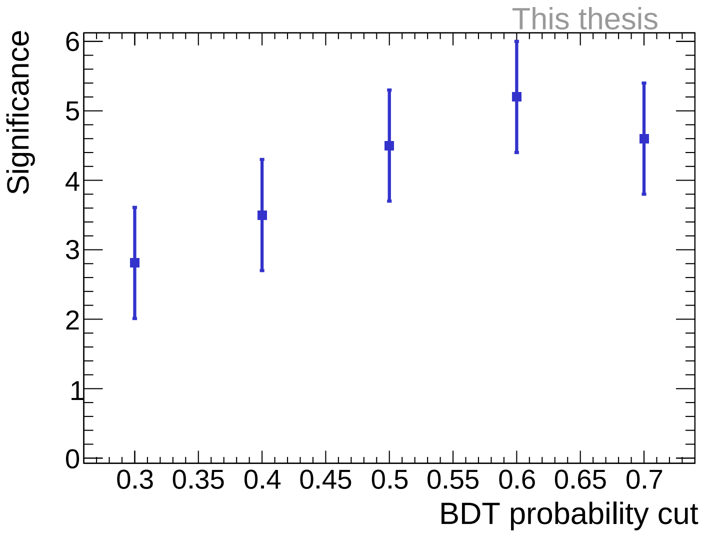
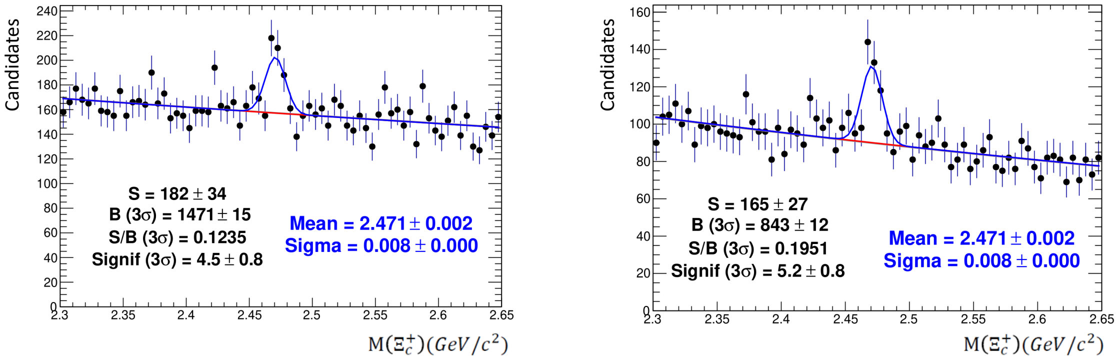


Figure 4.32: Signal significance as function of the BDT output probability.



(a) BDT probability > 0.5

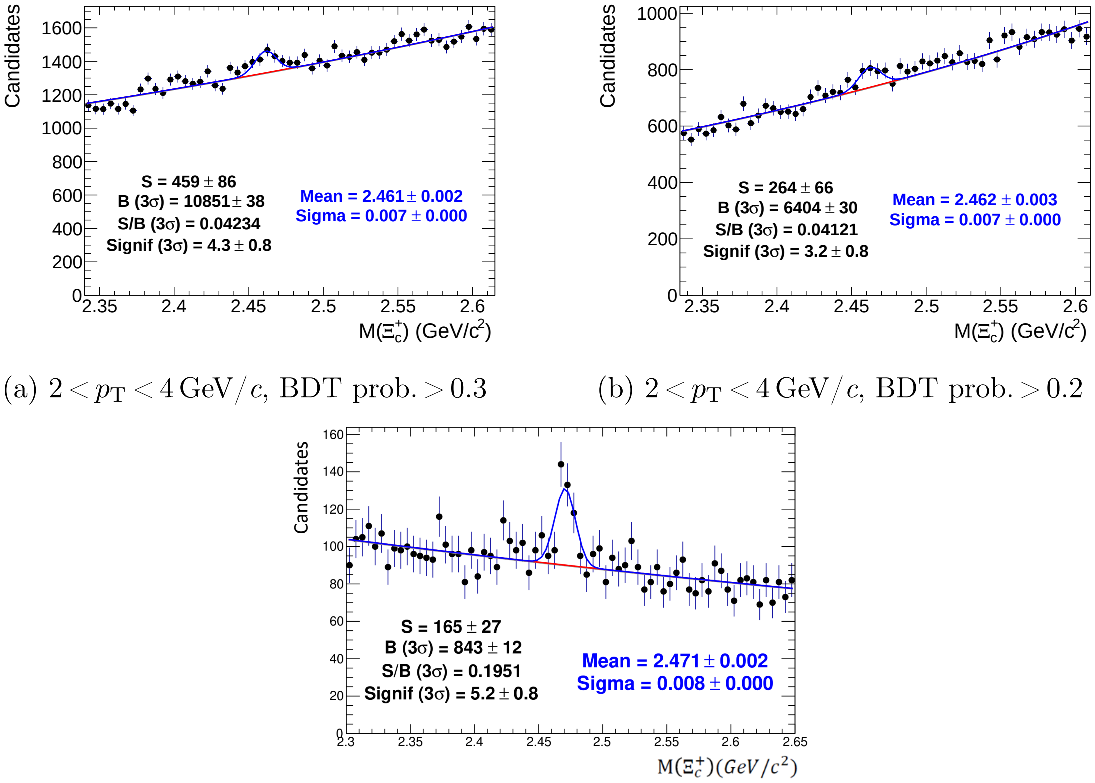
(b) BDT probability > 0.6

Figure 4.33: Invariant mass spectrum and fit of Ξ_c^+ candidates for the BDT model.

5 Conclusion and outlook

The goal of this analysis was to perform a feasibility study investigating the possibility of measuring the Ξ_c^+ baryon in p-Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV with ALICE. The Ξ_c^+ baryon was reconstructed via its hadronic decay into two π^+ and a Ξ^- baryon, decaying into a negatively charged pion and a Λ baryon, which further decays into a proton and a negatively charged pion. The analysis was performed using the KFParticle package for the reconstruction and a machine learning approach based on XGBoost for signal and background classification. Several Boosted-Decision-Tree models were trained with different training features and preselections and their performances were compared. Throughout all analysed p_{T} intervals a significant signal peak was found in the Ξ_c^+ invariant mass spectrum (Figure 5.1).
For the lowest p_{T} interval 2 - 4 GeV/c, the best signal significance is achieved with the 4th model trained with the variables and preselections reported in Table 4.5. A significance between 3 and 4 is reached for BDT probability selections of 0.3, 0.4 and 0.5 as shown in Figure 4.24 and Figure 5.1a. However, the mean of the Gaussian fit of the signal peak significantly differs by up to $3\,\sigma$ from the known mass of the Ξ_c^+, $M_{\Xi_c^+} = (2467.71 \pm 0.23)\,\mathrm{MeV}/c^2$ [7]. Even though these deviations are not significant, they indicate high statistical fluctuations and should further be investigated to understand any possible systematic effect on the final measurement.
For higher p_{T} the final models were trained with the variables shown in Figure 4.28a and Figure 4.31a. Instead of applying preselections on the transverse momentum of each of the decay pions, as done for low p_{T}, a preselection is applied on the p_{T} sum of both pions ($p_{\mathrm{T,sum}} = p_{\mathrm{T}}(\pi_0) + p_{\mathrm{T}}(\pi_1)$). Since the p_{T} spectra of both decay pions (Figure 4.12) indicate that pions with low and high p_{T} are stored separately, different preselections for the momentum of each of the two pions seem reasonable. This is implemented by applying preselections on the sum of the p_{T} values, which have been shown to be better than individual preselections on both p_{T} values. For the p_{T} interval 4 - 6 GeV/c only pion pairs with a p_{T} sum higher than 1.5 GeV/c were selected. A signal significance of about 3 was reached for a BDT probability selections of 0.2 and 0.3 in the p_{T} interval 4 - 6 GeV/c. The corresponding invariant

mass spectrum is shown in Figure 4.28 and Figure 5.1b. Again, the mean of the fits deviates from the known mass of the Ξ_c^+ [7], this time by about $2\,\sigma$. The mean also shows variations among the different BDT probability selections, indicating high statistical uncertainties. For the p_T interval 6 - 12 GeV/c the p_T sum of the two pions was selected to be greater than 1.3 GeV/c. A signal significance of about 5 was obtained for this p_T interval with BDT probability selection 0.6 (Figure 5.1c). The mean of the fit barely deviates from the known Ξ_c^+ mass. Surprisingly, the preselection on the p_T sum for the p_T interval 6 - 12 GeV/c is looser than the preselection for the p_T interval 4 - 6 GeV/c. This is unexpected since the momentum of both pions is shifted to higher values for high p_T. However, since the combinatorial background is generally smaller for high p_T, a looser preselection is apparently sufficient for the signal extraction. Overall, the highest signal significance was obtained for high p_T.



(a) $2 < p_\mathrm{T} < 4$ GeV/c, BDT prob. > 0.3

(b) $2 < p_\mathrm{T} < 4$ GeV/c, BDT prob. > 0.2

(c) $2 < p_\mathrm{T} < 4$ GeV/c, BDT prob. > 0.6

Figure 5.1: Invariant mass spectrum and fit of Ξ_c^+ candidates for the final BDT models.

The significant signal peaks found in all analysed p_{T} intervals give a strong indication that a full analysis of the Ξ_c^+ baryon in p-Pb collision is feasible. If a full analysis is performed in the future, the small deviation in the mass mean should be investigated to understand any possible systematic effect on the final measurement. Systematic uncertainties were not evaluated in this work and a more comprehensive uncertainty analysis would be important for future analyses. Furthermore, the selection on the BDT output probability by choosing the one with highest signal significance may lead to the amplification of statistical fluctuations. To avoid this possible bias, a blind optimisation of the selection on the output probability, the so-called working point procedure [11], should be used in further analyses.

Part I

Appendix

A Bibliography

[1] M. Thomson. “Modern Particle Physics”. In: *Cambridge University Press, New York* (2013).

[2] K. Rajagopal W. Busza and W. van der Schee. “Heavy Ion Collisions: The Big Picture and the Big Questions”. In: Annu. Rev. Nucl. Part. S 68.1 (2018), pp. 339–376. DOI: 10.1146/annurev-nucl-101917-020852.

[3] N. Cabibbo and G. Parisi. “Exponential hadronic spectrum and quark liberation”. In: *Physics Letters* B 95.1 (1975), pp. 67–69.

[4] K. Rajagopal W. Busza and W. van der Schee. “Heavy Ion Collisions: The Big Picture and the Big Questions”. In: *Annu. Rev. Nucl. Part.* S 68.1 (2018), pp. 339–376. DOI: 10.1146/annurev-nucl-101917-020852.

[5] R. Stock. “Relativistic Nucleus-Nucleus Collisions and the QCD Matter Phase Diagram”. In: (2008). DOI: 10.1007/978-3-540-74203-6_7.

[6] A. Kurkela et al. “Effective kinetic description of event-by-event pre-equilibrium dynamics in high-energy heavy-ion collisions”. In: Phys. Rev. C 99 (2019). DOI: 10.1103/PhysRevC.99.034910.

[7] P. A. Zyla et al. Particle Data Group Collaboration. “Review of Particle Physics”. In: *PTEP 2020.8* (2020). DOI: 10.1093/ptep/ptaa104.

[8] D. E. Soper J. C. Collins and G. Sterman. “Heavy particle production in high-energy hadron collisions”. In: Nucl. Phys. B 263.1 (1986), pp. 37–60. ISSN: 0550-3213. DOI: 10.1016/0550-3213(86)90026-X.

[9] ALICE Collaboration. “First measurement of Λ_c^+ production down to $p_{\mathrm{T}} = 0$ in pp and p-Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV”. In: (2022). DOI: 10.48550/arXiv.2211.14032.

[10] S. Acharya et al. ALICE Collaboration. “Measurement of the production cross section of prompt Ξ_c^0 baryons at midrapidity in pp collisions at $\sqrt{s} = 5.02\,TeV$”. In: (2021). DOI: 10.1007/JHEP10(2021)159.

[11] C. Reetz. “Measurement of Ξ_c^+ in proton–proton collisions at s = 13 TeV with the ALICE detector”. In: (2022).

[12] L. Evans and P. Bryant. “LHC Machine”. In: *Journal of Instrumentation* 3.08 (2008), S08001. ISSN: 1748-0221. DOI: 10.1088/1748-0221/3/08/S08001.

[13] The ALICE Collaboration. “The ALICE experiment at the CERN LHC”. In: *Journal of Instrumentation* 3.8 (2008), S08002. ISSN: 1748-0221. DOI: 10.1088/1748-0221/3/08/S08002.

[14] ALICE Collaboration. *First lead-ion collisions in the LHC at record energy.* November 2022. URL: https://alice-collaboration.web.cern.ch/Nov2022_leadtest (visited on 20/01/2023).

[15] ALICE Collaboration and F. Carminati et al. “ALICE: Physics Performance Report, Volume I”. In: *J. Phys. G30* 30 (2004). ISSN: 1517–1763. DOI: 10.1088/0954-3899/30/11/001.

[16] A. Tauro. *ALICE Schematics - General Photo.*

[17] L. Betev et. al. *CERN Engineering & Equipment Data Management Service.* URL: https://edms.cern.ch/document/406391/2 (visited on 05/01/2023).

[18] B. Abelev et al. ALICE Collaboration. “Performance of the ALICE Experiment at the CERN LHC.” In: *Int. J. Mod. Phys.* (2014). ISSN: 1430044. DOI: 10.1142/S0217751X14300440.

[19] M. Tanabashi et al. “Review of Particle Physics.” In: *Phys. Rev.* D 98 (2018). DOI: 10.1103/PhysRevD.98.030001.

[20] C. Lippmann. “Particle identification”. In: *Elsevier BV* 666 (2012). DOI: 10.1016/j.nima.2011.03.009.

[21] S. Gorbunov and I. Kisel. “Reconstruction of decayed particles based on the Kalman filter.” In: (2007).

[22] Ivan Kisel, Igor Kulakov and Maksym Zyzak. “Standalone First Level Event Selection Package for the CBM Experiment”. In: *IEEE Transactions on Nuclear Science* 60.5 (2013), pp. 3703–3708. DOI: 10.1109/TNS.2013.2265276.

[23] S. Gorbunov. *On-line reconstruction algorithms for the CBM and ALICE experiments.* 2013. URL: http://publikationen.ub.uni-frankfurt.de/frontdoor/index/index/docId/29538 (visited on 05/01/2023).

[24] M. Zyzak. “Online selection of short-lived particles on many-core computer architectures in the CBM experiment at FAIR.” In: (2015).

[25] A. S. Cornell et al. “Boosted decision trees in the era of new physics: a smuon analysis case study.” In: *JHEP* (2022). DOI: 10.1007/jhep04(2022)015.

[26] T. et al. Sjöstrand. “An Introduction to PYTHIA 8.2”. In: *Computer Physics Communications* (2015), pp. 159–177. ISSN: 00104655. DOI: 10.1016/j.cpc.2015.01.024.

[27] X. Wang and M. Gyulassy. “HIJING: A Monte Carlo model for multiple jet production in pp, pA, and AA collisions”. In: *Physical Review D* (1991), pp. 3501–3516. DOI: 10.1103/PhysRevD.44.3501.

[28] M. L. Miller et al. “Glauber modeling in high energy nuclear collisions.” In: *Ann. Rev. Nucl. Part. Sci.* (2007), pp. 468–471. DOI: 10.1146/annurev.nucl.57.090506.123020.

[29] C. Tsallis. “What are the Numbers that Experiments Povide.” In: *Química Nova 17* (1994), pp. 468–471.

[30] G. James et al. “An Introduction to Statistical Learning: with Applications in R”. In: *Springer Texts in Statistics* (2021). DOI: 10.1007/978-1-4614-7138-7.

Acknowledgements

I would like to thank Prof. Dr. Silvia Masciocchi for giving me the opportunity to join her group and write this thesis. Special thanks to Dr. Andrea Dubla, who guided me through this work, for his expertise, for proofreading and for always being available for questions and support. Many thanks also to Carolina Reetz, who generously shared her knowledge with me and helped with any arising problems and questions. Finally, I would like to thank my dear friends Mai, Rebekka and Julia for proofreading and my family.

Declaration of Authorship

Ich versichere, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg den 13.02.2023