Gradient Boosted Decision Tree Application to Muon Identification in the KLM at Belle II

Logan D. Benninghoff

Thesis submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Physics

Leo E. Piilonen, Chair

Camillo Mariani

Tatsu Takeuchi

May 3, 2024

Blacksburg, Virginia

Keywords: BDT, KLM, Decision Tree, Muon, Pion, Maximum Likelihood

Gradient Boosted Decision Tree Application to Muon Identification in the KLM at Belle II

Logan D. Benninghoff

(ABSTRACT)

We present the results of applying a Fast Boosted Decision Tree (FBDT) algorithm to the task of distinguishing muons from pions in K-Long and Muon (KLM) detector of the Belle II experiment. Performance was evaluated over a momentum range of $0.6 < p < 5.0\,\mathrm{GeV}/c$ by plotting Receiver Operating Characteristic (ROC) curves for $0.1\,\mathrm{GeV}/c$ intervals. The FBDT model was worse than the benchmark likelihood ratio test model for the whole momentum range during testing on Monte Carlo (MC) simulated data. This is seen in the lower Area Under the Curve (AUC) values for the FBDT ROC curves, achieving peak AUC values around 0.82, while the likelihood ratio ROC curves achieve peak AUC values around 0.98. Performance of the FBDT model in muon identification may be improved in the future by adding a pre-processing routine for the MC data and input variables.

Gradient Boosted Decision Tree Application to Muon Identification in the KLM at Belle II

Logan D. Benninghoff

(GENERAL AUDIENCE ABSTRACT)

An important task of a high-energy physics experiment is taking the input information provided by detectors, such as the distance a particle travels through a detector, the momentum, and energy deposits it makes, and using that information to identify the particle's type. In this study we test a machine learning model that sorts the particles observed into two categories—muons and pions—by comparing the particle's input values to a threshold value at multiple stages, then assigns a final identity to the particle at the last stage. This is compared to a benchmark model that uses the probabilities that these input variables would be seen from a particle of each type to determine which particle type is most likely. The ability of both models to distinguish muons and pions were tested on simulated data from the Belle II detector, and the benchmark model outperformed the machine learning model.

Dedication

This thesis is dedicated to my family and my teachers, who have supported me throughout my education.

Acknowledgments

The Belle II Analysis Software Framework [2] was developed by the Belle II collaboration, and the author acknowledges the continued work of the SuperKEKB group in operating the electron positron accelerator on behalf of the Belle II experiment. This work was supported by the Virginia Tech Department of Physics and its IT staff through the computing resources of the Tesla cluster. The author also acknowledges discussions about machine learning methods with Tommy Lam (Virginia Tech), and conversations about the Belle II software repository with Leo Piilonen (Virginia Tech).

Contents

List of Figures **viii**

1 Introduction **1**

1.1 Overview and Belle II 1

1.2 Belle II Detector 3

1.2.1 Vertex Detector (VXD) 4

1.2.2 Central Drift Chamber (CDC) 5

1.2.3 K-Long and Muon Detector (KLM) 6

2 Data and Evaluation Metrics **8**

2.1 Monte Carlo Simulated Data 8

2.1.1 Simulation Methodology 8

2.1.2 Simulation Parameters 9

2.2 Receiver Operating Characteristic (ROC) 11

3 Likelihood Method **14**

3.1 Likelihood Ratio Test 14

3.2 Likelihood Ratio Performance 16

4 Boosted Decision Tree Method **19**

4.1 Boosted Decision Tree Model 20

4.1.1 Decision Trees 20

4.1.2 Gradient Boosted Decision Trees 21

4.1.3 Fast Boosted Decision Trees 23

4.2 Fast Boosted Decision Tree Training 24

4.3 Fast Boosted Decision Tree Performance 25

5 Model Discussion **28**

6 Conclusion **30**

Bibliography **32**

Appendices **34**

Appendix A Particle Count Data **35**

Appendix B Full set of ROC Plots **38**

B.1 Likelihood Ratio ROC Curves 38

B.2 FBDT ROC Curves 49

List of Figures

1.1 Diagram of the Belle II detector given as Figure 3 in the Belle II Physics Book [1]. The horizontal (z) axis is approximately along the beam direction. The VXD which is comprised of the PXD and SVD is the innermost magenta section. The CDC is outlined in blue, the TOP detector is a thin strip between the CDC and ECL, and the ECL is outlined in purple. The outermost pink regions are the KLM, the primary detector of interest in this study. 3

1.2 Diagram of the coordinate system used in the Belle II detector, diagram courtesy of Belle II and KEK. The horizontal (z) axis is approximately along the beam direction (between the two beam paths), with $+z$ pointing in same direction as the electron beam. ϕ is the azimuthal angle, with $\phi = 0$ aligned with the $+x$ axis. θ is the polar angle, measured from the $+z$ axis. 5

2.1 Number of pions (red) and muons (blue) whose extrapolated track reached the innermost KLM layer. Particles are collected into momentum bins with width 0.1 GeV/c. This figure was generated for the likelihood ratio test evaluation data set, which had a total of 200,000 particles generated, split between muons and pions. Below a momentum of approximately 0.6 GeV/c we note that very few particles reach the KLM. 10

3.1 $\Delta = logL_\mu - logL_\pi$ test statistic plotted as an average for each momentum bin. One standard deviation is given above and below each Δ value. 17

3.2 Likelihood ratio model Receiver Operating Characteristic (ROC) curves for a select number of momentum bins within the full range. The Area Under the Curve (AUC) is calculated and given in the plot legend, and represents a measure of the overall classification performance, with an AUC close to 1 being optimal. The best choice of cutoff value Δ_{min} is also given for an FPR below certain thresholds. 18

4.1 FBDT model Receiver Operating Characteristic (ROC) curves for a select number of momentum bins within the full range. The Area Under the Curve (AUC) is calculated and given in the plot legend, and represents overall classification performance, with an AUC close to 1 being optimal. The best choice of cutoff value δ_{min} is also given for an FPR below certain thresholds. In comparison to the likelihood ratio model ROC curves in Figure 3.2 the performance of the FBDT model is worse across the whole momentum range. 27

A.1 The total number of muons correctly identified as muons (True Positive), muons incorrectly identified as pions (False Negative), pions correctly identified as pions (True Negative), and pions incorrectly identified as muons (False Positive) for the likelihood ratio model. The boxes shown in green are desirable classifications, while the number of tracks given in red boxes should be minimized to improve classifier performance. The cutoff value Δ_{min} chosen for these plots maximizes the combined number of True Positives and True Negatives. 36

A.2 The total number of muons correctly identified as muons (True Positive), muons incorrectly identified as pions (False Negative), pions correctly identified as pions (True Negative), and pions incorrectly identified as muons (False Positive) for the FBDT model. The boxes shown in green are desirable classifications, while the number of tracks given in red boxes should be minimized to improve classifier performance. The cutoff value δ_{min} chosen for these plots maximizes the combined number of True Positives and True Negatives. . . . 37

List of Abbreviations

AUC Area Under the Curve

BDT Boosted Decision Tree

CDC Central Drift Chamber

DT Decision Tree

ECL Electromagnetic Calorimeter

FBDT Fast Boosted Decision Tree

FPR False Positive Rate

GBDT Gradient Boosted Decision Tree

GEANT4e "GEometry ANd Tracking Extrapolator" software toolkit

KLM K-Long and Muon Detector

MC Monte Carlo

MLE Maximum Likelihood Estimate

PID Particle Identification

PXD Silicon Pixel Detector

ROC Receiver Operating Curve

SVD Silicon Vertex Detector

TOP Time-of-Propagation Counter

TPR True Positive Rate

VXD Vertex Detector

Chapter 1

Introduction

1.1 Overview and Belle II

A necessary component of many high energy physics experiments is the ability to correctly identify the final state particles produced in a collision. From the identification of the final state particles and the reconstructed trajectories of these particles, the full collision event can be mapped and analyzed. Using these reconstructed particles, a Particle Identification (PID) routine takes a set of values for the properties of each particle and outputs the most likely particle candidate. The primary goal of this analysis is to characterize, then improve upon, the performance of one such routine used for muon (μ) identification within the K-Long and Muon (KLM) detector at the Belle II experiment.

The Belle II experiment is a second generation B-factory at the SuperKEKB colliding beam accelerator. SuperKEKB collides electrons and positrons at a single interaction point with center-of-mass energies at or near the fourth radially excited state of the Υ resonance, a bound state of a b quark and a $\bar{b}$ antiquark [1]. The largest share of data is collected at this $\Upsilon(4S)$ resonance, which has a center-of-mass energy of $\sqrt{s} = 10.57$ GeV, produced using asymmetric beam energies of 7 GeV and 4 GeV for electrons and positrons respectively. The $\Upsilon(4S)$ resonance is slightly above the threshold for B-meson pair production, leading to a relatively clean signal of $B\bar{B}$ events. The decay products of the B mesons at these energies that are long lived enough to be tracked within the detector system are electrons (e), muons

(μ), photons (γ), pions (π), kaons (K), protons (p), and deuterons (d).

The particle identification (PID) routine of the Belle II reconstruction software [2] must be able to identify the nature of each reconstructed particle with high efficiency and purity. Within this task, one of the more difficult challenges is distinguishing charged pions (π^-, π^+) and muons (μ^-, μ^+). The two particles have relatively similar masses, with the muon having a mass of $m_\mu = 105.7\,\mathrm{MeV}/c^2$ and the charged pion mass being $m_{\pi^\pm} = 139.7\,\mathrm{MeV}/c^2$ [3]. This makes a muon difficult to distinguish from a pion with the same momentum but a slightly lower velocity in the Time-of-Propagation (TOP) or Aerogel Ring-Imaging Cherenkov (ARICH) subsystems, which determine the particle's mass using the ratio of momentum to velocity. The outermost detector of the Belle II experiment, the KLM, helps to address this shortcoming by providing a pattern of hits at longer ranges where muons are more likely than pions to penetrate deep into the detector due to the absence of strong (hadronic) interactions of the muon with the intervening material's nuclei.

The current method used for muon identification in the KLM is a likelihood ratio test performed for each possible pairwise particle hypothesis [1]. The test makes use of a set of probability distributions generated in advance for each particle by running a Monte Carlo (MC) simulation of a large number of particles of each type passing through the detector [2]. The performance of the likelihood ratio test method serves as the baseline against which the new method described in this thesis is compared. Final state particles such as kaons or protons are more easily distinguished using information from the TOP and ARICH subsystems, so we focus here on the distinguishability of muons and pions. In an attempt to improve muon identification within the KLM, the performance of this likelihood ratio test will be compared with that of a Fast Boosted Decision Tree (FBDT) machine learning model [4], which receives the same input variables as the likelihood ratio test. The FBDT is trained on a fresh collection of muon and pion MC simulated data. In principle, a FBDT model should be able to identify the input variables that have increased significance in distinguishing particles and weight the input variables accordingly to improve PID performance over the likelihood ratio test baseline.

1.2 Belle II Detector

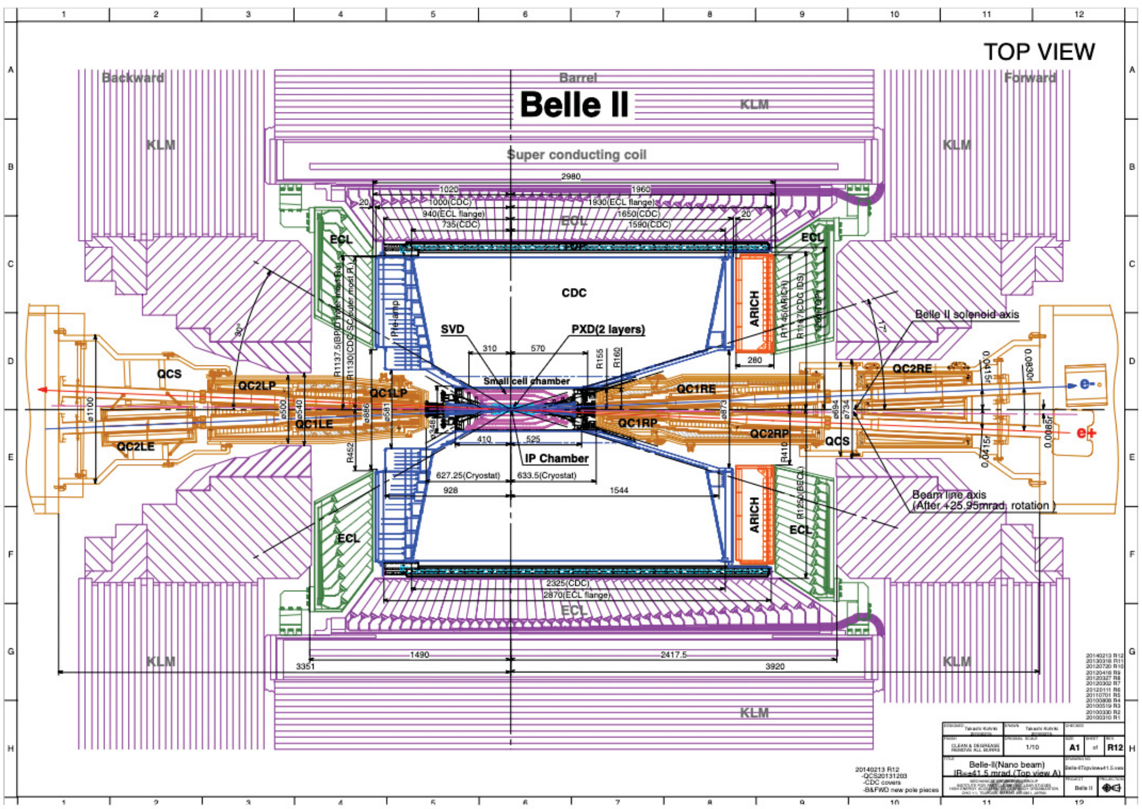


Figure 1.1: Diagram of the Belle II detector given as Figure 3 in the Belle II Physics Book [1]. The horizontal (z) axis is approximately along the beam direction. The VXD which is comprised of the PXD and SVD is the innermost magenta section. The CDC is outlined in blue, the TOP detector is a thin strip between the CDC and ECL, and the ECL is outlined in purple. The outermost pink regions are the KLM, the primary detector of interest in this study.

The Belle II experiment makes use of a series of concentric detectors surrounding the interaction point of the electron and positron beams, shown in Figure 1.1. The coordinate system of the experiment is shown in Figure 1.2, with the $+z$ axis approximately along the electron beam path. From the interaction point and moving outwards these detectors are the Vertex Detector (VXD), the Central Drift Chamber (CDC), the Time-of-Propagation counter (TOP), Aerogel Ring-Imaging Cherenkov (ARICH), the Electromagnetic Calorimeter (ECL), and the K-long and Muon detector (KLM) [1]. A superconducting solenoid outside the ECL provides a uniform magnetic field of 1.5 T within the VXD and CDC that bends charged particles along helical trajectories. The VXD and CDC detectors are essential to the reconstruction of charged particle tracks during an event; they translate the measured curvature of a helical trajectory into the particle's three-dimensional momentum. The tracks formed in these detectors are extrapolated out towards the KLM, and the expected intersections of extrapolated tracks with the KLM detector are input variables into the muon identification algorithm. The KLM is the primary detector of interest in this study, providing particle identification information for muons, charged pions, and other final state particles that reach this outermost detector.

1.2.1 Vertex Detector (VXD)

The innermost detector is the VXD, which is comprised of two sections, the Pixel Detector (PXD) and the Silicon Vertex Detector (SVD). The PXD is the closest to the interaction point, with two layers at radial distances of 14 mm and 22 mm respectively. It uses pixelated sensors to give two-dimensional position readings for particle hits. The SVD makes up an additional four layers located at radial distances of 38 mm, 80 mm, 115 mm, and 140 mm, and uses double-sided silicon strip detectors to give a pair of orthogonal one-dimensional hit readouts in each layer. Track finding begins in the VXD, which uses a cellular automaton based model [5] to assemble detector hits into track candidates. Pairs of hits in adjacent layers are combined into connected segments, then possible combinations of segments pass a

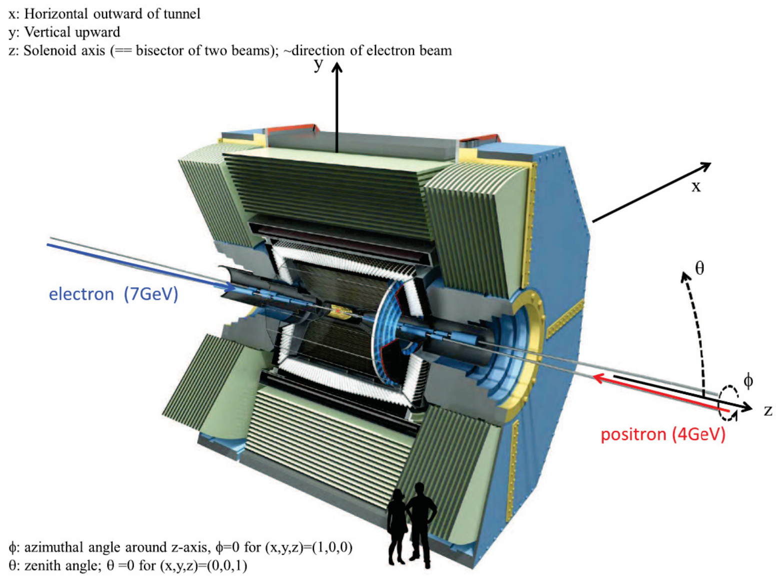


Figure 1.2: Diagram of the coordinate system used in the Belle II detector, diagram courtesy of Belle II and KEK. The horizontal (z) axis is approximately along the beam direction (between the two beam paths), with $+z$ pointing in same direction as the electron beam. ϕ is the azimuthal angle, with $\phi = 0$ aligned with the $+x$ axis. θ is the polar angle, measured from the $+z$ axis.

set of geometrical criteria to be combined into a track candidate. These track candidates are saved and used in the next stage of track fitting algorithm which occurs within the CDC.

1.2.2 Central Drift Chamber (CDC)

The Central Drift Chamber (CDC) contains 56 layers of sense wires arranged in axial or stereo orientations. The hits in the axial (stereo) layers provide the xy (z) coordinates of a track. Charged particles traveling through the drift chamber ionize the chamber gas (50% He, 50% C_2H_6), then the electrons and ions drift in opposite directions towards the anode and cathode wires. The lighter electrons are accelerated by the electric field surrounding the anode and induce an avalanche of additional ionizations, leading to a measurable pulse when this avalanche reaches the anode. From the drift times measured at the anode, a circle representing the nearest distance between the anode and the through-going track is constructed, and based on the synthesized helical path that is tangent to these circles in multiple CDC layers the trajectory can be constructed; in this way, the reconstructed track can be measured. The CDC also employs a cellular automaton model like the VXD to form track segments out of these hits in local regions, and uses a second more global algorithm that combines these segments into track candidates. The track candidates from the VXD are then extrapolated outwards to the CDC and merged with a matching track candidate in the CDC.

Once a series of hits in the VXD and CDC are combined into a track candidate, a fitting algorithm is run to reconstruct the full path of the particle. The particle is hypothesized to have the pion mass, then track curvature and perigee parameters are fit to the sequence of hits associated with the track candidate. The fitting algorithm is based on a standard Kalman filter fitting algorithm [6], which is a least squares method. The parameters of the reconstructed helix are converted to a three-dimensional momentum using the known solenoidal magnetic field of 1.5 T. The Kalman filter based fitting algorithm also outputs a χ^2 value for the fit which is used as input. Finally, the fit parameters determined from the VXD and CDC hits are used to extrapolate the particle track outwards into the KLM, providing multiple expected crossings of the KLM detector layers by the particle.

1.2.3 K-Long and Muon Detector (KLM)

Muon identification takes place primarily in the KLM, and it is the performance in this detector system we hope to improve. The KLM is comprised of three sections: a backward endcap, a forward endcap, and the barrel. Each section contains alternating layers of 4.7 cm thick iron layers and 4.4 cm thick gaps containing detector layers. There are 12 detector layers in the backward endcap, 14 detector layers in the forward endcap, and 15 detector layers in the barrel. (There is one iron layer before the innermost detector layer in each endcap, but not before the innermost detector layer in the barrel.) The endcap detector layers all use scintillator strips that are read out using silicon photomultipliers. The first two barrel detector layers also use scintillator strips, while the remaining 13 use glass-electrode resistive plate chambers. The length from the interaction point to the closest point in the first detector layer is roughly 200 cm, and the length to the closest point in the final detector layer is roughly 320 cm, as found in the collected hit position data of the innermost and outermost layers of the barrel used as input variables for this analysis. The octagonal barrel covers the full azimuthal angle ϕ in the xy plane, and covers a 45° to 125° range in the polar angle θ measured from the $+z$ axis. The endcaps extend the θ angle coverage range to 20° to 150°.

Both the RPC and scintillator layers are formed from overlapping a u-position strip detector and a v-position strip detector, where the u and v directions are orthogonal and tangent to the surface of the detector. Particle hits within 3.5σ of an extrapolated track are considered "matched hits", where σ is the combined measurement uncertainty of the 2D matched hit and the extrapolated-track position in the detector plane. Matched hits are used in the Kalman filter algorithm to update the track parameters. The χ^2 value of the Kalman filter, the hit pattern, and the number of degrees of freedom (two times the number of matched hits) are then used in the likelihood ratio method to determine particle identification, as described in detail in section 3.1.

Chapter 2

Data and Evaluation Metrics

2.1 Monte Carlo Simulated Data

2.1.1 Simulation Methodology

It is necessary to know the true identity of each particle in the event along with the identity assigned by the algorithm of choice for efficiency evaluations of both the maximum likelihood ratio method and the FBDT method. When using real data, this can be done by selecting events that produce only the target particle or decay to particles easily identified by other subsystems or kinematic constraints, such as a $J/\psi \to \mu^+\mu^-$ or $e^+e^- \to e^+e^-\mu^+\mu^-$ decay for muons, or a $K_S^0 \to \pi^+\pi^-$ decay for pions. The other option is to develop and test the algorithms on a Monte Carlo simulation of single-track events. For simulated particles, the identity of the particle is known, and the response of the detector can still be determined, though we have incomplete modeling of the actual detector response over time (such as temporal changes in localized detection efficiencies).

Monte Carlo simulation data is used for the likelihood model evaluation and for testing the FBDT model for this case, in the context of the Belle II Analysis Software Framework (basf2) [2]. The basf2 software package has an incorporated GEANT4 [7] simulation of the detector, the ability to generate particles of known type, starting position, and momentum, and is able to simulate the detector response to an event. Basf2 also provides access to more detailed information for each event; typically, the stored events contain only high level variables that would be used in a physics analysis, while many variables generated at run time describing the detector response are discarded to make the storage of a large amount of data feasible. Access to these extra variables is needed for the purposes of training an FBDT model within basf2, and these are easily stored in MC-simulated events. Finally, once the results of both methods applied on Monte Carlo simulated data are obtained, the program can be easily modified to run on real data samples if the FBDT method is effective.

2.1.2 Simulation Parameters

As discussed in section 1.1, two particle types are generated for all testing: muons and pions. All particles generated during this analysis possess the same initial parameters other than particle type, and are chosen to reflect the conditions of the real experiment. All generated particles start at the interaction point, then travel outwards into the detector. A uniform distribution of initial momentums is generated between 0.3 and 5.0 GeV/c. Below a momentum of approximately 0.6 GeV/c, muons and pions do not reach the KLM due to spiraling in the magnetic field except for when the polar angle of the hit is near the z axis (i.e., at the edge of either KLM endcap), and this low observed particle count can be seen in Figure 2.1. The upper bound on the generated particle momentum is due to the fact that the center of mass energy for the $\Upsilon(4S)$ resonance is 10.57 GeV, so a single muon in an $e^+e^- \to \mu^+\mu^-$ event would have a maximum energy of about half this value. Therefore an upper bound momentum of 5.0 GeV/c covers almost all possible momenta we expect to observe for these particles.

Each generated particle's momentum vector uses a uniform distribution to determine the azimuthal angle around the z axis, ϕ, while a uniform distribution in $\cos(\theta)$ is used for

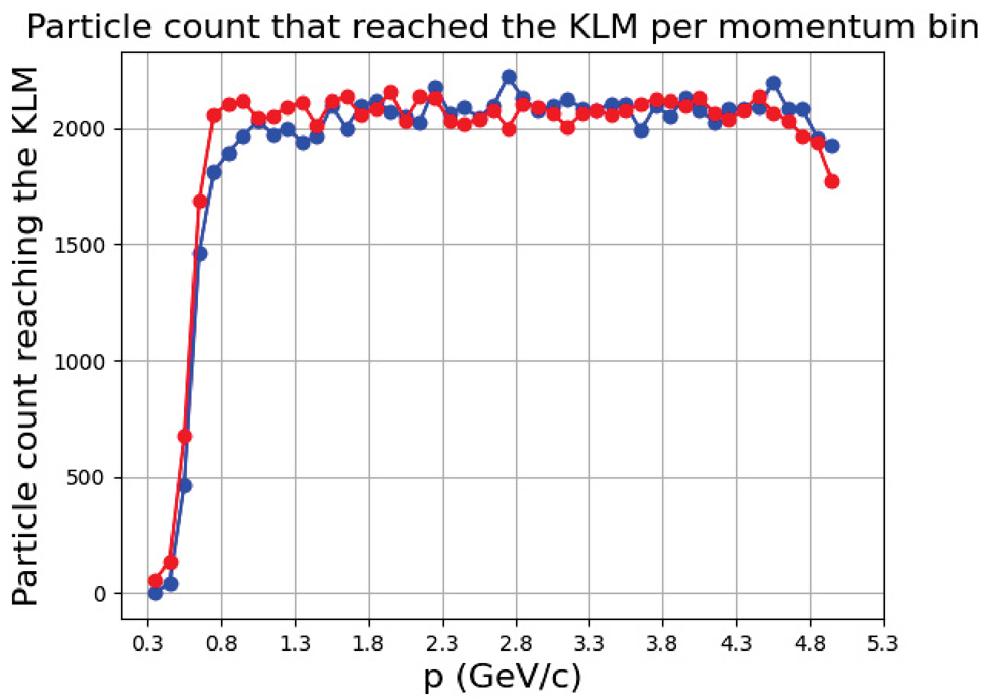


Figure 2.1: Number of pions (red) and muons (blue) whose extrapolated track reached the innermost KLM layer. Particles are collected into momentum bins with width 0.1 GeV/c. This figure was generated for the likelihood ratio test evaluation data set, which had a total of 200,000 particles generated, split between muons and pions. Below a momentum of approximately 0.6 GeV/c we note that very few particles reach the KLM.

the polar angle, θ. The uniform distribution in $\cos(\theta)$ results in more generated particles traveling towards the KLM barrel and regions covered by the endcaps.

Three data sets were produced for model testing and training in this analysis. Before discussing the newly generated data sets, we also note that the likelihood ratio model uses probability distributions for muons and pions that were measured from a prior Monte Carlo simulation of 75 million events for each particle type. The first newly generated MC data set was used in the likelihood ratio performance evaluation, and was generated using the initial conditions described in this section, along with a selection cut that removed tracks that did not reach the KLM from the data set. This data set contains approximately 100,000 muon events and 100,000 pion events before cuts, whose particle counts can be seen in figure 2.1. The other new data sets come from a collection of 2 million pion events and 2 million muon events generated for the purpose of training and evaluating the the FBDT model. Tracks in this data set that do not reach the KLM were not cut, but instead contain variables set to an outlier value to indicate this. This larger collection of 4 million events was split into two subsets: a training subset of 3.2 million events, and a test subset of 800,000 events.

2.2 Receiver Operating Characteristic (ROC)

Both the likelihood ratio test and the FBDT model chosen are examples of binary classifiers, selecting between either a pion or muon hypothesis. Before defining the evaluation metric, there is a necessary caveat to make about the pion and muon identifications. Particles within the MC simulation are capable of decaying; therefore, an observed track within the data set is not necessarily the parent particle that was generated. Such anomalies were removed from the data samples used to construct the likelihood ratio probability density functions, but not from the new FBDT training data sample. The effects of this discrepancy are discussed in the results in section 5. The data sets used for testing the likelihood ratio model and the FBDT model both contain these anomalies. Since the evaluation is constrained to the binary option of muon or pion, the result of not cutting these decay products from the data sets is that a "muon identification" refers to the particle being either a muon or a daughter particle of a muon, while a "pion identification" refers to the particle being either a pion or a daughter particle of a pion.

The effectiveness of a binary classifier can be represented graphically using a Receiver Operating Characteristic (ROC) curve, which plots a True Positive Rate (TPR) as a function of the False Positive Rate (FPR), at multiple thresholds of a test statistic value [8]. The TPR is the number of correctly labeled objects divided by the total number of objects of that type in the data set. For muon PID, this is the fraction of the correctly identified muon tracks (True Positive) divided by the sum of the correctly identified muon tracks (True Positive) and the muon tracks incorrectly labeled as pions (False Negative):

$$TPR = \frac{TruePositive}{TruePositive + FalseNegative} \tag{2.1}$$

The FPR in contrast gives the number of mistakenly labeled objects divided by the total number of objects of that type. Hence for muon PID, this is the fraction of pion tracks mistakenly labeled as muons (False Positive) divided by the sum of pion tracks incorrectly labeled as muons (False Positive) and correctly identified pion tracks (True Negative):

$$FPR = \frac{FalsePositive}{FalsePositive + TrueNegative} \tag{2.2}$$

For a binary classifier in the muon PID case whose output varies continuously between 0 (pion) and 1 (muon), a choice of a cutoff value must be defined, above which the particle is identified as a muon and below which the particle is given a pion identification. It is also possible to define two cutoff values that create an intermediate region corresponding to a non-muon, non-pion track, though this is not implemented here. Usually a higher cutoff decreases the chance that a pion is labeled as a muon, but this also results in more muons being incorrectly labeled as pions. Therefore it is best to calculate TPR and FPR at varying cutoff values to compare possible choices, and the plot of these points produces an ROC curve. A perfect classifier with the optimal cutoff value choice would produce a step-like curve with a vertical line from the origin to the point in the top left corner of the ROC plot, and a classifier that gives a completely random identification would produce a straight line between (TPR, FPR) = (0,0) and (1,1). Outside of the analysis of individual cutoff values and points, the Area Under the Curve (AUC) for an ROC plot can be calculated and used as a measure of the overall performance of the classifier. Another method of analyzing the classifier output is to record the number of particles in the True Positive, True Negative, False Positive, and False Negative categories to examine the effectiveness in each sector. Representative plots containing this information are included in appendix A.

Chapter 3

Likelihood Method

3.1 Likelihood Ratio Test

To understand the statistical method used to determine particle identifications in the KLM, we first describe our input data, then build up the likelihood ratio test method from there. From the Kalman filter algorithms used to fit tracks to observed hits we receive a χ^2 value and the number of degrees of freedom n. The probability of obtaining these values given the KLM regions crossed D (barrel, endcap, or both), and the particle hypothesis H would be $P(\chi^2, n|D, H)$. This probability is referred to as the transverse probability value, and represents the probability an extrapolated track would produce matched hits in the plane of the detector layers that result in the calculated χ^2 value. We can also define a bit pattern vector $\bar{c}$ with 15 elements that indicates whether there was a matched hit in each layer ($c_i = 1$ for a matched hit, $c_i = 0$ for no matched hit in layer i). The expected $\bar{c}$ vector varies with the outcome of the extrapolation O, the final layer containing a matched hit l, and the particle hypothesis H. Thus we have a second probability, $P(\bar{c}|H, l, O)$, for finding a given $\bar{c}$ vector. This is referred to as the longitudinal probability value since it describes a probability of a particle creating a specific hit pattern as it travels outwards through the KLM. The longitudinal and transverse probability distributions were both determined from a Monte Carlo simulation of a large number of particles of given type H passing through the detector, as mentioned in section 2.1.2.

Using these probability distributions we would like to determine H, the particle hypothesis. Thus we want to calculate $L(H|data) \simeq P(data|H)$, which is the likelihood that the particle hypothesis H is of a chosen type (muon, or pion here) given the input data [9]. We can use this relation to note that the desired likelihoods for the transverse and longitudinal probabilities would be $L(H|\chi^2, n, D) \simeq P(\chi^2, n|D, H)$ and $L(H|\bar{c}, l, O) \simeq P(\bar{c}|H, l, O)$ respectively. The total likelihood for H is then the product of these likelihoods, or written using the probability values,

$$L(H|\chi^2, n, \bar{c}, D, l, O) = P(\chi^2, n|D, H)P(\bar{c}|H, l, O) \tag{3.1}$$

To discriminate between the muon and pion particle hypotheses, we must choose a test statistic. The test with the highest discriminating power given that there are no unknown parameters is the likelihood ratio test [9]. While only considering the muon and pion hypotheses for H, the ratio of the likelihoods then has the form,

$$\Lambda(\chi^2, n, \bar{c}, D, l, O) = \frac{L(\mu^{\pm}|\chi^2, n, \bar{c}, D, l, O)}{L(\pi^{\pm}|\chi^2, n, \bar{c}, D, l, O)} \tag{3.2}$$

Taking the logarithm of this ratio we arrive at the final test statistic used,

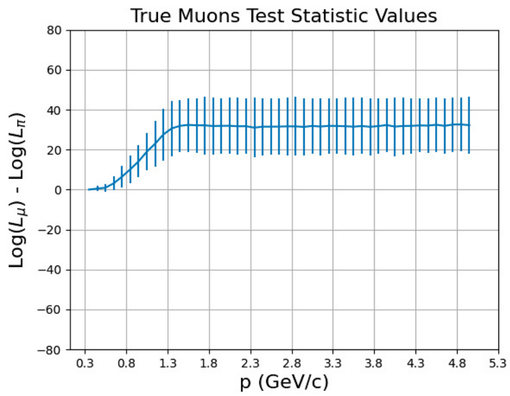
$$\Delta = log(L(\mu^{\pm}|\chi^2, n, \bar{c}, D, l, O)) - log(L(\pi^{\pm}|\chi^2, n, \bar{c}, D, l, O)) \tag{3.3}$$

Particle identification is then performed by calculating the log likelihood for the muon and pion hypotheses, finding the log likelihood difference Δ, then comparing it to a cutoff value Δ_{min}. In this analysis, particles with $\Delta > \Delta_{min}$ would then be identified as muons and particles with $\Delta < \Delta_{min}$ would be identified as pions, and possible values of the cutoff can be plotted and compared on an ROC curve. In general it is also possible to define an intermediate region between two cutoff values that identifies a non-muon, non-pion track, though this is not implemented here. It is important to note that unlike the probabilities, the likelihood values are not normalized, therefore the range of values possible for Δ should be measured before considering possible cutoff values.

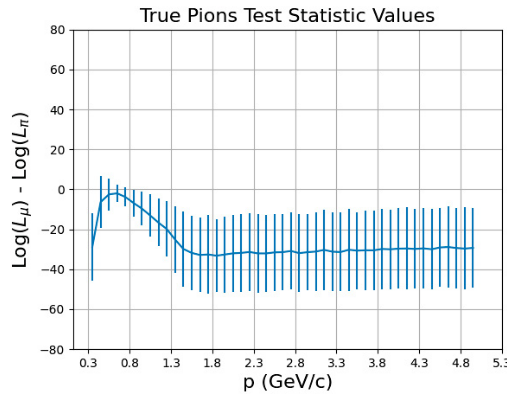
3.2 Likelihood Ratio Performance

The evaluation of the performance of the likelihood ratio method was done using the data set of 200,000 events split between muons and pions described in section 2.1.2, separated into 0.1 GeV/c bins. The average values of the log likelihood test statistic Δ for each bin are shown in two plots in Figure 3.1. For Figure 3.1a, Δ was calculated using only the subset of tracks from the 100,000 muon events, and Figure 3.1b contains only the subset of tracks from the 100,000 pion events. From the muon plot we see that within one standard deviation Δ reaches a maximum near 45, and for the pions within one standard deviation we see a minimum near -50. Therefore reasonable choices for the cutoff value Δ_{min} should lie between -40 and 40.

ROC curves of the likelihood ratio test results were generated for momentum bins between 0.6 and 5.0 GeV/c, with the cutoff value Δ_{min} increased from -40 to 40 in intervals of 2. The full set of ROC curve plots are given in Appendix B.1, while plots representative of the overall performance are used here. The ROC curve for the $0.6 < p < 0.7\,\mathrm{GeV}/c$ bin is given in Figure 3.2a, and is representative of the low momentum region $p < 1.1\,\mathrm{GeV}/c$. In the low momentum region the Area Under the Curve (AUC) of the plots remains below 0.96, suggesting that there is room for improvement. If we want a sample of muons in the $0.6 < p < 0.7\,\mathrm{GeV}/c$ momentum range, assuming we desire a False Positive Rate (FPR)



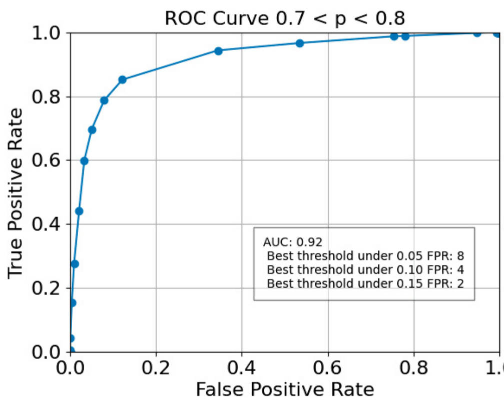
(a) Graph of Δ values calculated using only muon events.



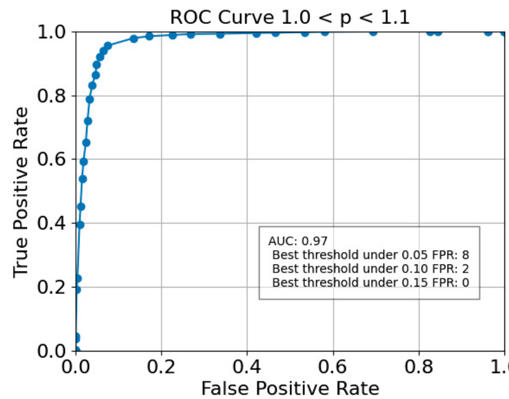
(b) Graph of Δ values calculated using only pion events.

Figure 3.1: $\Delta = logL_\mu - logL_\pi$ test statistic plotted as an average for each momentum bin. One standard deviation is given above and below each Δ value.

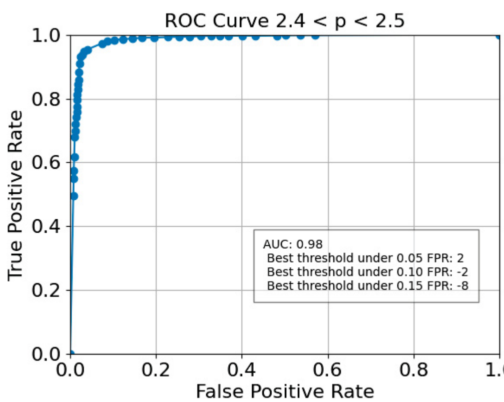
less than 10%, then the best possible True Positive Rate (TPR) is around 82%. The $1.0 < p < 1.1\,\mathrm{GeV}/c$ momentum bin plot shown in Figure 3.2b shows that as the momentum of the particles increases the efficiency of the likelihood ratio classifier significantly improves, and it is unlikely that another classifier will be able to improve upon this performance. The likelihood ratio classifier remains effective as momentum increases, as can be seen in the $2.4 < p < 2.5\,\mathrm{GeV}/c$ and $4.2 < p < 4.3\,\mathrm{GeV}/c$ momentum bins plotted in Figures 3.2c and 3.2d respectively.



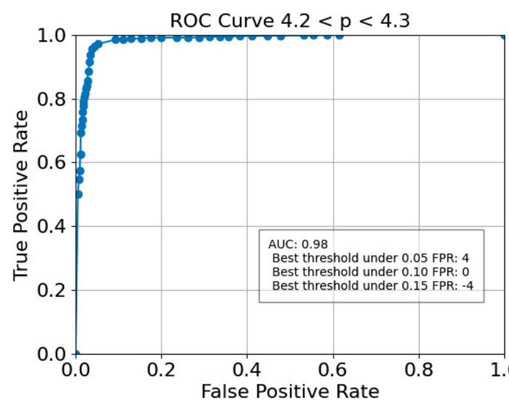
(a) ROC curve for $0.6 < p < 0.7\,\mathrm{GeV}/c$. In this lower momentum region, there is room for improvement in the classification model, which is evident from the AUC value of 0.92.



(b) ROC curve for $1.0 < p < 1.1\,\mathrm{GeV}/c$. Above this momentum range the classification model is very effective.



(c) ROC curve for $2.4 < p < 2.5\,\mathrm{GeV}/c$. Within the middle momentum ranges the performance of the log likelihood ratio model has little room for improvement.



(d) ROC curve for $4.2 < p < 4.3\,\mathrm{GeV}/c$. In the higher momentum ranges muons and pions are easily distinguished by the classifier model.

Figure 3.2: Likelihood ratio model Receiver Operating Characteristic (ROC) curves for a select number of momentum bins within the full range. The Area Under the Curve (AUC) is calculated and given in the plot legend, and represents a measure of the overall classification performance, with an AUC close to 1 being optimal. The best choice of cutoff value Δ_{min} is also given for an FPR below certain thresholds.

Chapter 4

Boosted Decision Tree Method

A Boosted Decision Tree (BDT) is a machine learning model that is useful for classification tasks [10][11]. In order to produce a functional classifier with machine learning methods, we begin with a set of input variables $\bar{x}$ and a target variable y that contains a known classification (a muon or pion identification in this case). Therefore we want to generate a final classification function $y = F(\bar{x})$ that maps the input variables $\bar{x}$ to the desired target variable y, so that true muons are assigned a value of $y \simeq 1$ while true pions are assigned a value of $y \simeq 0$. We choose a "loss function" Ψ, which is a function of the target variable y and the current output of an intermediate classification function $\tilde{F}(\bar{x})$. The loss function is chosen such that the minimum of its expectation value is the best possible classification function $F(\bar{x})$,

$$F(\bar{x}) = argmin_{\tilde{F}(\bar{x})} E(\Psi(y, \tilde{F}(\bar{x}))) \tag{4.1}$$

We must use the expectation value $E(\Psi)$ here because the loss function is dependent on the the random input variables $\bar{x}$, so it is also a random variable. The process of minimizing the loss function over multiple updates of the intermediate classification function $\tilde{F}(\bar{x})$ is called "training". The process of training is described more in depth in sections 4.1.2, 4.1.3, and 4.2. We note here that the final classification function is constructed in an additive manner out of smaller functions called "base learners". Therefore the output of a training is a file containing weights which are the final parameter values of these base learners.

The Belle II Analysis Software Framework (basf2) provides a built in package that implements a Fast Boosted Decision Tree (FBDT) training algorithm and programs to apply the trained model to data. In the following sections we outline the theoretical framework behind Boosted Decision Trees to better understand the structure and output of an FBDT.

4.1 Boosted Decision Tree Model

4.1.1 Decision Trees

Decision Trees (DTs) are a classification model that use a series of cuts to split the data set and lead each subsection to a final classification [4]. The full data set can be represented at the top of a tree by an initial node. The terminal nodes at the bottom of the tree represent values of the target variable y. Intermediate nodes are arranged in layers, and each intermediate node contains a bifurcation on one of the input variables x_i. At each of these cuts or "decision nodes" the data set is split into two sets that head to different decision nodes or into a terminal node. The variable and value used to cut on is determined from the separation gain at each decision node. The DT must have a fixed finite depth, otherwise the DT being trained will continue to create subdivisions and produce a final tree that is effective only for the training data set, referred to as overfitting in machine learning terminology. In the context of muon identification, we want to train a DT that separates muons and pions as the input data set passes through the tree, then assigns the pions a classifier value near 0 and assigns the muons a classifier value near 1 at the terminal nodes.

4.1.2 Gradient Boosted Decision Trees

For a Decision Tree (DT), ”boosting” refers to a method of approximating the function $F(\bar{x})$ given in equation 4.1, which maps the input variables $\bar{x}$ to the output $y = F(\bar{x})$. With the boosting method we use the following expansion to approximate $F(\bar{x})$,

$$F(\bar{x}) \approx \sum_{m=0}^{M} \beta_m h(\bar{x}, \bar{a}_m) \tag{4.2}$$

Where $h(\bar{x}, \bar{a}_m)$ is a base learner, which is a simple function of the input variable vector $\bar{x}$ and a set of parameters $\bar{a}_m$. The values of β_m are the expansion coefficients. The index m can be thought of as the index of the training stages, since at each successive stage of training we add the $(m+1)$-th expansion term to the series. (Usually in machine learning this is called the m-th stage and the prior stage is denoted as $m-1$.) For a Boosted Decision Tree (BDT), the base learner function is a smaller regression DT with L terminal nodes. The regression DT splits the joint space covered by possible values of the x_i variables into L regions. These regions are denoted as $\{R_l\}_{l=1}^{L}$, and are a parameter for the regression DTs. Each region can be thought of as a path from the initial node down the decision nodes to a terminal node, thus there are L total terminal nodes. The performance of each region R_l is calculated for the prior $m-1$ step of training and used to improve the DT performance in the following stage of training m. The equation for the base learner regression trees in each training step m and region l is similar to equation 4.2, since it is another DT model, and it has the form,

$$h(\bar{x}, \{R_l\}_{l=1}^{L}) \approx \sum_{l=1}^{L} \bar{y}_{lm} G(\bar{x} \in R_{lm}) \tag{4.3}$$

Here $G(\bar{x} \in R_{lm})$ is a function that equals 1 when the argument is true and 0 otherwise. The expansion coefficient β_m is replaced by $\bar{y}_{lm}$ for decision tree base learners. The $\bar{y}_{lm}$ value is the mean value of the pseudo-residuals of the loss function, in the l partition region, at the m stage of training, and is calculated from the previous stage of training $(m-1)$. The $\tilde{y}_{lm}$ psuedo-residuals replace the initial y values in training stages after the initial training stage, thus changing the parameter and value that optimizes information gain at each decision node of the DT. Thus for $\bar{y}_{lm}$ and $\tilde{y}_{lm}$ we have the following,

$$\bar{y}_{lm} = mean_{\bar{x}_i \in R_{lm}}(\tilde{y}_{lm}) \tag{4.4}$$

$$\tilde{y}_{lm} = -\left(\frac{\partial \Psi(y_i, F(\bar{x}_i))}{\partial F(\bar{x}_i)}\right)_{F(\bar{x})=F_{m-1}(\bar{x})} \tag{4.5}$$

From this point we are able to update the approximation of the desired function $F(\bar{x})$ for each separate region indexed by l, so instead of using β_m we use γ_{lm} as coefficients, defined by,

$$\gamma_{lm} = argmin_{\gamma} \sum_{\bar{x}_i \in R_{lm}} \Phi(y_i, F_{m-1}(\bar{x}_i) + \gamma) \tag{4.6}$$

Therefore γ_{lm} is a value that when added to the $F(x)$ approximation from the previous training step $(m-1)$, denoted as $F_{m-1}(\bar{x}_i)$, minimizes the cost function Φ. Each stage of training, the regions R_{lm} that are summed over are redefined by the retraining of the DT. Thus each new DT generated alters γ_{lm}, the result of minimizing the loss function. We then want to update the approximation of the previous step by adding this new term, giving the update rule:

$$F_m(\bar{x}) = F_{m-1}(\bar{x}_i) + \nu\gamma_{lm}G(\bar{x} \in R_{lm}) \tag{4.7}$$

Here we have defined a shrinkage parameter ν that is chosen at the start of training that controls the "learning rate" of the algorithm, which scales how large an effect each training stage has. This defines the full procedure of gradient boosting, after we start from an initial approximation function $F_0(\bar{x})$. The initial approximation chosen is a DT which is created using the initial labels y_i for quantifying separation gain at each decision node. The extension to this procedure is stochastic gradient boosting [10], which is used in the algorithm built in to the Belle II Analysis Software Framework (basf2). The alterations to the Gradient Boosted Decision Tree (GBDT) model will not be fully elaborated on here, but in summary, the stochastic modification performs these update rules using a randomly selected subset of the total data set. This incorporates randomness into the model which improves the performance of the training over time, and reduces computation time.

4.1.3 Fast Boosted Decision Trees

Within the Belle II Analysis Software Framework (basf2) a Fast Boosted Decision Tree (FBDT) Model [4] is included and is used for this analysis. FBDT's are stochastic Gradient Boosted Decision Trees (GBDTs) modified by changing the variable access pattern to optimize the computation speed. The loss function used for the training procedure is a negative binomial log-likelihood [4][11], which has the following form,

$$\Psi = \log(1 + e^{-2yF}) \tag{4.8}$$

for y values in the range $[-1, 1]$, which is modified to $[0, 1]$ in the output of the FBDT model.

(There is also an alternative form of this loss function for the interval $[0, 1]$ that could be used directly.)

4.2 Fast Boosted Decision Tree Training

The FBDT model was trained on a total of 138 input variables. For each layer there is an input variable for the extrapolated hit position (radius, azimuthal angle, and polar angle in separate variables), the matched hits (radius, azimuthal angle, and polar angle), the hit time measured from the particle being generated, and the Kalman filter fit χ^2. This set of variables is given for 15 layers. The final variables passed to the FBDT model are the track momentum magnitude, and its polar and azimuthal angles. This collection of input variables is the same as the set of variables used in the likelihood ratio model, with slight differences in the way the values are processed (the likelihood model uses a bit pattern of layer hits, as opposed to the individual layer variables used in the FBDT model).

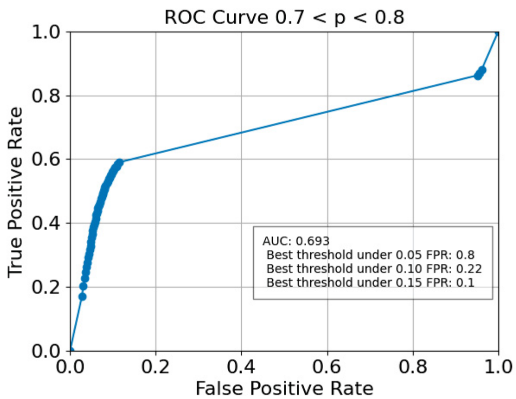
Training of the FBDT model was performed on the Virginia Tech physics department's computing cluster Tesla using the FBDT training data set of 3.2 million events split between muons and pions described in section 2.1.2. Six training runs were performed, varying the tree depth, number of trees constructed in each stage of training, learning rate, and training on a subset of the data in the low momentum region. No appreciable improvement in the ROC curves was observed for tree depths above 3 levels, with altering the learning rate, or with an increased number of trees generated per stage. The FBDT model trained only on a low momenta subset of the data had decreased performance due to an insufficient amount of training data, thus this specialization towards the momentum region of interest was not productive for improving the ROC curves. The final results presented are for a FBDT model with 200 trees created at each training stage, 8 cuts made per stage to divide the input variables into separate regions, a learning rate of 0.1, and a maximum tree depth of 3.

4.3 Fast Boosted Decision Tree Performance

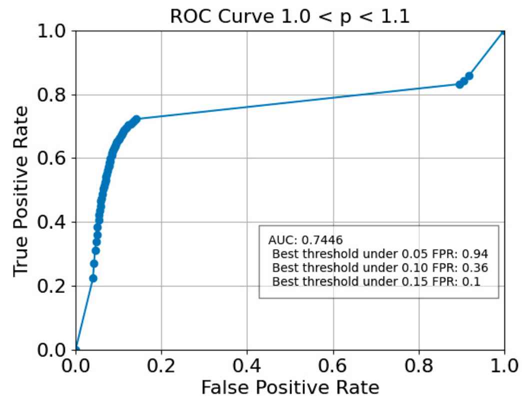
The performance of the trained FBDT model is evaluated by using a test data set of 800,000 events split between muons and pions that was not used during training (described in section 2.1.2). ROC curves were generated for 0.1 GeV/c momentum bins between 0.6 and 5.0 GeV/c, with the cutoff value δ_{min} increasing from 0 to 1 in 0.02 increments. The full list of ROC curves is given in Appendix B.2, and we discuss examples shown in Figure 4.1 that are representative of the performance for specific regions of interest.

The ROC curve for the $0.7 < p < 0.8$ momentum bin is given in Figure 4.1a, which is within the low momentum region where improvement over the likelihood ratio model may be possible. We see from Figure 4.1a that performance is notably worse than the corresponding likelihood ratio ROC given in Figure 3.2a, with an Area Under the Curve (AUC) of 0.693 for the FBDT ROC curve, compared to the AUC of 0.92 for the likelihood ROC curve. Within the FBDT ROC curve for this momentum bin, we also note that there is a sharp horizontal jump between two points (cutoff value choices). This indicates that the FBDT creates a large subset of tracks with a similar value for the classification variable. Since this drastic shift occurs around 0.15 FPR, there is a large cluster of tracks given a classification value of 0.1 as indicated from the threshold value given for this FPR. This feature is observed for the ROC curves of all momentum bins for the FBDT model.

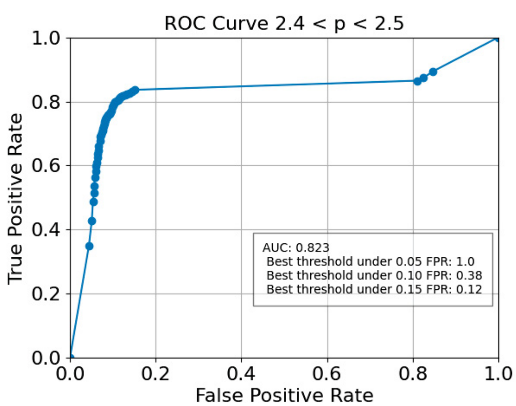
As the momentum increases, the improvement in classifier performance is not as fast for the FBDT model as it is for the likelihood ratio model. This can be seen in the comparison between the FBDT ROC curve for the $1.0 < p < 1.1$ momentum bin, shown in Figure 4.1b, and the likelihood ratio ROC curve for the same bin shown in Figure 3.2b. The likelihood ratio ROC curve has already reached near its maximum efficiency, with an AUC of 0.96, while the FBDT ROC curve quality is rising more slowly with momentum increases, at an AUC of 0.745 for this momentum bin. The increase in quality of the FBDT ROC curves stops around the ROC curve for the $2.4 < p < 2.5$ momentum bin, shown in Figure 4.1c. At higher momenta the FBDT ROC curves do not reach the performance level shown in the likelihood ratio ROC curves, as can be seen in Figure 4.1d.



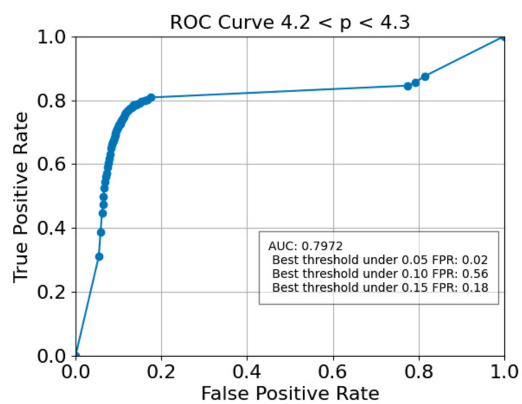
(a) ROC curve for $0.6 < p < 0.7\,\mathrm{GeV}/c$. In the low momentum region there is room to improve on the likelihood ratio model, but the AUC value of 0.693 and the ROC curve found for the FBDT model in this range does not accomplish this goal.



(b) ROC curve for $1.0 < p < 1.1\,\mathrm{GeV}/c$. The FBDT classifier performance does improve going to higher momentum, but does not surpass the likelihood ratio classifier performance.



(c) ROC curve for $2.4 < p < 2.5\,\mathrm{GeV}/c$. At higher momentums the FBDT model approaches the performance the likelihood ratio model achieved at low momentum.



(d) ROC curve for $4.2 < p < 4.3\,\mathrm{GeV}/c$. In the higher momentum ranges the increase in quality of the ROC curves stagnates.

Figure 4.1: FBDT model Receiver Operating Characteristic (ROC) curves for a select number of momentum bins within the full range. The Area Under the Curve (AUC) is calculated and given in the plot legend, and represents overall classification performance, with an AUC close to 1 being optimal. The best choice of cutoff value δ_{min} is also given for an FPR below certain thresholds. In comparison to the likelihood ratio model ROC curves in Figure 3.2 the performance of the FBDT model is worse across the whole momentum range.

Chapter 5

Model Discussion

From the analysis of the ROC curves generated for both the FBDT and likelihood ratio models, the FBDT model performance as a classifier was worse across the entire momentum range. Alterations to the FBDT training and data set were made to attempt to improve performance, but were unsuccessful. As mentioned briefly in section 4.2, a training run of the FBDT model was attempted for low momenta (below 1.3 GeV/c) in an effort to identify features more significant at lower momentum. However, the data set suffered from a lack of training examples when subdivided in this way, so a larger MC data set would be needed to attempt training focused on this momentum region. Changes to the FBDT model parameters led primarily to the same outcome, suggesting that the trained model converges on the result produced for the FBDT ROC curves when using this data set.

A primary cause of the FBDT model's poor performance is the lack of pure muon and pion signals in the training data set due to particle decays occurring during the MC simulation for the FBDT model. Since all particle tracks that reached the KLM were accepted into the data set, if a muon or pion decayed and produced daughter particles that reached the KLM layers it would produce a non-muon, non-pion track. This affects the training since the model must categorize these decay product tracks. (This is also the reason why the two final classifications of the FBDT model are "muon or muon decay product" and "pion or pion decay product" as discussed in section 2.2.) These daughter particles were not present in the data used to develop the probability density functions for the likelihood ratio model, unfairly biasing the training of the FBDT model. In future work these daughter particles can be removed from the data set by performing pre-processing when the MC simulation is performed to generate the data set. With a data set of pure muon and pion tracks, the classification task becomes easier in training and we could expect to see an increase in performance for the FBDT classifier.

A final consideration for improving the FBDT model's performance is that the variable choice may be sub-optimal. Machine learning models are sensitive to the format in which input variables are presented, so changes like giving the FBDT model the distance between a matched hit and an extrapolated hit as input instead of the individual matched hit and extrapolated hit positions could affect overall performance. Using a bit pattern vector of matched hits for the FBDT similar to the one used for the likelihood ratio test is another possible alteration. In addition, the FBDT model is not constrained to using only the variables used in the likelihood ratio model. The natural choice is training on the same variables, but the FBDT model could be trained using lower level detector variables, such as the energy deposition of a hit in the KLM, in addition to the variables used for the likelihood ratio model.

Chapter 6

Conclusion

Efficient particle identification classifiers for muons and pions using the KLM detector are necessary to the research done by the Belle II collaboration. Using a Monte Carlo (MC) simulated data set of 200,000 events, the performance of the likelihood ratio classifier in use within the Belle II analysis software (basf2) was characterized for momenta between 0.6 and 5.0 GeV/c. Below momenta of 1 GeV/c, the likelihood ratio test model achieves True Positive Rates (TPRs) less than 90% for False Positive Rates (FPRs) no greater than 10%. In this region the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve is also under 0.95, indicating room for improvement in muon identification. Above momenta of 1.2 GeV/c, the likelihood ratio model performs very well, with AUCs of 0.97 or higher for every ROC curve.

It was possible that muon classification performance could be improved in the low momentum region through the implementation of a machine learning model trained on the same input variables as the likelihood ratio test. A Fast Boosted Decision Tree (FBDT) model was chosen for this purpose since it is designed for use in classification tasks. The FBDT model was trained on a data set of 3.2 million events split between muons and pions, with the same input variables as the likelihood ratio model. The performance of the trained FBDT model was then evaluated using a MC data set of 800,000 events split between muons and pions. The resulting FBDT ROC curves performed worse than the likelihood ratio ROC curves for all momenta considered, having lower AUC values in all cases. The FBDT ROC curves also have a sharp horizontal shift as they reach their maximum TPR, making it difficult to select good cutoff values past this point. Worse performance of the FBDT model is likely due to the lack of pre-processing of events to remove muon and pion decay products before training the FBDT.

Bibliography

[1] Kou E. et al. The Belle II Physics Book. *Progress of Theoretical and Experimental Physics*, 2019(12), 2019. ISSN 2050-3911. doi: 10.1093/ptep/ptz106.

[2] Kuhr T. et al. The Belle II Core Software. *Comput Softw Big Sci*, 3(1), 2019. doi: https://doi.org/10.1007/s41781-018-0017-9.

[3] Workman R. L. et al. Review of Particle Physics. *PTEP*, 2022:083C01, 2022. doi: 10.1093/ptep/ptac097.

[4] Keck T. FastBDT: A Speed-Optimized Multivariate Classification Algorithm for the Belle II Experiment. *Comput Softw Big Sci*, 1(2), 2017. doi: https://doi.org/10.1007/s41781-017-0002-8.

[5] Lueck T. et al. The track finding algorithm of the Belle II vertex detectors. *EPJ Web Conf.*, 150(00007), 2017. doi: https://doi.org/10.1051/epjconf/201715000007.

[6] Fruhwirth R. Application of Kalman filtering to track and vertex fitting. *Nucl. Instrum. Meth.*, 262:444–450, 1987. doi: https://doi.org/10.1016/0168-9002(87)90887-4.

[7] Agostinelli S. et al. Geant4: A simulation toolkit. *Nucl. Instrum. Meth.*, A506152: 250–303, 2003. doi: https://doi.org/10.1016/S0168-9002(03)01368-8.

[8] Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874, 2006. doi: https://doi.org/10.1016/j.patrec.2005.10.010.

[9] Bove D. S. Held L. *Applied Statistical Inference - Likelihood and Bayes.* Springer Berlin, Heidelberg, 2014. ISBN 978-3-642-37887-4. doi: https://doi.org/10.1007/978-3-642-37887-4.

[10] Friedman J. Stochastic Gradient Boosting. *Comput. Stat. Data Anal.*, 38:367–378, 2002. doi: https://doi.org/10.1016/S0167-9473(01)00065-2.

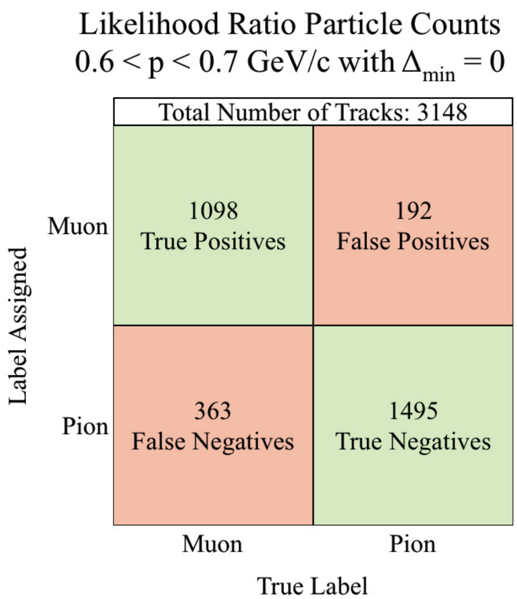
[11] Friedman J. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

Appendices

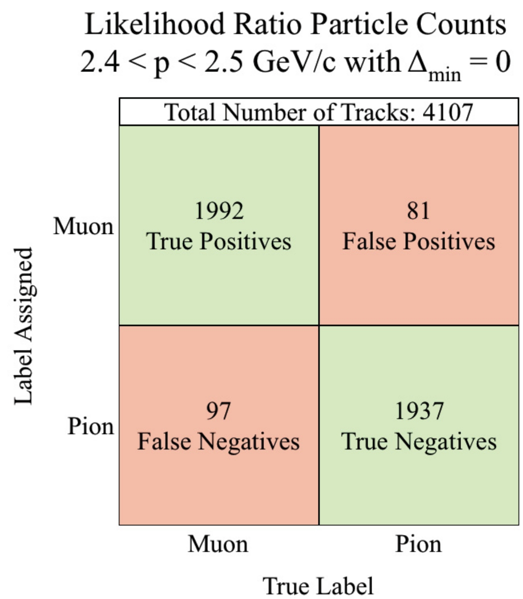
Appendix A

Particle Count Data

These particle count plots show an alternate method of characterizing binary classifier performance. ROC curves were used in place of these because the ROC plots summarize results for multiple cutoff choices, whereas these plots require a choice of cutoff value Δ_{min} or δ_{min}. The cutoffs chosen for these plots maximizes the total number of True Positives and True Negatives, which is a reasonable choice for general use cases, but is not necessarily the best cutoff for all applications. The advantage of these representations of the data is the ability to consider performance in specific categories, such as pions incorrectly identified as muons (False Positives).

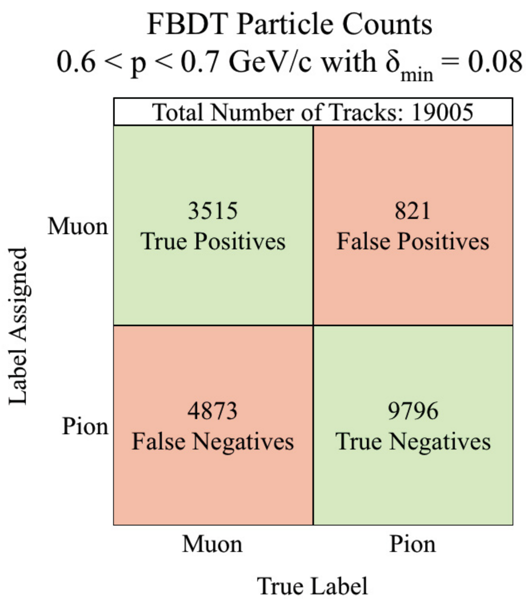


(a) Particle counts in the low momentum bin $0.6 < p < 0.7$ for the likelihood ratio model show that a muon is about twice as likely to be misidentified as a pion in comparison to pions misidentified as muons.

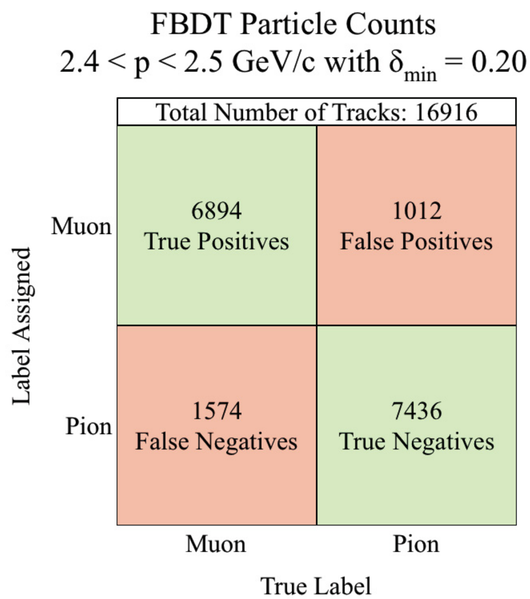


(b) At higher momentums, the misidentification rates are roughly equivalent as seen in the $2.4 < p < 2.5$ bin. The effectiveness of the likelihood ratio model as a classifier in this region is reflected by the numbers of particles in green boxes in the particle count plots. The optimal choice of cutoff value based on our criteria remains the same at higher momenta.

Figure A.1: The total number of muons correctly identified as muons (True Positive), muons incorrectly identified as pions (False Negative), pions correctly identified as pions (True Negative), and pions incorrectly identified as muons (False Positive) for the likelihood ratio model. The boxes shown in green are desirable classifications, while the number of tracks given in red boxes should be minimized to improve classifier performance. The cutoff value Δ_{min} chosen for these plots maximizes the combined number of True Positives and True Negatives.



(a) Particle counts in the low momentum bin $0.6 < p < 0.7$ for the FBDT model reiterate that a muon is much more likely to be misidentified as a pion in comparison to pions misidentified as muons. However when compared to the likelihood ratio model this discrepancy is far larger, and the number of False Negatives outpaces the number of True Positives.



(b) Similarly to the likelihood ratio method, at higher momentums the misidentification rates for the FBDT model are much closer as seen in the $2.4 < p < 2.5$ bin. The likelihood ratio model still outperforms the FBDT model at higher momenta when looking at particle counts. We note that the optimal choice of cutoff value based on our criteria increases at higher momenta.

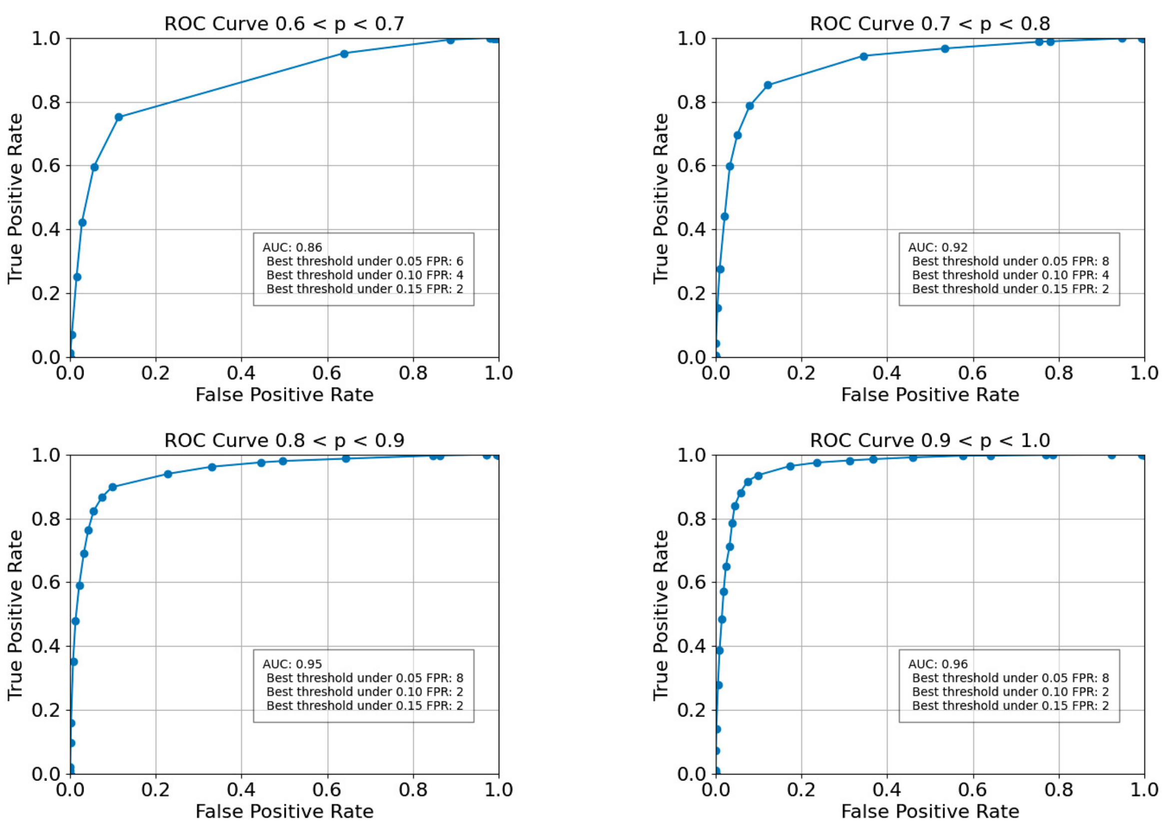
Figure A.2: The total number of muons correctly identified as muons (True Positive), muons incorrectly identified as pions (False Negative), pions correctly identified as pions (True Negative), and pions incorrectly identified as muons (False Positive) for the FBDT model. The boxes shown in green are desirable classifications, while the number of tracks given in red boxes should be minimized to improve classifier performance. The cutoff value δ_{min} chosen for these plots maximizes the combined number of True Positives and True Negatives.

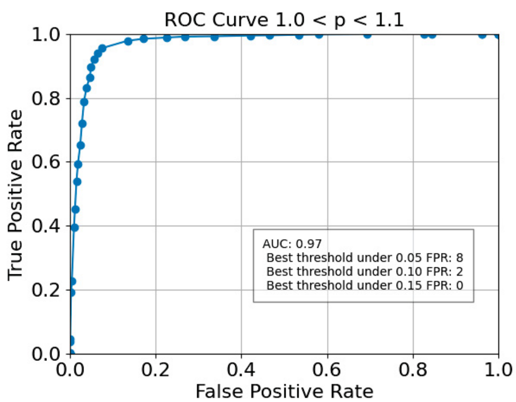
Appendix B

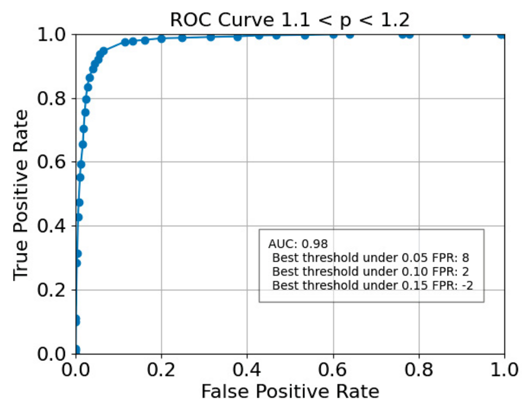
Full set of ROC Plots

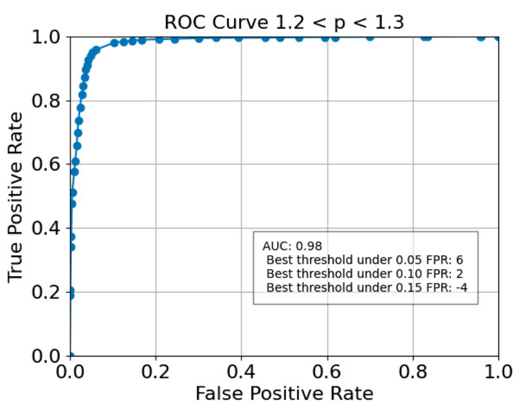
B.1 Likelihood Ratio ROC Curves

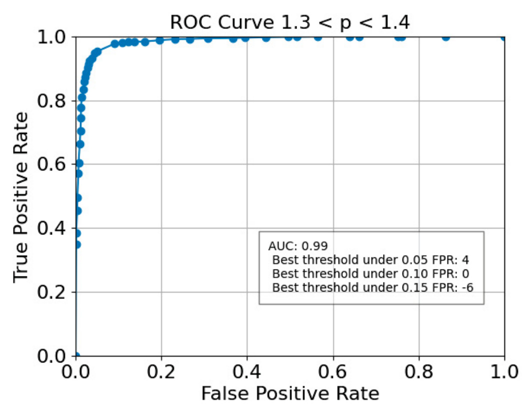
Full set of likelihood ratio ROC curves generated for the momentum range $0.6 < p < 5.0\,\mathrm{GeV}/c$.

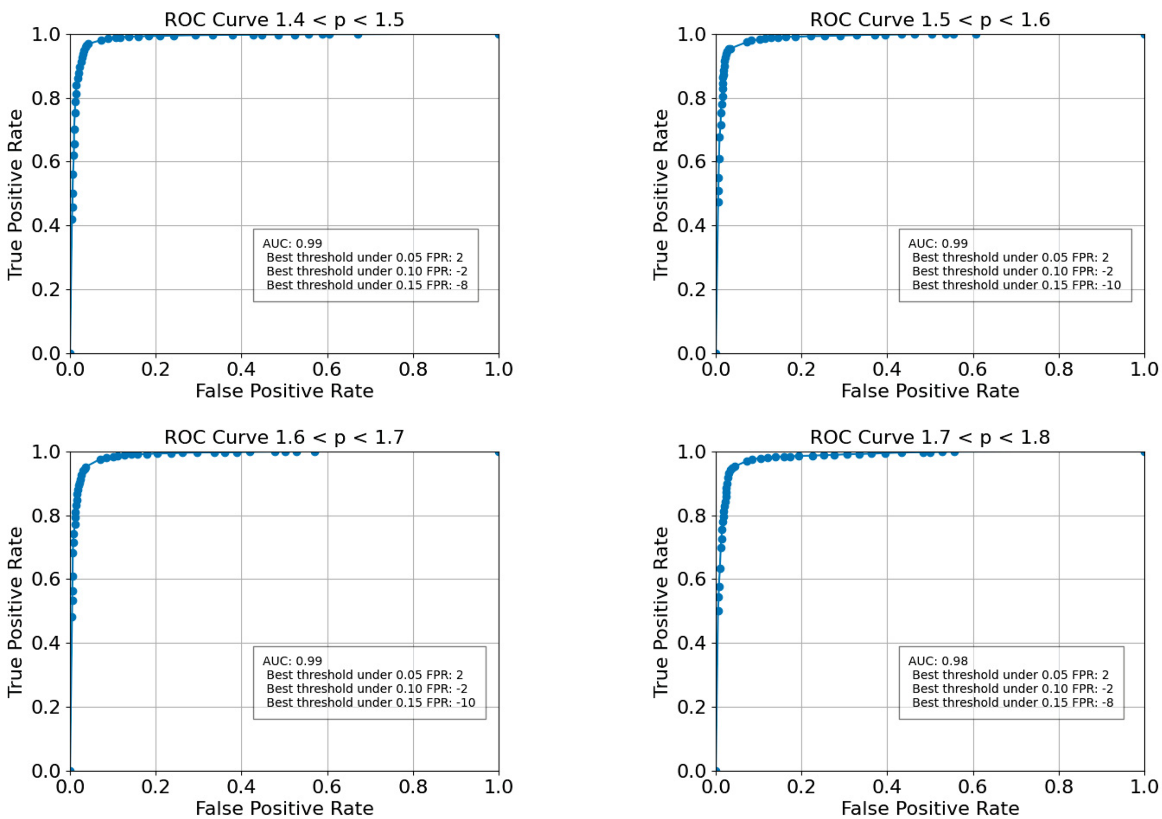


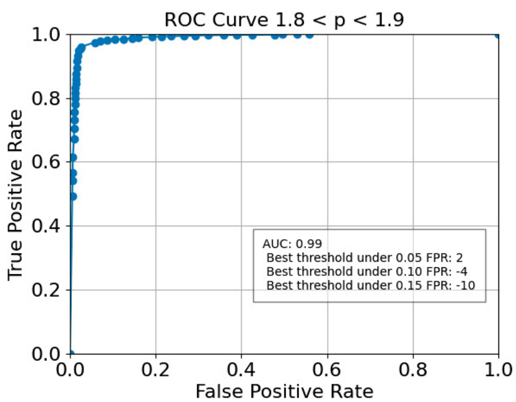

ROC Curve 1.0 < p < 1.1
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC: 0.97
Best threshold under 0.05 FPR: 8
Best threshold under 0.10 FPR: 2
Best threshold under 0.15 FPR: 0

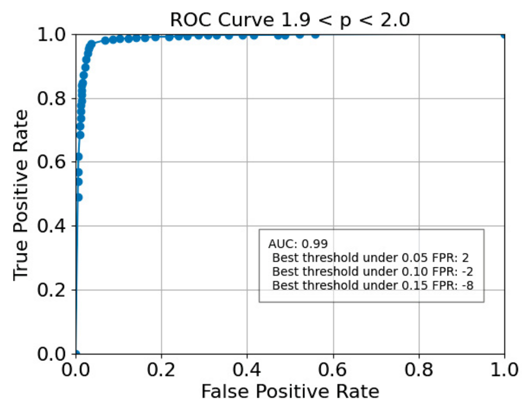

ROC Curve 1.1 < p < 1.2
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC: 0.98
Best threshold under 0.05 FPR: 8
Best threshold under 0.10 FPR: 2
Best threshold under 0.15 FPR: -2

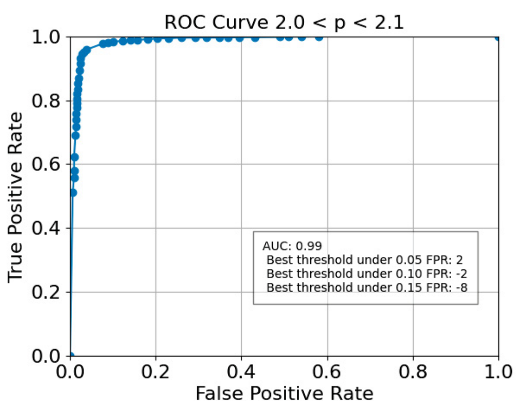

ROC Curve 1.2 < p < 1.3
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC: 0.98
Best threshold under 0.05 FPR: 6
Best threshold under 0.10 FPR: 2
Best threshold under 0.15 FPR: -4

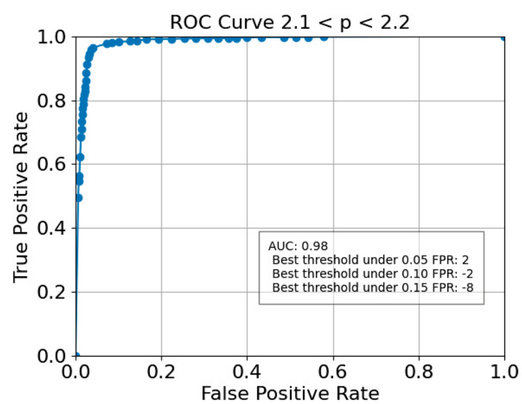

ROC Curve 1.3 < p < 1.4
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC: 0.99
Best threshold under 0.05 FPR: 4
Best threshold under 0.10 FPR: 0
Best threshold under 0.15 FPR: -6

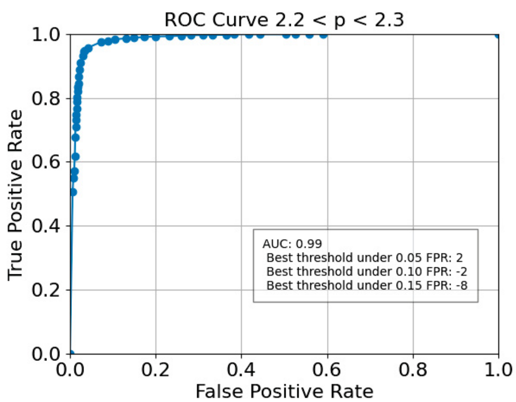

ROC Curve 1.4 < p < 1.5
AUC: 0.99
Best threshold under 0.05 FPR: 2
Best threshold under 0.10 FPR: -2
Best threshold under 0.15 FPR: -8
ROC Curve 1.5 < p < 1.6
AUC: 0.99
Best threshold under 0.05 FPR: 2
Best threshold under 0.10 FPR: -2
Best threshold under 0.15 FPR: -10
ROC Curve 1.6 < p < 1.7
AUC: 0.99
Best threshold under 0.05 FPR: 2
Best threshold under 0.10 FPR: -2
Best threshold under 0.15 FPR: -10
ROC Curve 1.7 < p < 1.8
AUC: 0.98
Best threshold under 0.05 FPR: 2
Best threshold under 0.10 FPR: -2
Best threshold under 0.15 FPR: -8
True Positive Rate
False Positive Rate

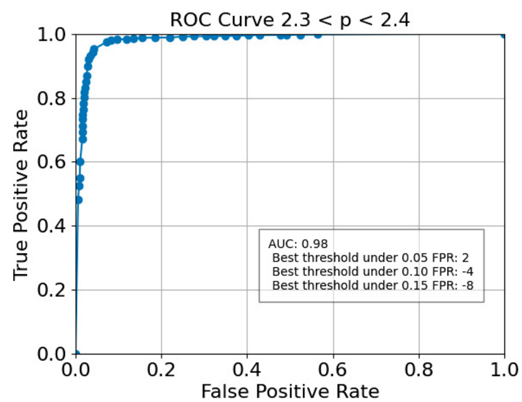

ROC Curve 1.8 < p < 1.9
True Positive Rate
False Positive Rate
AUC: 0.99
Best threshold under 0.05 FPR: 2
Best threshold under 0.10 FPR: -4
Best threshold under 0.15 FPR: -10

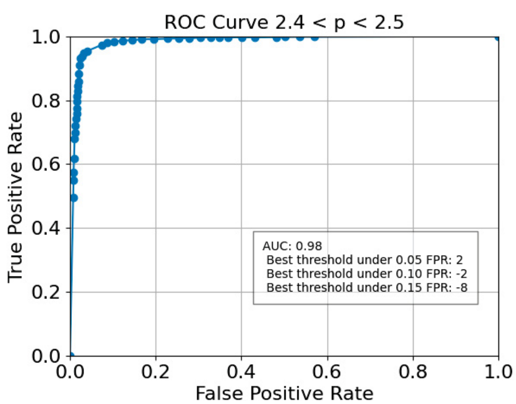

ROC Curve 1.9 < p < 2.0
True Positive Rate
False Positive Rate
AUC: 0.99
Best threshold under 0.05 FPR: 2
Best threshold under 0.10 FPR: -2
Best threshold under 0.15 FPR: -8

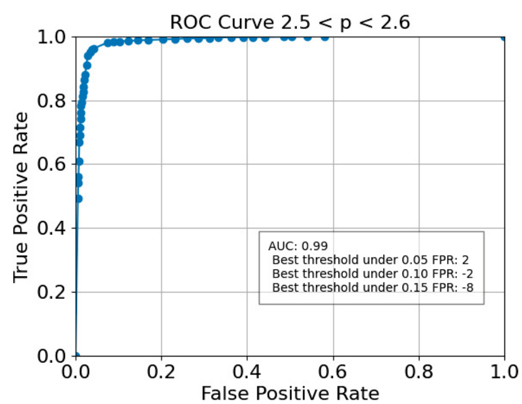

ROC Curve 2.0 < p < 2.1
True Positive Rate
False Positive Rate
AUC: 0.99
Best threshold under 0.05 FPR: 2
Best threshold under 0.10 FPR: -2
Best threshold under 0.15 FPR: -8

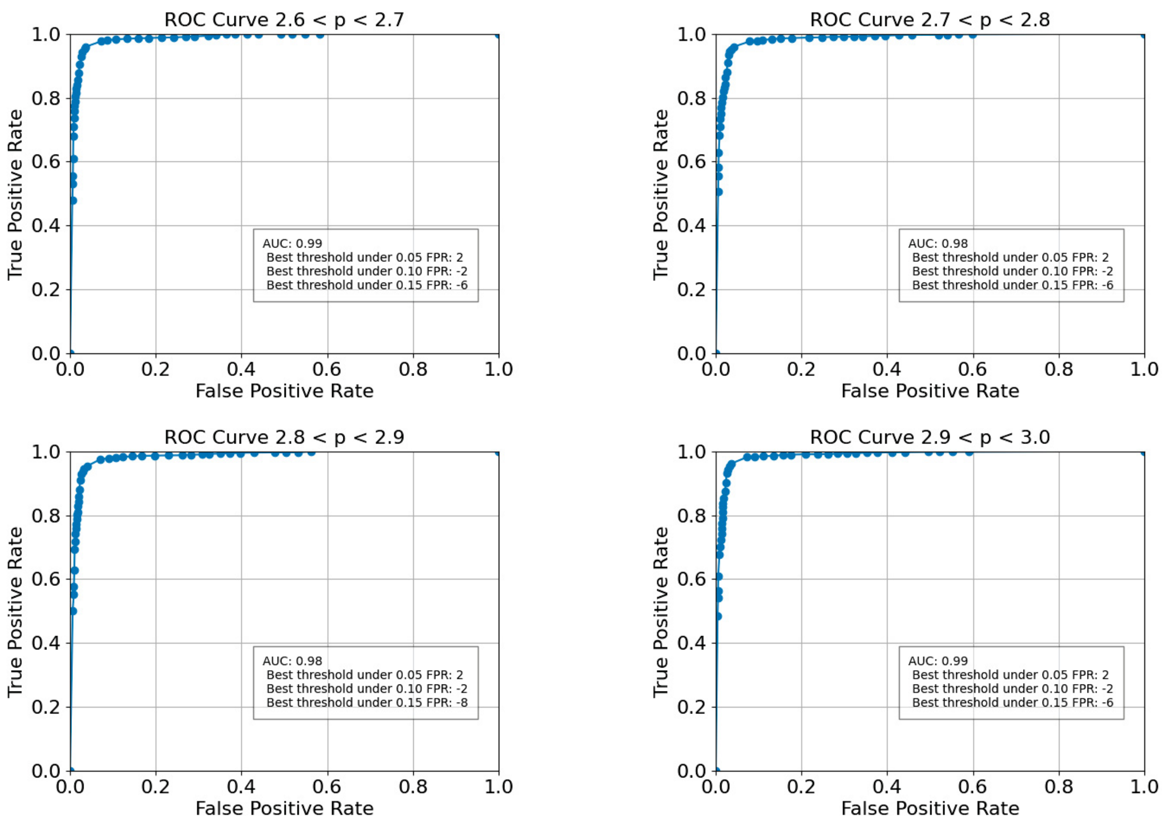

ROC Curve 2.1 < p < 2.2
True Positive Rate
False Positive Rate
AUC: 0.98
Best threshold under 0.05 FPR: 2
Best threshold under 0.10 FPR: -2
Best threshold under 0.15 FPR: -8

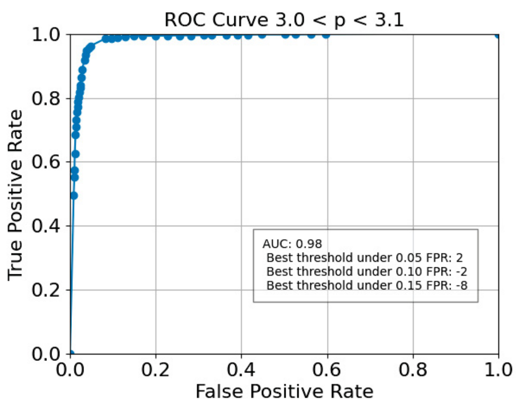

ROC Curve 2.2 < p < 2.3
True Positive Rate
False Positive Rate
AUC: 0.99
Best threshold under 0.05 FPR: 2
Best threshold under 0.10 FPR: -2
Best threshold under 0.15 FPR: -8

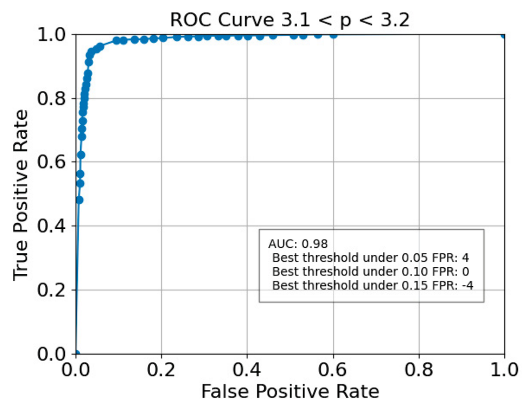

ROC Curve 2.3 < p < 2.4
True Positive Rate
False Positive Rate
AUC: 0.98
Best threshold under 0.05 FPR: 2
Best threshold under 0.10 FPR: -4
Best threshold under 0.15 FPR: -8

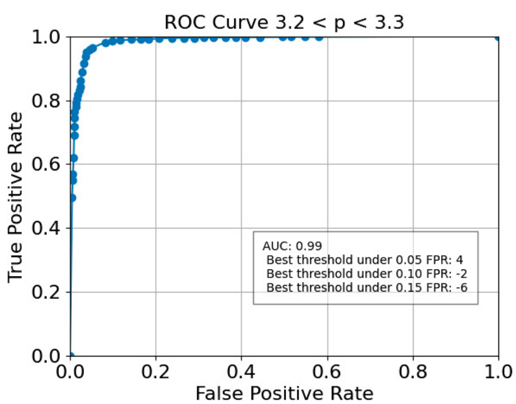

ROC Curve 2.4 < p < 2.5
True Positive Rate
False Positive Rate
AUC: 0.98
Best threshold under 0.05 FPR: 2
Best threshold under 0.10 FPR: -2
Best threshold under 0.15 FPR: -8

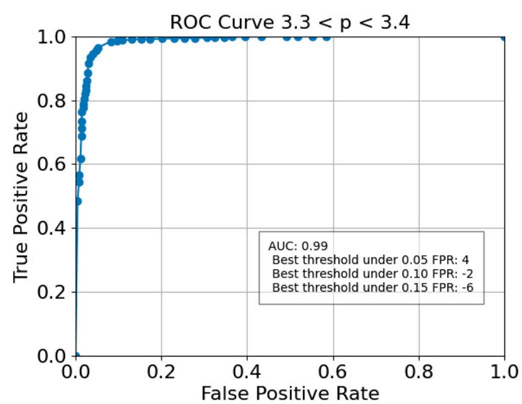

ROC Curve 2.5 < p < 2.6
True Positive Rate
False Positive Rate
AUC: 0.99
Best threshold under 0.05 FPR: 2
Best threshold under 0.10 FPR: -2
Best threshold under 0.15 FPR: -8

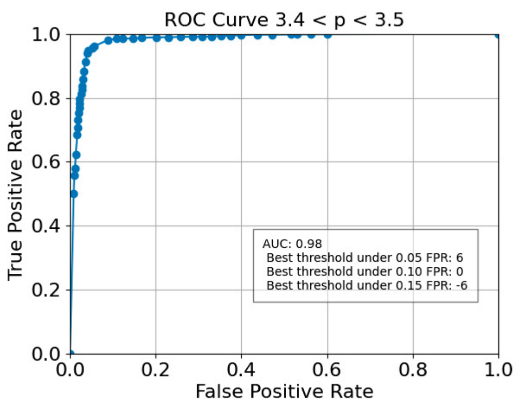

ROC Curve 2.6 < p < 2.7
AUC: 0.99
Best threshold under 0.05 FPR: 2
Best threshold under 0.10 FPR: -2
Best threshold under 0.15 FPR: -6
ROC Curve 2.7 < p < 2.8
AUC: 0.98
Best threshold under 0.05 FPR: 2
Best threshold under 0.10 FPR: -2
Best threshold under 0.15 FPR: -6
ROC Curve 2.8 < p < 2.9
AUC: 0.98
Best threshold under 0.05 FPR: 2
Best threshold under 0.10 FPR: -2
Best threshold under 0.15 FPR: -8
ROC Curve 2.9 < p < 3.0
AUC: 0.99
Best threshold under 0.05 FPR: 2
Best threshold under 0.10 FPR: -2
Best threshold under 0.15 FPR: -6
True Positive Rate
False Positive Rate

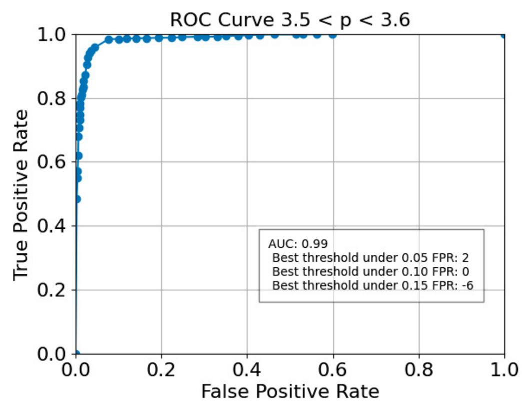

ROC Curve 3.0 < p < 3.1
True Positive Rate
False Positive Rate
AUC: 0.98
Best threshold under 0.05 FPR: 2
Best threshold under 0.10 FPR: -2
Best threshold under 0.15 FPR: -8

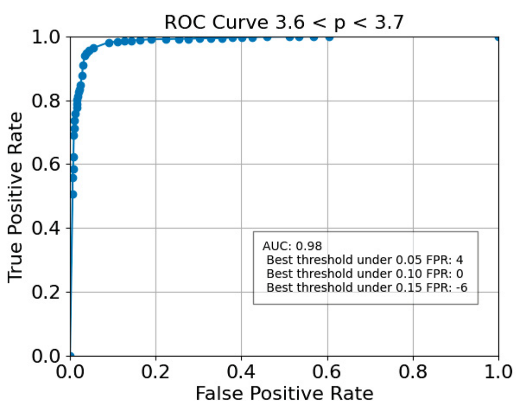

ROC Curve 3.1 < p < 3.2
True Positive Rate
False Positive Rate
AUC: 0.98
Best threshold under 0.05 FPR: 4
Best threshold under 0.10 FPR: 0
Best threshold under 0.15 FPR: -4

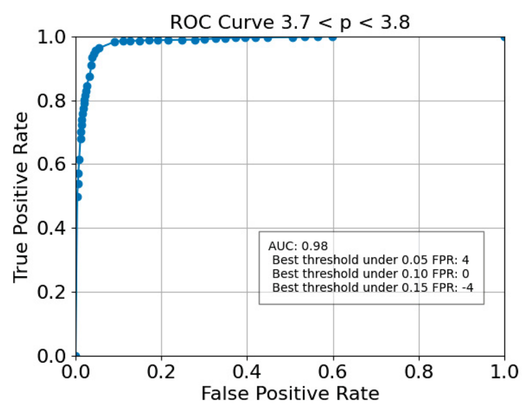

ROC Curve 3.2 < p < 3.3
True Positive Rate
False Positive Rate
AUC: 0.99
Best threshold under 0.05 FPR: 4
Best threshold under 0.10 FPR: -2
Best threshold under 0.15 FPR: -6

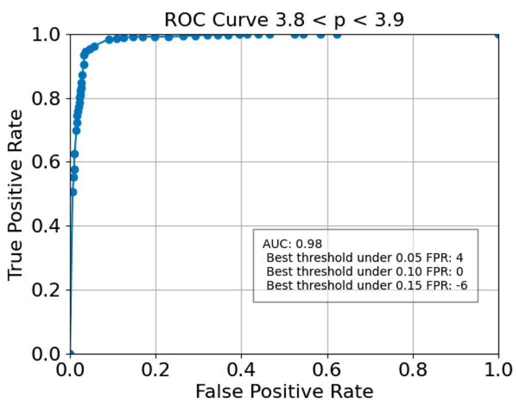

ROC Curve 3.3 < p < 3.4
True Positive Rate
False Positive Rate
AUC: 0.99
Best threshold under 0.05 FPR: 4
Best threshold under 0.10 FPR: -2
Best threshold under 0.15 FPR: -6

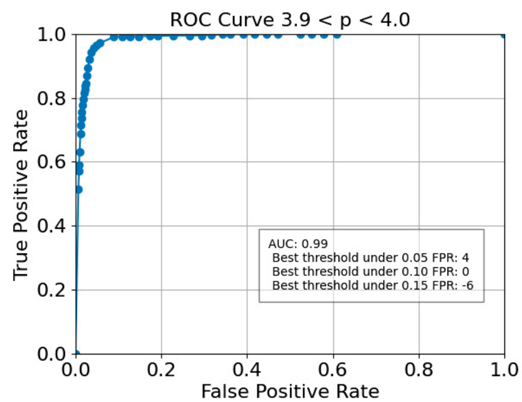

ROC Curve 3.4 < p < 3.5
True Positive Rate
False Positive Rate
AUC: 0.98
Best threshold under 0.05 FPR: 6
Best threshold under 0.10 FPR: 0
Best threshold under 0.15 FPR: -6

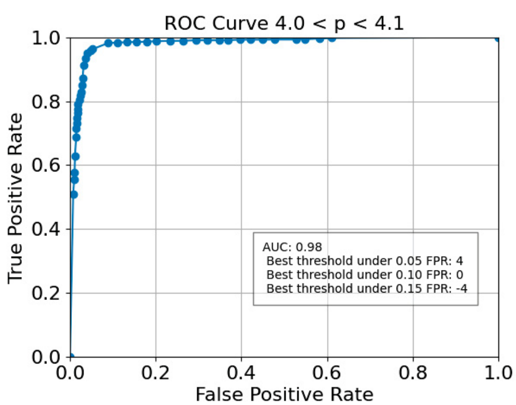

ROC Curve 3.5 < p < 3.6
True Positive Rate
False Positive Rate
AUC: 0.99
Best threshold under 0.05 FPR: 2
Best threshold under 0.10 FPR: 0
Best threshold under 0.15 FPR: -6

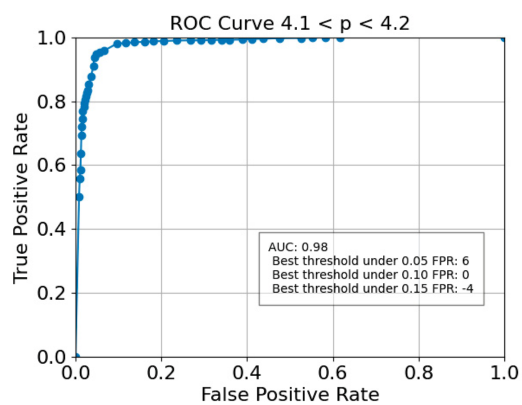

ROC Curve 3.6 < p < 3.7
True Positive Rate
False Positive Rate
AUC: 0.98
Best threshold under 0.05 FPR: 4
Best threshold under 0.10 FPR: 0
Best threshold under 0.15 FPR: -6

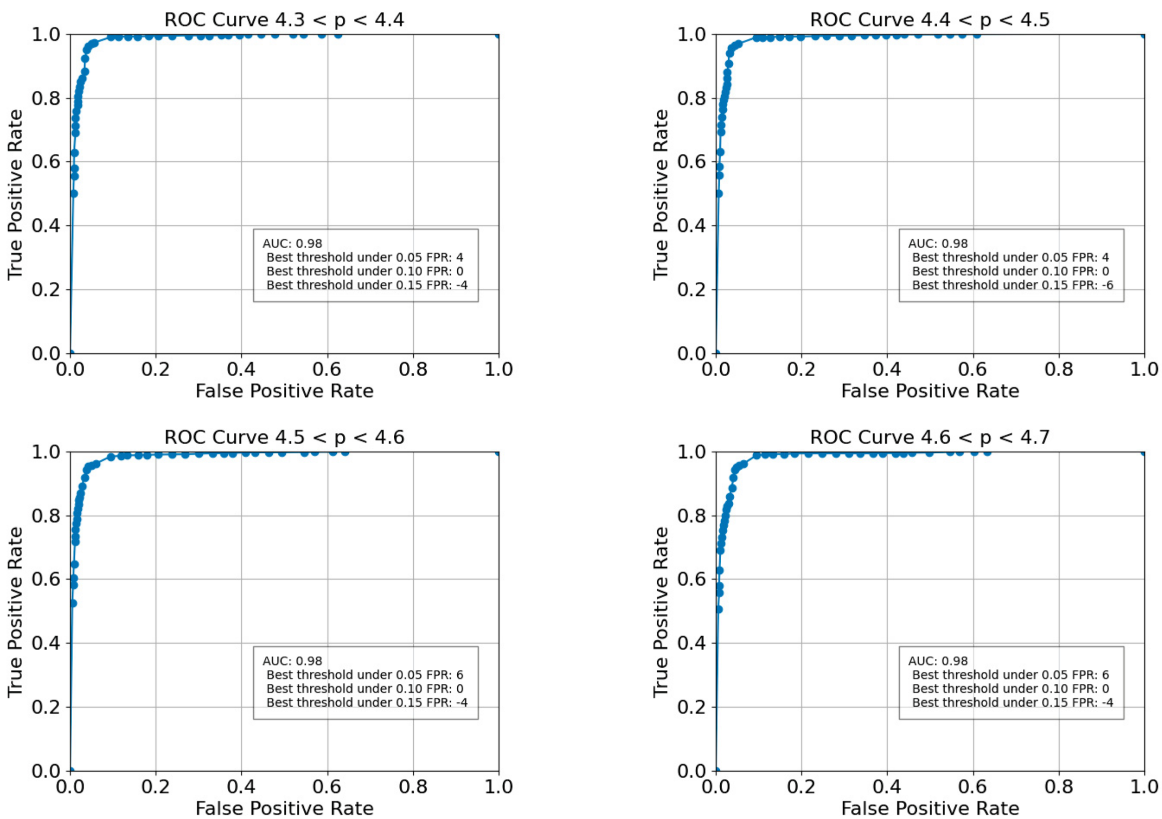

ROC Curve 3.7 < p < 3.8
True Positive Rate
False Positive Rate
AUC: 0.98
Best threshold under 0.05 FPR: 4
Best threshold under 0.10 FPR: 0
Best threshold under 0.15 FPR: -4

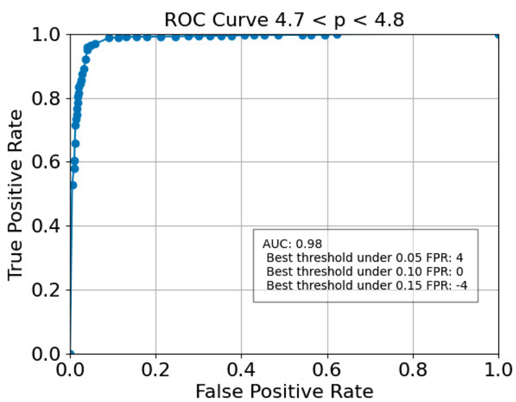

ROC Curve 3.8 < p < 3.9
AUC: 0.98
Best threshold under 0.05 FPR: 4
Best threshold under 0.10 FPR: 0
Best threshold under 0.15 FPR: -6
True Positive Rate
False Positive Rate

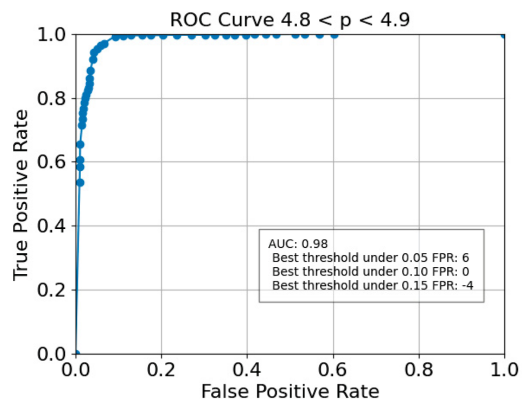

ROC Curve 3.9 < p < 4.0
AUC: 0.99
Best threshold under 0.05 FPR: 4
Best threshold under 0.10 FPR: 0
Best threshold under 0.15 FPR: -6
True Positive Rate
False Positive Rate

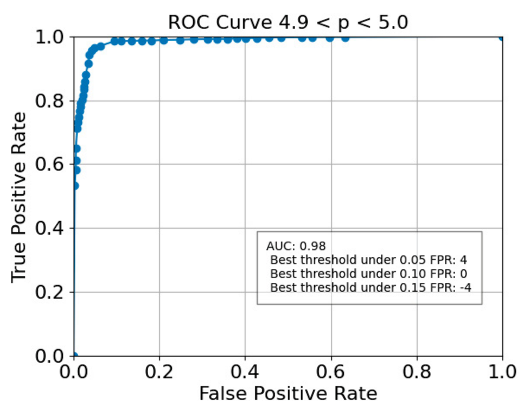

ROC Curve 4.0 < p < 4.1
AUC: 0.98
Best threshold under 0.05 FPR: 4
Best threshold under 0.10 FPR: 0
Best threshold under 0.15 FPR: -4
True Positive Rate
False Positive Rate


ROC Curve 4.1 < p < 4.2
AUC: 0.98
Best threshold under 0.05 FPR: 6
Best threshold under 0.10 FPR: 0
Best threshold under 0.15 FPR: -4
True Positive Rate
False Positive Rate


ROC Curve 4.3 < p < 4.4
True Positive Rate
False Positive Rate
AUC: 0.98
Best threshold under 0.05 FPR: 4
Best threshold under 0.10 FPR: 0
Best threshold under 0.15 FPR: -4
ROC Curve 4.4 < p < 4.5
AUC: 0.98
Best threshold under 0.05 FPR: 4
Best threshold under 0.10 FPR: 0
Best threshold under 0.15 FPR: -6
ROC Curve 4.5 < p < 4.6
AUC: 0.98
Best threshold under 0.05 FPR: 6
Best threshold under 0.10 FPR: 0
Best threshold under 0.15 FPR: -4
ROC Curve 4.6 < p < 4.7
AUC: 0.98
Best threshold under 0.05 FPR: 6
Best threshold under 0.10 FPR: 0
Best threshold under 0.15 FPR: -4

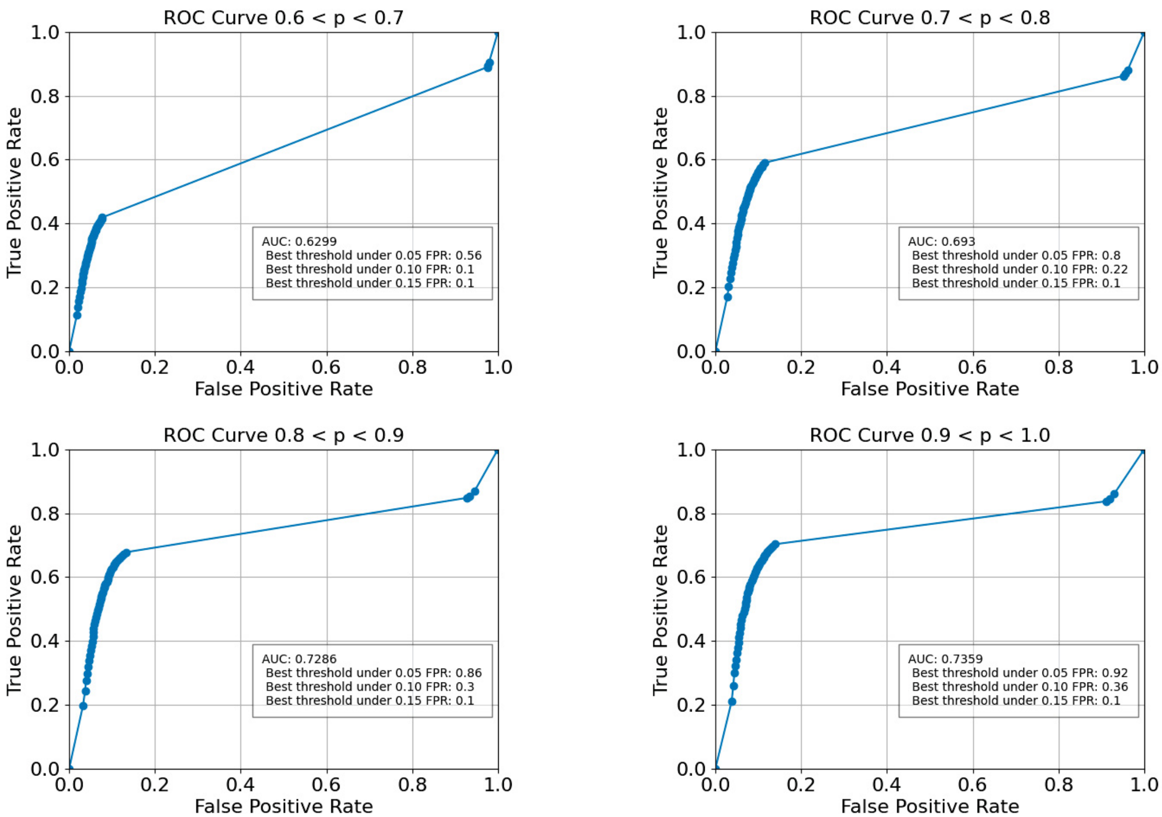

ROC Curve 4.7 < p < 4.8
True Positive Rate
False Positive Rate
AUC: 0.98
Best threshold under 0.05 FPR: 4
Best threshold under 0.10 FPR: 0
Best threshold under 0.15 FPR: -4

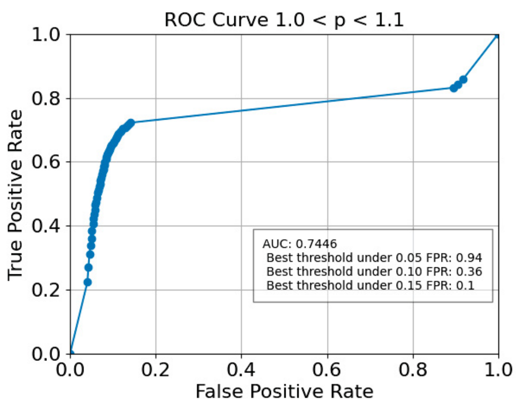

ROC Curve 4.8 < p < 4.9
True Positive Rate
False Positive Rate
AUC: 0.98
Best threshold under 0.05 FPR: 6
Best threshold under 0.10 FPR: 0
Best threshold under 0.15 FPR: -4

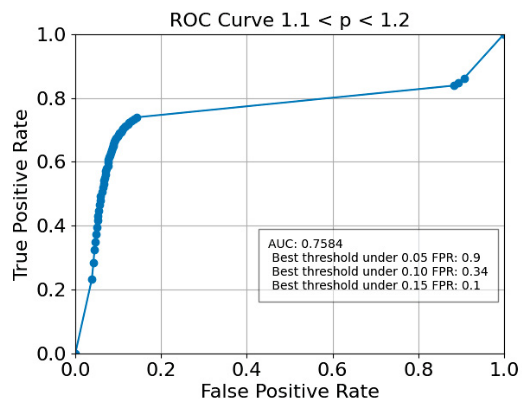

ROC Curve 4.9 < p < 5.0
True Positive Rate
False Positive Rate
AUC: 0.98
Best threshold under 0.05 FPR: 4
Best threshold under 0.10 FPR: 0
Best threshold under 0.15 FPR: -4

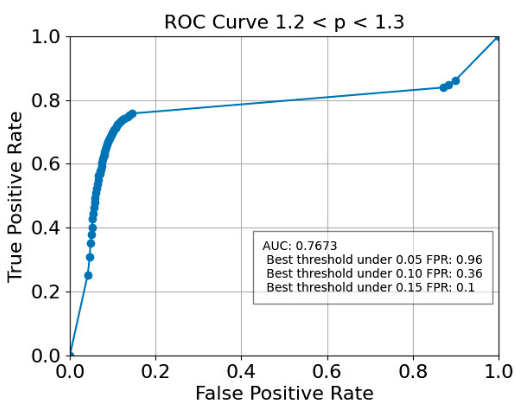
B.2 FBDT ROC Curves

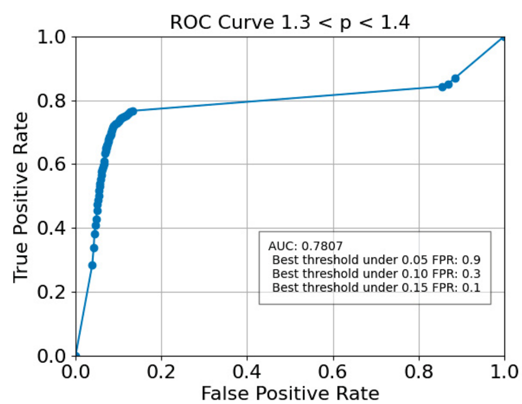
Full set of FBDT ROC curves generated for the momentum range $0.6 < p < 5.0\,\mathrm{GeV}/c$.

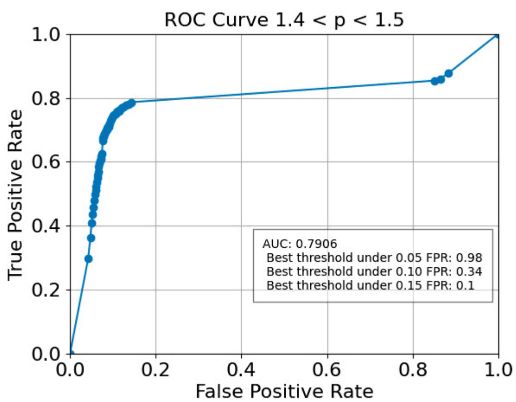


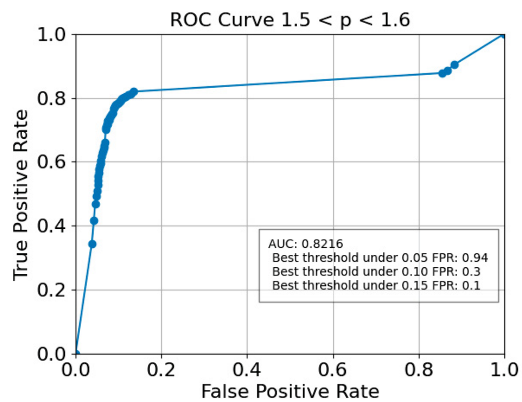

ROC Curve 1.0 < p < 1.1
True Positive Rate
False Positive Rate
AUC: 0.7446
Best threshold under 0.05 FPR: 0.94
Best threshold under 0.10 FPR: 0.36
Best threshold under 0.15 FPR: 0.1

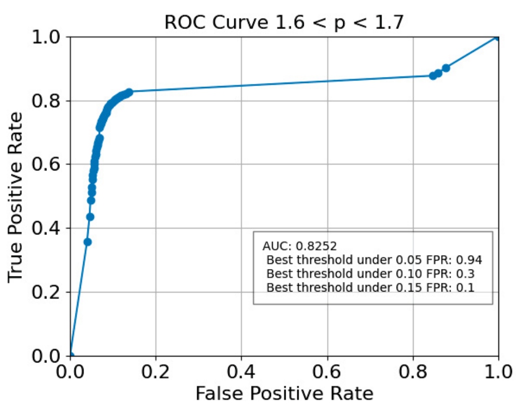

ROC Curve 1.1 < p < 1.2
True Positive Rate
False Positive Rate
AUC: 0.7584
Best threshold under 0.05 FPR: 0.9
Best threshold under 0.10 FPR: 0.34
Best threshold under 0.15 FPR: 0.1

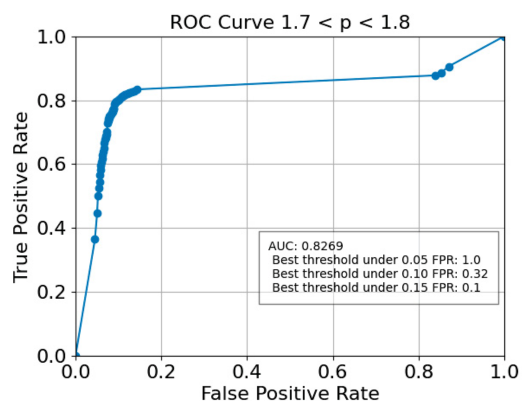

ROC Curve 1.2 < p < 1.3
True Positive Rate
False Positive Rate
AUC: 0.7673
Best threshold under 0.05 FPR: 0.96
Best threshold under 0.10 FPR: 0.36
Best threshold under 0.15 FPR: 0.1

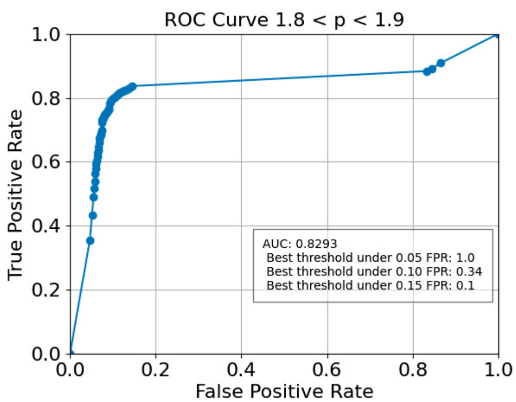

ROC Curve 1.3 < p < 1.4
True Positive Rate
False Positive Rate
AUC: 0.7807
Best threshold under 0.05 FPR: 0.9
Best threshold under 0.10 FPR: 0.3
Best threshold under 0.15 FPR: 0.1

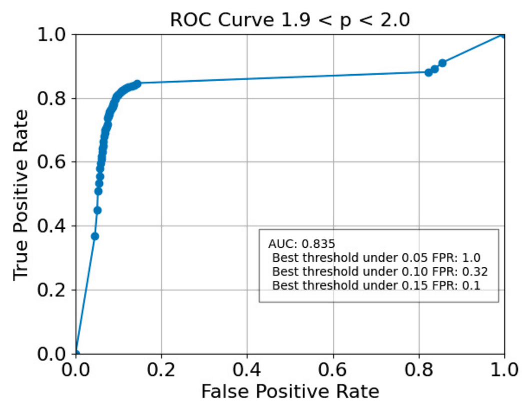

ROC Curve 1.4 < p < 1.5
True Positive Rate
False Positive Rate
AUC: 0.7906
Best threshold under 0.05 FPR: 0.98
Best threshold under 0.10 FPR: 0.34
Best threshold under 0.15 FPR: 0.1

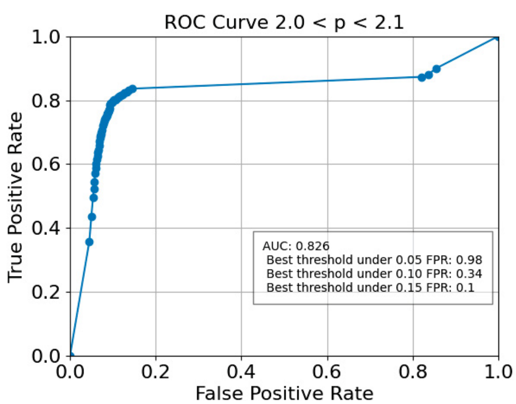

ROC Curve 1.5 < p < 1.6
True Positive Rate
False Positive Rate
AUC: 0.8216
Best threshold under 0.05 FPR: 0.94
Best threshold under 0.10 FPR: 0.3
Best threshold under 0.15 FPR: 0.1

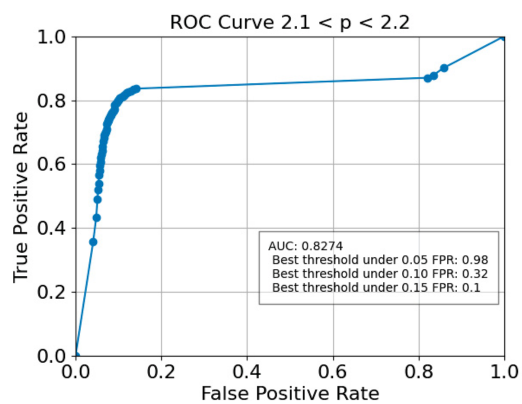

ROC Curve 1.6 < p < 1.7
True Positive Rate
False Positive Rate
AUC: 0.8252
Best threshold under 0.05 FPR: 0.94
Best threshold under 0.10 FPR: 0.3
Best threshold under 0.15 FPR: 0.1

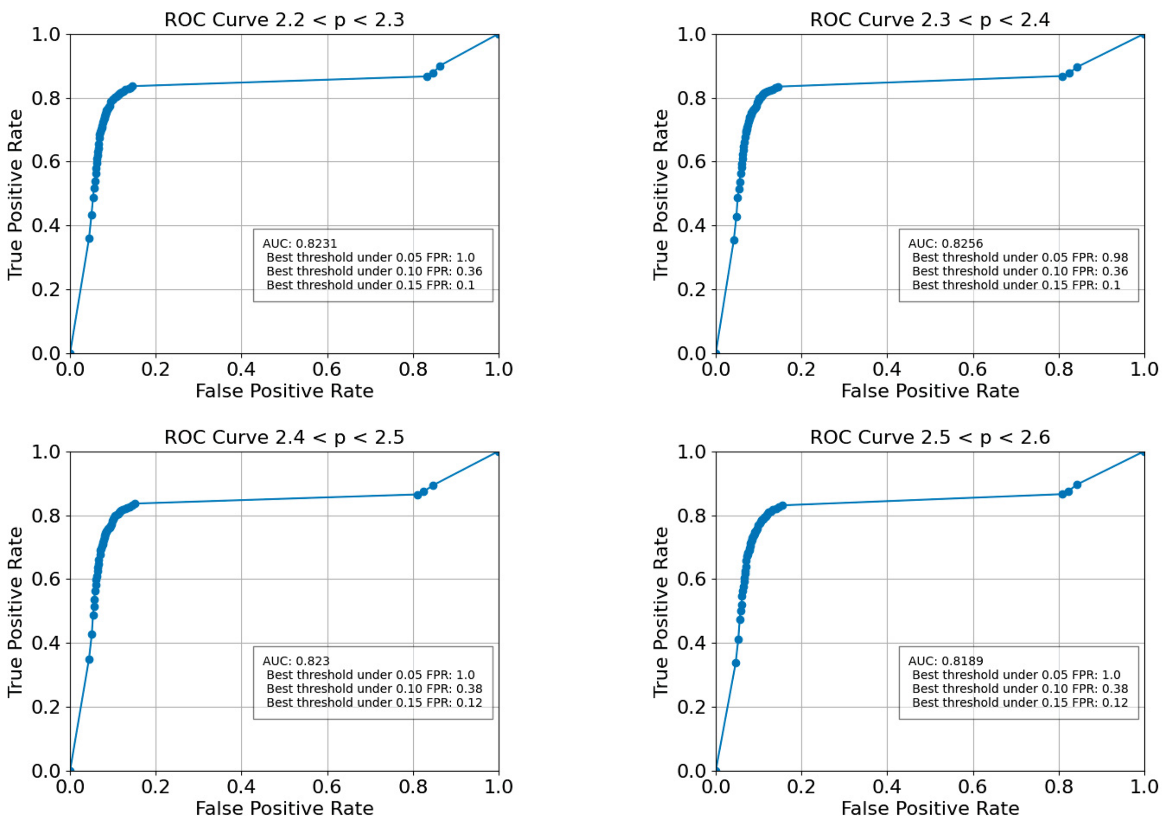

ROC Curve 1.7 < p < 1.8
True Positive Rate
False Positive Rate
AUC: 0.8269
Best threshold under 0.05 FPR: 1.0
Best threshold under 0.10 FPR: 0.32
Best threshold under 0.15 FPR: 0.1

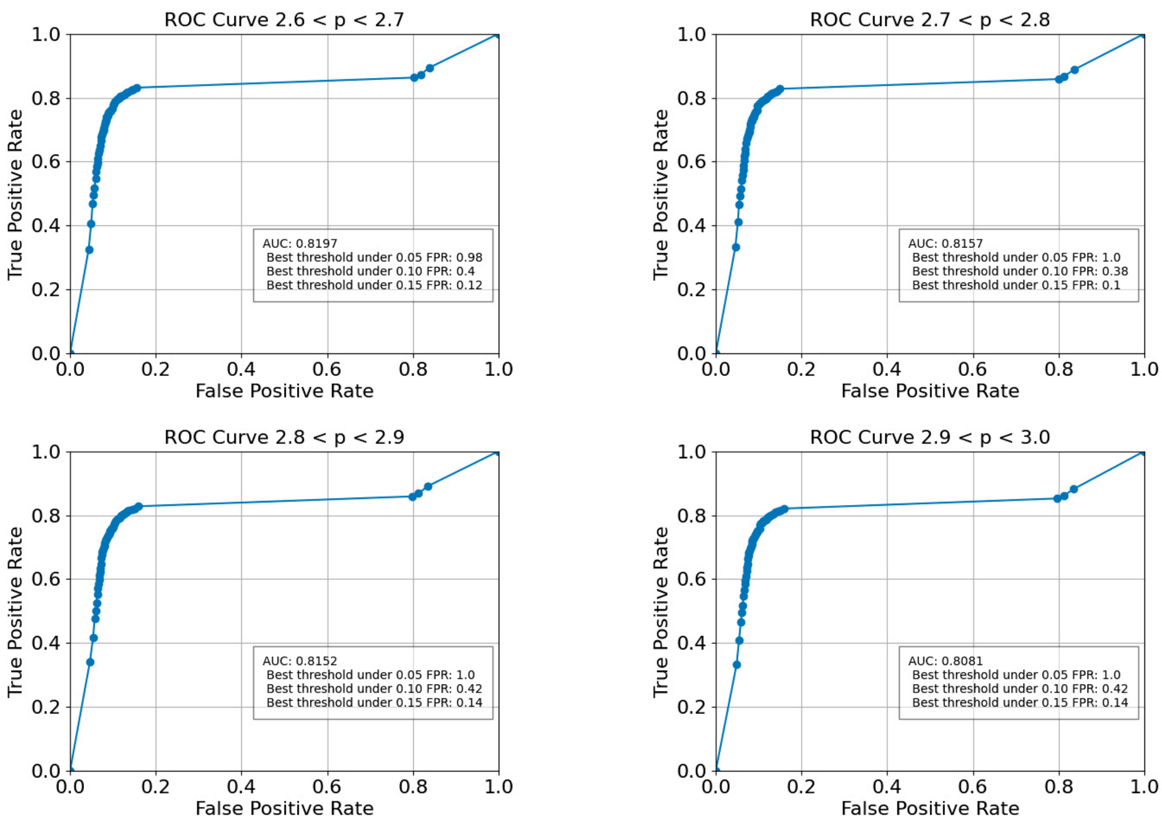

ROC Curve 1.8 < p < 1.9
AUC: 0.8293
Best threshold under 0.05 FPR: 1.0
Best threshold under 0.10 FPR: 0.34
Best threshold under 0.15 FPR: 0.1
True Positive Rate
False Positive Rate

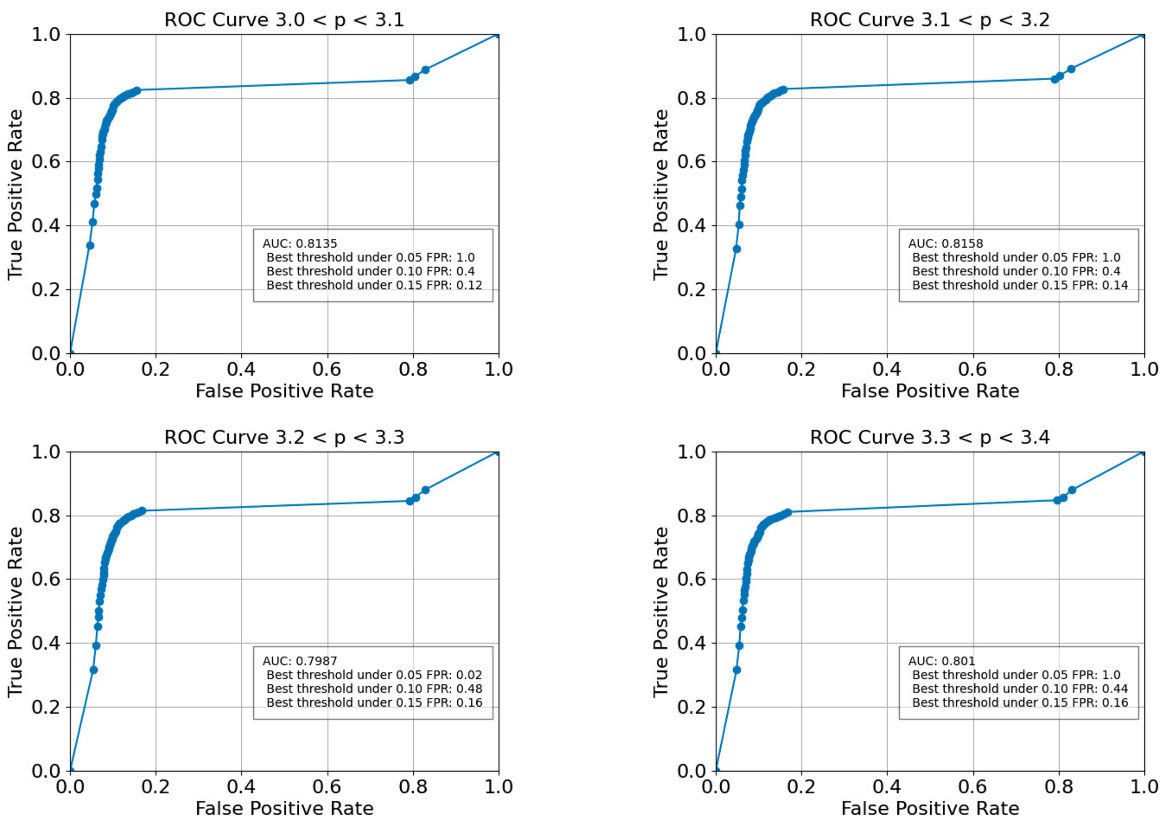

ROC Curve 1.9 < p < 2.0
AUC: 0.835
Best threshold under 0.05 FPR: 1.0
Best threshold under 0.10 FPR: 0.32
Best threshold under 0.15 FPR: 0.1
True Positive Rate
False Positive Rate

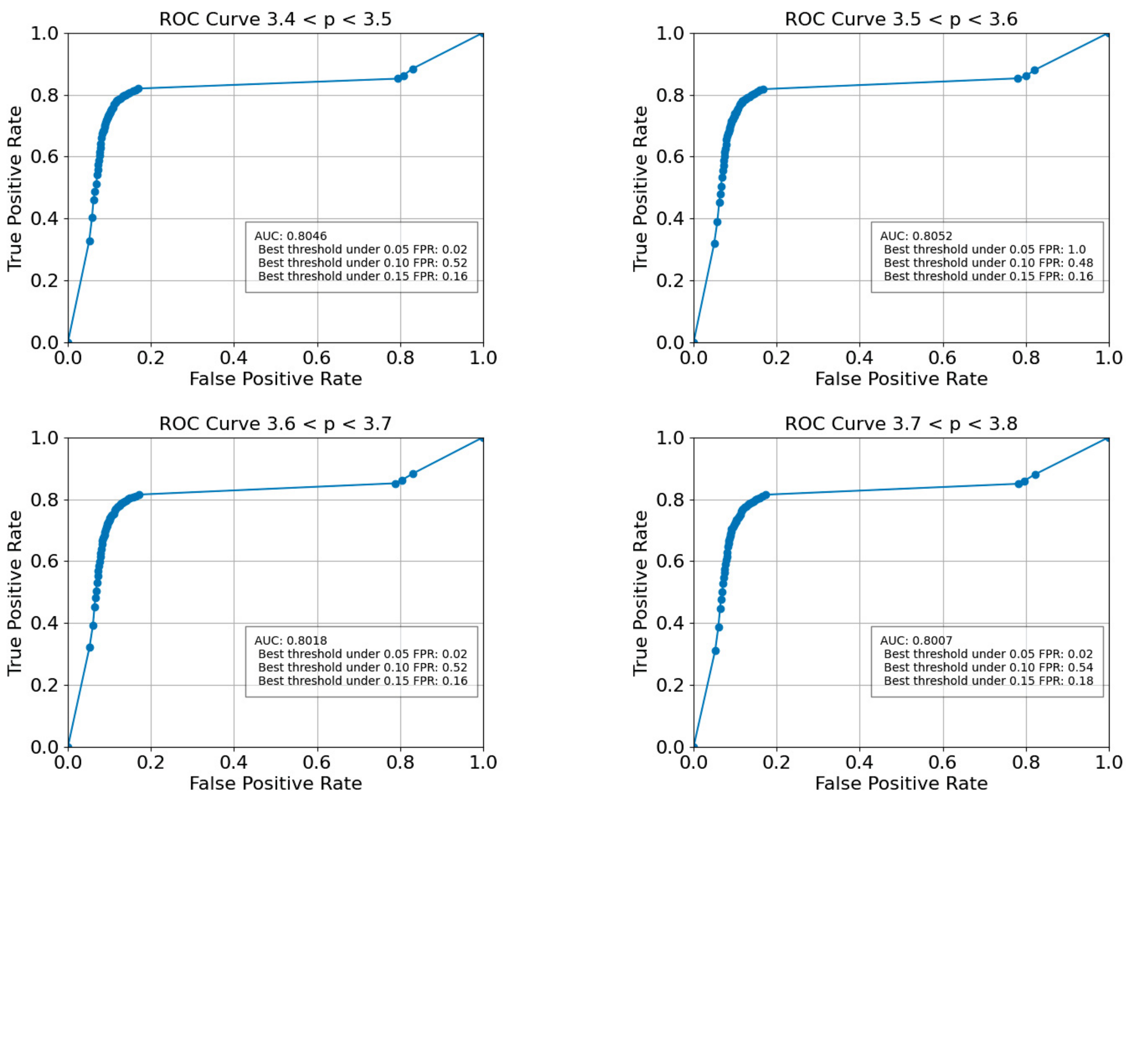

ROC Curve 2.0 < p < 2.1
AUC: 0.826
Best threshold under 0.05 FPR: 0.98
Best threshold under 0.10 FPR: 0.34
Best threshold under 0.15 FPR: 0.1
True Positive Rate
False Positive Rate

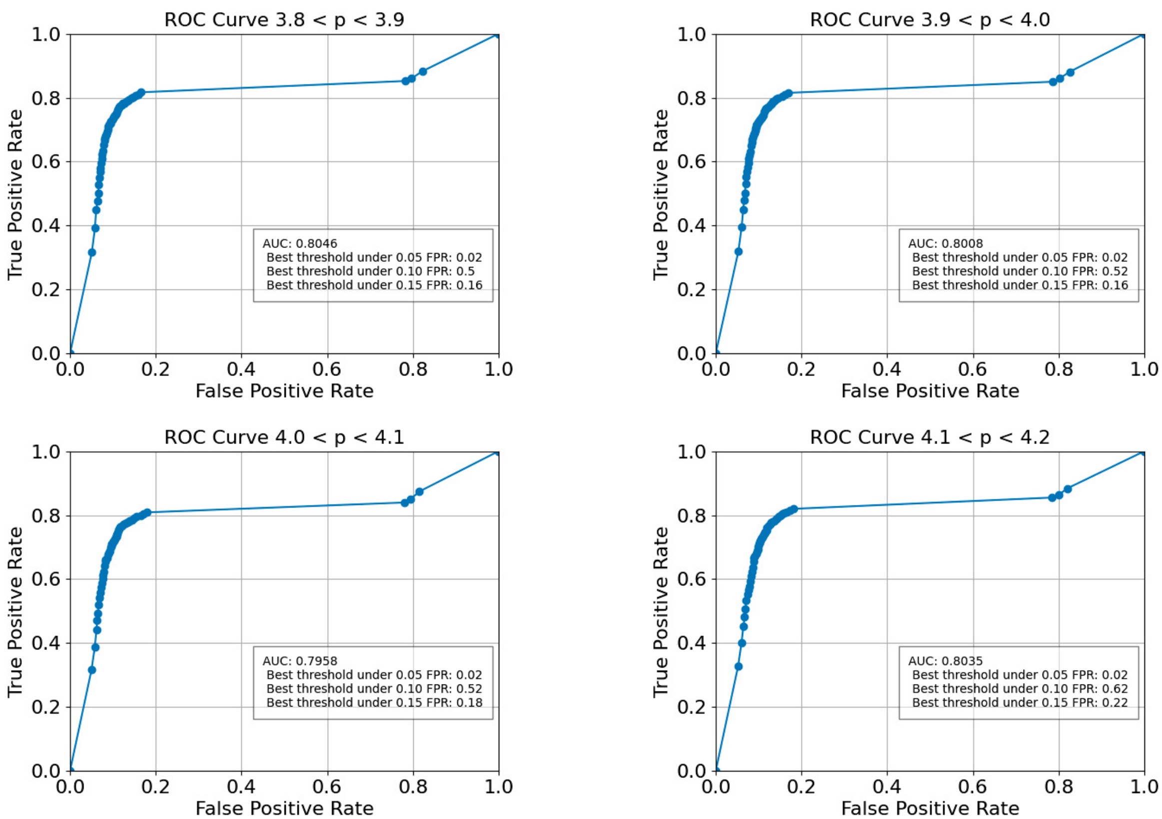

ROC Curve 2.1 < p < 2.2
AUC: 0.8274
Best threshold under 0.05 FPR: 0.98
Best threshold under 0.10 FPR: 0.32
Best threshold under 0.15 FPR: 0.1
True Positive Rate
False Positive Rate

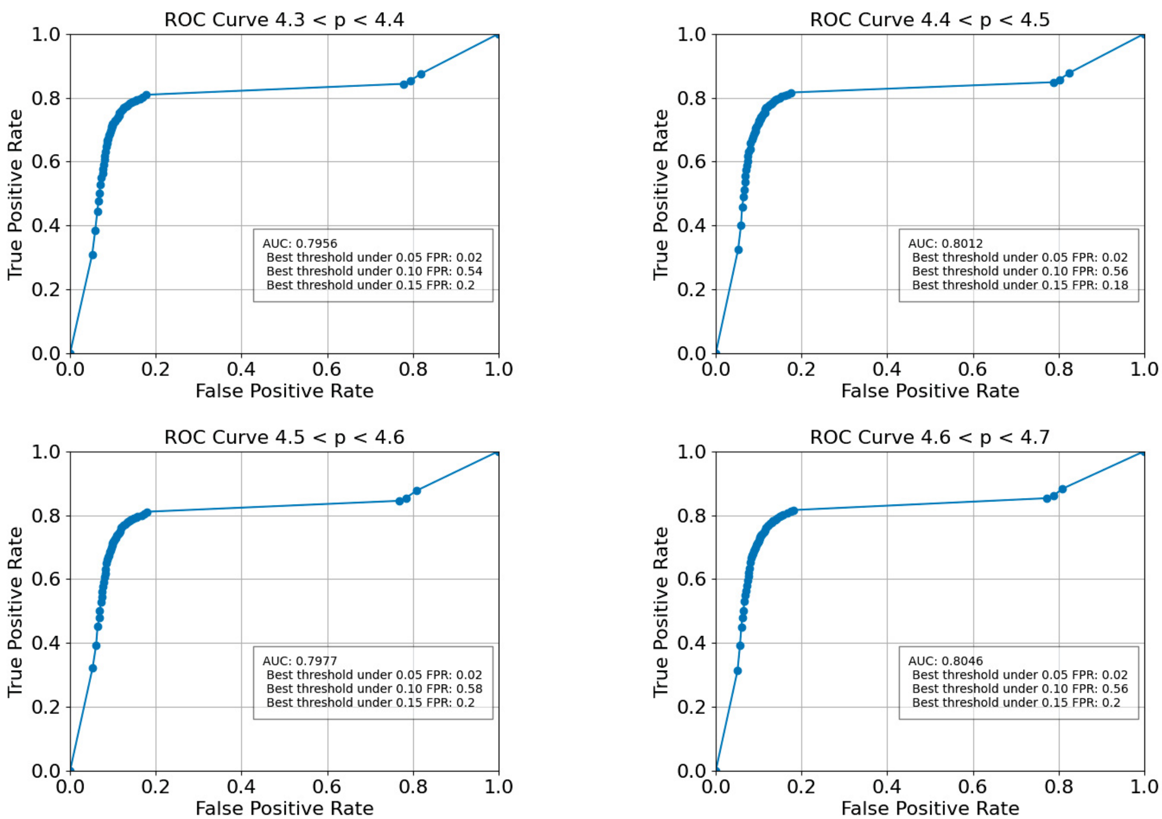

ROC Curve 2.2 < p < 2.3
AUC: 0.8231
Best threshold under 0.05 FPR: 1.0
Best threshold under 0.10 FPR: 0.36
Best threshold under 0.15 FPR: 0.1
ROC Curve 2.3 < p < 2.4
AUC: 0.8256
Best threshold under 0.05 FPR: 0.98
Best threshold under 0.10 FPR: 0.36
Best threshold under 0.15 FPR: 0.1
ROC Curve 2.4 < p < 2.5
AUC: 0.823
Best threshold under 0.05 FPR: 1.0
Best threshold under 0.10 FPR: 0.38
Best threshold under 0.15 FPR: 0.12
ROC Curve 2.5 < p < 2.6
AUC: 0.8189
Best threshold under 0.05 FPR: 1.0
Best threshold under 0.10 FPR: 0.38
Best threshold under 0.15 FPR: 0.12
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

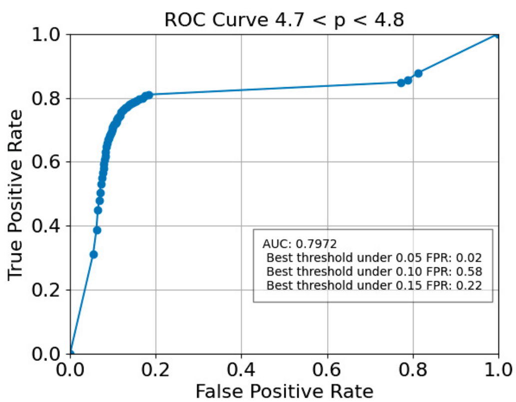

ROC Curve 2.6 < p < 2.7
AUC: 0.8197
Best threshold under 0.05 FPR: 0.98
Best threshold under 0.10 FPR: 0.4
Best threshold under 0.15 FPR: 0.12
ROC Curve 2.7 < p < 2.8
AUC: 0.8157
Best threshold under 0.05 FPR: 1.0
Best threshold under 0.10 FPR: 0.38
Best threshold under 0.15 FPR: 0.1
ROC Curve 2.8 < p < 2.9
AUC: 0.8152
Best threshold under 0.05 FPR: 1.0
Best threshold under 0.10 FPR: 0.42
Best threshold under 0.15 FPR: 0.14
ROC Curve 2.9 < p < 3.0
AUC: 0.8081
Best threshold under 0.05 FPR: 1.0
Best threshold under 0.10 FPR: 0.42
Best threshold under 0.15 FPR: 0.14
True Positive Rate
False Positive Rate

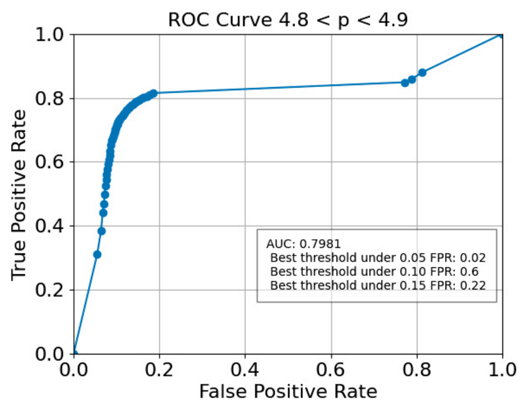

ROC Curve 3.0 < p < 3.1
AUC: 0.8135
Best threshold under 0.05 FPR: 1.0
Best threshold under 0.10 FPR: 0.4
Best threshold under 0.15 FPR: 0.12
ROC Curve 3.1 < p < 3.2
AUC: 0.8158
Best threshold under 0.05 FPR: 1.0
Best threshold under 0.10 FPR: 0.4
Best threshold under 0.15 FPR: 0.14
ROC Curve 3.2 < p < 3.3
AUC: 0.7987
Best threshold under 0.05 FPR: 0.02
Best threshold under 0.10 FPR: 0.48
Best threshold under 0.15 FPR: 0.16
ROC Curve 3.3 < p < 3.4
AUC: 0.801
Best threshold under 0.05 FPR: 1.0
Best threshold under 0.10 FPR: 0.44
Best threshold under 0.15 FPR: 0.16
True Positive Rate
False Positive Rate

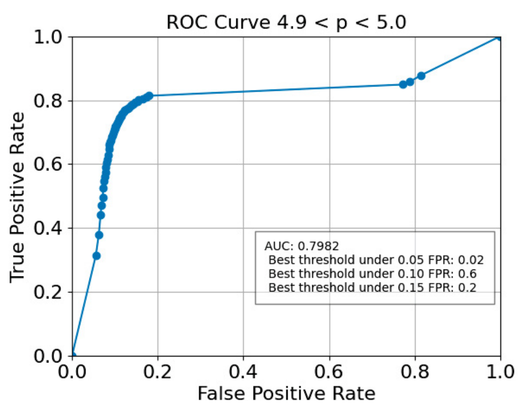

ROC Curve 3.4 < p < 3.5
True Positive Rate
False Positive Rate
AUC: 0.8046
Best threshold under 0.05 FPR: 0.02
Best threshold under 0.10 FPR: 0.52
Best threshold under 0.15 FPR: 0.16
ROC Curve 3.5 < p < 3.6
True Positive Rate
False Positive Rate
AUC: 0.8052
Best threshold under 0.05 FPR: 1.0
Best threshold under 0.10 FPR: 0.48
Best threshold under 0.15 FPR: 0.16
ROC Curve 3.6 < p < 3.7
True Positive Rate
False Positive Rate
AUC: 0.8018
Best threshold under 0.05 FPR: 0.02
Best threshold under 0.10 FPR: 0.52
Best threshold under 0.15 FPR: 0.16
ROC Curve 3.7 < p < 3.8
True Positive Rate
False Positive Rate
AUC: 0.8007
Best threshold under 0.05 FPR: 0.02
Best threshold under 0.10 FPR: 0.54
Best threshold under 0.15 FPR: 0.18


ROC Curve 3.8 < p < 3.9
True Positive Rate
False Positive Rate
AUC: 0.8046
Best threshold under 0.05 FPR: 0.02
Best threshold under 0.10 FPR: 0.5
Best threshold under 0.15 FPR: 0.16
ROC Curve 3.9 < p < 4.0
True Positive Rate
False Positive Rate
AUC: 0.8008
Best threshold under 0.05 FPR: 0.02
Best threshold under 0.10 FPR: 0.52
Best threshold under 0.15 FPR: 0.16
ROC Curve 4.0 < p < 4.1
True Positive Rate
False Positive Rate
AUC: 0.7958
Best threshold under 0.05 FPR: 0.02
Best threshold under 0.10 FPR: 0.52
Best threshold under 0.15 FPR: 0.18
ROC Curve 4.1 < p < 4.2
True Positive Rate
False Positive Rate
AUC: 0.8035
Best threshold under 0.05 FPR: 0.02
Best threshold under 0.10 FPR: 0.62
Best threshold under 0.15 FPR: 0.22


ROC Curve 4.3 < p < 4.4
AUC: 0.7956
Best threshold under 0.05 FPR: 0.02
Best threshold under 0.10 FPR: 0.54
Best threshold under 0.15 FPR: 0.2
ROC Curve 4.4 < p < 4.5
AUC: 0.8012
Best threshold under 0.05 FPR: 0.02
Best threshold under 0.10 FPR: 0.56
Best threshold under 0.15 FPR: 0.18
ROC Curve 4.5 < p < 4.6
AUC: 0.7977
Best threshold under 0.05 FPR: 0.02
Best threshold under 0.10 FPR: 0.58
Best threshold under 0.15 FPR: 0.2
ROC Curve 4.6 < p < 4.7
AUC: 0.8046
Best threshold under 0.05 FPR: 0.02
Best threshold under 0.10 FPR: 0.56
Best threshold under 0.15 FPR: 0.2
True Positive Rate
False Positive Rate


ROC Curve 4.7 < p < 4.8
True Positive Rate
False Positive Rate
AUC: 0.7972
Best threshold under 0.05 FPR: 0.02
Best threshold under 0.10 FPR: 0.58
Best threshold under 0.15 FPR: 0.22


ROC Curve 4.8 < p < 4.9
True Positive Rate
False Positive Rate
AUC: 0.7981
Best threshold under 0.05 FPR: 0.02
Best threshold under 0.10 FPR: 0.6
Best threshold under 0.15 FPR: 0.22


ROC Curve 4.9 < p < 5.0
True Positive Rate
False Positive Rate
AUC: 0.7982
Best threshold under 0.05 FPR: 0.02
Best threshold under 0.10 FPR: 0.6
Best threshold under 0.15 FPR: 0.2