# Optimization of Particle Identification


Bachelor's Thesis at the Faculty of Physics
of the
Ludwig-Maximilians-University Munich

submitted by
**Gordian Edenhofer**
born in Frankfurt am Main on the 13th July, 1997

Munich, Germany, 29th June, 2018

Supervisor:
**Prof. Dr. Thomas Kuhr**

# Optimierung der Teilchenidentifizierung


Bachelorarbeit an der Fakultät für Physik
der
Ludwig-Maximilians-Universität München

vorgelegt von
**Gordian Edenhofer**
geboren in Frankfurt am Main am 13. Juli 1997

München, den 29. Juni 2018

Betreuer:
**Prof. Dr. Thomas Kuhr**

# Abstract

This study aims at evaluating particle identification approaches.

First, the goodness of the detector yield is measured. Flaws are revealed and possible causes evaluated. In addition, current and future techniques for combining detector variables are outlined.

Next, a Bayesian approach to particle identification is discussed. It aims to produce probabilities of a track belonging to a particle species depending on the received signals. The process of obtaining conditional probabilities is described in detail. Furthermore, some extensions to the Bayesian approach are presented and evaluated. Flaws and benefits are compared using a generic decay.

Finally, a neural network is used to label particle tracks. Different methods to adapt the weights and various pre-processing steps are evaluated for a simple network. Hereby, tools from machine learning and statistics are discussed and their application is outlined. Last but not least, the accuracy of the network on a generic decay is determined and a comparison with the Bayesian approaches is performed.

# Contents

# Chapter 1

# Introduction

One of the most fundamental challenges in physics is to find the elemental constituents of matter. In this regard, the Standard Model has proven to be extremely useful. It postulates six quark species, three charged kinds of leptons, a neutral neutrino for every charged lepton species and four species of gauge bosons in addition to the Higgs boson. Besides its achievement in describing the very basic principals of interactions and its experimental predictions being very precise, there are significant flaws; namely the lack of a description of gravitation, matter & anti-matter asymmetry and the existence of dark matter and dark energy.

The violation of $CP$-symmetry in $B$ mesons had already been observed at the predecessor experiment Belle. The upgrade of the experiment, named Belle II is designed to search for new physics phenomena with a massive volume of data in the flavour sector. The volume is important as rare decays with only small corrections to the Standard Model are difficult to analyse properly and require high statistics. At the hardware frontier, the detector and accelerator system were updated. In addition, the software side required adaption to cope with the massive amount of data.

In order to recognize such small deviations in such large sets of data, a reliable particle identification is an essential requirement. Its role is to assign particle species labels to tracks which are identified by the detector system. By doing so, it allows further analysis of events to better focus on the particles relevant to them.

# Chapter 2

# Belle II

## 2.1 Experiment

The Belle II experiment is performed at the SuperKEKB accelerator located in Tsukuba, Japan. It is mainly designed to study B mesons. In the experiment, asymmetric electron-positron-beams are collided with a center of mass energy of $\sqrt{s} = 10.58$ GeV, exactly on the $\Upsilon(4S)$ resonance. The two beams, positrons at 4 GeV and electrons at 7 GeV, are focussed to a narrow crossing section. The additional boost in one direction is used to measure the $B$ meson lifetimes.

In comparison to the predecessor experiment Belle, the integrated luminosity will be 50 $ab^{-1}$ and hence 50 times higher. The instantaneous luminosity will be $8 \cdot 10^{35}\text{cm}^{-2}\text{s}^{-1}$ which represents a 40-fold increase.

## 2.2 Detector system

### 2.2.1 Overview

The Belle II detector system [1, 2, 3] is a composition of multiple detectors, each measuring a subset of a particle's properties. Its design is depicted in Figure 2.1. The inner three detectors – **PiX**el **D**etector (PXD), **S**ilicon **V**ertex **D**etector (SVD) and **C**entral **D**rift **C**hamber (CDC) – record the position of traversing charged particles. Hence they are also called tracking detectors. They are located in a homogeneous magnetic field of 1.5 T. The innermost detector is the PXD. Together with the SVD which surrounds the PXD, it is used to reconstruct decay vertices and identify tracks belonging to particles with low-momenta. The CDC measures the momentum and charge of particles via their curvature in the magnetic field. Next, the **T**ime **O**f **P**ropagation (TOP) counter ('Barrel PID')

and the **A**erogel **R**ing-**I**maging **CH**erenkov (ARICH) counter ('Endcap PID') are used to identify charged particles via their emission of Cherenkov radiation in the detector. However, there is no such installation for the backwards-facing endcap of the detector due to the asymmetric beams. The **E**lectromagnetic **CaL**orimeter (ECL) identifies photons and electrons. The outermost detector called $K_L^0/\mu$ (KLM) is used to identify kaons and muons.[1]

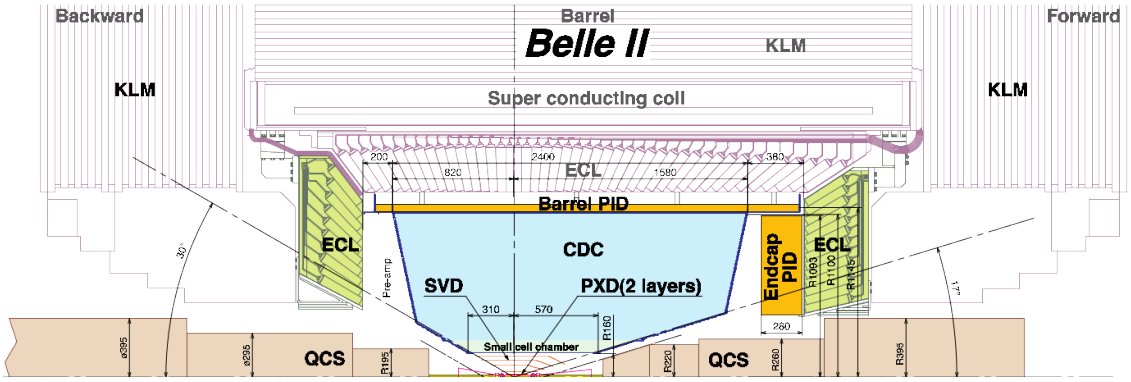



**Figure 2.1:** Side view of the upper half of the Belle II detector. Adapted from [1].

### 2.2.2 Silicon detectors

The PXD and SVD consist of tiny doped silicon chips which yield the location of electron-holes created by particles passing through them. The PXD detector uses small pixels while the SVD detector uses strips of detector material. Therefore, the PXD detector is able to further differentiate multiple simultaneous tracks while the SVD allows for a faster readout and is less prone to noise.

### 2.2.3 Central drift chamber

The CDC which surrounds the PXD and SVD, consists of a collection of field wires and sense wires located in a volume filled with gas. The sense wires are used to measure the current produced by electromagnetic showers. The latter is caused by particles passing through the gas. The wires are close to being parallel to the beampipe but have a slight twist. This allows the detector to not only have an excellent estimation of the transverse distance to the beam pipe but also provides information about the longitudinal position.

[1] If not specifically stated otherwise, the charge conjugate of a particle is implied.

### 2.2.4 Barrel and endcap PID

The Cherenkov effect is used to measure the velocity of particles in the TOP and ARICH detector. Charged particles which travel faster than the speed of light in the medium – Quartz in case of the TOP detector and aerogel for the ARICH detector – produce light. The velocity can be calculated by measuring the time of propagation and the angle of the emitted light.

### 2.2.5 Electromagnetic calorimeter

The main purpose of the ECL detector is to determine the energy of photons and electrons. Both particle species excite the medium and create electromagnetic showers. The light of the de-excitation can subsequently be measured.

### 2.2.6 $K_L^0/\mu$ detector

Last but not least, the KLM detector identifies kaons and muons which have passed through the previous layers of the system. When traversing this detector, the particles pass through plates serving as electrodes separated by layers of gas in between them. Ionized particles created by the incident particle are accelerated in this field and subsequently produce a spark picked up by the detector.

## 2.3 Interaction with matter

### 2.3.1 Charged particle interaction

Particles with non-zero charge mainly interact with the medium electromagnetically. In general, an interaction occurs either by scattering in the electric field of the atom, polarization of the medium, ionization or excitation. Besides, hadrons may scatter at the atom itself. Particles and their anti-particles additionally have the ability to annihilate.

The polarization of the medium causes Cherenkov radiation to be emitted. At velocities below the speed of light in the medium ($v < c/n$), only boundaries of the particle's velocity may be determined. However, at $v > c/n$ the Cherenkov effect can be observed. The effect occurs due to the information about the charge of the traversing particle not reaching the medium in front of it soon enough. Hence, the medium behind the particle is already aligned with the electric field while the medium in front is not. The result is an electromagnetic wave. The angle between the normal vector of the wave and the track of

the particle is given by

$$\cos(\Theta_c) = \frac{1}{v/c \cdot n} = \frac{1}{\beta \cdot n}. \tag{2.1}$$

This is the effect which the TOP and ARICH detectors exploit.

Particles with low energy traveling through a medium interact predominantly with atomic electrons. The average energy loss of a particle is described by the Bethe-Bloch formula. It is given by $\langle \mathrm{d}E/\mathrm{d}x \rangle \propto 1/\beta^2$ for velocities of up to about 90% of the speed of light and is minimal at $\beta\gamma \approx 4$. It describes the momentum dependency of the average energy loss. However, the actual shape of $\mathrm{d}E/\mathrm{d}x$ is modelled by the Landau distribution. Note that the initial assumptions needed for this formula are not met for electrons. This is because both participants of the interaction belong to the same species and have identical masses. A $\mathrm{d}E/\mathrm{d}x$-measurement is performed at the silicon detectors and the CDC.

The interaction with the electromagnetic field of the nucleus is the dominant cause for the energy loss of high-energy particles. Energy is radiated away via so called Bremsstrahlung. The leftover energy decreases exponentially with the distance traversed and is inversely proportional to the square root of the mass. Therefore it is mainly important for particles with a low mass, e.g., electrons. The radiation due to this effect is predominantly measured by the ECL.

### 2.3.2 Particle identification

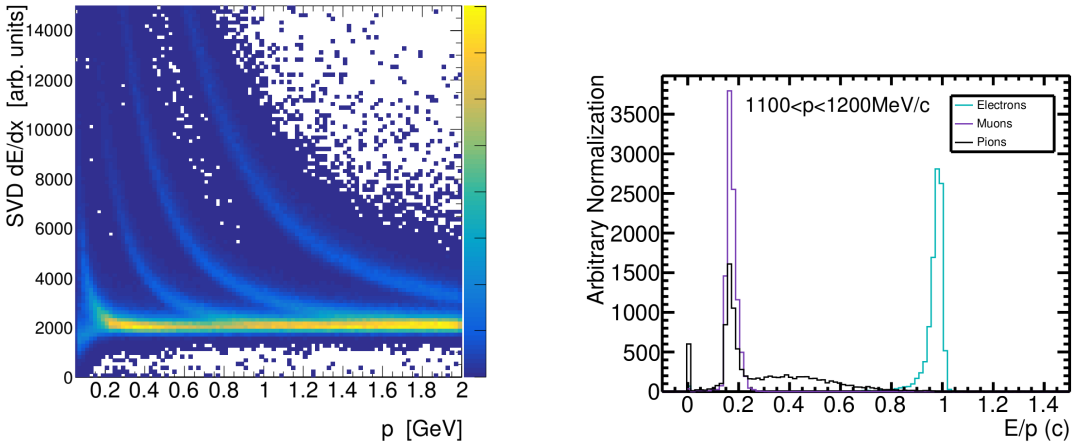
At Belle II the detector system differentiates among six long living particle species $K, \pi, e, \mu, p$ and deuteron.

The $\mathrm{d}E/\mathrm{d}x$-measurement from the silicon detectors and the CDC are one of the most useful measurements. Figure 2.2a showcases this for one of the tracking detectors for momenta below 1 GeV/c. Distinct patterns may be observed for various particle species below this momentum threshold. New tracks can now be assigned a likelihood of producing the measured detector signal given they belonging to a certain particle species. This is done by postulating a hypothesis for the loss of energy for each such species.

ARICH, TOP and CDC furthermore extend the identification and are able to differentiate among $K, \pi, p$ and deuteron but also contribute to $e$ and $\mu$ identification. They provide likelihoods for each signal given a particle hypothesis.

Further out the ECL detector provides a good separation of electrons from other charged particles above 1 GeV/c. It is able to do so via measuring $E/p$ of the shower. The detector response is provided by estimating the degree of agreement for different particle hypothesis with the signal. An exemplary $E/p$ curve is shown in Figure 2.2b. It demonstrates the observable difference for electrons compared to other particle species but also shows that no clear separation of pions and muons is possible.

The KLM detector provides a good separation between muons and non-muons and contributes to the discrimination in the form of different likelihoods as well.




**(a)** $\mathrm{d}E/\mathrm{d}x$ **(b)** $E/p$

**Figure 2.2:** Separation of different particle species for various simulated signals. Taken from [4].
**Figure a** shows the $\mathrm{d}E/\mathrm{d}x$ means as a function of momentum in the SVD with the color encoding the number of hits. In order to reduce outliers the lowest 5% and highest 25% of the measurements of each track are not used in the estimation.
**Figure b** shows the $E/p$ distribution for different particle species.

# Chapter 3

# Statistics for Particle Analysis

## 3.1 Classification functions

The main concepts to compare the goodness of identification methods which are used throughout the thesis are based on statistical classification functions. However, their use is not limited to physics, let alone particle physics, but can be found in all fields containing some form of (binary) classification problem. A classification function is a tool which separates elements which do not have the desired feature from those which have it. In the following examples the classifier assumes the role of a discriminator between kaons and non-kaons.

The most important classification functions are:

- **T**rue **P**ositive **R**ate (**TPR**): *proportion of accepted elements which are correct relative to all positives*

  Hence, in the example it is the ratio of identified kaons which actually are kaons in proportion to the number of kaons in the data.

- **T**rue **N**egative **R**ate or Specificity (**TNR**): *proportion of rejected elements which are incorrect relative to all negatives*

  It is the ratio of non-kaon particles being identified as non-kaons in proportion to the number of all non-kaon particles.

- **F**alse **P**ositive **R**ate (**FPR**): *proportion of accepted elements which are incorrect relative to all negatives*

  This rate represents the fraction of non-kaon particles identified as kaons over the number of all non-kaons.

- **F**alse **N**egative **R**ate (**FNR**): *proportion of rejected elements which are correct relative to all positives*

  It is the fraction of kaons classified as being non-kaons over the number of all non-kaons.

- **P**ositive **P**redicted **V**alue (**PPV**): *proportion of accepted elements which are correct relative to all accepted*

  The definition represents the fraction of kaons classified as such over the number of all tracks classified as kaons but not necessarily actually being a kaon.

## 3.2 Receiver operating characteristic curve

The **R**eceiver **O**perating **C**haracteristic (**ROC**) curve is the TPR plotted over the FPR. The values on the $x$- and $y$-axis go from zero to unity. Each point on the curve represents an applied selection criterion on the data or a so called *cut*.

A straight diagonal line connecting the point $(0,0)$ with $(1,1)$ would be the result of a classifier which is merely guessing the classes of two equally likely yields. A curve below this diagonal is worse than guessing and anything above is some degree of good. An optimal curve achieves a high TPR value at a very low FPR. Multiple methods can therefore be compared by assessing the value and the slope of each method's TPR in dependence on the FPR. Figure 3.1 visually underlines the above described relations.

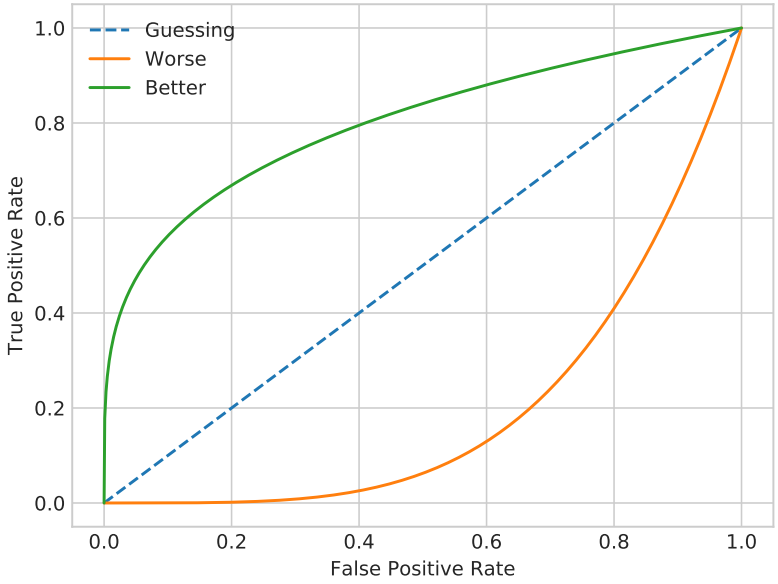



**Figure 3.1:** ROC curves for a binary classification problem with each outcome being equally likely.

Usually the points on the left are of most interest as they represent a selection with only few false elements contaminating the sample.

## 3.3 Identification efficiencies

The identification efficiency is defined as the proportion of correctly classified particles of a class relative to all of the available particles belonging to it. Hence, it directly represents the TPR. Both terms will be used as synonyms throughout the thesis.

The $\varepsilon_{PID}$-matrix is the confusion matrix normalized by row for an exclusive particle classification. The term 'exclusive' in this context denotes that each track is labeled with exactly one particle hypothesis. Such a classification can be achieved by, e.g., assigning the track the label of the highest identification variable. This idea is used throughout the analysis.

The values of the matrix are given by the fraction of particles $i$ classified as $j$ over the true abundance of particle $i$. Hence, its values are

$$\varepsilon_{ij} = \frac{N_{i \text{ classified as } j}}{A_{i \text{ true}}}. \tag{3.1}$$

The matrix has the shape of a $6 \times 6$ matrix when listing the confusion probabilities for all six particle species of interest:

$$\begin{pmatrix} \varepsilon_{KK} & \varepsilon_{K\pi} & \varepsilon_{Ke} & \varepsilon_{K\mu} & \varepsilon_{Kp} & \varepsilon_{Kd} \\ \varepsilon_{\pi K} & \varepsilon_{\pi\pi} & \varepsilon_{\pi e} & \varepsilon_{\pi\mu} & \varepsilon_{\pi p} & \varepsilon_{\pi d} \\ \varepsilon_{eK} & \varepsilon_{e\pi} & \varepsilon_{ee} & \varepsilon_{e\mu} & \varepsilon_{ep} & \varepsilon_{ed} \\ \varepsilon_{\mu K} & \varepsilon_{\mu\pi} & \varepsilon_{Ke} & \varepsilon_{K\mu} & \varepsilon_{Kp} & \varepsilon_{\mu d} \\ \varepsilon_{pK} & \varepsilon_{p\pi} & \varepsilon_{pe} & \varepsilon_{p\mu} & \varepsilon_{pp} & \varepsilon_{pd} \\ \varepsilon_{dK} & \varepsilon_{d\pi} & \varepsilon_{Ke} & \varepsilon_{K\mu} & \varepsilon_{Kp} & \varepsilon_{dd} \end{pmatrix}. \tag{3.2}$$

The definition generalizes to non-normalized matrices, e.g., resulting from non-exclusive cuts. Although, reading the matrix is less intuitive. Comparing matrices in this case becomes ambiguous as a particle might belong to multiple classes.

The diagonal of the matrix contains the identification efficiencies of each particle species. In general, its values should be close to unity while non-diagonal entries should vanish for a good classification approach. The efficiency of a particle classification is always normalized by the abundance of the particle and as such each row may have a different normalization. This is especially important when calculating the overall efficiency which is the fraction of all correctly classified tracks relative to all available tracks. In this case, each efficiency on the diagonal has to be weighted with the abundance of the particle.

## 3.4 Likelihood

### 3.4.1 Likelihood ratio

The ratio of likelihoods is commonly used for comparisons of the goodness of different models. For each hypothesis a likelihood of event $\boldsymbol{x}$ occurring is calculated under the assumption the hypothesis is indeed true. The ratio of the likelihoods of two hypothesis $H_0$ and $H_1$

$$\frac{\mathcal{L}(\boldsymbol{x}|H_0)}{\mathcal{L}(\boldsymbol{x}|H_1)} \tag{3.3}$$

denotes how many times more likely the event $\boldsymbol{x}$ is under hypothesis $H_0$ compared to $H_1$.

However, the event $\boldsymbol{x}$ need not necessarily take the form of a simple one dimensional value. It may very well be a composition of, e.g., multiple detector responses. In case the components $x_i$ are independent from one another, the overall likelihood of $\boldsymbol{x}$ may be constructed by multiplying the separate likelihoods of each $x_i$. Hence, $\mathcal{L}(\boldsymbol{x}|H_0)$ is composed out of multiple likelihoods each assuming $H_0$ to be true:

$$\mathcal{L}(\boldsymbol{x}|H_0) = \prod_i \mathcal{L}_i(x_i|H_0). \tag{3.4}$$

In case of event $\boldsymbol{x}$ being a detector response, the likelihood $\mathcal{L}(\boldsymbol{x}|H_0)$ is the probability of measuring a signal given a particle hypothesis is true. Its value is constructed by multiplying the likelihoods of $\mathcal{L}_i(x_i|H_0)$ for each detector $i$.

### 3.4.2 Neyman-Pearson

The Neyman-Pearson lemma is useful for evaluating the goodness of separating two models which have no unknown parameters. It states that a test on the likelihood ratio has the highest probability of correctly rejecting the original hypothesis at a given significance level. In other words: A test on the likelihood ratio provides the highest purity at a given efficiency.

The purity of a selection is defined as the proportion of correctly classified particles relative to all the identified ones. Its definition is identical to the PPV and as such will be used synonymously throughout the thesis.

Hence, by plotting the purity over the likelihood ratio, a monotonically increasing function is to be expected. An idealized version of such a graph is depicted in Figure 3.2. Since the underlying data may not be assumed to be a continues stream, the likelihood ratio is binned as it better represents the actual expected shape.

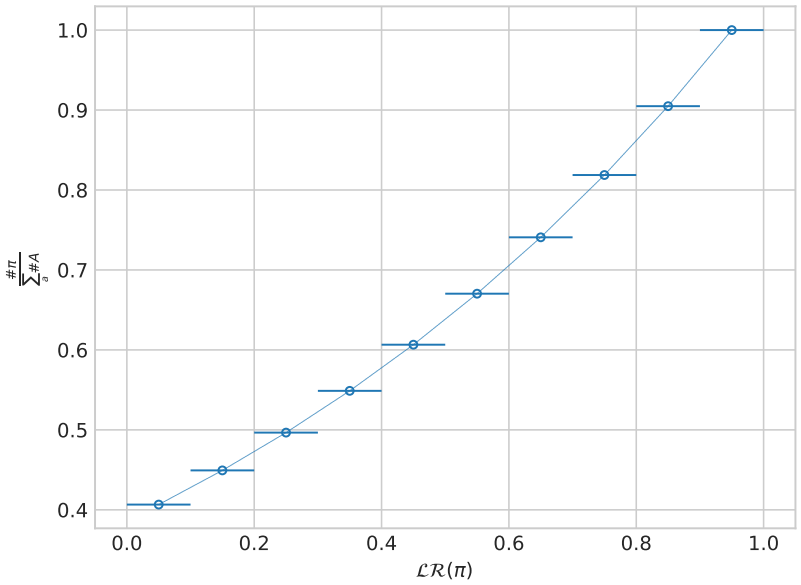



**Figure 3.2:** Visualization of a test on the likelihood ratio. A monotonically increasing function should be expected on the basis of the Neyman-Pearson lemma. The small horizontal lines indicate likelihood ratio bins, while the curve represents the overall trend. The pion purity and likelihood ratio is merely used to emphasize the connection to particle physics.

## 3.5 Neural network

An artificial neural network or simply neural network is a class of algorithms inspired by the central nervous system of biological beings. Instead of electrical signals passing from neuron to neuron with complex biochemical processes involved, an artificial neural network passes on numbers with functions representing neurons.

Despite only employing simplistic building blocks, a neural network is able to model any continuos function arbitrarily well using one layer and an infinite number of neurons [5]. It is used in hopes of discovering hidden relations among variables and to utilize high dimensional correlations not otherwise obvious.

A simple approach is to stack multiple layers of neurons (*nodes*) on top of each other and to connect the outputs of the previous layer with inputs of the new layer (*feed-forward neural network*). A network can be designed arbitrarily deep and provide a multitude of additional feedback loops (*recurrent neural network*) and further binning restrictions on node-inputs (*convolutional neural network*).

A simple feed-forward network is depicted in Figure 3.3. Each line between two nodes represents a connection. In other words, the output of the node at the bottom is passed to the node at the top. The function used for calculating the various values of $z_i$ is called

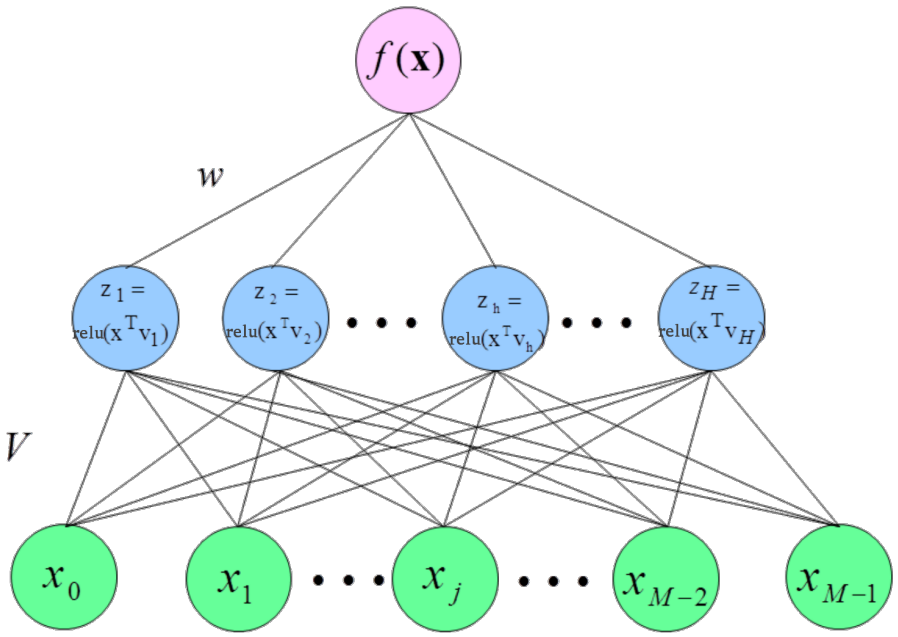



**Figure 3.3:** Design of an artificial neural network with three layers, $\boldsymbol{x}$ as input, $z_i$ as activation function, $V$ and $w$ as weights and $f(\boldsymbol{x})$ as prediction. Adapted from [6].

ReLU [7] and takes the form of $relu(h) = \max(0,h)$. In terms of a biological system it can be thought of as a boundary which has to be overcome prior to a signal being passed on.

Layers not representing the input or output are called hidden layers (blue nodes from Figure 3.3). The dimensions of the input (green nodes) are also referred to as features. A function of a node is called *activation function* ($z_i$). *Learning* or *training* in the context of neural networks refers to the process of adapting parameters or so called *weights* ($V$ and $w$) of a node. The process of adapting them is performed in batches. The *batch size* describes the number of individual data points contained within a batch.[1] Each weight is altered according to a gradient which optimizes the desired function. Often weights include a *bias* which is a constant offset not influenced by any previous neuron. The desired function which is to be optimized is referred to as *loss function*. It measures the predictive power of the classification. The duty of the *optimizer* is to adapt the weights in a way which minimizes the function, a task usually done via propagating the error back through the network in a schema called *back propagation*.

It is important to avoid making the network too dependent on the specific characteristics of the training data. Otherwise it will simply *over-fit* the given events without learning the more general concept. Hence, the neurons of an over-fitted network are perfectly adjusted to the input which it has already seen. However, the system fails upon receiving anything

[1] The last batch may be smaller if the total number of samples is not divisible by the batch size.

it has not already seen in this exact form.

Slight amendments are to be made for a multi-class classification problem. Namely a different activation function than the previously mentioned ReLU is used in the final layer. In this thesis the softmax algorithm is employed as last activation. Its response is given by

$$P(c|x) = \frac{exp(x \cdot w_c)}{\sum_{c'=1}^{\#\text{classes}} exp(x \cdot w_{c'})}, \tag{3.5}$$

with $c$ representing a class, $w_c$ the weights of a class and $x$ the input. The function assigns values between zero and unity to an element of belonging to class $c$. Note that the final output of the network is an exclusive classification into the class with the highest softmax value.

Additionally the loss function must be adapted to reflect the existence of more than two classes. In this study the categorical cross entropy is chosen. In information theory, it puts a measure on the additional information needed to describe the data if deviating from the true underlying distribution.

# Chapter 4

# Analysis

## 4.1 Data sample

In order to validate a particle identification approach and to make sure it is behaving as expected, it is mandatory to measure the performance on Monte Carlo simulated data. Events are generated in accordance with the current understanding of the Standard Model. The software framework EvtGen [8] is used for the purpose of simulating the production and the decay of the $\Upsilon(4S)$. After the simulation of the particles and their various properties, the detector responses are emulated using GEANT4 [9]. Hits in the various components are simulated and finally the veracity of a track identification is calculated. The process of matching the identification with the truth is not possible for real data, as the true properties of a particle are not known in the experiment. However, for testing purposes it is a valuable tool as it allows to compare the performance of new approaches.

Throughout the thesis several decays are discussed, most notably the decay of the charged $B$ mesons. The initially created $\Upsilon(4S)$ decays in $(51.4 \pm 0.6)\%$ of cases into the charged $B^+$ and $B^-$. It represents a good sample of the overall particle species which are to be expected. Observations seen in this generic charged decay are validated using data of the mixed decay of the $B$ meson into $B^0$ and $\bar{B}^0$ which has a branching ratio of $(48.6 \pm 0.6)\%$. Both samples are generic decays and not specific to one single analysis only. A complete list of possible decays as well as the above mentioned branching fractions can be found in [10].

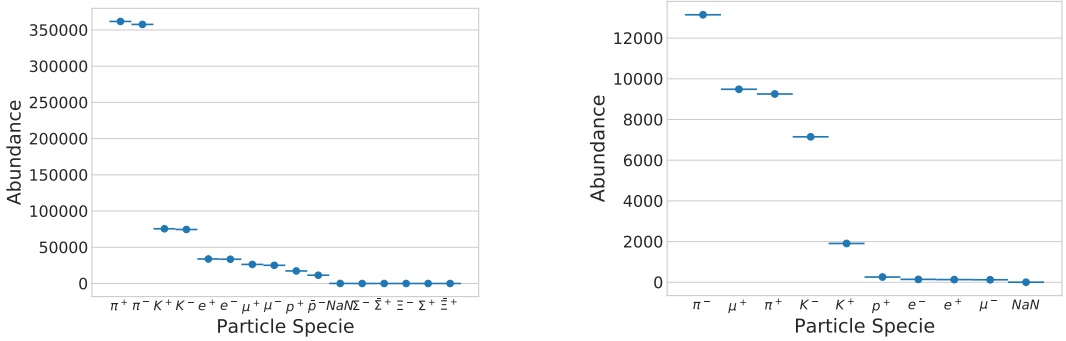
Additionally a decay of the $B^+B^-$ with non-generic properties is simulated. Its parameters are outlined in Table 4.1. It allows a fast processing of tests due to its simplicity and helps in differentiating between decay specific observations and generic ones. The data from this decay is not used for visualizations in this thesis as it does not translate to an appropriate application.

As seen in Figure 4.1 the decays are dominated by kaons and pions. However, the overall

| Decay | | | Branching ratio |
|---|---|---|---|
| $\Upsilon(4S) \rightarrow$ | $B^+B^-$ | | 100% |
| | $B^+ \rightarrow$ | $\mu^+\nu_\mu\gamma$ | 100% |
| | $B^- \rightarrow$ | $D^0\pi^-$ | 100% |
| | | $D^0 \rightarrow$ $\pi^+K^-$ | 20% |
| | | $D^0 \rightarrow$ $\pi^0\pi^+K^-$ | 20% |
| | | $D^0 \rightarrow$ $\pi^+\pi^+\pi^-K^-$ | 20% |
| | | $D^0 \rightarrow$ $K^-K^+$ | 20% |
| | | $D^0 \rightarrow$ $\pi^0\pi^+$ | 20% |

**Table 4.1:** Simulated non-generic decay of the $\Upsilon(4S)$ with charge conjugated decays implied.

distribution is much more peaked for the charged generic decay. Furthermore the non-generic sample decay features a lot more $\mu^+$ relatively speaking in comparison to the charged generic decay. The abundances of the mixed generic decay are not explicitly shown in the figure due to their distinct similarity to the charged decay.




**(a)** Charged-generic **(b)** Non-generic

**Figure 4.1:** True particle abundances in various simulated decays with particle species by charge on the $x$-axis. The particles are sorted by their abundances. *NaN* stands for an invalid conversion[1] from the unique identification code of the particle species to an actual particle.

Falsely identified and poorly defined tracks are removed prior to evaluating the data. First the transverse momentum is limited to the range of 0.05 GeV/c to 5.29 GeV/c. This removes very slow and very fast particles because both types provide insufficiently described tracks. Next, the transverse distance of the closest point of the track to the point of interaction is limited to 2 cm, respectively 5 cm for the longitudinal distance. Finally, tracks with no Monte Carlo truth information assigned to them are pruned. Such tracks are considered falsely reconstructed as a majority of detector hits of the track do

[1] The error occurs due to the Particle Data Group code in the ROOT file being saved as `float32`. However, some codes for particles exceed this limitation. Notably this effects the deuteron and its anti-particle.

not belong to a simulated particle. Including such false tracks would require the particle identification process to identify non-particles tracks. Anyway, the effect of tracks with no truth information does not influence the final result in any significant way. The tracks are merely excluded for the purpose of clarity.

Each generic decay data set features $100,000$ initial $B\bar{B}$ events and about ten times as many identified tracks. The non-generic decay contains $10,000$ initial events and about $40,000$ tracks.

In the following analysis, the generic decay data sets are used. The number of events in one such set is sufficiently high for the purpose of this study while still providing acceptable performance.

## 4.2 Particle identification variables

### 4.2.1 Legacy PID

The current particle identification approach consists of variables calculated via ratios of likelihoods. A particle identification is performed by applying a selection criterion on the variables via a cut.

Prior to version 2 of the Belle II software, the default particle identification approach is to take the likelihood of the desired particle and divide it by itself plus the likelihood of the pion. To construct the ID of the pion, the kaon likelihood is used as second summand in the denominator. Table 4.2 shows the definition of the IDs for each of the six particle species of interest. In the future it will be replaced by the global PID approach described in subsection 4.2.2.

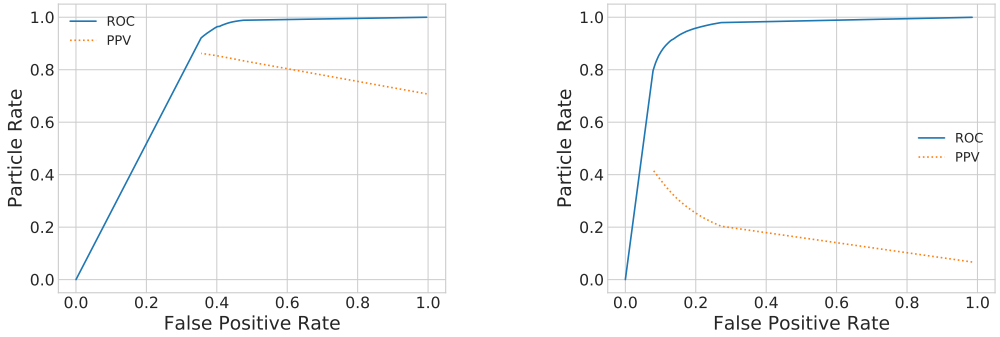
| | |
|---|---|
| pionID | $\mathcal{L}_\pi/(\mathcal{L}_\pi+\mathcal{L}_K)$ |
| kaonID | $\mathcal{L}_K/(\mathcal{L}_K+\mathcal{L}_\pi)$ |
| protonID | $\mathcal{L}_p/(\mathcal{L}_p+\mathcal{L}_\pi)$ |
| electronID | $\mathcal{L}_e/(\mathcal{L}_e+\mathcal{L}_\pi)$ |
| muonID | $\mathcal{L}_\mu/(\mathcal{L}_\mu+\mathcal{L}_\pi)$ |
| deuteronID | $\mathcal{L}_d/(\mathcal{L}_d+\mathcal{L}_\pi)$ |

**Table 4.2:** Definition of the legacy PID variables currently used by default for particle identification.

The identification efficiencies for pions and kaons are high as the approach is able to properly differentiate between both classes. However, it has obvious limitations in identifying rare particles as the fraction is dominated by the high abundance of pions.

The differences in the TPR for identifying the pion and the electron can be seen in Figure 4.2. The pion is used as a representative of a particle species with a high abundance,

while the electron is used as one of the low abundance particle species. A high efficiency in selecting particles is achieved quickly. However, the purity of the sample clearly shows the preference towards the pion identification.




**(a)** Pion identification

**(b)** Electron identification

**Figure 4.2:** Particle identification rates for the pion and electron using the legacy PID approach, each showing the True Positive Rate (ROC curve) and the Positive Predicted Value depending on the False Positive Rate.

This unbalanced classification is further demonstrated by analyzing the identification efficiencies. Figure 4.3 shows the row-wise normalized confusion matrix for the legacy PID approach (see section 3.3). The matrix clearly highlights the bias of the classification towards kaons and pions. Regardless of the actual particle's identity, there is a high chance of it being identified as kaon or pion. Particle species of low abundance such as the electron, muon, proton and deuteron each have an identification efficiency below 40%.

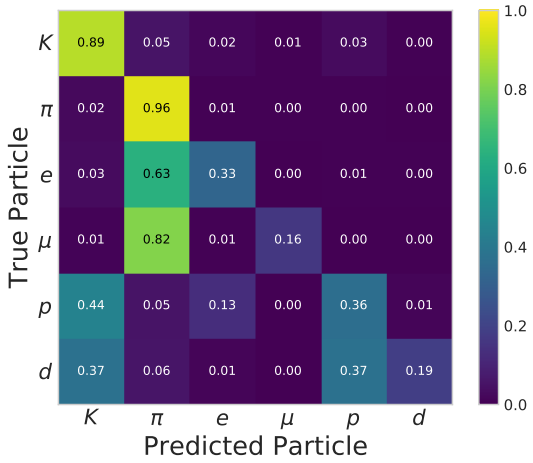



**Figure 4.3:** Matrix of ε values for the legacy particle ID approach. Its values represent the row-wise normalized confusion matrix. Each particle species is assigned a probability of being identified as another kind of particle.

### 4.2.2 Global PID

The global PID approach is the upcoming approach to be used as new default variable with which to select particle samples. Instead of having the likelihood of the pion fixed in the denominator of every particle's ID, it is replaced by the sum over all the other particle likelihoods. Hence, the global PID of the kaon is now represented by the likelihood of the kaon divided by the sum of all the likelihoods of every other particle including the kaon itself. The complete list of the definition for each particle can be seen in Table 4.3.

| | |
|---|---|
| globalPionID | $\mathcal{L}_\pi / \mathcal{L}_{all}$ |
| globalKaonID | $\mathcal{L}_K / \mathcal{L}_{all}$ |
| globalProtonID | $\mathcal{L}_p / \mathcal{L}_{all}$ |
| globalElectronID | $\mathcal{L}_e / \mathcal{L}_{all}$ |
| globalMuonID | $\mathcal{L}_\mu / \mathcal{L}_{all}$ |
| globalDeuteronID | $\mathcal{L}_d / \mathcal{L}_{all}$ |
| $\mathcal{L}_{all} = \sum_{A \in \{\pi, K, p, e, \mu, d\}}$ | $\mathcal{L}_A$ |

**Table 4.3:** Definition of the globalPID variables which is to be used by default for particle identification in the future.

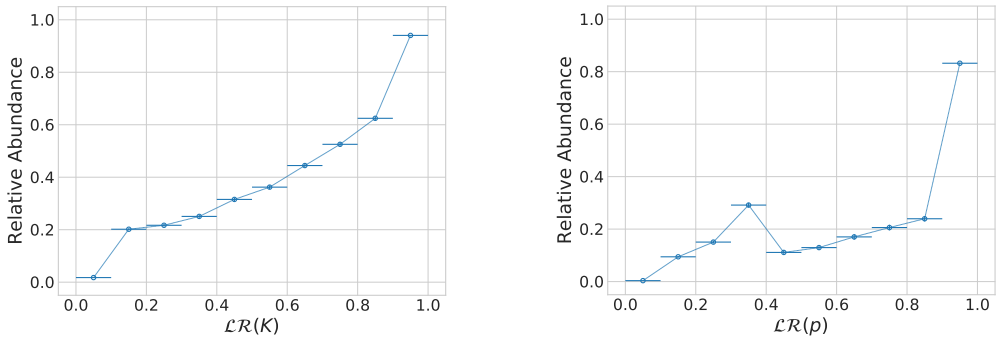
The approach does not favor kaons and pions but has a more balanced classification. Every likelihood of a particle species has the same weight in the denominator.

### 4.2.3 Goodness of the global PID variables

In order to ensure that the global PID variables are properly defined, a likelihood ratio test based on the Neyman-Pearson lemma as outlined in subsection 3.4.2 is performed. The lemma states that the highest purity for a given efficiency is to be expected for each selection on the likelihood ratio as seen in Figure 3.2.

For the following analysis the data is sampled into ten bins of equal width. The error is calculated via gaussian error propagation under the assumption that the counting of the events follows a Poisson distribution. Thereby the assumption is made that the number of desired particles in a bin and the number of undesired particles are independent. The error is underestimated for a purity of zero and unity as gaussian error propagation does not apply.

When applying the previously discussed approach to the data, the goodness of the likelihoods can be validated as seen in Figure 4.4a. The purity of the kaon sample in the bins increases with a stricter likelihood cut and the uncertainty is low due to the high statistics. Figure 4.4b shows a completely different picture. The purity peaks at a likelihood ratio of about 0.35, while the following values are far lower. However, this effect is unique to the proton and can not be observed in any other of the six particle species of interest. This means that the likelihoods for the kaon, pion, electron, muon and deuteron are properly defined and actually behave like probabilities under this test.

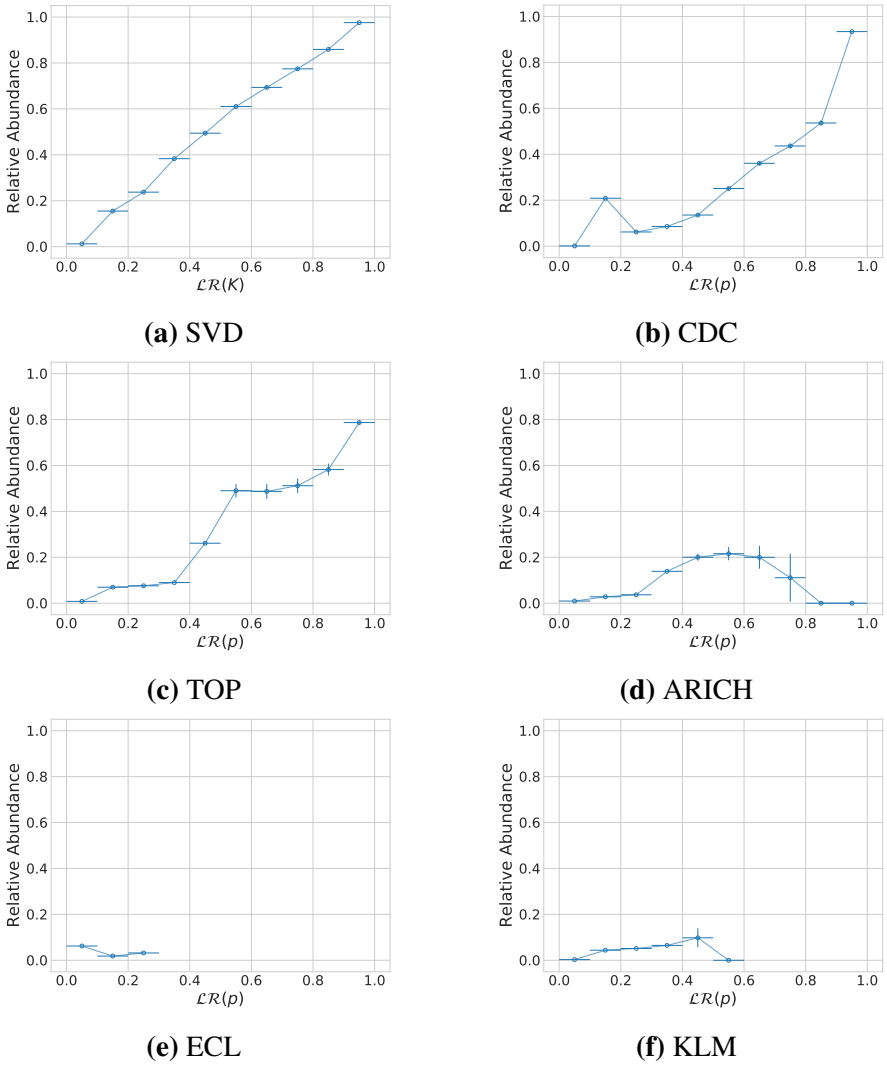



**(a)** Kaon **(b)** Proton

**Figure 4.4:** Relative abundance (purity) of various particle samples in equal width likelihood ratio bins.

In order to understand the observed effect in Figure 4.4b it is important to pin down its cause. Since the likelihoods are values which are returned by each detector, a natural conclusion might be that it is caused by one poorly defined detector response. Figure 4.5 shows the relative abundance of the proton in likelihood ratio bins for various detector components.

The response of the SVD is in perfect agreement with the expectations. However, the CDC shows the same kink, merely shifted a little to lower values. The TOP is in agreement




**(a)** SVD **(b)** CDC

**(c)** TOP **(d)** ARICH

**(e)** ECL **(f)** KLM

**Figure 4.5:** Relative abundance (purity) of the proton in likelihood ratio equal width bins for all available detectors. The error in a bin is given by a vertical line.

again. The ARICH, ECL and KLM have insufficient statistics and/or the likelihoods do not fill the whole range. The error is underestimated at multiple points close to a purity of zero. Hence, the CDC seems like the only viable cause of the unexpected kink in the plot. The observed shift to the left is due to the likelihood of a single detector being lower in general than the likelihood of all detectors combined.

In addition, the $p_t$ dependency in the CDC detector is analysed as depicted in Figure 4.6. The transverse momentum is chosen as it is easily measurable and has a significant influence on the identification process. As sampling method for $p_t$ three equal height bins are chosen. Using few bins with the same number of particles in each one, provides good

statistics and allows for balanced comparisons. Analyzing the results, it becomes obvious that especially low to medium transverse momentum protons contribute to the effect.

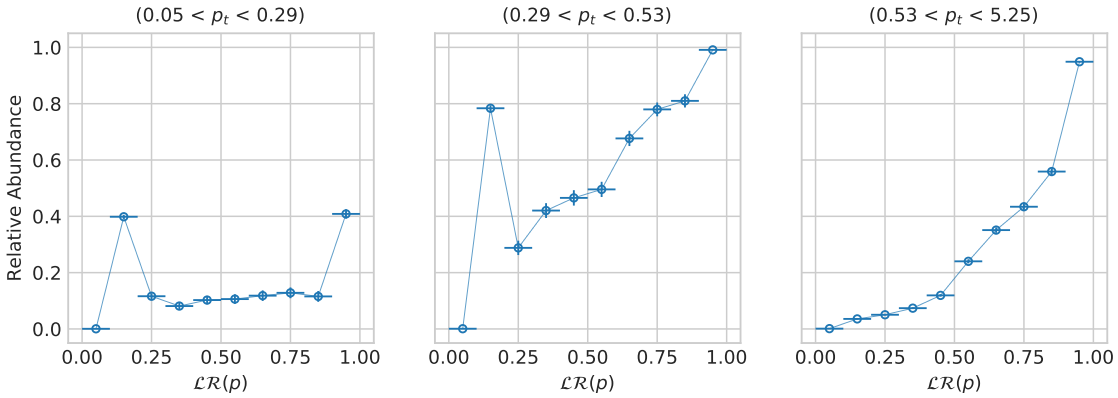



**Figure 4.6:** Relative abundance of the proton in likelihood ratio bins for different transverse momentum equal height bins in units of GeV/c for the CDC detector.

Analyzing the dependency of the likelihood on the angle between the beampipe and track reveals nothing of interest. Although, the angle is important for the identification process as well.

In conclusion, there seem to be a problem with the proton likelihood returned by the CDC detector as it does not behave like a likelihood. As part of this work the findings were communicated to the experts but have yet to be resolved. Overall, it does not affect the results presented in the following sections as the proton does not play a significant role.

## 4.3 Bayesian approach

### 4.3.1 Simple Bayes

The idea for this approach is inspired by the work of the ALICE collaboration as outlined in [11] and [12].

The goal of the simple Bayesian approach is to weight the probability of a particle by the abundance of the species in the sample. This process increases the likelihood of a particle being identified as belonging to a species with a higher abundance and decreases the likelihood of being identified as belonging to one with a lower abundance.

Bayes theorem provides the mathematical foundation:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}, \qquad \text{e.g.,} \quad P(e|Signal) = \frac{P(Signal|e) \cdot P(e)}{P(Signal)}. \tag{4.1}$$

The variable $P(e|Signal)$ is the probability of the track being from an electron given that *Signal* is measured. The term of most interest in the equation is the a priori probability $P(A)$, respectively $P(e)$. As a first simple approach, this variable is now dependant on the absolute abundance of the particle, e.g., the electron. The probability of detecting the signal is modeled using

$$P(Signal) = \sum_{A \in \{\pi, K, p, e, \mu, d\}} P(Signal|A) \cdot P(A) \tag{4.2}$$

with $P(Signal|A)$ being the previously discussed likelihood of the signal using the particle hypothesis $A$.

The normalization of the a priori probabilities $P(A)$ is not important as it appears in both the nominator and the denominator of Equation 4.1. Only the proportions between $P(A)$ of the particle species matter as any common factor within the variables cancels itself out. Therefore, it makes no difference if the value of the abundance of particle species $A$ is used directly in place of $P(A)$ in the calculation.

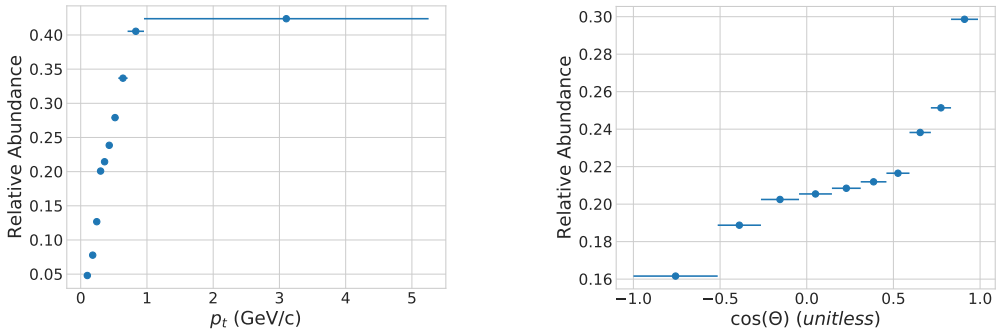
The absolute particle abundance of a sample taken from the Monte Carlo simulation of the charged decay of the $B$ mesons can be seen in Figure 4.1a. Both particle and anti-particle abundances for each species are summed up when calculating the values for $P(A)$. In this example the bias towards pions and kaons is obvious.

The approach depends on the detector yielding decay-agnostic results. Hence, the detector shall be assumed to always output the likelihood of measuring the received signal given a specific particle hypothesis regardless of prior probabilities. Additionally, the approach assumes a bias in the abundance towards one or a few particle species in the data. Otherwise the a priori probabilities would be flat and result in the same values as those given by the global PID.

Thankfully, both of those assumptions are fulfilled: The detector can be assumed to behave independently of the relative particle abundance and the measured data shows a clear predominance of one or a few particle species. This is not surprising in itself as the branching fractions are not equally distributed.

### 4.3.2 Univariate Bayes

The univariate Bayesian approach adds a further dependency in the form of a detector variable to the a priori probabilities of the Bayesian approach. Hence, instead of having the a priori probability depend only on the overall abundance of the particle, the univariate approach additionally varies the values. Namely the abundance is made dependant on, e.g., the transverse momentum and the angle between the beampipe and the track $\Theta$. Those two variables play a significant role in the track fitting process and are dominant factors for the particle identification process outlined in subsection 2.3.2. Instead of just $\Theta$, the cosine of it used in order to even out the distribution.




**(a)** $p_t$ **(b)** $\cos(\Theta)$

**Figure 4.7:** Kaon abundance relative to the pion abundance in equal height bins of $p_t$ and $\cos(\Theta)$ with horizontal lines indicating the size of each bin.

The relative abundance of one particle species in comparison to the pions is shown in Figure 4.7. Equal height bins are chosen to enforce good statistics across all bins. The abundance relative to the pion is used in order to introduce a point of reference. The pion as reference point is chosen arbitrarily. As previously discussed, the absolute normalization of the a priori probabilities does not matter. The graphs underline the motivation for using such an approach as the abundance varies among different bins. Especially the dependency on $p_t$ reveals drastic changes in the relative abundances of the kaon. Overall, binning by $p_t$ reveals a higher contrast and therefore is used as default binning method.

Due to its additional dependency, the univariate Bayesian approach is able to adapt to the underlying data slightly better in comparison to the Bayesian approach.

### 4.3.3 Multivariate Bayes

The multivariate approach extends the univariate one by further increasing the number of free parameters on which the a priori probabilities may depend on. As previously discussed, both $p_t$ and $\cos(\Theta)$ represent an excellent choice as default variables for said dependency. Those variables provide a good separation between different particle species for certain pockets.

Figure 4.8 shows the $p_t$ and $\cos(\Theta)$ of particles. It is important to note the sickle shape of the distribution. Higher values of the cosine are slightly preferred over lower ones due to the asymmetry of the beams. Furthermore, the plots highlight the fact that angles of about $90^\circ$ ($\cos(\Theta) = 0$) correspond to the highest transverse momenta. Overall, the yellow sickle dominates the picture (pions). Only at momenta beyond 2 GeV/c does their abundance visually thins out. On top of the large pion sickle, a slight separation between the cyan sickle in the middle (kaons) and the violet sickle on the right (electrons) can be seen. The separation between kaons and electrons intuitively makes sense since the

production of a heavier particle[2] consumes more phase space which in turn reduces the amount of momentum it can carry. Analog, the electron has a small mass and therefore may carry a lot of momentum (not necessarily transverse momentum though). Obviously this interpretation disregards the fact that particles are not created one by one, but rather in a complex decay and the phase space being distributed across all daughters. Nevertheless, it illustrates the observed effect.

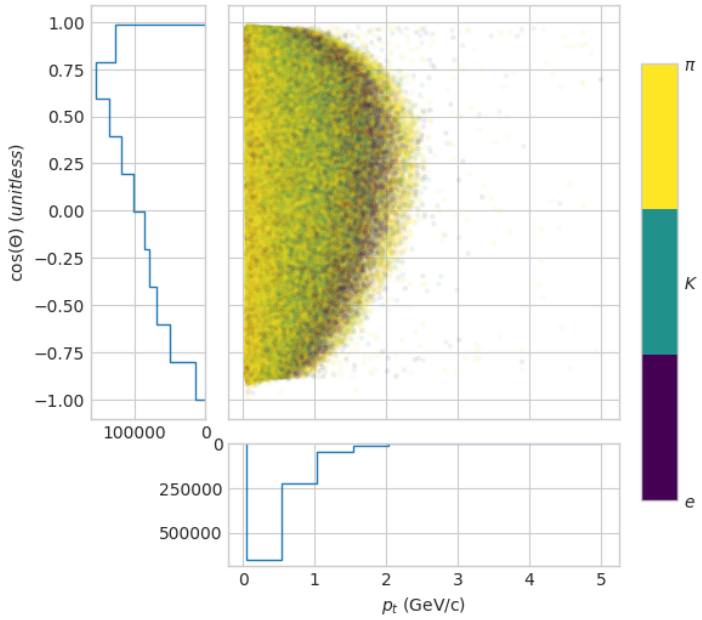



**Figure 4.8:** Scatter plot with tracks represented as opaque points depending on $p_t$ and $\cos(\Theta)$ which is accompanied by two histograms showing the distribution in transverse momentum and the cosine of the angle between the beampipe and the track. The color encodes the species of the particle.

Enabling the a priori probability to depend on those measurements allows the multivariate Bayesian approach to pick more fine grained priors. The a priori probabilities are estimated in accordance with the Monte Carlo information for each combination of variables. In order to achieve this the variable $p_t$ and $\cos(\Theta)$ are each distributed independently across equal height bins. Next, the abundances of particle species relative to the pion are calculated for every combination of $p_t$- and $\cos(\Theta)$-bins.

As an example, the abundance of kaons relative to the abundance of pions is shown in Figure 4.9. The graph demonstrates the fact that particle abundances are unevenly distributed across the $p_t$-, $\cos(\Theta)$-plane.

[2]The mass of the kaon is about a thousand times higher than the mass of the electron.

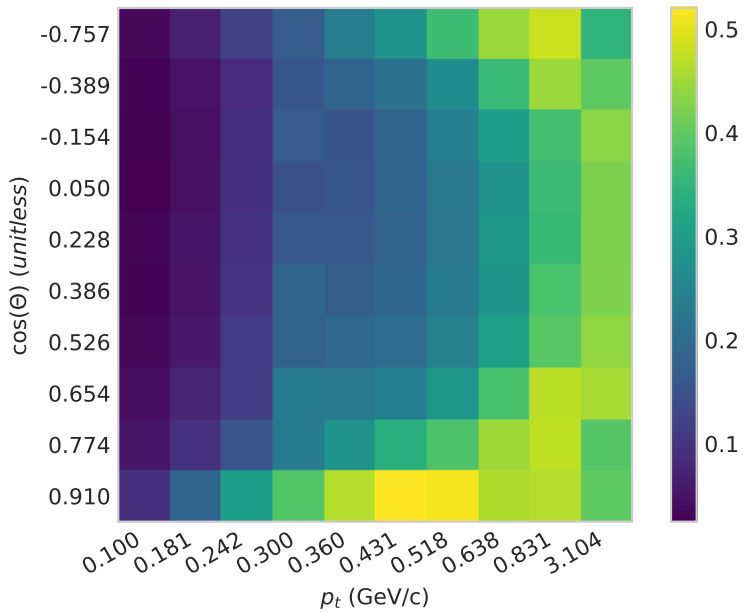



**Figure 4.9:** Relative kaon abundance in intersected bins of the transverse momentum and cosine of Θ. The bins are of equal height for each variable separately. The tick labels at the axes denote the centers of the bins.

### 4.3.4 Comparison

Figure 4.10 shows a comparison of the **legacy PID** approach to the **simple Bayesian** one for identifying pions. The simple Bayesian approach is able to achieve a higher efficiency at a low FPR. In addition, the new approach provides a very high purity at a very low FPR in a range where the legacy PID approach is not even able to identify pions. For a high FPR both approaches converge into similar shapes with rate ratios close to unity. The described effect can be seen for all six particle species studied and is not limited to the pion. However, the comparison of the pion identification is a conservative choice as the legacy PID approach favors high abundance particles and put it at an advantage.

The improvements in the identification efficiencies are less obvious for an exclusive cut on the identifying variables. However, in general the simple Bayesian approach is less prone to confusing particles with one another as seen in Figure 4.11.

The differences are less pronounced in comparison to the **global PID** variables. Both approaches misidentify only a low percentage of each particle species. The global PID approach classifies muons with a higher efficiency while the simple Bayesian approach is better at identifying pions. The efficiencies for the identification of the kaon, electron and proton remain more or less unchanged. Both have a lower efficiency for the kaon and pion identification than the legacy PID approach. Especially the pion identification takes a

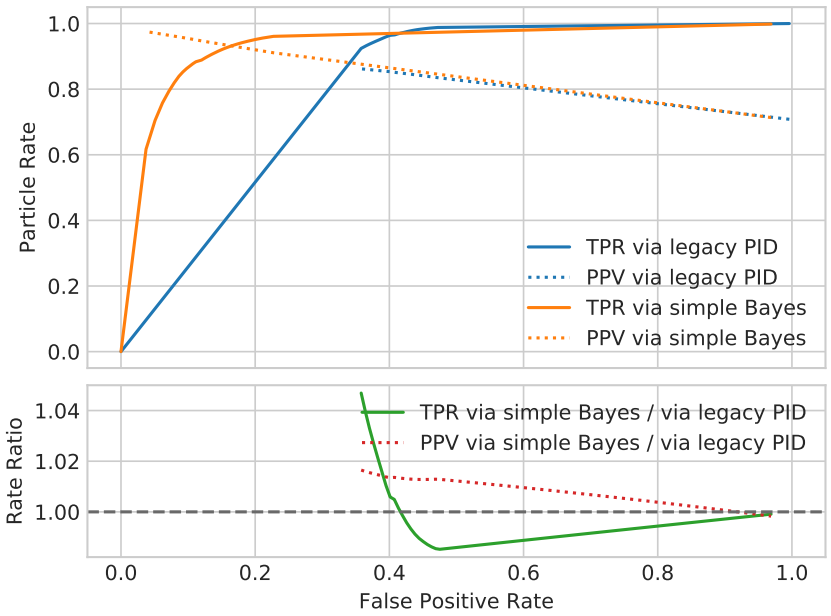



**Figure 4.10:** Comparison of PPV, TPR and FPR for the simple Bayesian with the legacy PID approach for identifying pions. The upper graph shows both rates of each approach separately using different colors, while the lower visualizes the ratio between the PPVs respectively the TPRs. In order to get a smooth curve for the rate ratios, more values than actually present are interpolated for both approaches. This interpolation is vetted for points with a high FPR as with sufficiently fine grained cuts the interpolated values could be achieved. However, the gap between the two leftmost values can often not be filled with additional cuts. Hence, no interpolation is performed between the first and second point.

steep decline. Note that the pion is the particle of highest abundance.

Overall, simple Bayes achieves a higher efficiency than the legacy and global PID approaches. Hence, in general it classifies more tracks correctly as seen in Figure 4.12. The good agreement between predicted and true particle abundances for the simple Bayesian approach is of special interest. It emphasizes that the idea behind simple Bayes is taking effect.

Introducing more degrees of freedom in the form of the **univariate Bayesian** approach reveals merely small differences in comparison to the simple Bayesian approach. Figure 4.13 shows a comparison of the global PID approach with the univariate Bayesian one. The observed changes are predominantly inherited from the simple Bayesian one. The slight loss in efficiencies of low abundance particles is more than compensated by the gain in the pion identification efficiency. Between the simple Bayesian approach and the univariate one, only changes of one percentage point are seen in the particle identification efficiencies. The univariate Bayesian approach boosts the kaon and proton identification

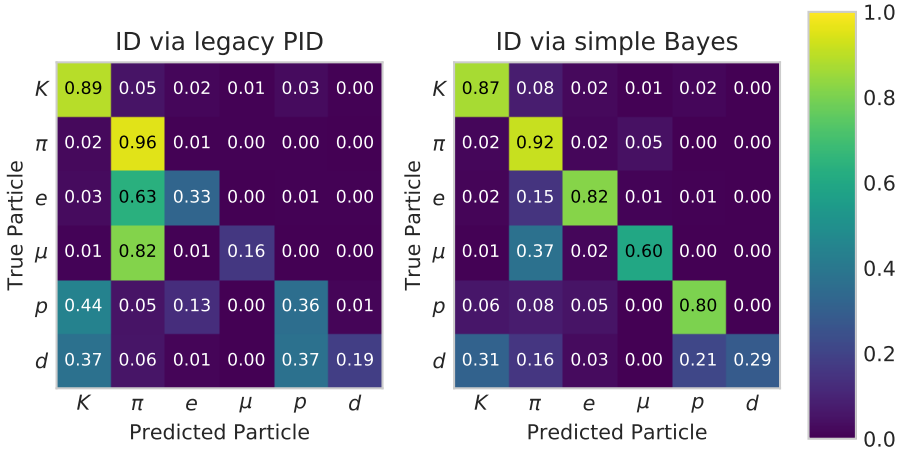
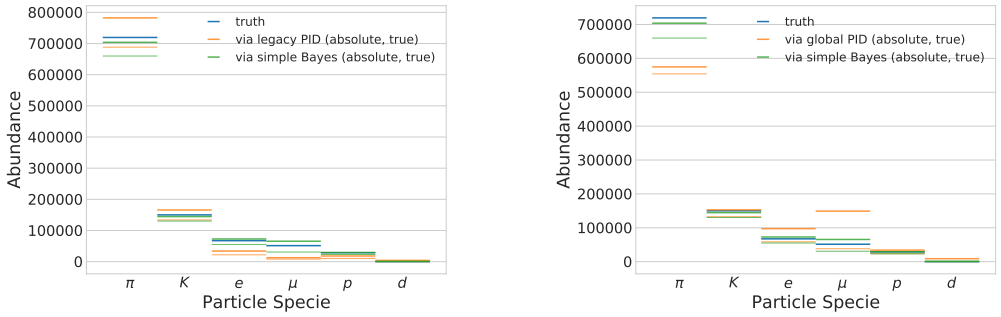



**Figure 4.11:** Comparison of the row-wised normalized confusion matrix for the legacy PID approach with the simple Bayesian one.




**(a)** PID, simple Bayes

**(b)** global PID, simple Bayes

**Figure 4.12:** Counting from the top the first line of each approach (color coded) represents the abundance according to the approach, while the lower line represents the number of correctly classified particles. The particle classification is exclusive.

by this amount at the cost of a reduced efficiency for the muon. A comparison between the univariate Bayesian approach and the legacy PID is redundant due to the similarity between the simple Bayesian and univariate Bayesian approach.

Lastly, the **multivariate Bayesian** approach is tested. The TPR and PPV for every particle species are very similar to the ones from the univariate approach – neglecting effects below a maximal relative change of 2%. Negative effects are always below a threshold of 0.4% while the PPV for the muon identification is slightly boosted by 2% at one point. The similarity of the multivariate and univariate approach is also reflected in the efficiencies of the row-wised normalized confusion matrix being virtually identical. Comparing the approach to the legacy PID, global PID or simple Bayes would duplicate the comparison

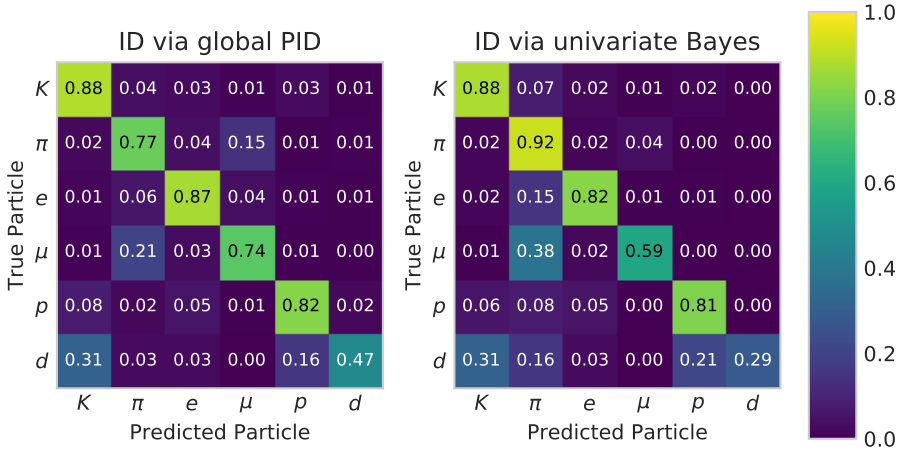



**Figure 4.13:** Comparison of the row-wised normalized confusion matrix for the global PID approach with the univariate Bayesian one.

which are outlined for the univariate Bayesian approach.

The global efficiencies over all particle species are displayed in Table 4.4. The efficiencies underline the quasi non-existing performance improvement of the multivariate Bayesian approach relative to the univariate Bayesian one. Furthermore, it highlights the previously discussed overall improvements in the various other approaches.

| Particle identification approach | Overall identification efficiency |
|---|---|
| legacy PID | $(84.4 \pm 0.4)\%$ |
| global PID | $(79.2 \pm 0.2)\%$ |
| simple Bayes | $(88.2 \pm 0.2)\%$ |
| univariate Bayes | $(88.5 \pm 0.3)\%$ |
| multivariate Bayes | $(88.5 \pm 0.3)\%$ |

**Table 4.4:** Mean overall efficiencies of the discussed particle identification approaches with their statistical error. Errors are estimated using four charged and four mixed generic decay data sets. The result of the univariate Bayesian approach is calculated using $p_t$-bins and the multivariate Bayesian approach utilizes $p_t$- and $\cos(\Theta)$-bins.

### 4.3.5 Summary and outlook

In comparison to the legacy PID approach the positive effects introduced by the Bayesian approaches are mainly due to the inclusion of further particle likelihoods. Both the global PID approach as well as the Bayesian one decrease the rate at which particles of high abundance are identified. However, both kinds of approaches in turn increase the

identification efficiencies of particles with a low abundance significantly. Hereby, the Bayesian approaches are able to improve upon the global PID as they correctly classify more pions in the charged generic decay at the cost of a reduced identification rate for some particles of lower abundance.

The differences among the simple Bayesian, the univariate Bayesian and the multivariate Bayesian approach are slim. Neither the overall identification efficiencies differ significantly, nor does the identification rate of certain particles get boosted. Slight overall gains are measurable only for the univariate Bayesian approach.

As an outlook, further studies concerning the cell structure of the a priori probabilities are worth considering. The current approach of selecting equal height bins is shown to be useful. Nevertheless, it has its limitations. Especially if it is used in conjunction with multiple other variables. Two intervals on separate variables, each containing a certain percentage of the data, are not guaranteed to contain the same amount of data if intersected. A possible extension could be to implement a proper two dimensional clustering using, e.g., Voronoi cells instead of interval boundaries. Density based clustering algorithms such as DBSCAN [13] and OPTICS [14] might also be worth considering. Those approaches could further extend the a priori probabilities of the two dimensional cells of the multivariate Bayesian approach to density connected clusters.

Additionally, overlapping boundaries either from the current approach or introduced by new clustering algorithms could be used to weight multiple a priori probabilities. Instead of relying on one exclusive boundary, the mean of several a priori probabilities could be used instead.

## 4.4 Neural network approach

### 4.4.1 Design

A neural network consisting of sequential layers is chosen for the following discussion. As software library, Keras [15] is used with TensorFlow [16] as back-end.

It is found that eight layers perform reasonably well without over-fitting the data. As for the layers themselves, two different concepts are used. On the one hand a so called *dense* layer connects all inputs with each node within the layers, as seen in Figure 3.3. It usually has a high number of free parameters depending on the function used in a node. On the other hand a so called *dropout* layer is employed which – as its name suggests – drops a certain percentage of inputs by chance. It enforces the weight adaption on only the remaining nodes. It is commonly used to counteract the effect of having a lot of free parameters [17] as it randomly disregards values and therefore possibly drops nodes which would otherwise start to over-fit.

Three different approaches for the choice of input parameters are evaluated. First the unaltered global PID variables for each detector are used as a baseline. As second approach a rich mixture of variables is used, containing information on the momentum, the angle between the beampipe and the track, the distance of the vertex to the interaction point, the energy, the curvature of the track, the charge, the legacy PID as well as the global PID variables for each detector. The third approach uses the same initial inputs as the second one. However, prior to passing the values to the network, a **P**rincipal **C**omponent **A**nalysis [18] (PCA) on the standardized data is performed. The standardization step centers the data by removing the mean and scales the value so that the standard deviation becomes unity. This is an important pre-processing step as it counteracts the effect of having different units and scales for each variable.

The output, i.e., the actual classification, is done in the final step of the network via the previously discussed softmax algorithm. The gradient of the cross-entropy is used for weight adaption. In order to get a unique classification, the class with the highest softmax value is selected. The six long-living particle species are used as targets for the classification. In addition, a zeroth class is added. It is used for classifying particles which do not belong to the six other classes. Although this class merely contains a fractional amount of particles, it is important nevertheless as each track on which the network is supposed to learn is required to have a target associated with it.

Table 4.5 shows the parametrization of the network. Overall, it has between $1,186$ and $2,258$ free parameters which are to be adapted, depending on the dimension of the input. The number is the lowest for the global PID variables as input and the highest if given all variables.

| Layer number | Type | Activation | Bias | #Nodes |
|---|---|---|---|---|
| 1 | Dense | ReLU | True | 14 |
| 2 | Dropout (20%) | | | |
| 3 | Dense | ReLU | True | 21 |
| 4 | Dropout (20%) | | | |
| 5 | Dense | ReLU | True | 14 |
| 6 | Dropout (20%) | | | |
| 7 | Dense | ReLU | True | 10 |
| 8 | Dense | Softmax | True | 7 |

**Table 4.5:** Design parameters of the neural network.

### 4.4.2 Performance

The accuracy is used as performance measure. It is the fraction of correctly classified tracks relative to all available ones. In the discussed case it is synonymous to the overall efficiency.

Testing and validation is done using a fixed split between training and validation data (*holdout method*). It is important to separate those two samples in order to spot networks which over-fit the data instead of actually 'learning'. The training data is a random sample containing 80% the number of tracks as the original data. This does not necessarily need to actually be 80% of the data as some tracks can be picked multiple times. The remaining tracks are used for validating the actual learning process. Both training set and validation set are used to compute the accuracy, however, only the training sample is used to adapt weights. If the network does not over-fit, the validation accuracy fluctuates around the training accuracy.

Furthermore, two approaches for sampling the data are analysed. One approach picks all tracks at random with equal weights and hence in theory has a bias towards particle species with a higher abundance (*biased sampling*). The other approach weights tracks in a way which makes every particle species equally likely to be picked (*upsampling*). The idea behind this approach is to force the network to focus less on the particles of high abundance.

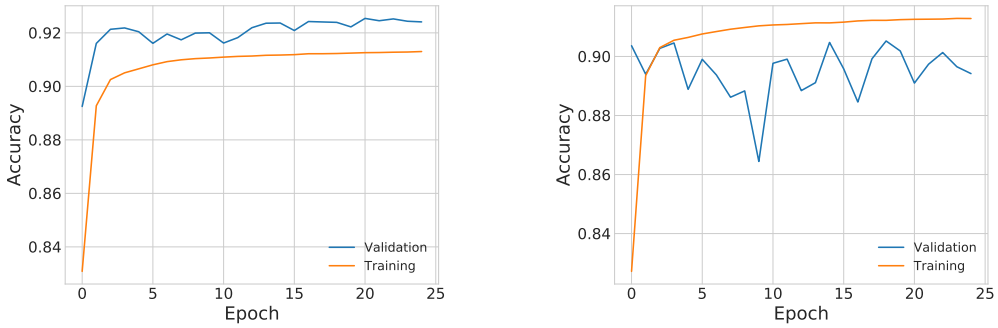
Each training is performed with a batch size of 256. Smaller batch sizes turn out to perform poorly as the weights are adapted too frequently and the accuracy fluctuates without increasing overall. The batch size is found via iteratively picking higher values until finding a rate at which it changes only gradually.

Furthermore, several optimizers – algorithms describing the process of weight adaption – are tested. In the end Adadelta [19] and Adamax [20] prove to have the best performance with the least fluctuations in the validation accuracy. Hence, the following discussion will focus only on those two weight adaption approaches.

Figure 4.14 shows the two optimizers in comparison. Both yield similar results, especially considering the scaling of the accuracy. It is important to note that both are able to 'learn' the classification problem quickly. Within the first epoch they are able to achieve accuracies of nearly 90%. However, they level off fast. After approximately five to ten iterations no noticeable change in the training accuracy can be observed. The validation accuracy follows the general trend of the training accuracy but as expected it shows a lot more fluctuations. The performance of the training accuracy of the Adadelta algorithm is slightly better. The validation accuracy on the other hand is almost the same for the two approaches after about ten epochs.

Figure 4.15 compares the two sampling approaches using the global PID variables as input. The graph employs the Adamax optimizer, however Adadelta behaves similarly. The network which receives the tracks of the particles in proportion to their abundance is able to outperform the network which is trained on data using upsampling. The final difference in validation accuracy is about three percentage points. Using only the global PID improves the results slightly for the biased approach in comparison to selecting everything as feature.

In Figure 4.16 the PCA feature selection approach is shown next to the 'all' approach for



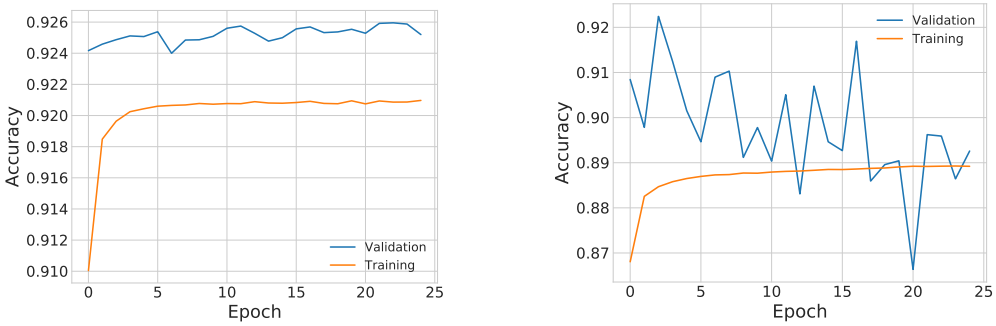

**(a)** Adadelta **(b)** Adamax

**Figure 4.14:** Accuracy of the neural network using the Adadelta and Adamax optimizer with everything selected as features. The values are displayed for a biased sampling approach.
Note that the sample on which the network is training does not represent the same sample which is used for the comparison of the various Bayesian approaches as tracks can be picked multiple times. Hence, the accuracies given here are not comparable to the ones presented for the other particle identification approaches.

upsampling. The term 'all' in this context denotes that every available variable is selected as feature. As seen in the graphs, the accuracy increases with the number of components and easily surpasses the selection of all components. However, further increasing the number of inputs beyond the depicted value does not reveal significant changes. The 'all' approach in combination with upsampling performs poorly. Its training accuracy is consistently lower than that of any other approach. In general, the accuracy of the networks which use upsampling are lower in comparison to the biased selection.

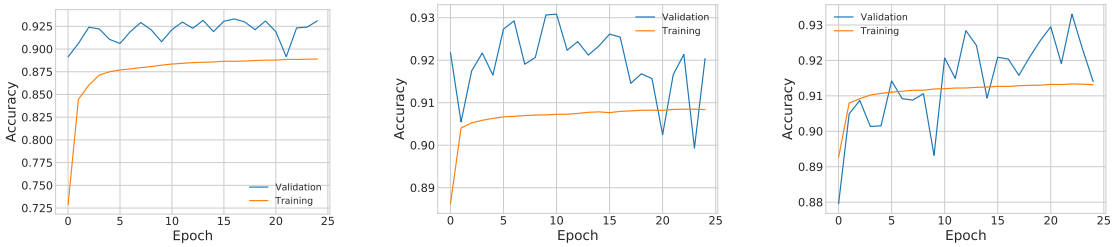
Using the PCA components as feature for the biased sampling approach yields very similar results to just using all variables as feature. In the end the training and validation accuracy is higher by about one percentage point.

Independent of the sampling method and the feature selection, it can be observed that the slope at which the accuracy initially increases is high. All approaches are able to structure the data in a way to improve the classification while training. The number of samples which the training process requires before the training accuracy levels off is similar for all approaches. Overall, selecting about 90 principal components yields the highest accuracies independently of the sampling approach. The best performers for both sampling methods are close in training accuracy but the validation accuracy is superior for the biased sampling approach.




**(a)** Biased **(b)** Upsampled

**Figure 4.15:** Accuracy of the neural network using the Adamax optimizer with the global PID as features. The values are displayed for a biased sampling approach (left) and the upsampling approach (right).




**(a)** 'All' **(b)** PCA, 50 components **(c)** PCA, 90 components

**Figure 4.16:** Accuracy of the neural network using the Adamax optimizer with selecting all features (left), the 50 most principal components (center) and the 90 most principal components (right). The values are displayed for a set using upsampling.

### 4.4.3 Comparison

The following figures are based on data from a different charged generic decay than the one which is used for training. By doing so, basically the same distribution of particles is achieved without reusing training data. The previously trained model is applied to this new decay using the same parameters.

Figure 4.17 shows the comparison of the identification efficiencies for the best performing network using biased sampling. Of note is the predominance of particles of high abundance. The network focusses on kaons and pions. Additionally, protons are identified properly as well. Electrons, muons and deuterons on the other hand are disregarded completely. By comparing the neural network to the multivariate Bayesian approach, a harsh comparison is made as the multivariate Bayesian approach is able to achieve a good classification for particles of high abundance without disregarding particles of lower abundance. Except for the proton, the neural network is not able to classify particle of low abundance at all. In

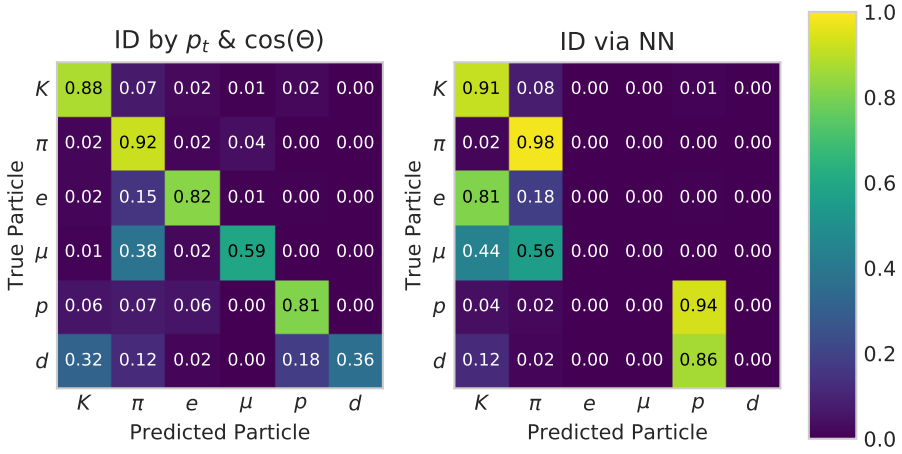
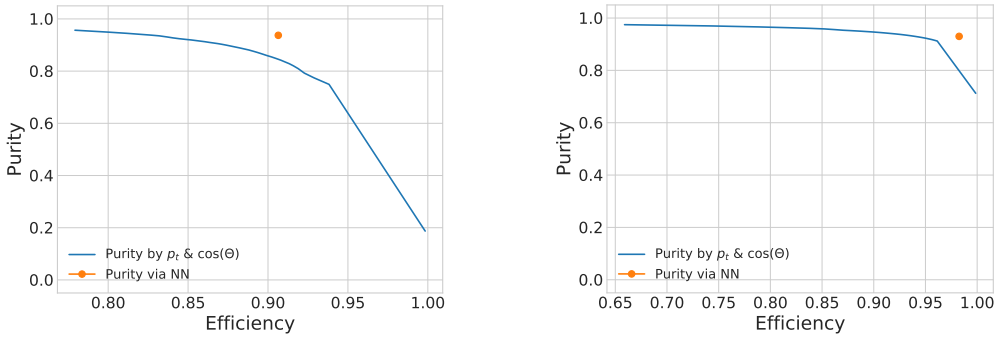



**Figure 4.17:** Comparison of the row-wised normalized confusion matrix for the multivariate Bayesian approach by $p_t$ and $\cos(\Theta)$ and a neural network using the 90 most principal components as features and employing biased sampling.

this regard, the network is worse in comparison to any other approach.

The identification efficiency of the proton behaves antithetically to the efficiencies of other particles of low abundance. Of special note is that the deuteron is classified as a proton as well. Hence, the heavier particles are predominantly categorized as protons. Further analysis shows that the purity of the selected sample is low. The performed selection categorizes nearly all true protons as such but it also categorizes a lot of other particles as being a proton as well. However, in comparison to the Bayesian approach the purity is higher nevertheless.




**(a)** Kaon **(b)** Pion

**Figure 4.18:** Purity over efficiency comparison for the kaon and pion identification using the multivariate Bayesian approach by $p_t$ and $\cos(\Theta)$, and a neural network using the 90 principal components as feature and employing biased sampling.

The classification in Figure 4.18 shows the purity over the efficiency for the kaon and pion identification. The network is able to outperform the multivariate Bayesian approach as the point representing the network lies above the line for the Bayesian approach. However, the improvements are limited to those particle species plus the proton as discussed previously.

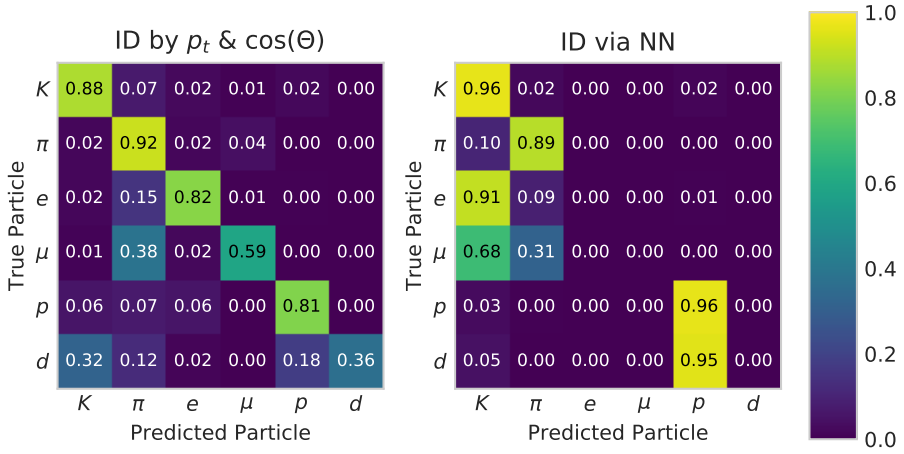



**Figure 4.19:** Comparison of the row-wised normalized confusion matrix for multivariate Bayesian approach by $p_t$ and $\cos(\Theta)$, and a neural network using the 90 most principal components as features and employing fair sampling.

Using upsampling, slightly decreases the pion identification efficiency and increases the efficiency in identifying kaons as seen in Figure 4.19. Hence, the approach is successful in adapting the network to treat particle species more evenly but is unsuccessful in enabling it to identify electrons, muons or deuterons.

A possible cause for the complete disregard of the electron, muon and proton might be that the network is over-fitting the particles of low abundance. However, this would not explain the high identification efficiency for the proton. One possible explanation as to why it still identifies it and classifies the deuterons as proton as well, might be due to them being comparably heavy. The mass of a particle significantly influences the detector responses and might be easy to 'learn'. Furthermore, the batch size might be an additional cause for the indifference towards pions and kaons as particles of low abundance are not represented properly in each weight adaption step.

Overall, the classification efficiency of the network using biased sampling is 85.6% compared to 80.1% for the network using upsampling. Those efficiencies can be compared with further approaches from Table 4.4. Both neural network approaches perform significantly worse than any of the Bayesian approaches. In comparison to the global PID approach the neural network approaches perform slightly better at the cost of loosing the ability to identify anything but kaons, pions and protons. The network which uses biased sampling additionally outperforms the legacy PID approach in terms of overall efficiency.

### 4.4.4 Summary and outlook

The overall efficiencies of the neural network are promising as they reach 80% by just using a simple design in combination with a high batch size and an appropriate optimizer. However, a bias in the form of different abundances for different particle species complicates the learning process significantly. Upsampling does not provide the necessary boost in performance for particles of low abundance, although it helps against the dominance of the pion. Certain particles like the proton are classified with a high efficiency despite being of low abundance. In general, both approaches heavily favor particles with a high abundance. This is not surprising in itself as the training accuracy and therefore the function which is to be optimized can be easily improved by disregarding particles of low abundance.

As an outlook, further research in this field seems promising as even a simple neural network is able to achieve an overall acceptable efficiency. Instead of using a generic decay, it might make more sense to generate events which produce particles with equal abundance while otherwise sharing the characteristics of a generic decay. This would completely circumvent the need for upsampling the data. In addition to an overall good classification of generic events, such a network could be further trained briefly on decay specific parameters. In an optimal scenario both approaches could be combined: a decay specific network being employed whenever possible and a generic classifier being used as fallback. Furthermore, the unique classification could be extended to a probability assignment for each particle species. Hereby, the full output of the softmax activation could be utilized.

# Chapter 5

# Conclusion

Three main approaches for the particle identification at Belle II are studied.

First, the current and future particle identification variables are described. Hereby, the goodness of the global PID variables is discussed. Altogether, the likelihood ratios are properly defined, however, a flaw in the definition of the global PID of the proton is revealed. The unexpected kink can be traced back to the CDC detector and is especially dominant for particles with a low to medium transverse momentum.

On the basis of the particle likelihoods, further extensions of the particle identification utilizing Bayes theorem are discussed. The likelihood of different particles are weighted by species and are assigned a priori probabilities. At first, these a priori probabilities only depend on the abundance of a species and show similar results to the global PID approach. The new approach increases the overall identification efficiencies by increasing the weights of particles with a high abundance. The introduction of further dependencies of the a priori probabilities increases the performance only marginally.

Finally, an alternative approach in identifying particles using a neural network is analysed. Using a simple network provides acceptable performance with overall accuracies in the range of the previously discussed global PID. Nevertheless, it is heavily biased by particles of high abundance and upsampling the data shows only limited improvements.

In conclusion, the particle identification can be improved by using a Bayesian approach. However, particle identification rates for specific species highly depend on the identification algorithm and may be chosen depending on the decay of interest. Currently, the neural network does not provide a sufficiently good classification to be used right away, but is an especially promising candidate for further research.

# Appendix A

# Computer Code

The repository for the programs and the thesis is hosted at GitHub:Edenhofer/PID-boost. It includes the slides for weekly updates, all required graphics for building the PDFs, and some documentation.

The core of the program is divided into three main parts: the program for generating figures `event_metrics.py`, the program `event_nn.py` used for training the neural network and the all encompassing library `lib.py`. Further scripts are relatively small compared to those three ones. The smaller programs were used, e.g., to simulate a decay, generate decay data sets or to visualize parameters of a neural network. All important features are documented via an entry in the program's `help` page. Functions in the code itself which are deemed not self-explanatory are described via a Python `docstring`.

Probably of most interest is the program `event_metrics.py`. It was used for all plots which compare two (or possibly more) approaches. Its parameters are further divided into groups called `actions`, `sub-options` and `utility-options`. An `action` induces some form of calculation or visualization which can be further parameterized via `sub-options`. The `utility-options` group allows for controlling the input and output behavior of the script.

For example, the parameter `--multivariate-bayes` would be an action. It can be further configured by, e.g., specifying the number of cuts to perform on the output of the classification functions via `--ncuts`. The number of bins for the variables on which the a priori probability may depend on, can be set via `--nbins`. By using `-i` and `-o`, the input and the output location can be configured.

Several of the previously described parameters are valid for other programs in the repository as well. Sensible default parameters alongside multiple examples can be found in the readme of the repository.

# Bibliography


[1] T. Abe et al. “Belle II Technical Design Report”. In: (2010). arXiv: 1011.0352 [physics.ins-det].

[2] Christian Pulvermacher. *dE/dx particle identification and pixel detector data reduction for the Belle II experiment*. Dipl. ; Dipl. 2012. URL: https://ekp-invenio.physik.uni-karlsruhe.de/record/48263.

[3] Christian Pulvermacher. *Analysis Software and Full Event Interpretation for the Belle II Experiment*. PhD dissertation. 2015. URL: https://docs.belle2.org/record/336/files/BELLE2-PTHESIS-2016-003.pdf.

[4] Belle II Theory Interface Platform. “The Belle 2 Physics Book”. In: *Progress of Theoretical and Experimental Physics* (to be published), p. 721.

[5] George Cybenko. *Approximations by superpositions of sigmoidal functions*. Feb. 1989. URL: https://web.eecs.umich.edu/~cscott/smlrg/approx_by_superposition.pdf.

[6] Volker Tresp. *Lecture notes in Machine Learning: Neural Networks*. Apr. 2017. URL: http://www.dbs.ifi.lmu.de/Lehre/MaschLernen/SS2017/Skript/03-NeuralNetworks2017.pdf.

[7] R. H. R. Hahnloser, H. S. Seung, and J. J. Slotine. “Permitted and Forbidden Sets in Symmetric Threshold-Linear Networks”. In: *Neural Computation* 15.3 (Mar. 2003), pp. 621–638. DOI: 10.1162/089976603321192103.

[8] David J. Lange. “The EvtGen particle decay simulation package”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 462.1 (2001), pp. 152–155. DOI: 10.1016/S0168-9002(01)00089-4.

[9] S. Agostinelli et al. “GEANT4 – a simulation toolkit”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250–303. DOI: 10.1016/S0168-9002(03)01368-8.

[10] C. Patrignani et al. “Review of Particle Physics”. In: *Chin. Phys.* C40.10 (2016), p. 100001. DOI: 10.1088/1674-1137/40/10/100001.

[11] Jaroslav Adam et al. “Particle identification in ALICE: a Bayesian approach”. In: *Eur. Phys. J. Plus* 131.5 (2016), p. 168. DOI: 10.1140/epjp/i2016-16168-5. arXiv: 1602.01392 [physics.data-an].

[12] I. Belikov et al. “Bayesian approach for combined particle identification in ALICE experiment at LHC”. In: *Computing in high energy physics and nuclear physics. Proceedings, Conference, CHEP'04, Interlaken, Switzerland, September 27-October 1, 2004*. 2005, pp. 423–426. URL: http://doc.cern.ch/yellowrep/2005/2005-002/p423.pdf.

[13] Martin Ester et al. “A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD'96. Portland, Oregon: AAAI Press, 1996, pp. 226–231. URL: http://www.dbs.ifi.lmu.de/Publikationen/Papers/KDD-96.final.frame.pdf.

[14] Mihael Ankerst et al. “OPTICS: Ordering Points To Identify the Clustering Structure”. In: ACM Press, 1999, pp. 49–60. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.129.6542&rep=rep1&type=pdf.

[15] François Chollet et al. *Keras*. https://keras.io. 2015.

[16] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. URL: https://www.tensorflow.org/.

[17] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. Adaptive computation and machine learning. MIT Press, 2016. ISBN: 9780262035613. URL: https://books.google.de/books?id=Np9SDQAAQBAJ.

[18] Karl Pearson F.R.S. “LIII. On lines and planes of closest fit to systems of points in space”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572. DOI: 10.1080/14786440109462720.

[19] Matthew D. Zeiler. “ADADELTA: An Adaptive Learning Rate Method”. In: *CoRR* abs/1212.5701 (2012). arXiv: 1212.5701.

[20] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *CoRR* abs/1412.6980 (2014). arXiv: 1412.6980.

# Acknowledgments

This thesis would not have been possible without the continues support of numerous people. First of all, I would like thank my supervisor Prof. Dr. Thomas Kuhr for the chance to work in his research group as a Bachelor's student. By providing helpful and constructive feedback at every opportunity, he gave me inspiring insights into various topics of experimental particle physics. Of cource, I would also like to thank James Kahn and Dr. Martin Ritter for providing profound answers to all of my questions. Specifically, I would like to thank Dr. Michael Bender, Dr. Martin Ritter and Jakob Roth as well for proofreading my thesis and giving valuable feedback.

Furthermore, thank you to all the students at the working group of Prof. Dr. Kuhr for their welcoming attitude and for making my stay an enjoyable time. Thank you, Christoph Ames, Anna Bertolini, Nan-Hee Kang, Yasin Silyanoglu, Daniel Moritz and Jakob Roth.

Last but not least, I would like to thank my family which has always supported me in countless ways.

Thank you all!

## Erklärung zur Bachelorarbeit

Hiermit erkläre ich, die vorliegende Arbeit selbständig verfasst zu haben und keine anderen als die in der Arbeit angegebenen Quellen und Hilfsmittel benutzt zu haben.

## Bachelor's thesis statement of originality

I hereby confirm that I have written the accompanying thesis by myself, without contributions from any sources other than those cited in the text and acknowledgments.

| | |
|---|---|
| ..... | ..... |
| Date and location | Signature |