Improving upstream electron identification with bremsstrahlung information

by

N.C. Kruse

In partial fulfilment of the requirements for the degree of Master of Science
University of Groningen, July 2020

university of groningen

faculty of science and engineering

Master's Project for the MSc. Physics

July 2020

Student: N.C. Kruse

First supervisor: Dr. ir. C.J.G. Onderwater

Daily supervisor: Dr. M. C. van Veghel

Second assessor: Dr. J. G. Messchendorp

Acknowledgements

I wish to extend my sincerest gratitude to Maarten van Veghel, who has guided me with unending patience throughout this project. My thanks also go out to Gerco Onderwater, who has provided me with the opportunity to experience the most fulfilling year of my academic career.

Many thanks to my parents, who are the embodiment of parental love and care, and who have always inspired me to seek out new knowledge and skills.

Abstract

Lepton universality is a property which is embedded in the Standard Model of Particle Physics. A violation of this universality would indicate the presence of physics beyond the Standard Model. In this work, the search for lepton universality violation is aided by investigating the possibility of using bremsstrahlung-related information, to increase the efficiency of particle identification of electrons in the LHCb detector. An ensemble of variables is created that, combined with machine learning tools, increases false positive rejection rates by 58.4% for upstream electrons.

Contents

1 Introduction

In the 5^{th} century BC, Greek philosopher Leucippus and his pupil Democritus first coined the word *atomos*, or 'indivisible', when talking about the fundamental building blocks of matter. This pursuit of the smallest pieces of nature has been a constant in human history. At several points in time, the end of this search has been wrongfully proclaimed. The discovery of the atom was followed by that of the electron and the nucleus. The nucleus revealed itself to be composed of protons and neutrons, which we now know are also not indivisible.

With the introduction of the Standard Model of Particle Physics, true elementary particles are once again within our grasp. It is now up to the scientific community to find flaws and shortcomings in the Standard Model, and to devise possible alterations or extensions of the model that better allow us to understand what is happening in nature. One avenue of approach is probing the universality of leptons, which is a property that is present in the Standard Model. A violation of this universality would imply a direct violation of the Standard Model in its current form. This project has been performed to aid the search for lepton (non)universality, by attempting to answer the following question:

Can bremsstrahlung-related information enhance the particle identification of upstream electrons in the LHCb detector?

By applying machine learning tools on available data, the amount of electrons that are identified correctly can be increased. This improved identification in turn leads to an effective increase in usable data that can be used in lepton universality analyses.

Chapter 2 will go into detail about the implications of lepton universality violation, as well as work that has already been performed in this area. The LHCb detector itself will be described in chapter 3, along with a description of the detection of electrons within it. Following this will be the methods used in acquiring the data for this work, and how results were obtained from this data. This can be found in chapter 4. The results themselves are located in chapter 5, followed by a discussion in chapter 6.

2 Probing Lepton Universality

The main scope of this thesis is the improvement of the identification of a subclass of electrons within the LHCb detector, which in itself is only one of the experiments aimed towards investigating lepton universality. To mitigate the effect of tunnel vision which comes naturally when investigating such a comparatively small cog in a larger ensemble, the current chapter will paint a picture of the overarching end goal of understanding the properties of leptons.

Several facets of this subject will be covered, from a brief introduction of the Standard Model and the place of lepton universality within it, to the possible approaches used in probing this universality. The decay channel under investigation in this work will also be described, along with the motivation for its choice. With the possibility in mind that the Standard Model of Particle Physics in its current form may not be the end-all theory for physics, two alternative theories are discussed briefly that could explain lepton non-universality.

2.1 Lepton Universality in the Standard Model of Particle Physics

The apex of predictive models in physics for several decades, the Standard Model (SM) provides unrivalled insights in the nature and interactions of elementary particles. The model provides a unified description of all fundamental forces, with the exception of gravity. It comprises all currently known elementary particles, grouped in half-integer spin fermionic matter and integer spin bosons, seen in Figure 1. For detailed data on the properties of these fundamental particles, one can access the biennial Review of Particle Physics [31]. Excellent works on the Standard Model of Particle Physics include those by Zee [35] and Peskin & Schroeder [28]

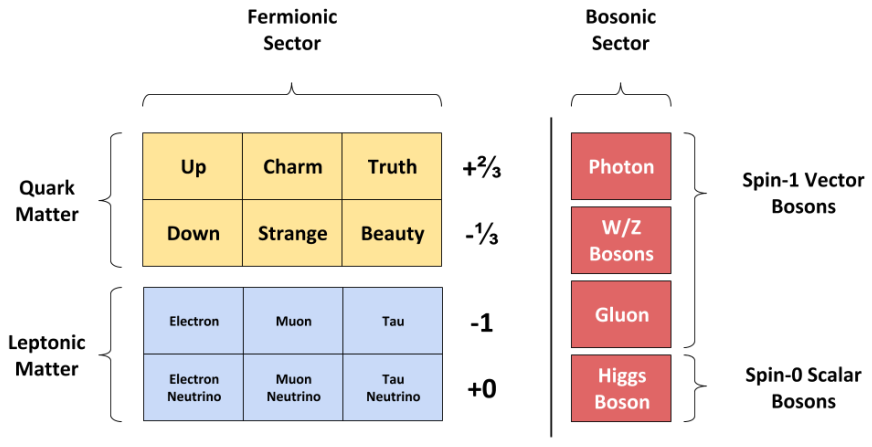

Figure 1: The elementary particles currently part of the Standard Model, with charges listed for the fermionic sector. Every particle additionally has a corresponding antiparticle with opposite charge.

In the unbroken form, the electroweak symmetry describing the unified forces of electromagnetism and the weak interaction is given by $SU(2) \times U(1)$, acting on the massless weak isospin fields W_1, W_2, and W_3, and on the weak hypercharge field B. This symmetry is however broken at the energy where the Higgs field acquires a nonzero vacuum

expectation value (VEV), due to the coupling of the Higgs field to the electroweak fields. This symmetry breaking, among other things, gives rise to the existence of the photon and the Z_0 boson. One way of describing this is shown in Figure 2, where the electroweak mixing angle θ_W induces linear combinations of the B and W_3 fields to represent themselves as the photon and Z_0 fields.

The coupling of the Higgs field to the electroweak fields additionally allows for them to acquire a nonzero mass, with the exception of the photon field. This mass term would be equal for the W and Z bosons, were it not for the electroweak mixing angle mentioned above. The quarks and charged leptons in the Standard Model Lagrangian also have a nonzero mass due to the Higgs field, albeit through their Yukawa coupling to the Higgs field and not directly through electroweak symmetry breaking. For neutrinos the matter is more complicated, as they are experimentally observed to have mass, but the Standard Model provides no explanation for this.

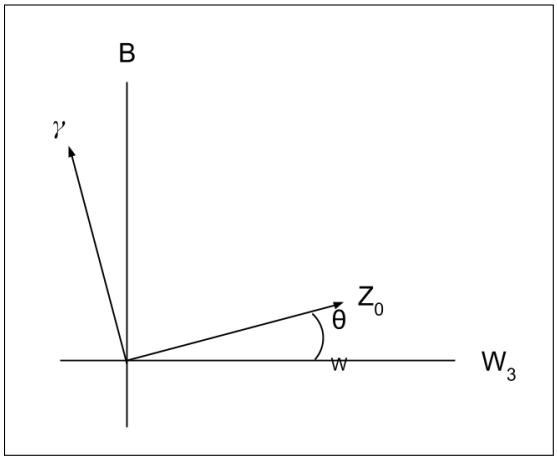

Figure 2: The photon and Z_0 boson fields described as linear combinations of two of the fields present in the unbroken electroweak theory. The mixing angle θ_W nonzero due to the Higgs field acquiring a nonzero VEV.

The mechanism outlined above describes how the charged leptons and gauge bosons acquire mass, but does not touch upon the respective charges of these particles, nor does it explain why the masses are different for the three generations. In the theory of electroweak symmetry breaking, there is no process that alters the respective gauge charges of the leptons, or the coupling strengths in the weak and electromagnetic interactions. The three generations of leptons are thus identical, save for their mass. This is known as lepton universality.

Since a violation of lepton universality would thus imply a violation of the Standard Model in its current form, extensive research is performed in this area. This research takes the form of investigations into the branching ratios of decays which involve the different leptons. Once compensated for mass, an anomaly in these branching ratios would indicate lepton universality violation.

2.1.1 Ongoing Lepton Universality Research

The possible approaches for the study of lepton universality are manifold, and include the leptonic decays of gauge bosons, as well as semileptonic decays involving quarks, and purely leptonic decays such as $\tau^- \to e^- \overline{\nu}_e \nu_\tau$. As example, the decay of the Z boson into leptons has been measured in multiple experiments and has shown to be consistent with the SM prediction [14]. Additionally, the leptonic decay of the W boson also is found to be in accordance with the SM, as shown in Figure 3 [4].

Currently, the strongest deviations from the Standard Model are found in $b \to c l^- \overline{\nu}_l$ transitions in B hadron decays, at a level of 3σ, as seen in Figure 4. The R(D) and R(D^*) values are ratios that express the relative occurrences of decays into the different generations of leptons, and will be described in more detail in the next section. The

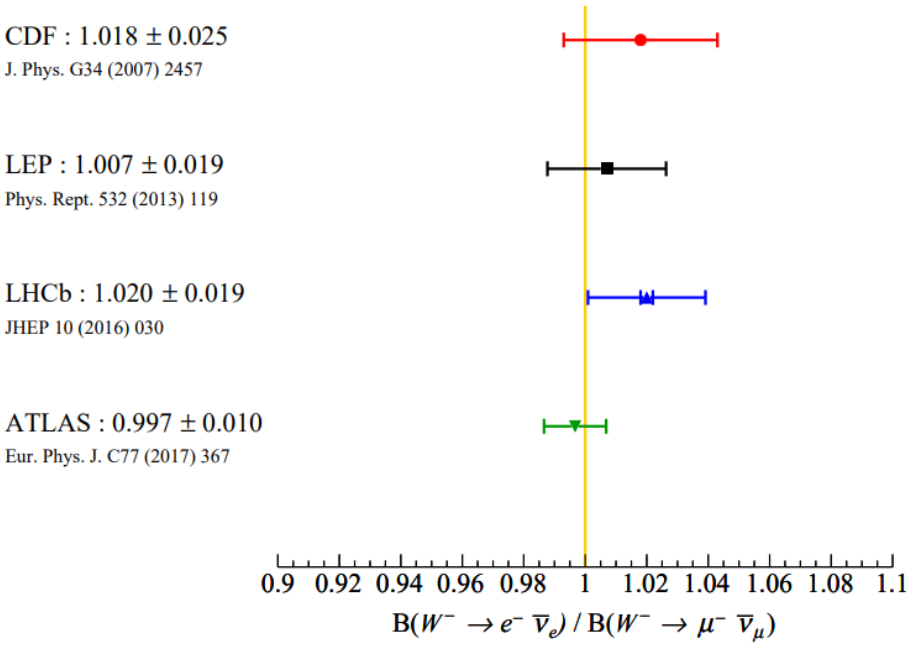

Figure 3: Branching ratios for the leptonic decay of the W boson into electrons and muons, with results for all four collaborations showing accordance with SM predictions. Taken from [4]

D and D^* modes indicate decays to ground state and excited state charmed D mesons respectively.

2.2 Channel Selection

In the previous section, the many channels that allow LU probing to be accomplished were outlined. The current section will describe the channel selected for this research, and why it is such a promising channel.

When searching for the possibly negligible presence of New Physics (NP) in experiments, the use of a channel with a high branching fraction may well drown out the signal that is present. If one is however to investigate a comparatively rare channel, the effects of physics beyond the standard model will have a larger impact on the discrepancy between the expected and measured outcome. This is the primary reason for the selection of the type of channel in this work: $b \to sl^+l^-$. It should be noted that these rare decays are analysed in a much broader context than just lepton universality, being also extensively investigated in the search for charge parity violation.

In the SM, these types of decay feature a Flavour Changing Neutral Current, and are forbidden at tree level. The lowest order diagrams for these transitions are a penguin diagram featuring a quark loop and a Z boson or photon, or a W boson box diagram. The Feynman diagrams for both these transitions are shown in Figure 5. For the penguin diagram, a loop consisting of an up quark will be suppressed, due to the small corresponding CKM matrix [11] element. The transition to a truth quark is energetically unfavourable, leaving the charm loop to be the most probable.

The quantities that are used to gauge lepton universality in these decay types are the R(K) and R(K^*) ratios. These are the ratios of the branching fractions defined as

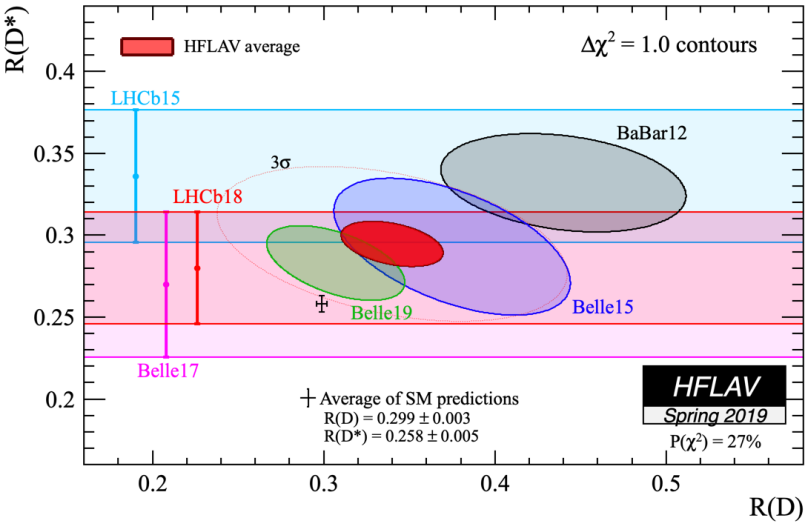

Figure 4: Results for various R(D) and R(D) measurements, averaged by the Heavy FLAVour averaging group. The standard model predictions for R(D) and R(D*) are not unity, due to the influence of form factors that are negligible in other LU measurements. Taken from [1] (online update 2019)*

follows:

$$R(K) = \frac{\mathcal{B}^+ \to K^+\mu^+\mu^-}{\mathcal{B}^+ \to K^+e^+e^-}, \quad R(K^*) = \frac{\mathcal{B}^0 \to K^{*0}\mu^+\mu^-}{\mathcal{B}^0 \to K^{*0}e^+e^-} \tag{1}$$

Prior results for the measurement of these ratios can be found in Figure 6. The measurements performed by the LHCb collaboration show values of $R(K)$ and $R(K^*)$ that "... are found to be 2.6 σ and 2.1-2.5 σ lower than SM expectations." [4] These recent 'B-anomalies' are discussed in e.g. reference [17].

2.2.1 Normalisation Channel

An additional property of the $b \to sl^+l^-$ channel is the presence of the J/ψ decay mode. In this mode, the penguin diagram features the J/ψ resonance in the form of a charmonium loop. This resonance mode is heavily favoured over other modes and can be used as a normalisation channel to facilitate the determination of R(K) and R(K*). Since the branching ratios of the J/ψ resonance into electrons and muons is measured to be 0.9983 ± 0.0065 [31], a double ratio can be used to calculate R(K) and R(K*) (shown here only for R(K)):

$$R(K) = \frac{\mathcal{B}^+ \to K^+\mu^+\mu^-}{\mathcal{B}^+ \to K^+e^+e^-} \times \frac{\mathcal{B}^+ \to K^+J/\psi(e^+e^-)}{\mathcal{B}^+ \to K^+J/\psi(\mu^+\mu^-)} \tag{2}$$

By using this double ratio, many of the unknown variables regarding the detection efficiency of muons and electrons cancel out, and systematic effects can also be better kept under control.

In the remainder of this work, the channel that will be used is:

$$\mathcal{B}^0 \to K^{*0}(K^+\pi^-)J/\psi(e^+e^-) \tag{3}$$

All data and descriptions will refer to this channel, unless explicitly mentioned.

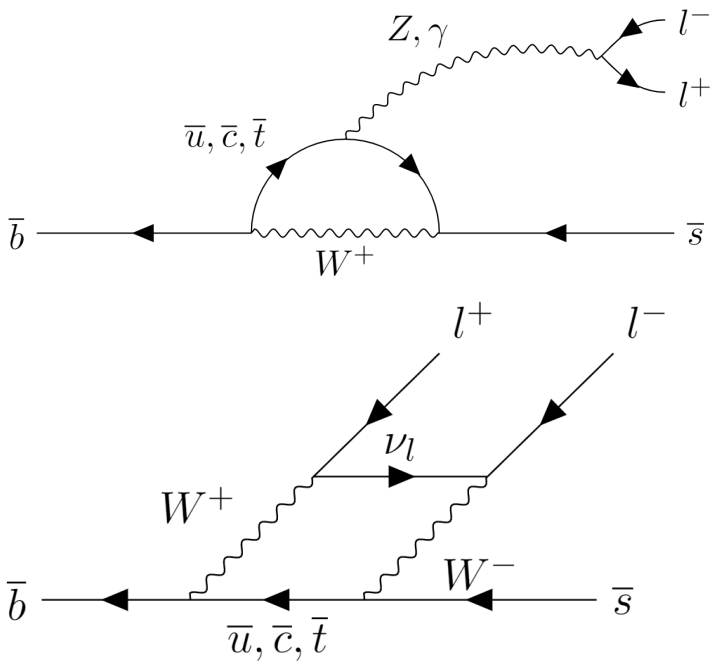

Figure 5: Feynman diagrams for the lowest order $b \to sl^+l^-$ transitions. At the top is a penguin diagram with a quark loop and a Z boson or photon. At the bottom a W boson box diagram.

2.3 Alternative Theories

The most probable cause for NP to reveal itself in the aforementioned channels would be a tree level diagram featuring a previously undiscovered particle. Being of lower order than the decay channels previously discussed, a tree level diagram would be less suppressed and would thus have a higher chance of discovery. Two competing theories that attempt to explain flavour changing neutral currents are the leptoquark and heavy boson models.

Leptoquark Model The leptoquark model solves the issue of FCNC with the addition of a particle that carries both leptonic and baryonic quantum numbers. The particle would be created at high energies, with a mass comparable to that of a lead nucleus. In Figure 7, a beauty antiquark decays into a positively charged lepton and a leptoquark, which subsequently decays into a negatively charged lepton and strange antiquark. An effective field theory incorporating this leptoquark is discussed in references [5, 6]. This particular model is interesting as it is able to explain the natural hierarchical structure of fermions, as well as being UV complete. The latter being a requirement for any theory describing physical interactions between particles.

Heavy Boson Model Heavy boson models assume the existence of so-called Z' and W' bosons, through an extension of the weak interaction. The exact properties of these bosons vary per theory, but they all share the requirement of a higher mass than the known bosons. Any Z' boson would differ from the known Z boson in that it would be able to decay into a quark-antiquark pair of different generations, in order to account for FCNC. Papers detailing the search for a possible Z' boson can be found in references [13, 10, 18]

With the importance of lepton universality measurements emphasised, the next chapter will describe the LHCb experiment and how it can be used to collect relevant data.

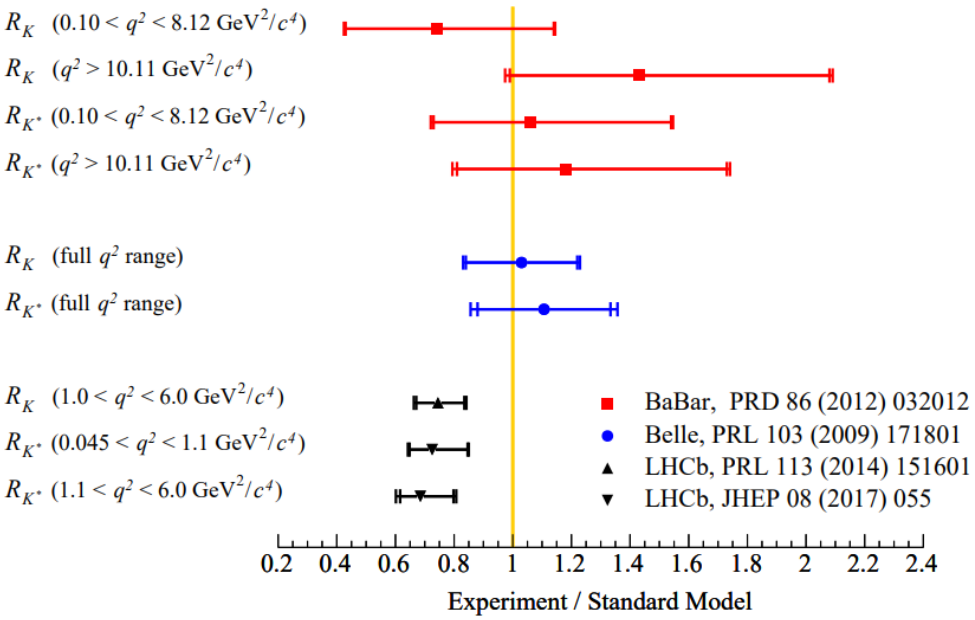

Figure 6: Prior results for the branching ratios R(K) and R(K). The quantity q^2 refers to the invariant mass of the dilepton pair. Integration is performed over the full q^2 range to obtain R(K) and R(K*). Taken from [4]*

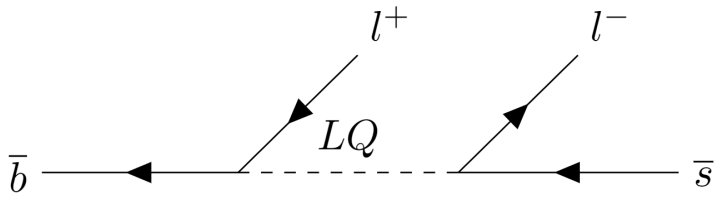

Figure 7: Feynman diagram showing tree level FCNC through a theorised leptoquark particle.

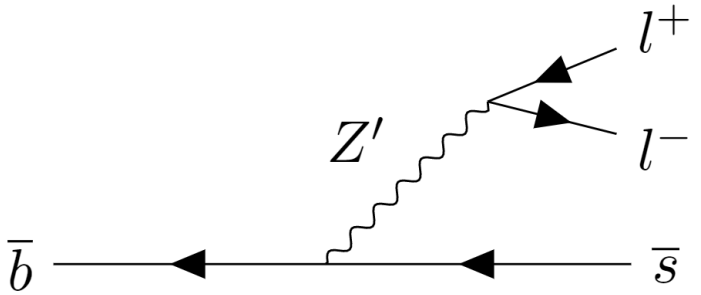

Figure 8: Feynman diagram showing tree level FCNC through a theorised heavy boson.

3 Large Hadron Collider Beauty Experiment

The LHCb experiment is one of the major experiments situated on the Large Hadron Collider (LHC) machine, which can provide an abundance of b-flavoured hadrons through proton-proton collisions. It is located at the CERN facility near Geneva, beneath the France-Switzerland border and consists of a 27km diameter circular particle collider capable of achieving 14TeV collision energies. The primary colliding particles are protons, but experiments are also conducted using heavy ions. The protons are fed into the machine through a series of prior accelerators, both linear and circular, at an energy of several hundred GeV. Particles are injected in so called 'bunches', consisting of more than 100 million protons per bunch, with a minimal bunch spacing of 25ns. They are subsequently accelerated to the required centre-of-mass energy through the use of Radio Frequency or RF cavities. The cavities serve two purposes: The first is the aforementioned acceleration, and the second is to maintain a tight clustering of the protons in a bunch. This is achieved through the shape of the potential in the cavity, which accelerates slower protons more than faster ones, causing all off-centre particles to oscillate in the longitudinal direction. The beams cross each other at four points along the ring, and experiments are located at and around these stations.

The LHCb detector is a single-arm forward spectrometer at the LHC, designed for the study of heavy flavour physics [15]. For the current LHCb setup, there is on average a single collision per bunch crossing. The main production process for the B-mesons studied at LHCb is gluon fusion, described in reference [2]. The resulting B-mesons are predominantly boosted in either the forward or backward beam direction and the detector geometry reflects this. The maximum acceptance of the subdetector modules is 300mrad in the (horizontal) magnet bending plane, and 250 in the (vertical) non-bending plane. The choice for these acceptances is based on encompassing as much of the particle momentum range as possible, at acceptable cost. The minimum acceptance is 10mrad, and is governed by the beam pipe having to pass through the subdetectors.

The LHCb collaboration employs a typical right-handed coordinate system for the detector, with the z-axis facing along the beam line, the x-axis laying in the magnet bending plane, and the y-axis fixed near-vertically.

A large magnet with an integrated field of roughly 4 Tm is used as the primary means of particle momentum reconstruction. Combining the equation for the Lorentz force with an infinitesimal deflection angle, one can obtain the following:

$$\frac{d\theta}{ds} = \frac{q}{p_{\parallel}} B(\hat{v}_s \times \hat{B}) \tag{4}$$

Here, ds is an infinitesimal displacement along the direction of travel and $p_{\parallel}$ is the momentum component in the direction of travel. From this equation it is apparent that the deflection angle per unit of traversed distance is dependent only on the charge, momentum, and field strength. Therefore, an evaluation of the direction of travel of a particle before and after the magnet is used as a means of determining the particle's momentum. As a whole, the reconstruction of the particle trajectories, or tracks, making up a hadron decay is primarily based around the determination of the constituent particles' four-momenta.

The next few sections will cover the different subdetectors in detail. A clear distinction is made between subdetectors before the bending magnet, dubbed the 'upstream' detec-

tors, and those behind the magnet, the 'downstream' detectors. Much of the information in this chapter can be found in more detail in the LHCb design document [15].

3.1 Upstream Detection

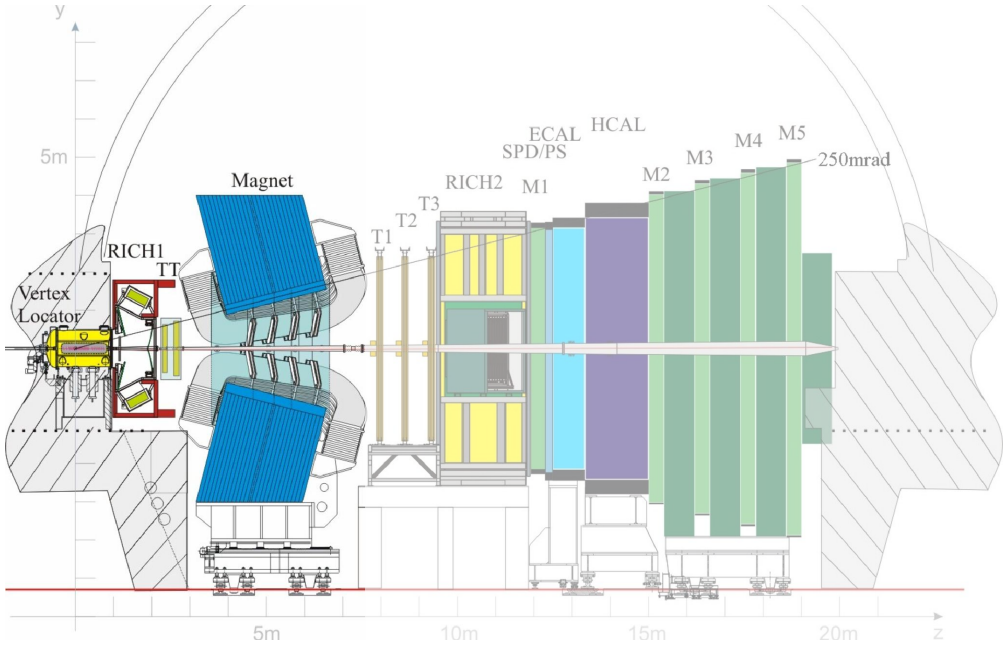

Figure 9: Schematic view of the LHCb detector in the y-z plane. Particles traverse the detector from left to right. View is focused on the upstream subdetectors and the magnet, downstream components are made transparent [15].

In this section the upstream detectors will be discussed. These subdetectors are particularly important in this work as they are the primary source of information on the particles that are bent out of the detector acceptance by the magnet. As these detectors are located prior to the magnet, the particles in these detectors will mostly travel straight from their origin vertices, save from some bending due to the magnetic field permeating outside of the magnet.

3.1.1 VELO

The VErtex LOcator is the first detection instrument, located closest to the interaction zone where the protons collide. It consists of a series of silicon strip detectors, arranged in planar modules and mounted perpendicular to the beam axis. Silicon strip detectors are the one of the primary means of particle detection in the detector. Charged particles travelling through the detector strip bulk will free up charges that are collected and amplified to obtain a 'hit' signal, as illustrated in Figure 11. Alternating VELO modules respectively measure the radial distance r and the azimuth angle ϕ from the beam axis, demonstrated in Figure 10. The spacing between subsequent modules is kept small and the distance to the beam is kept to a minimum to allow for an accurate interpolation between detector hits. This accuracy is required to detect the displacement of secondary decay vertices from primary decay vertices, which is of the order of a centimetre due to the comparatively short lifetimes of B-mesons. The choice for an (r, ϕ) coordinate system in favour of a Cartesian coordinate system is due to the increased computing efficiency that it provides.

When the beam is first injected into the LHC, the width of the beam is larger than during nominal data taking. To prevent radiation damage to the VELO and to avoid

interactions with the beam, the VELO modules can be retracted from their data-taking position close to the beam, to a safer distance of 35mm. The VELO modules are placed in a separate vacuum from the primary beam vacuum, with a thin aluminium shielding ensuring this separation.

The VELO is the first significant amount of material that particles will encounter and, as discussed in section 3.3.2, is thus the location of many of the Bremsstrahlung emissions. These emissions will be used later in this work for improved particle identification. Additionally, the vacuum in which the VELO operates is kept separate from the vacuum of the beam pipe, which is higher. As the VELO needs to be retractable, so does this separation, which is called the 'RF foil'. This foil also contributes considerably to the amount of material present at the VELO subdetector.

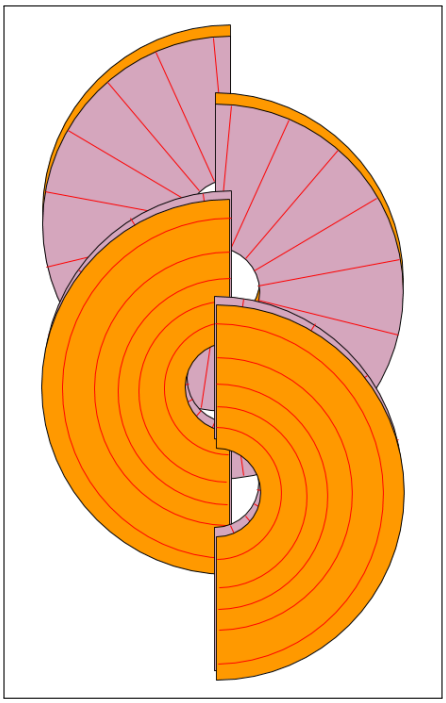

Figure 10: Schematic view showing the mounting scheme of the r-sensors (orange) and the ϕ-sensors (pink). This situation represents the data-taking position, where the modules are closest to the beam.

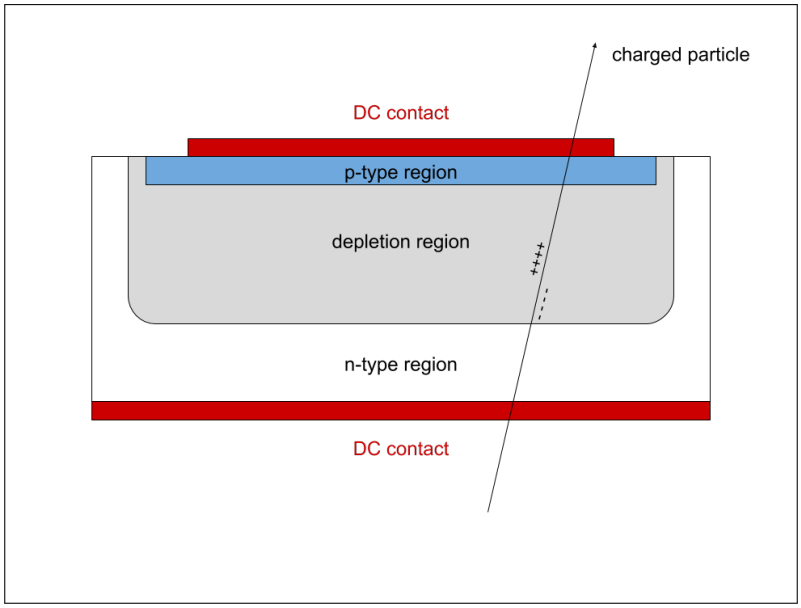

Figure 11: Schematic view of a silicon microstrip detector. Passing particles create electron-hole pairs that are collected in the electrodes. Adapted from [3].

3.1.2 RICH1

The Ring Imaging CHerenkov detector is one of the primary means of particle identification. The principle of operation is the emission of a cone of Cherenkov radiation when a charged particle passes through a medium where it has superluminal phase velocity. The emission angle of the radiation with respect to the direction of travel increases with increasing velocity, and pairing this information with momentum and charge information from other detectors allows for discrimination between particles of varying masses and thus identities.

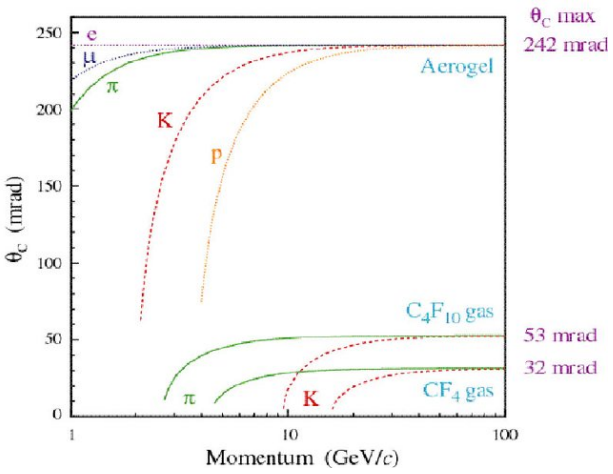

Figure 12: Relation between Cherenkov emission angle and particle momentum for various common particles in the LHCb detector for the three media used in the RICH1 and RICH2 detectors [15].

RICH1 is located upstream of the magnet and covers the full angular acceptance range of the LHCb detector. The medium used in the detector is C_4F_{10}, covering a momentum range of about 10-65GeV/c, suitable for the particles with a higher off-beam angle that generally carry lower momentum. Lightweight spherical mirrors are used to reflect the radiation to photon detectors whilst minimising possible interactions with event particles.

As the Cherenkov emission angle follows $sin(\theta_c) = \frac{c}{nv}$, particles with increasing momentum become increasingly hard to distinguish. This effect is demonstrated in Figure 12 for various particles for the media in both the RICH1 and RICH2 detectors. The lower momentum upstream particles thus benefit from a more accurate RICH1 reading, which

aids in particle identification.

3.1.3 TT

The Tracker Turicensis, or Trigger Tracker, is a subdetector containing four layers of silicon strip detectors. The four layers are split in two closely spaced pairs, with a larger distance of 27cm between the pairs. In each pair, one of the layers' detector strips are oriented vertically, with the other layer rotated by 5 degrees. Combining the hits in both layers allows for the reconstruction of the location of a particle traversal in the x-y plane with a high resolution in the bending plane. The design for this specific type of detector was chosen in order to achieve a resolution of roughly 50μm, which was deemed sufficient for the experiment.

It should be noted that due to the proximity of the TT station to the magnet, there is a residual field permeating the detector, resulting in some bending for charged particles travelling through this station.

3.2 Downstream Detection

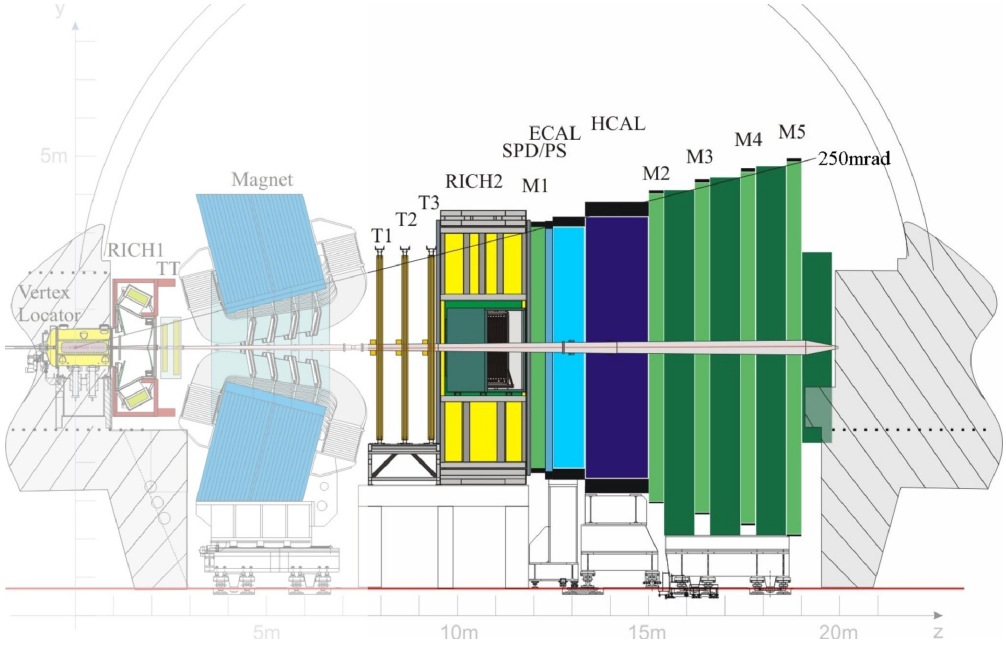

Figure 13: Schematic view of the so-called downstream portion of the LHCb detector [15].

The following subsections cover the downstream detectors, located behind the bending magnet. Particles traversing both the upstream and downstream detectors can have their four-momentum measured and benefit from increased particle detection efficiency due to crossing both RICH stations. Their tracks are hereafter referred to as 'long' tracks. Since the primary focus of this work is on upstream particle tracks, the descriptions will be brief except where applicable to upstream particles. Detailed information can be acquired from the LHCb design document [15].

3.2.1 T1-T3

Three tracking stations are located downstream of the bending magnet, their inner sections (∼120cm x 40cm) consisting of the same silicon strip detector technology as the TT. The outer sections span up to 600cm x 490cm and employ so-called straw tube detector technology. Vertically mounted tubes of 5mm diameter contain a mixture of

argon and carbon dioxide that can be ionised by passing particles. The freed charges drift towards a centrally mounted anode wire where they are collected, creating the signal. By also timing the drift of the charges, a resolution higher than the 2.5mm tube radii can be achieved, upward of 0.2mm.

The tracking stations provide information about particle momentum through the mechanism covered earlier in this chapter.

3.2.2 RICH2

The RICH2 detector provides particle identification capability for the 15GeV/c-100GeV/c momentum range. The acceptance for this detector is 120mrad from the beam axis in the bending plane, and 100mrad in the non-bending plane. The use of CF_4 which has a lower refractive index, allows for the discrimination of the high momentum particles that are expected to fall within the angular acceptance of this detector.

3.2.3 SPD/PS

The Scintillating Pad Detector and PreShower modules are the first of the calorimeter modules, which additionally comprise the Electromagnetic calorimeter and the Hadronic calorimeter, both of which will be discussed in the next section. All calorimeter modules are based on the same scintillating material detectors that emit photons when hit by a charged particle. The photons are collected by an optic fibre and guided to a photo multiplier tube to obtain a signal.

The primary task of the SPD and PS detectors is to differentiate between different particles. The SPD is the first module to be encountered by particles and simply provides a single bit signal to indicate a hit. Since neutral particles will rarely activate the scintillator material, this provides a means of distinguishing neutral from charged particles. The SPD is followed by a 12mm lead plate, the purpose of which is to initiate a shower of secondary particles when hit by incoming particles. The presence and shape of this shower depends a great deal on the instigating particle, which is used in particle identification. Photons and electrons will readily interact with the lead and produce electromagnetic showers that are detected by the PS module, pions and other hadrons have much longer shower development lengths and can thus be discerned. The separation of the most common particles is summarised in Figure 14.

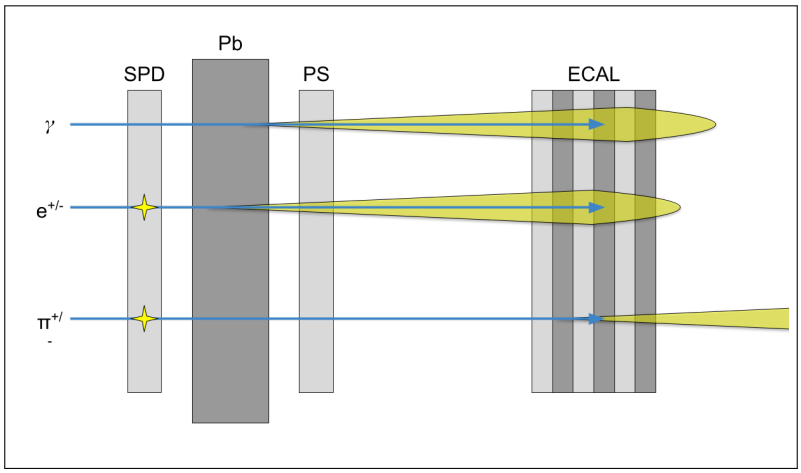

Figure 14: Illustration of the method for separating the most common particles in the LHCb detector through their characteristic particle shower.

3.2.4 ECAL/HCAL

Located downstream of the SPD and PS are the Electromagnetic CALorimeter and the Hadronic CALorimeter. These detectors employ scintillator detector technology to measure the energy of incident particles through the creation of electromagnetic and hadronic secondary particle showers respectively. In the ECAL, so-called shashlik technology is used, where 4mm layers of scintillator are alternated with 2mm layers of lead. In order to encompass as much of an electromagnetic shower as feasible, a detector thickness of 25 radiation lengths was chosen, resulting in the ECAL module being 42cm thick.

Hadronic showers are generally much longer than their electromagnetic counterparts and will thus have excess energy when leaving the ECAL. For this reason, an additional calorimeter is installed after the ECAL that is aimed towards the energy reconstruction of hadrons. The HCAL has a different design than the ECAL, featuring a higher granularity. The scintillating pads are mounted in the y-z plane as opposed to the x-y plane. The pads are alternated with iron plates to form a submodule, and submodules are in their turn separated by iron sheets, the large iron nuclei acting as absorbers to initiate hadronic showers.

The upstream particles at the focus of this work will not reach the calorimeter systems. There is however important data to be gathered through their emission of bremsstrahlung photons, which can be detected by the calorimeters. As will be described in chapter 4, the energy deposits, or 'clusters', that are left by the bremsstrahlung photons can in certain cases be associated with their respective particles, which can aid in their identification.

3.2.5 M1-M5

The five muon stations are tasked with detecting the barely interacting muons produced in some of the key decays in the LHCb detector. The primary detection components are compartments filled with a mixture of argon, carbon dioxide, and CF_4 that can react electromagnetically with the passing muons. The compartments provide a singular signal when hit and the resolution is thus limited to the size of the compartments. One of the muon stations is located before the SPD detector, the other four are downstream of the HCAL. The four downstream stations are interspersed with 80cm thick iron absorber slabs, to increasingly filter out all but the highest momentum muons. This primarily entails the filtering of hadrons. A momentum of 6GeV/c is required for a muon to traverse all the stations.

A particle track passing through all muon stations is a telltale sign of a real muon, making particle identification easy in that particular case. Other particles, such as electrons, are oftentimes not so easily identified due to their different properties. The next section will cover the behaviour of electrons in the LHCb detector and the associated difficulties in their identification and detection.

3.3 Electrons in the LHCb Detector

The range of particles that are directly detected by the LHCb detector is small in comparison to the amount of particles that are produced in the studied decays, whose existence is inferred through the detection and measurement of secondary decay particles. Electrons are one class of particle that are detected directly, but have some properties that make them not ideal for this purpose. The lightest of the charged leptons, the electron is produced in abundance in many weak and electromagnetic interactions. The low

mass however means that the electron is prone to losing significant fractions of its momentum in interactions with the detector material. Whilst this is desirable for example in the calorimeter system, where the aim is measure the total momentum of the particle, it can throw off track reconstruction if it occurs in non-monitored material.

Figure 15 shows the loss of energy in the form of bremsstrahlung of electrons prior to reaching the calorimeter systems, which averages 38% for electrons with associated bremsstrahlung clusters. The bremsstrahlung emissions will cause the electrons to deviate from their mostly straight trajectories, increasing the uncertainty in their direction of motion and, in turn, their momentum. Even more troublesome are the electrons that are either created with relatively low momentum, or that lose a significant fraction of their momentum before the bending magnet. These upstream electron will not be detected in the downstream detectors and will lack information critical to their correct identification, primarily the ratio $\frac{E}{P}$ which, for electrons, will be often close to unity due to their tendency to deposit a lot of their energy in material.

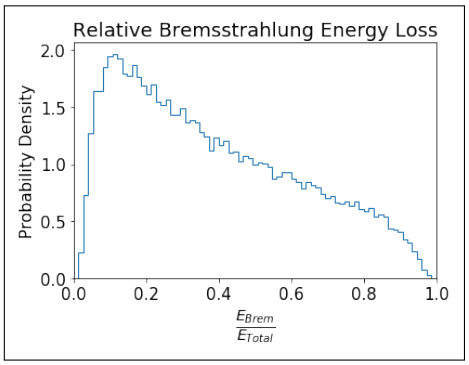

Figure 15: Typical fractional loss of energy in the form of bremsstrahlung emission for electrons in the LHCb detector. $B^0 \to J/\psi(e^+e^-)K^*(K^+\pi^-)$ Channel, simulated data.

3.3.1 Channel Backgrounds

There are many types of backgrounds that hinder accurate measurements in the LHCb detector. As an example, if the pion is not detected in the decay $\mathcal{B}^0 \to K^{*0}(K^+\pi^-)J/\psi(e^+e^-)$, it would then clutter the $\mathcal{B}^+ \to K^+e^+e^-$ channel as a partially reconstructed event. The relevant types of background for this work are single and double misidentifications, and ghosts. Single misidentifications occur when one particle in the decay has been incorrectly identified. Double misidentifications include 'swaps' where for example the electron and the pion in $\mathcal{B}^0 \to K^{*0}(K^+\pi^-)J/\psi(e^+e^-)$ have been swapped, leading to incorrect decay reconstruction. Ghosts refer to particle trajectories that do not originate from a real particle, but are instead collections of observations that by chance resemble a real particle.

To indicate the extent of background pollution of a sample of upstream tracks, Table 1 shows particles that have been identified as being electrons. The electron row shows true positives, whereas the other rows show false positives. This table applies to simulated upstream tracks from the data set used for this work, and the total of 1 345 events is typical for a 10 000 event total sample size. The table demonstrates that misidentification of electrons is a considerable issue for upstream tracks, with less than half of the electrons being 'true' electrons. Furthermore, it can be seen that the contribution of ghost tracks is sizeable, revealing the troublesome nature of combinatorial backgrounds.

Table 1: Simulated particles identified as being electrons. The electron row shows true positives, the other rows show false positives. Sample is representative for upstream electrons in a 10 000 event total sample size, i.e. upstream and long tracks.

Particle	**Quantity**	**Fractional**
Electron	567	42.2%
Pion	440	32.7%
Ghost	201	14.9%
Kaon	121	9.0%
Proton	12	0.9%
Muon	4	0.3%
Total	**1345**	**100%**

3.3.2 Bremsstrahlung

Bremsstrahlung, literally "braking radiation", is the type of radiation that is emitted during the acceleration of electric charges. In the context of the LHCb experiment, bremsstrahlung mostly refers to charged decay products interacting with the static nuclei and electrons of the detector material. In these interactions, the free particle is typically slowed down, hence the name of the radiation.

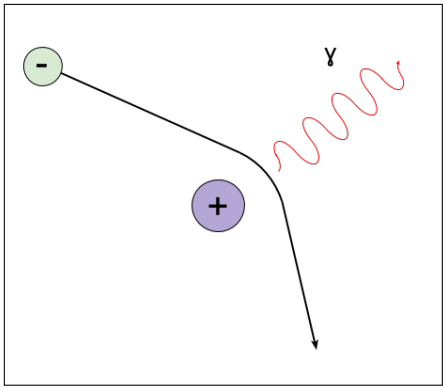

Figure 16: Bremsstrahlung occurs when the Coulomb interaction accelerates an electric charge, the acceleration is paired with the emission of electromagnetic radiation.

The power emitted in a bremsstrahlung process can be calculated both classically and relativistically. The classical equation for bremsstrahlung power emission is given by the Larmor formula [21], and a relativistic generalisation can be derived by treating moving point charges as Liénard–Wiechert potentials [34]. The latter being a relativistic description of Maxwellian potentials.

Both equations listed above however assume a vacuum, and the average bremsstrahlung-related interaction of a particle within any physical detector is more complex. An in-depth description of bremsstrahlung, both general and specific to the LHCb detector, can be found in reference [33].

In a work by W. Heitler [23], an estimate is made for the cross section of bremsstrahlung photon emission for highly relativistic particles:

$$\sigma \sim \frac{Z^2}{137}\left(\frac{e^2}{mc^2}\right)^2 \tag{5}$$

Here, Z stands for the nuclear charge of the particle and m for the mass. From this equation it becomes immediately apparent why electrons, being roughly four times lighter than even the up quark, are the most affected by bremsstrahlung. As the cross section scales with the inverse mass squared, the electron will be most prone to bremsstrahlung emission, due to it being the lightest charged particle.

Bremsstrahlung in the LHCb detector is generally considered a nuisance, especially for electrons. This is mainly due to the increased difficulty in accurate reconstruction of the particles' tracks. In what follows, the information contained in bremsstrahlung emissions will be put to use as a tool for better identifying electrons.

4 Electron Identification

This chapter will cover the methods employed in improving the particle identification of electrons. The goal is to find signature properties of an electron traversing the detector that are captured in the outputs of the detectors. These properties can then be used to train a machine learning classifier to better discern electrons from non-electrons. The first section will elaborate on the required data and how it is obtained from the detector. After this, section 4.2 will elaborate on how extra bremsstrahlung-related information is acquired from the data. Finishing the chapter are two sections that dive into the search for high-performance classifier variables and the application of machine learning.

4.1 Data Gathering

The LHCb collaboration conducts research on many different decay channels, and in order to attain a manageable quantity of data, this work focuses on a specific decay channel. As mentioned earlier, the $B^0 \longrightarrow J/\psi(e^+e^-)K^*(K^+\pi^-)$ normalisation channel is chosen for this purpose, due to its relevance for lepton universality tests as well as it being a relatively well-studied channel. Furthermore, this channel contains a representative sample of electrons, kaons, and pions, from a kinematic standpoint. A drawback to this selection is that an extension of the methods used in this work to other channels may not be warranted, but steps are taken to attempt to improve generalisability.

4.1.1 LHCb Software

It is informative to examine the interface between the physical LHCb detector and the data that is most commonly used by researchers in their analyses. A variety of software frameworks and packages are used in this process, which will be outlined in this section. In Figure 17 the flow of detector output and simulation output is shown.

There are, on average, about 30 million proton-proton collisions occurring per second in the LHCb interaction point. If all events would be recorded, a constant stream of 1TB of data would have to be stored. As this is currently not feasible, several stages of filtering are applied. Currently, the first filtering step is the so-called hardware trigger, which is built into the detector equipment. The hardware trigger is required to accept or reject a certain event on a 4 μs basis [22]. The only subdetectors that are capable of achieving such a high rate of data acquisition are the muon stations and the calorimeter system, and the trigger conditions are thus based on these systems. Particles with high

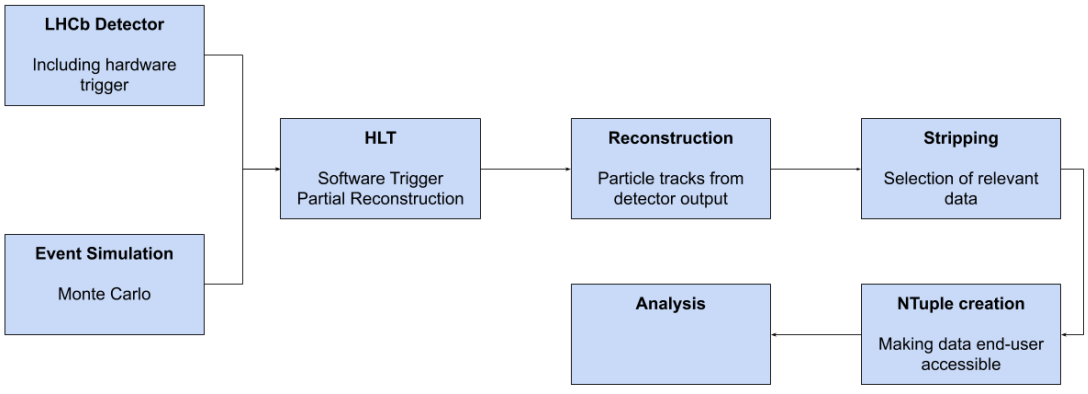

Figure 17: Simplified scheme illustrating the process of converting the raw detector output and Monte Carlo simulated events into data that can be used in analyses.

off-beam momenta are often signatures of interesting events, and trigger conditions are thus based on specific transverse momenta (p_T).

The High Level Trigger (HLT) is a software based trigger system consisting of two stages. Events that are accepted by the hardware trigger will be partially reconstructed in the first stage, HLT1. The partial reconstruction entails the creation of tracks from hits in the VELO that are extrapolated to the tracking stations if there are matching hits in the muon station or if the VELO track is substantially displaced from the primary vertex. This displacement, or impact parameter (IP), is a sign of a high-energy secondary particle and thus reason for accepting the event. The HLT1 runs in real-time, meaning that as long as there are events occurring in the detector, the trigger will be active and recording events to storage. Accepted events are subsequently evaluated by the HLT2 trigger, which can run both in real-time, using excess computational resources, and when the beam is offline. For the HLT2 stage a more thorough reconstruction is performed, where all tracks above a p_T of 300 MeV/c are reconstructed [22]. It should be noted that at this stage, a track is simply a collection of hits in the detector which is consistent with originating from a single charged particle. This consistency is based on properties such as the χ^2 value describing the displacement between a hit and the suspected particle trajectory. Whilst information from e.g. the RICH systems can be associated to the track at this point, there is not a definitive identity assigned to the particle yet. Events accepted by HLT2 are saved to storage and are ready for offline processing. Trigger conditions for both the hardware and software triggers are subject to frequent changes and allow researchers a degree of control in the types of decays that can be investigated.

The offline processing stage is where full event reconstruction takes place. The first step in the full reconstruction of all the tracks in a given event is to determine which of the hits in the subdetector systems belong to a track. To this end, pattern recognition software is employed that attempts to match straight sections of tracks in e.g. the VELO to hits in the tracking stations. The various types of tracks are shown in Figure 18. If enough hits are compatible with the hypothesised track, it is accepted. Since not all particles will traverse all of the subdetectors, different pattern recognition methods are used to exhaust as many of the detector hits as possible. For upstream particles, straight track segments from the VELO are extrapolated to the TT station, assuming a transverse momentum of 400 MeV/c, chosen as an appropriate average of simulated events [7]. Using the extrapolated track and accompanying momentum assumption an evaluation of the compatibility of TT hits is performed. If three of the four TT stations have hits within a given deviation, the track and constituent hits are accepted as an upstream track. On average, every event will see the reconstruction of about 72 tracks, mainly long tracks and VELO-only tracks [15]. Note that this figure is based on simulated $b\bar{b}$ events. For an in-depth description of the track-fitting procedure for charged particle tracks, refer to the work of J. van Tilburg [32].

Once pattern recognition is complete, a track-fitting procedure is performed that finds the best possible track parameters given the associated hits, and which also calculates χ^2 values to show the consistency of the resulting track. Tracks are stored to memory as a collection of track states: Vectors consisting of an x and y coordinate, $\frac{dx}{dz}$ and $\frac{dy}{dz}$ derivatives, and the scalar quantity $\frac{q}{p}$. The entire vector is a function of the z coordinate. The track as a whole is embedded in a so-called protoparticle object, which is also used to store additional information from e.g. the calorimeters and the RICH detectors. Bremsstrahlung-related information, if available, can also be assigned to this protoparticle container, detailed in section 4.2. An artificial neural network (ANN) is supplied with the protoparticle information and calculates pseudo-probabilities for the

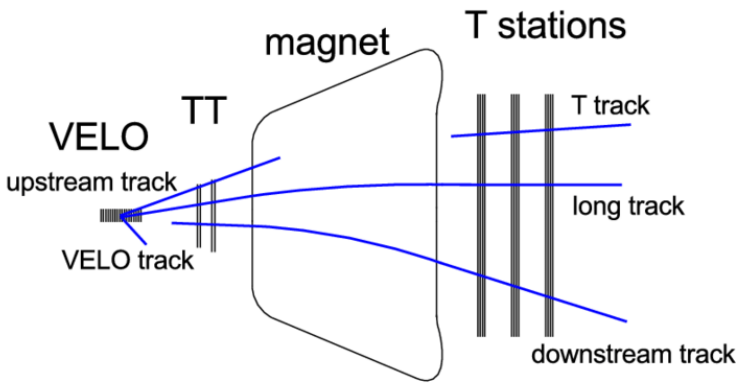

Figure 18: Schematic representation of the different track type definitions used for the LHCb experiment. Taken from [30].

protoparticle being one of the particles listed in Table 2. A 'ghost' particle is a result of detector noise, ambient particles, and other effects creating hits in the subdetectors that resemble a single, real particle as described in section 3.3.1.

Table 2: The variable names used for the assignment of ANN outputs. The ANN takes protoparticle reconstruction data and gives a pseudo-probability for the protoparticle to be one of the listed particles. A 'ghost' particle refers to a track that is wrongly attributed to have originated from a real particle.

Particle	**Associated ANN output variable**
Electron	ProbNNe
Muon	ProbNNmu
Pion	ProbNNpi
Kaon	ProbNNk
Proton	ProbNNp
Ghost	ProbNNghost

Up until this point the protoparticle has had no identity assigned to it yet, and the event as a whole has not been shown to contain any specific decay. In the 'selection' phase, all events are searched for so-called 'candidates'. In the broadest sense, this selection takes a signature of a decay, consisting of e.g. mass and momentum constraints, impact parameter requirements, or whatever else is required by the researchers, and looks for this signature in all the stored events. For example, the decay investigated in this work features a J/ψ meson with an invariant mass of 3.1 GeV/c^2, which subsequently decays into two electrons. This translates into a constraint on the invariant mass of the combination of electrons to lie between 2.1 GeV/c^2 and 4.3 GeV/c^2, to fully encompass the width of the J/ψ resonance plus resolution-related tails. It is entirely possible for an event to satisfy the conditions of signatures for different decays, or to even satisfy the same signature multiple times. The latter is possible when a different assignment of particle identities also satisfies the signature conditions. It is obvious that at most one arrangement can be correct, but all candidates are tagged for analysis nonetheless. The selection process is computationally intensive and is only performed at most a handful of times each year.

The final step in the data-chain is to collect all information into a convenient format. The ROOT software kit has been developed by the High Energy Physics community to deal with large amounts of data in an efficient way [12]. A core object in ROOT is the NTuple, a data structure specialised for fast data access. An NTuple can be filled by so-called TupleTools, classes that contain the necessary methods to access relevant information in the event and to calculate derived quantities if needed. The researcher will write a configuration file containing references to the required TupleTools, extra cuts

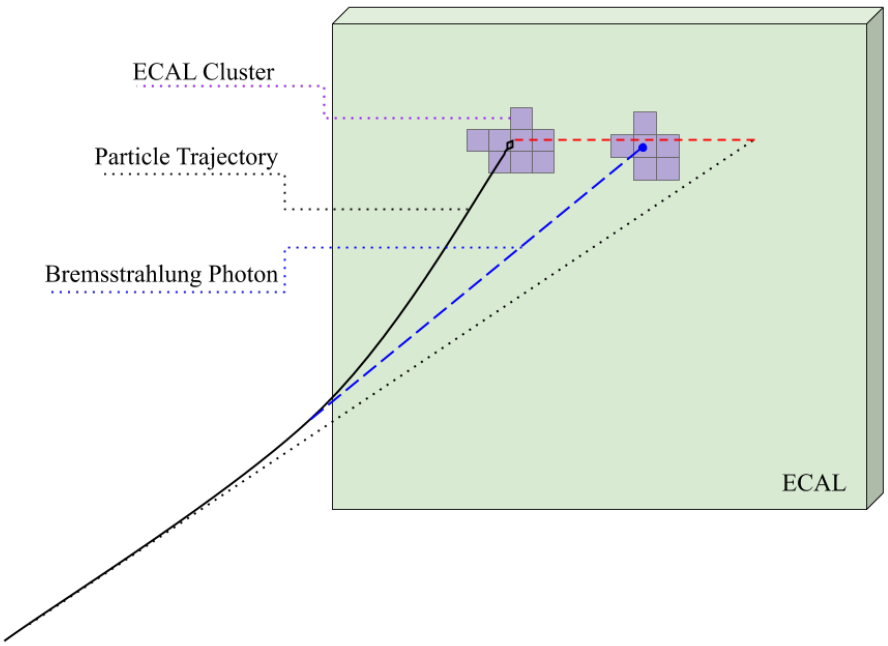

Figure 19: The BremAdder algorithm attempts to find clusters of photons between the impact point of a particle and an extrapolated section of VELO track

and selections that may be desired, or customised information that must be extracted from the event. This configuration file is then run by the LHCb analysis software: DaVinci, run on the Gaudi framework [16]. DaVinci evaluates the stored data and outputs the requested NTuple, ready for analysis.

The description of the data-chain above concerns the physical data taking at LHCb, the simulated data follows mostly the same route, though with obviously different origins. The simulation process is mainly governed by Gauss, the LHCb simulation framework. At first, proton-proton collisions are simulated, commonly using the Pythia application. The generated events subsequently will see the simulated decay of the constituent particles through the EvtGen [25] application. The Geant4 framework then propagates the resulting particles through the LHCb detector, attempting to replicate the intricate material and magnetic field interactions that the particles may undergo. The final step in the simulation process is the conversion of the simulated hits to digital detector output, this is accomplished by the Boole application. The digitised output is made to mimic the physical detector output as closely as possible, and is fed into the HLT trigger to follow the same data-chain as the real data.

The procedures described in this section culminate in the acquisition of a data set that contains primarily the requested decays. In the case of this work these are simulated events, which come with extra information that would normally not be revealed by nature.

4.2 BremAdder Algorithm

As detailed in section 3.3, electrons will often lose a significant fraction of their energy in bremsstrahlung processes. If this energy is not accounted for, electron momentum reconstruction will be severely hampered. The BremAdder algorithm is designed to mitigate this issue as well as possible by adding compatible clusters of photons in the ECAL to the particle track. The BremAdder algorithm is run in HLT2, but is compu-

tationally intensive. As an alternative, a rudimentary form of bremsstrahlung matching is performed, which will be discussed first.

The CaloBremMatch algorithm is fast enough to be run during online reconstruction, but sacrifices accuracy and capabilities. The matching method takes an ECAL photon cluster and a particle track and produces a χ^2 value for the match between the two. To accomplish this, the function finds the track state closest to the TT and extrapolates in the direction of travel to the z-value of the cluster. The χ^2 value is then calculated based on the distance between the extrapolated impact point and the cluster location. After performing these operations for all tracks and photon clusters, the cluster with the best χ^2 matching is assigned to each track. This means that a particle track will always have an assigned cluster, even though the matching may be very poor. An obvious drawback to this method is that bremsstrahlung emission can occur at any point in the detector and not solely in the TT station, which is only chosen due to its proximity to the bending magnet.

In an effort to improve the momentum reconstruction of primarily electrons, the BremAdder algorithm was developed. As illustrated in Figure 19, the first available track state of the particle is extrapolated in a straight line to the ECAL and if a photon cluster is found between the particle impact cluster and the extrapolation, the photon cluster momentum can be added to the particle. To illustrate the effectiveness and necessity of the BremAdder algorithm, Figure 21 shows the deviation of the reconstructed electron three-momentum from the true momentum (simulated). It is apparent from the Figure that the bremsstrahlung recovery procedure increases the reconstruction precision, with the average percentual deviation shifting from -33% to +15%, demonstrating a slight over-correction. In order to achieve the highest precision in the momentum correction, the entire track is scanned at intervals of 50 mm in order to find the point where the direction of motion most closely matches the location of the bremsstrahlung cluster, as illustrated in Figure 20. A χ^2 can also be calculated for the match between the bremsstrahlung origin point and the CALO cluster. This calculation is not implemented in the BremAdder code at the time of writing, but a custom version of BremAdder authored by M. van Veghel was used which includes this χ^2 calculation. A final note is that the BremAdder algorithm will add a 'HasBremAdded' flag to a particle, which will be used extensively in this work.

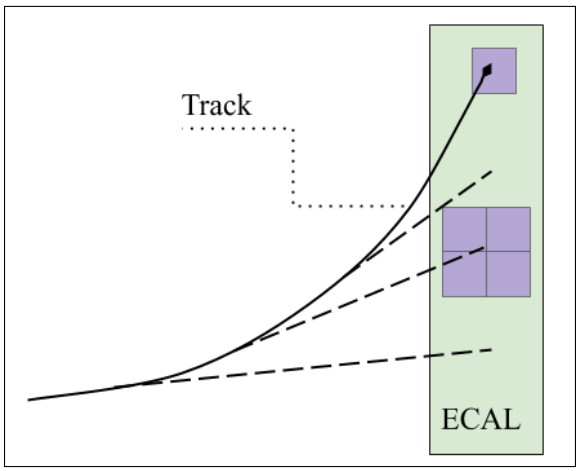

Figure 20: The track scanning procedure employed by the BremAdder algorithm. Linear extrapolations (dashed) of the track (solid) are performed every 50 mm in order to find the best matching origin state for a given bremsstrahlung cluster.

4.3 High-Performance Discriminators

With the data taking terminology and procedures described in the previous sections, the current section will cover the specific tools that are used and the information that they provide. Also described will be the methods employed to pursue the goal of better electron identification.

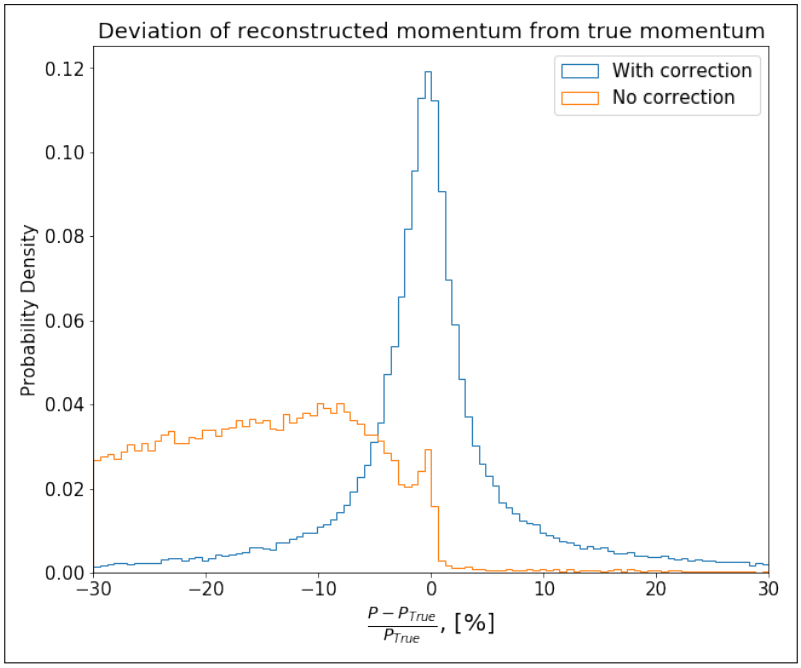

Figure 21: The percentage deviation of the reconstructed electron three-momentum from the true momentum, with and without the BremAdder correction. Clear features in this graph are the trend of underestimating momentum and the improvement that BremAdder provides in the accuracy of the reconstruction.

As mentioned earlier, the overarching strategy is to evaluate the detector variables that are accessible for the $B^0 \longrightarrow J/\psi(e^+e^-)K^*(K^+\pi^-)$ decay and find a set of variables that allow for the discrimination of electrons from the backgrounds discussed in section 3.3.1. A machine learning algorithm is trained to perform the classification. While it may seem tempting to add all available variables to the training set to maximise the available information, this proves to be counterproductive for the given task. The reason for this is twofold, on the one hand the increase in computation time is evident and on the other hand there is the issue of overtraining the classifier. Overtraining will be discussed in sections 4.3.2 and 5.2.4, and refers to a parametrisation of the classifier that fits the training sample very well, but fails to generalise to new data. Aside from this, part of the variables that are available can be shown to have no physical correlation to other variables, but a machine learning algorithm can find correlations in stochastic variations nonetheless. These found correlations are thus non-physical and cannot be used for particle identification. One could consider training to these coincidental correlations to be a form of overtraining as well. Finally, only variables that are not decay-specific will be considered, in order to be able to apply the achieved results to other channels as well.

The practice of using machine learning and multivariate analyses is not new in LHCb, and neither is the use of Bremsstrahlung-related information. The novel approach in this work is the expansion of the amount of variables in the most efficient way possible, and more importantly the application to upstream electrons. Some basic statistics concerning the data used for this work are summarised in Table 3.

4.3.1 Variable Selection

What follows will be a description of the variables and why they were selected. The 'em_' prefix signifies that the quantity is affiliated with the electron of the candidate,

Table 3: Basic statistics concerning the data set of simulated events.

Electrons (long)	518 131
Electrons (upstream)	81 869
True electrons (long)	304 297
True electrons (upstream)	34 664
HasBremAdded flag (long)	184 241
HasBremAdded flag (upstream)	24 379
Total	600 000

but everything described holds for opposite charges as well. The 'em_' prefix will be omitted when no confusion is deemed possible.

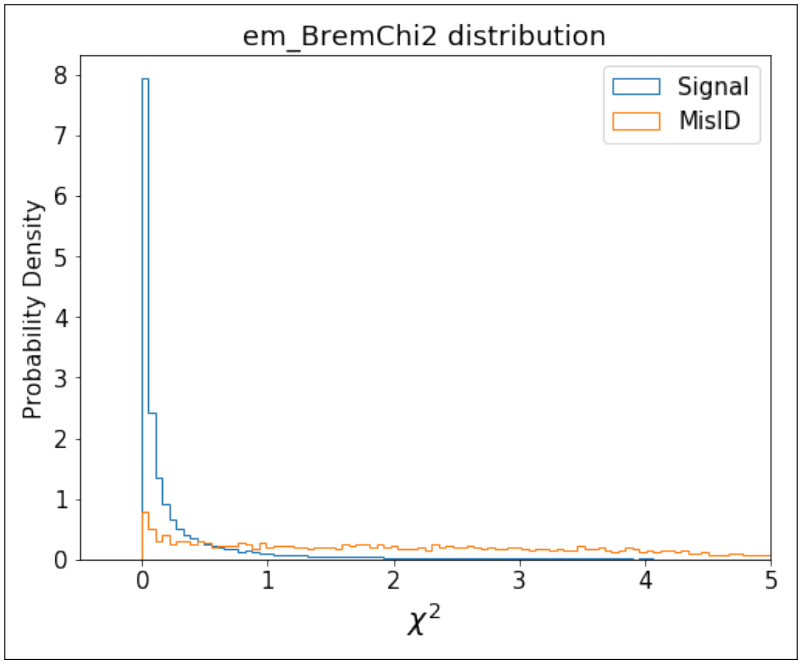

Figure 22: Distributions for the variable 'em_BremChi2', calculated in the BremAdder Programme. Distributions for both signal electrons and background are plotted.

BremChi2 Mentioned earlier in section 4.2, this χ^2 value is calculated in the BremAdder programme by scanning the track of an electron for a point where the emission of a bremsstrahlung photon would fit with a given calorimeter cluster. Subsequently the χ^2 is calculated based on the match between the linear extrapolation of the track at this point, and the actual cluster location. A match between a random photon cluster and particle will generally give a higher χ^2 value, as can be seen in Figure 22. This notable difference can be used for electron identification.

CaloBremMatch This variable is a 3D χ^2 value which originates from the calorimeter software during the (online) reconstruction. Also described in section 4.2, the calculation entails linearly extrapolating a particle track at the state closest to the TT and matching this extrapolation to the nearest photon cluster. This variable is inferior to BremChi2 in that the matched photon cluster may not be compatible with the particle at all, it is simply the cluster with the lowest χ^2 value. Also, taking the bremsstrahlung emission location to be fixed at the TT is generally inaccurate. However, CaloBremMatch is available for all electrons, not just the ones to which the BremAdder algorithm has been

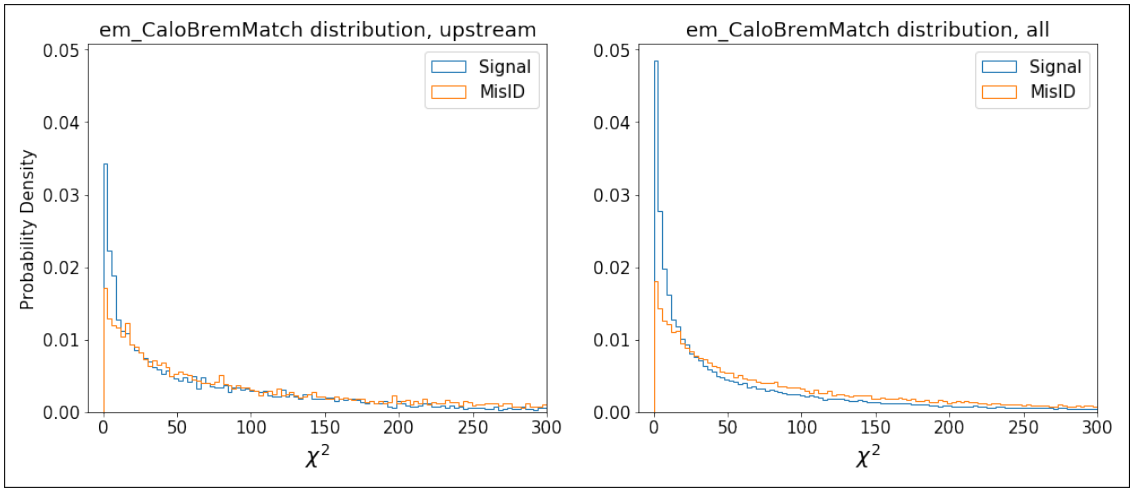

Figure 23: Distributions for the variable 'em_CaloBremMatch', plotted for both signal electrons and background. The left graph shows the distributions solely for upstream tracks, the right graph shows all tracks.

applied. Figure 23 reveals that there is a definite distinction between the variable for real electrons and misidentified background, albeit less so for upstream particles. The discriminating power of this variable will be in the availability for all events.

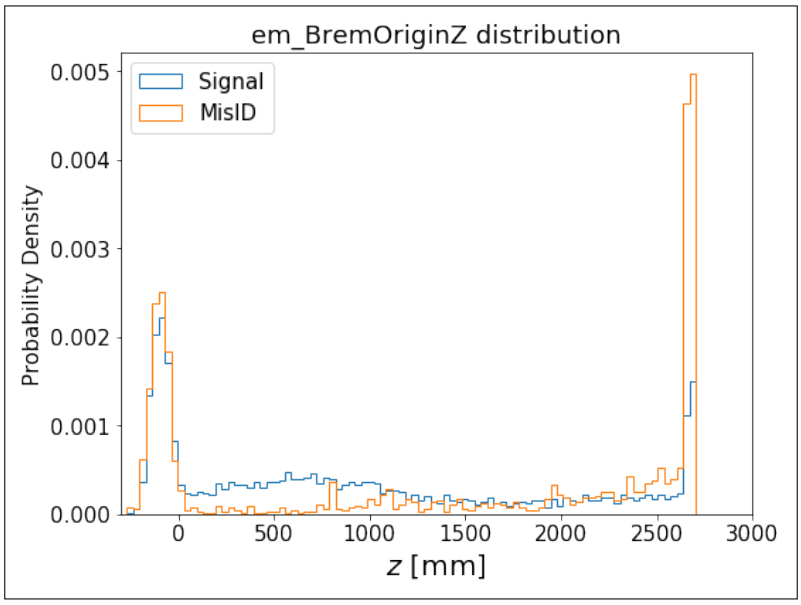

Figure 24: Distributions for the variable 'em_BremOriginZ', calculated in the BremAdder Programme. Distributions for both signal electrons and background are plotted.

BremOriginZ The origin point of bremsstrahlung emission is one of the key pieces of information by which electrons may be distinguished from other particles. The tendency of electrons to emit bremsstrahlung in material interactions will lead to a large portion of bremsstrahlung photons coming from the first sizeable amount of material, which is the VELO. This effect can clearly be seen in Figure 24 where the relative size of the left signal electron peak is much larger than the right peak than for the MisID particles. Additionally, non-electrons have a higher chance to be matched to a random photon in areas where their tracks have more curvature, which is at higher z-values due to the increasing influence of the leaking magnetic field. These peaks incidentally correspond

to the material at the VELO and the TT respectively. In between the VELO and TT stations there is a noticeable bump for signal electrons, the exact cause for this is unknown but it is possible that the electrons interact with the beam pipe.

CaloNeutralECAL Although CALO information from direct observations of particles are unavailable for upstream tracks, bremsstrahlung photons can still be measured. CaloNeutralECAL is a measure of the energy of a photon cluster associated with a particle track, expressed in MeV. Similarly to the CaloBremMatch variable, this quantity is available for all events in the data set. Although there is a noticeable difference between signal electrons and MisID background, it should be noted that this variable is not a ratio with respect to e.g. the total energy, a different data set or channel could thus not show this difference.

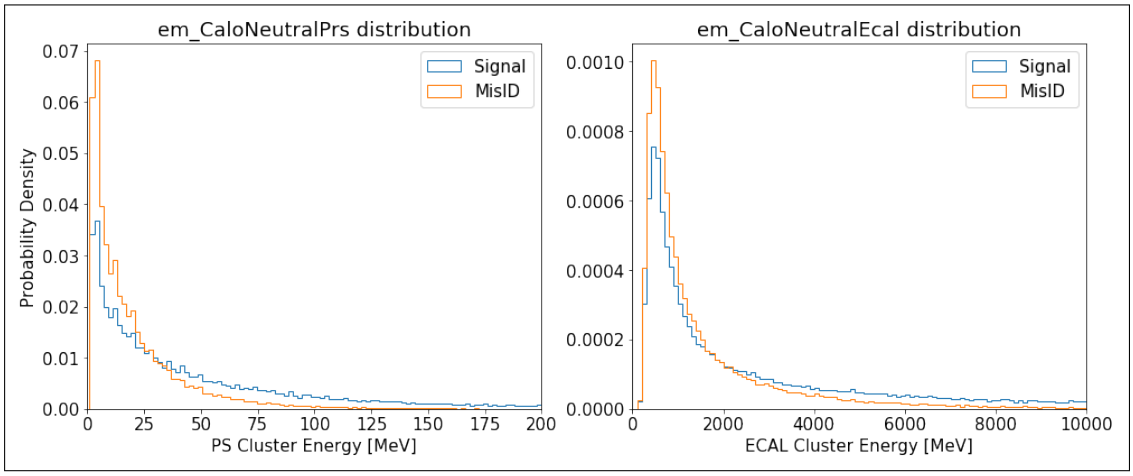

Figure 25: Distributions for the variables 'em_CaloNeutralPrs' and 'em_CaloNeutralECAL', plotted for both signal electrons and background.

CaloNeutralPrs The PreShower detector can be shown to output a distinctly different distribution for signal and background, as is displayed in Figure 25. It follows a similar trend as CaloNeutralECAL in that the average photon cluster energy is higher than that of the background particles. Just like CaloNeutralECAL, this variable is not normalised to the total energy and could thus be solely applicable to this particular decay.

Angular Variables The angular quantities η and ϕ represent the pseudorapidity and azimuth angle of a particle. The pseudorapidity distribution will generally vary per decay, and should therefore not be used directly to train a classifier. However, the direction of motion of a particle will greatly influence the material that said particle will encounter and is thus strongly correlated to the probability and location of bremsstrahlung emission. For this reason it was decided to include these angular variables as, in conjunction with the other variables, they can provide an extra means of true bremsstrahlung recognition. The data set is however resampled as described in section 4.4, to not include any decay-specific pseudorapidity features. To illustrate the correlation between material and angle, Figure 26 shows the angular distribution of all electrons with the HasBremAdded flag. The 'horn' shapes at 0° and 180° azimuth angle represent the RF foil, and it is thus apparent that many electrons that emit bremsstrahlung will pass through this region of extra material. Finally, the x and y component of the bremsstrahlung origin point may be approximated reasonably well by combining BremOriginZ with η and ϕ,

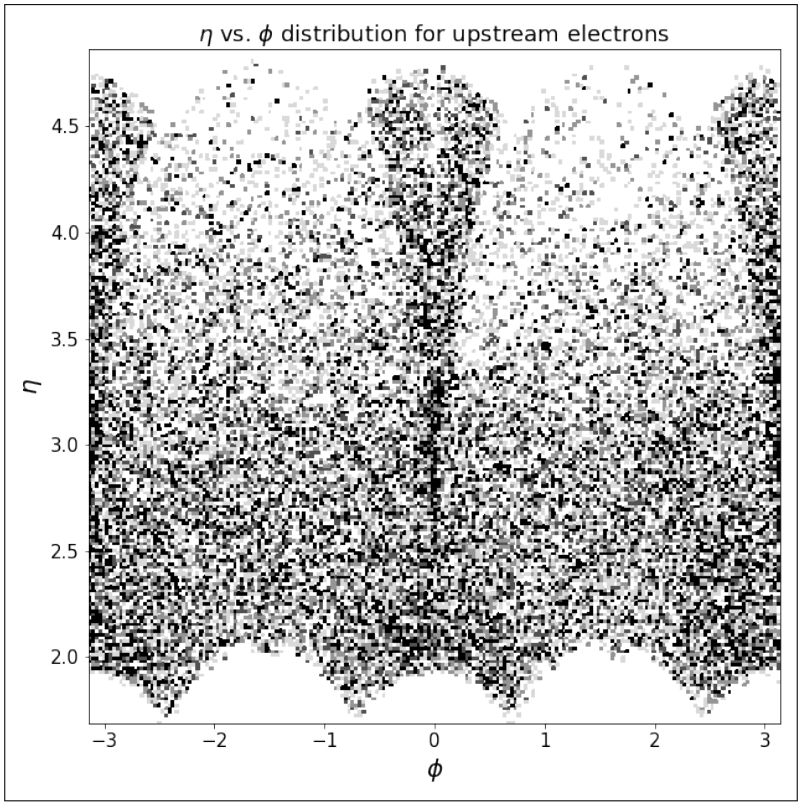

Figure 26: An η vs. ϕ graph for upstream electrons. Clearly visible are the rectangular acceptances of the detector at the top and bottom (deformed due to polar coordinate system) and the horn-shaped influence of the RF foil.

assuming that the electron will travel mostly in a straight line before bremsstrahlung emission.

4.3.2 Machine Learning Classifier

The process of electron identification is performed by a machine learning algorithm. Artificial neural networks (ANN) are already used in the LHCb software for the purpose of particle recognition, and due to the efficiency, performance, and speed of machine learning, this work will make use of machine learning too.

For the purposes of this project, the goal is to obtain a test statistic that has the greatest separation power between electrons and other charged particles. Machine learning is very useful for this task, as it can determine the signature properties of an electron much faster than a human possibly could. In general, the machine learning algorithm will be fed a user-defined set of variables, or 'features', and a set of training samples. The training samples will have flags indicating to which class a specific sample belongs. In the case of this work, those classes are simply 'correctly classified' and 'incorrectly classified'. After the training process, the classifier algorithm can be used to attempt to classify a set of testing samples. If class flags are also available for this testing sample, as is the case in the simulated data used for this work, the performance can be evaluated.

Many of the figures in this and following sections will contain so-called Receiver Operating Characteristic (ROC) curve plots. The true positive rate (TPR), also called sensitivity, is on the vertical axis, with the false positive rate (FPR) on the horizontal. The TPR and FPR for the classifier are plotted for a number of decision thresholds, resulting in a curve. A perfect classifier will have a TPR of 1 and an FPR of 0. A

commonly used metric for performance is the area under the ROC curve (AUC), and will be used extensively in this work. A perfect classifier will have an AUC of 1, whereas a random guess will result in an AUC of 0.5.

As the exact software, configuration, and training data sets of the LHCb ANN's are either hard to access, scarcely documented or difficult to use in isolation from the rest of the LHCb software, it was decided to emulate the ANN behaviour as well as possible and to instead look at relative performance increase rather than absolute. To achieve this, it is necessary to train on the same sets of variables, which are listed in the appendices in section 8.1.

The selection of what type of machine learning algorithm to use is based on classification performance, stability of the training result, and computation speed. Three types of algorithm from the scikit-learn Python package are considered for the classification task: A Multilayer Perceptron (MLP), an ADABoost algorithm, and a Gradient Boosting Classifier (GBC). The MLP [19] represents perhaps the most well-known form of ANN, with a number of nodes or 'neurons' organised in layers and interconnected in various ways. The output of one neuron is used as an input for all connected neurons in the next layer, weighted differently for different connections. The weighted sum of all these inputs is then used as the argument of a model-dependent function, the output of which becomes the output of this neuron. This continues until the last layer of neurons is reached. The machine learning aspect of this is in the self-adaptation of the network: The initial user input is fed into the network, and in an iterative process the node connections are weakened or strengthened based on the error in the output nodes. In other words, the weights that are used to modify neuron inputs are modified until a given input gives a satisfactory output.

ADABoost is an early example of a boosted decision tree algorithm [20]. The basic working principle is to add weak learners to a predictive model in a step by step fashion, with the weak learners consisting of a simple, two-option decision tree, called a 'stump' [24]. The process is outlined in Figure 27. In the first training step, all samples will receive an equal weighting and a weak learner is initialised that performs the best on this equally weighted set. The learner is now weighted based on its performance, receiving a larger weight if it performed well. The samples then receive a new weighting based on the performance of the learner, with misclassified samples having increased weight and vice versa for correctly classified samples. This way, when the majority of samples are classified correctly by a weak learner, the focus shifts strongly to the samples that it could not classify well enough. A resampling is subsequently performed so that large-weight samples have a higher chance of appearing in the new set and the process repeats with the addition of a new weak learner.

Lastly, the GBC algorithm expands on the ADABoost algorithm by not only attempting to maximise the number of correctly classified samples, but by also minimising the prediction error, which can be thought of as increasing the certainty that the sample belongs to the predicted class [27]. The added benefit of GBC over ADABoost is that the latter can only perform binary classification, and is less accurate. GBC is however more computationally intensive.

To determine which machine learning algorithm is best suited for the classification, several tests are performed with the three algorithms, varying the parameters and evaluating their computation time, performance, and stability. The first of these tests is displayed in Figure 28. This test is performed on 31 260 samples and plots the ROC curves for the three classifiers for mostly default parameters. The variables that are trained on are the same as the variables that the LHCb ANN uses for its upstream

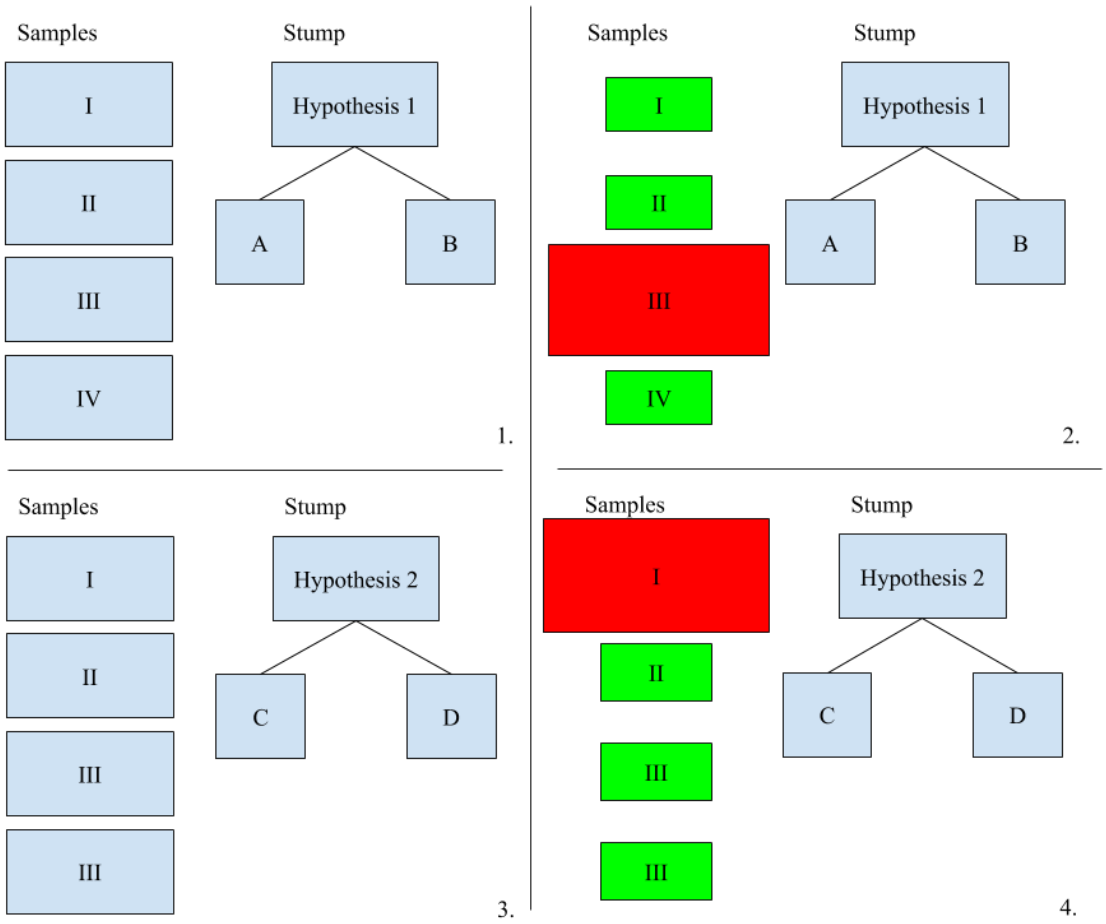

Figure 27: The ADABoost algorithm adds weak learners to a predictive model in an iterative fashion, emphasising samples that previous iterations classified incorrectly. In step 1, a hypothesis is formed and used to classify the training samples. In step 2, wrongly classified samples (red) obtain a high weight. Step 3 features a resampled training set, based on the sample weights, as well as a new hyptothesis, with the whole process being repeated in step 4.

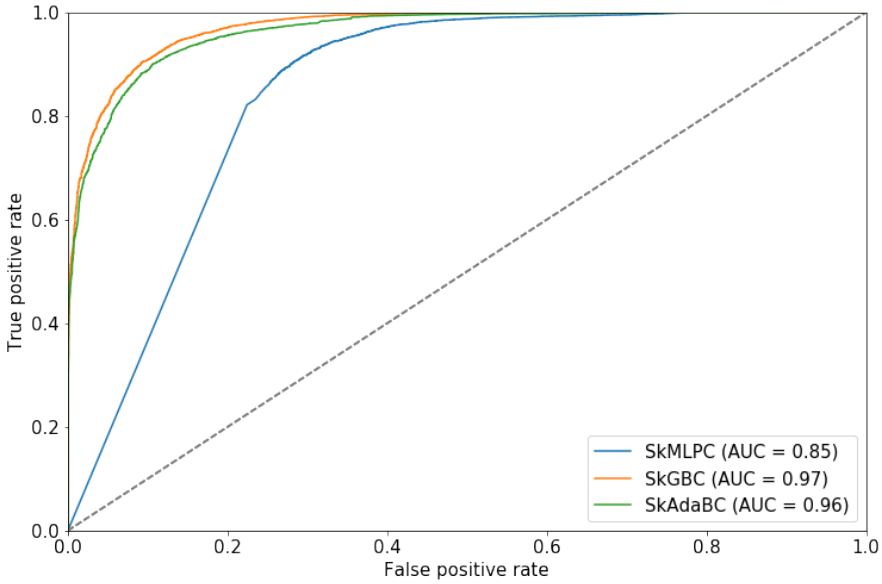

Figure 28: ROC plot for the three classifiers under consideration. Parameters for the ADABoost and Gradient Boosting classifiers are in table 4. The MLP has one layer consisting of 17 neurons, which is equal to the amount of input variables. Area Under Curve (AUC) values are listed for convenience.

Table 4: The variables used for the plots in Figure 28. Since ADABoost by construction relies on decision tree stumps, max_depth is always 1. Tolerance is only applicable to GBC, as it is capable of more advanced regression analysis than ADABoost

Classifier	**ADABoost**	**GBC**
max_depth	1 (fixed)	3
n_estimators	100	100
learning_rate	0.1	0.1
tol	N/A	1E-3

electron classification, and are listed in the appendix in section 8.1.3. The MLP consists of a single layer with 17 neurons, the same amount as input variables. The parameters for ADABoost and GBC that are different from default are listed in Table 4, they were changed to provide a better direct comparison between the two. From the Figure it is apparent that both boosting classifiers outperform the MLP classifier to a large extent, with the GBC performing slightly better than ADABoost. In terms of computation time, ADABoost performed best, with a 2.0 s average for the sample size of 31 260, followed by GBC with a 4.5 s average. The MLP cost more than 6 times that time with an average of 28.2 s, and would often not converge to a good training fit. These computations are performed on a single, 1.6GHz laptop CPU core.

As a second test, the performance of the MLP is analysed under varying configurations of layer sizes, in an attempt to increase its stability and its classifying efficiency. The results are shown in Figure 29. The stability of the training is poor compared to the boosting algorithms, with many of the training runs not converging to a solution. In the Figure this is visible as the (3,2,1) layer network, which in this specific case has not converged at all, with the ROC curve on the diagonal and being no better than a random guess. At this point, the MLP is no longer considered, due to its comparatively long computation time, poor stability, and mediocre performance. The choice is made to continue with the GBC algorithm over the ADABoost algorithm, for the following

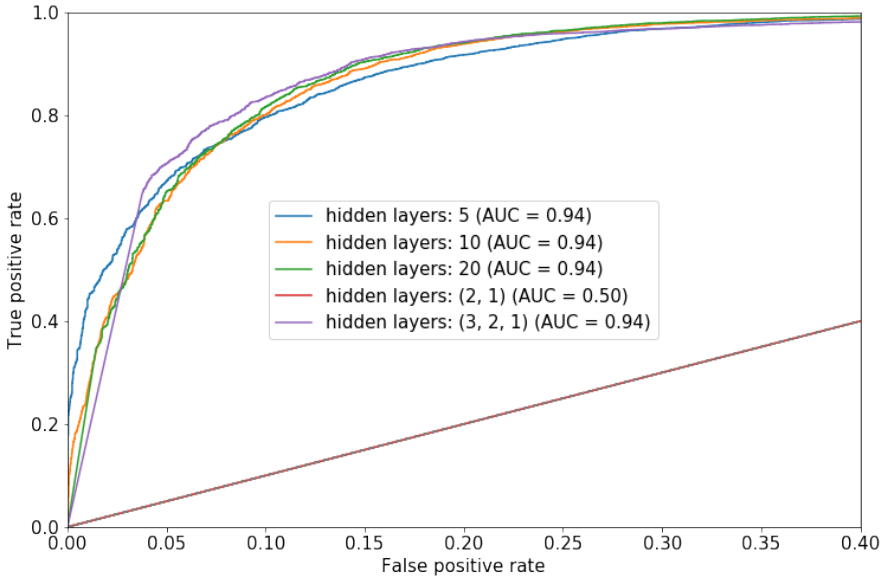

Figure 29: ROC plot for the MLP classifier for various configurations of the layer structure. Results varied between training runs, this plot is a representation of a typical outcome.

reasons:

1. Better performance than ADABoost
2. Acceptable increase in computation time
3. Possibility of non-binary classification problems

The 'acceptable' increase in the second point is based on a single training run taking no longer than 15 minutes for the entire data set of 600 000 events, for the variables listed in section 8.1.1. The third point is included as the possibility of classifying multiple types of backgrounds or specific particles was considered at the start of the project. To keep consistency, the classifier was not changed during the project.

With GBC selected as the machine learning algorithm to be used, the parameter space for the algorithm is explored. This is done by evaluating the performance under variations of the most important parameters, namely the amount of estimators or 'stumps', the depth of each estimator, and the learning rate that diminishes the influence of an estimator in a new iteration. The results are displayed in the form of ROC curves in Figures 30, 31, and 32. From these graphs one can draw the conclusion that the product of the amount of estimators and the learning rate is of primary importance; a low learning rate increases precision but requires more estimators to arrive at a good training solution, and vice versa. Since only binary signal/non-signal classification is considered in this work, a maximum depth of 1 would suffice, however, due to the reason listed above, multiple depths are considered. The results are as expected, with no specific depth seemingly having a particular benefit over another, and performances vary per training run. In the end a choice is made to use the default depth of 3, to limit computation impact but also allow for non-binary classification. With regards to the number of estimators and learning rate, a value of 200 and 1.0 are chosen respectively for the following reasons:

1. The given learning rate is not so large as to 'overshoot' the optimum solution, for the given data
2. The particular combination of estimators and learning rate always converges for

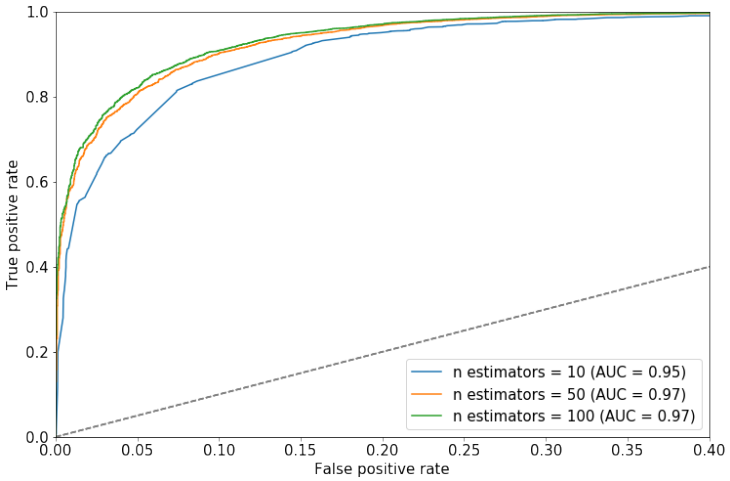

Figure 30: ROC plot showing varying amounts of estimators used in the GBC algorithm (default 50).

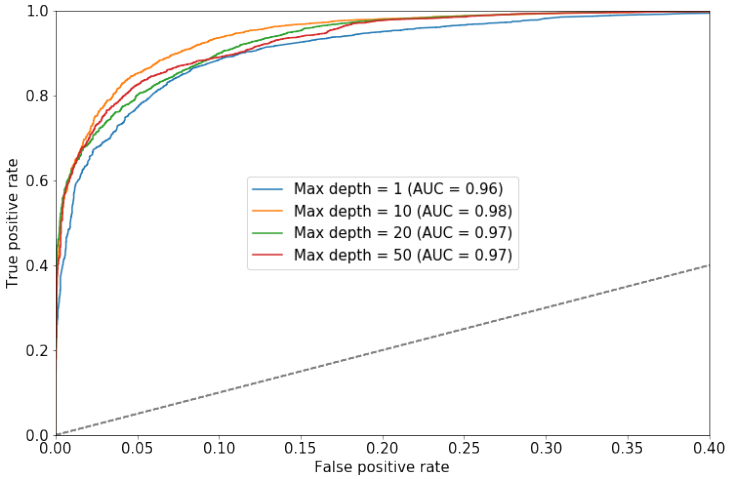

Figure 31: ROC plot showing varying estimator tree depths. Results vary between runs, with no visible optimum depth (default 3).

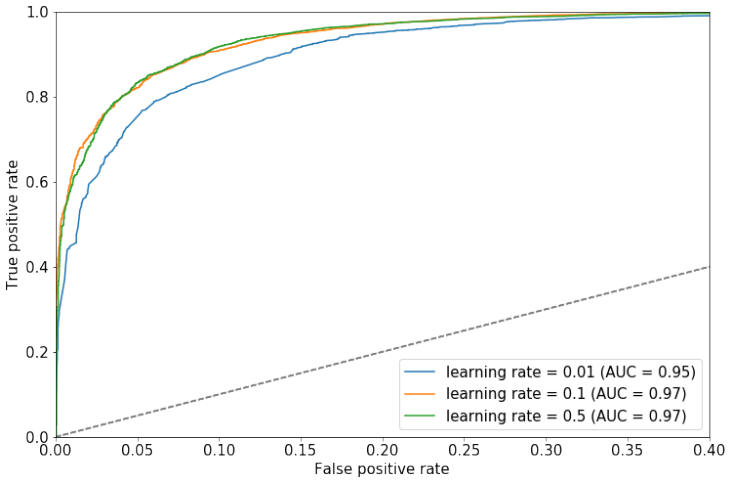

Figure 32: ROC plot showing results for varying learning rates. The lowest rate of 0.01 terminates too early most of the time, with the number of estimators fixed in this graph (default 1.0).

the given data

3. The amount of estimators can be taken larger than required, due to the algorithms resistance to overtraining

The third point is made based on the commonly accepted [8, 9] theory that gradient boosting is not sensitive to overtraining, meaning that a longer algorithm run time will not increase the out-of-sample error by a large factor. It should further be noted that, with over and undertraining in mind, the choice for the split between training and testing sample size is 50%/50% unless stated otherwise. This choice is arguably arbitrary (within reason), and it can be seen in section 5.2 that a similar classification efficiency can be achieved for a small training sample size, for the given data.

4.4 ProbNNBrem Algorithm

ProbNNBrem is the main software component that has been created for this project. It is written in Python 3 and consists of two parts: A main file where data handling and computation is performed, and a library containing auxiliary functions and parameters.

1. data import
2. calculation of derived quantities
3. truth flagging
4. resampling
5. a main training body consisting of:
    - a *default* mode
    - a *split_tracktype* mode
    - an *analyse_significance* mode
6. result analysis and visualisation

In the import step, ROOT files are loaded and converted to Pandas DataFrames[1]. The derived quantities refer to variables that are not present in the original data set, such as the ratio of electron bremsstrahlung energy over its total energy. In the truth flagging step, the distinction between signal and background is made, based on user-defined conditions. It is for example possible to filter ghost tracks out of the data set, to emphasise the ability of the classifiers to differentiate between real particles. The resampling step is necessary to avoid a training bias in the sample. For this work, resampling has only been performed on the pseudorapidity of electrons, to ensure a similar distribution for signal and background. The reason for this is that a noticeable difference is present between the distributions, which a classifier will pick up on. This is undesirable, as different decays will generally have different pseudorapidty distributions and this would thus influence the classifiers performance. The three modes of operation in the main body each serve their own purpose. The *default* mode trains classifiers in the most general way, using the most extensive lists of variables and not differentiating between upstream and long tracks. It can also be used to train smaller subsets of variables, as in section 5.2.1. The second mode, *split_tracktype*, trains classifiers on either long or upstream tracks, as the name implies. The training variables for long and upstream tracks are different, and can be found in the appendix in sections 8.1.2

[1] pandas.pydata.org

and 8.1.3 respectively. The last mode is used to produce the results found in section 5.1.2. It creates a list of all possible pairs in a given set of variables, and trains classifiers on the pairs in order to analyse the impact of specific variables.

Implementation In future work, the ProbNNBrem algorithm can be modified to serve as a TupleTool, described in section 4.1.1. In this format, an analyst could use it to acquire more accurate information on the particle identification of electrons.

Now that all the tools are collected, constructed, and ready for use, the next step is to evaluate how to best apply them. The next chapter will detail the experimental results that were obtained with the tools described in this chapter.

In the remainder of this work, ProbNNBrem will refer primarily to the set of seven variables that are used in the algorithm, detailed in section 4.3.1. In other cases, the context is believed to be sufficiently clear.

5 Validation and Results

With the concepts and methods of operation developed in the previous chapter, the current chapter will turn towards the obtained results. Since grading the performance of a classifying task is non-trivial, section 5.1 will introduce the types of metric used to quantify the performance of the classifiers. The section that follows contains the actual results obtained during the project, approached from various perspectives.

5.1 Quantifying Performance

The task of classifying a sample as being either an electron or background can be graded in a variety of ways, depending on the application. If one is for example looking for an exceedingly rare signal electron, the requirements for a positive identification could be restricted somewhat, as to not include a background particle. In this case the performance metric would emphasise the reduction of false positives. In other cases it might be desirable to maximise the signal to noise ratio. As there is no generally preferred method of grading performance [29], the next sections will cover the choices made for this work.

5.1.1 Performance Metric

For most of the classification tasks, a choice is made to present the results in the form of a ROC-curve (see page 32) with an accompanying Area Under Curve (AUC) value. This is based on three major considerations:

1. Applicable for all classification tasks under investigation and thus allows direct comparison
2. Performance is easily gauged at a glance
3. Some training peculiarities are evident when inspecting the curves

To illustrate the third point, it is for example at times possible to gauge whether information from one variable is contained in another variable, by looking for an overlap in their respective ROC curves. This effect can be seen later in Figure 39.

It is worth restating that a good classifier will have an AUC close to 1 and a high True Positive Rate (TPR) for a low False Positive Rate (FPR)

5.1.2 Cross-Variable Performance Validation

Although a ROC curve grants insight in the performance of a classifier trained on a single set of variables, it is considered desirable to be able to gauge the effectiveness of any given set of variables. There seems to be no straightforward way to accomplish this, aside from using brute computational force to train the classifier on all possible combinations. This task was performed to first order, on a subset of all available variables, by training classifiers on all possible pairs of variables. The subset of variables consists of those used for training the LHCb ANN (Found in the appendix in section 8.1.3), and those selected for this work (section 4.3.1). The resulting AUC values are plotted in a matrix-like fashion in Figure 33.

From the figure, some variables immediately stand out as having a high AUC. Notable examples are the 'RichDLL' variables [26], which represent the differences in the Log-Likelihood for the given particle with respect to that of a pion. In other words, a high RichDLLe value would mean that this particle has a much higher likelihood of being an

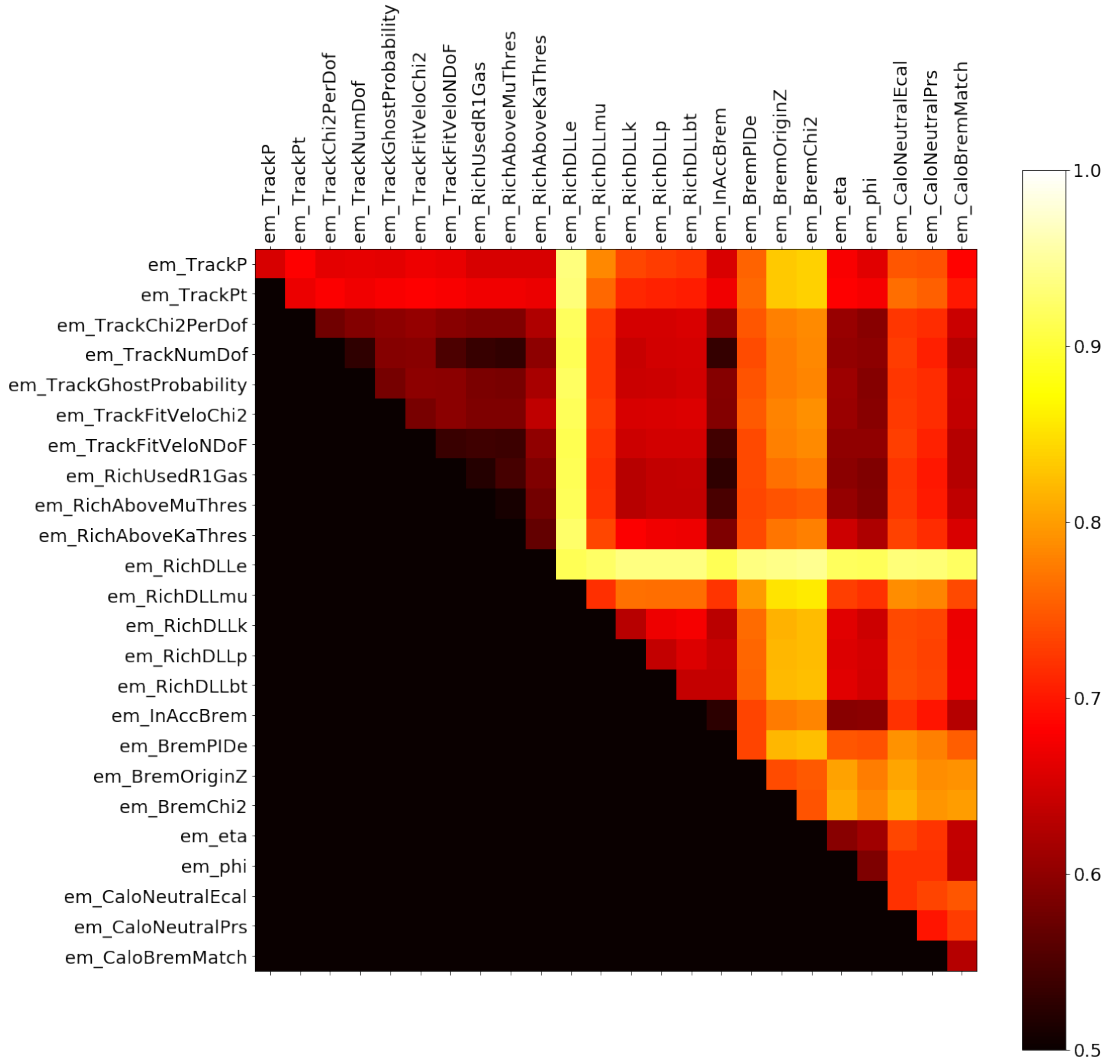

Figure 33: AUC values for classifiers trained on all possible pairs within a given set of variables. Variable list can be found in the appendix in section 8.1.3.

electron than a pion, based on particle identification info from the reconstruction. Of the variables added in this work, em_BremOriginZ and em_BremChi2 variables perform comparatively well.

Figure 33 shows how well a given pair of variables performs in absolute terms. In order to determine the added combinatorial effect, a formula was derived. This combinatorial effect may arise through correlations between variables, as well as separate, uncorrelated effects. For a given variable 'A' and 'B', the combinatorial multiplier 'C' is defined as:

$$C = \frac{AUC_{A,B}}{max(AUC_A, AUC_B)} \tag{6}$$

The idea behind this is that if the AUC of a pair of variables is not better than the best performing variable in the pair, there is no added benefit of that particular combination. Note that this combinatorial multiplier is limited to the range [0.5,2], as AUC is always taken to be in the range [0.5,1]. The resulting multipliers are shown in Figure 34.

With the goal of 'scoring' the usefulness of any given variable, the AUC of a pair is multiplied with the combinatorial multiplier for that pair, to arrive at an adjusted performance, displayed in Figure 35. Possible values range from 0.25 to 2. This final

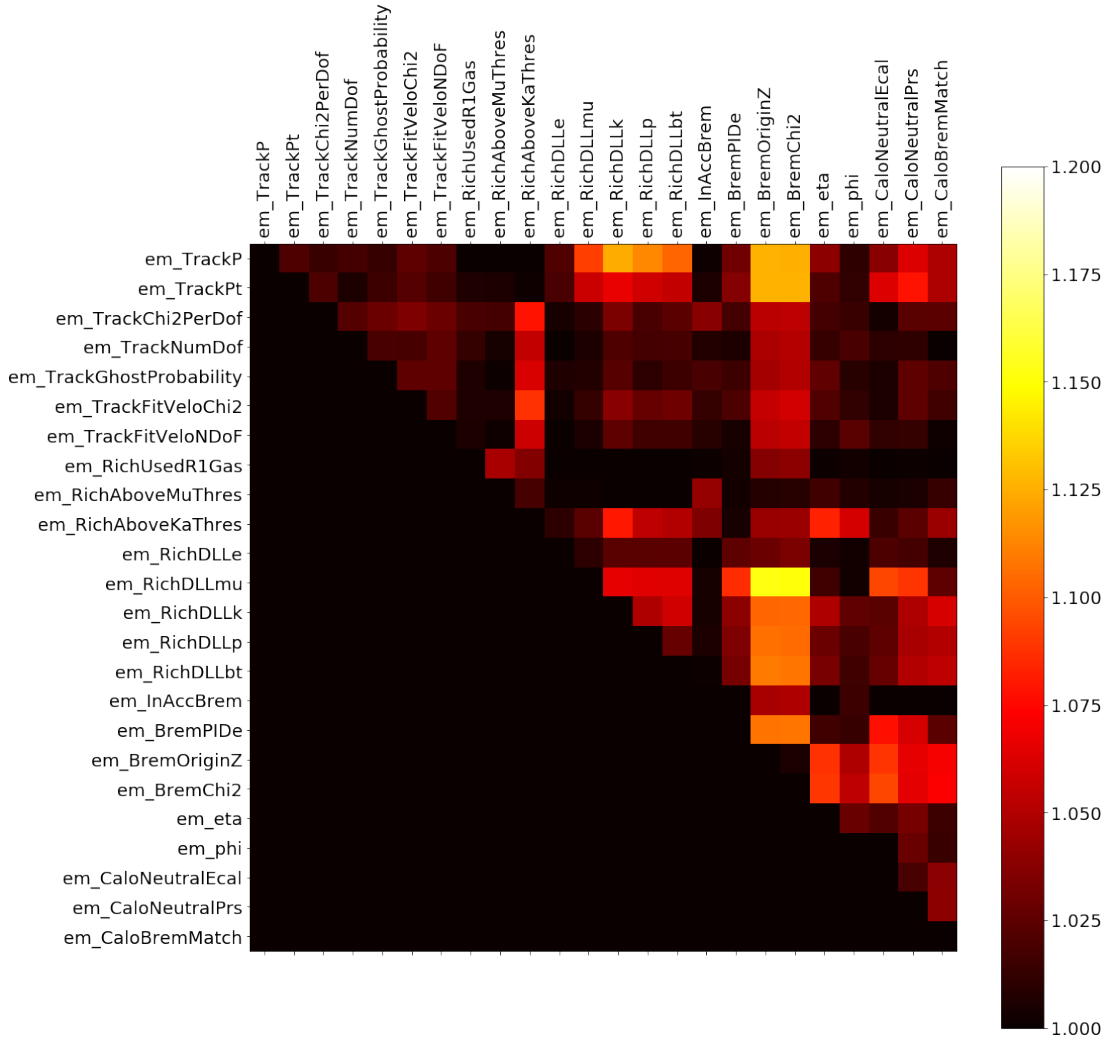

Figure 34: Combinatorial multiplier values derived from Equation 6. Note that the colour scale does not encompass the full range of possible values, in order to emphasise differences between variables.

performance Figure will be used in section 5.2.3 to establish the order importance of the selected set of variables.

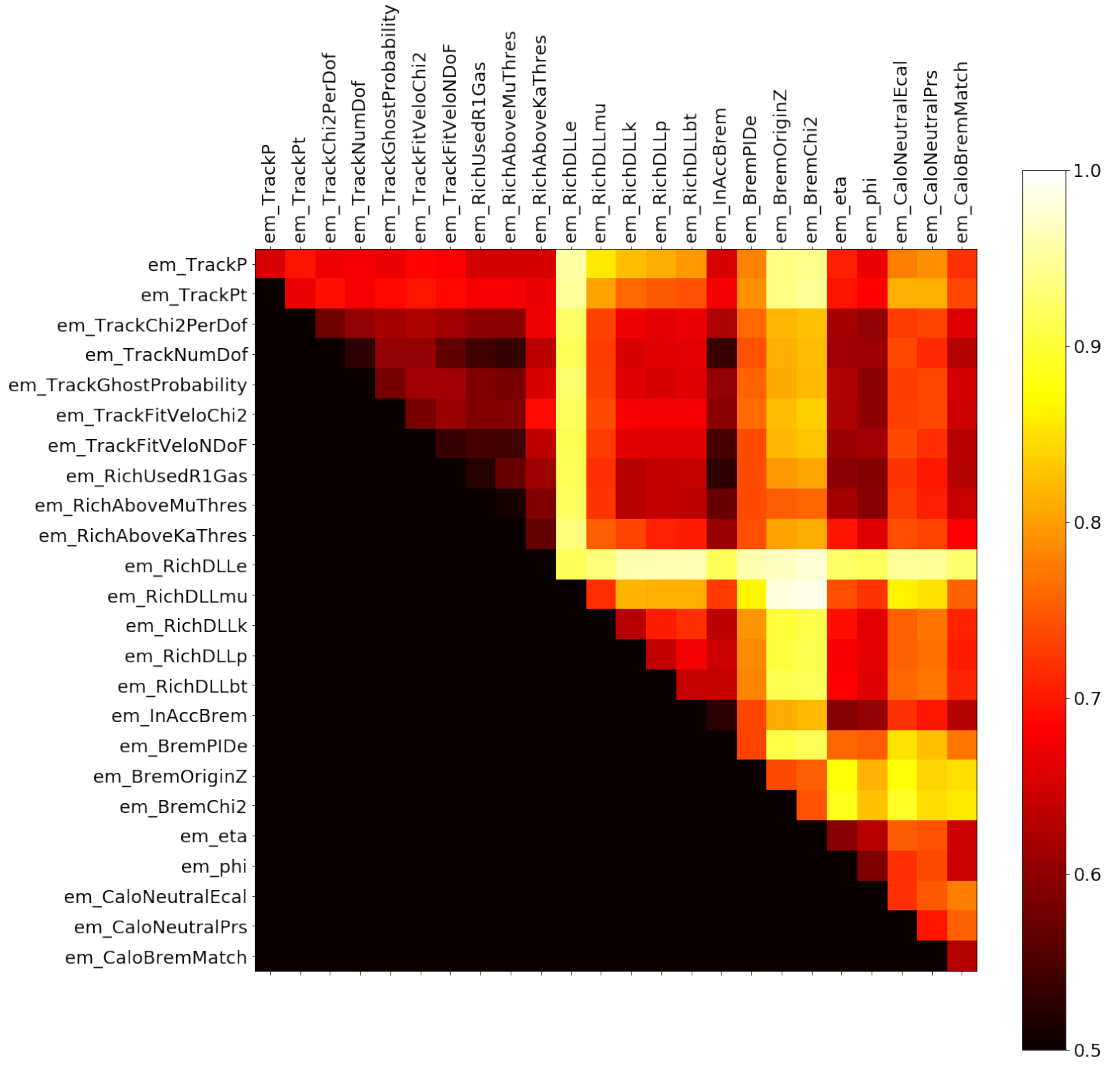

Figure 35: Adjusted, pair-wise performance, obtained by combining the AUC of a pair with the combinatorial multiplier from Figure 34. Note that the colour scale again does not encompass the full range of possible values.

5.2 Results

Presented in this section will be overviews of classifier performances for various configurations of variables. Plots will be shown that visualise performances for separate groupings of variables, performance based on electron track type (upstream or long), and general performance per variable. Finally, the effects of overtraining and sample size will be displayed and discussed.

Figure 36 shows arguably the most complete picture attainable to summarise the work. ROC curves are shown for the full data set, consisting of 600 000 samples, for a number of different classifier configurations. The *Current Training* classifier has the master list of section 8.1.1 as its training variable set. The bremsstrahlung category of this set is itself used to train another classifier, resulting in the *Old Brem Category* in red. Replacing this bremsstrahlung category with the variables added in this work results in the *Proposed Training* classifier, with the replacing variables being used to train the *ProbNNBrem* classifier in turn. Finally, the single variable em_ProbNNe is used to train a last classifier by the same name, in order to check for consistency and to have a benchmark for measuring improvements.

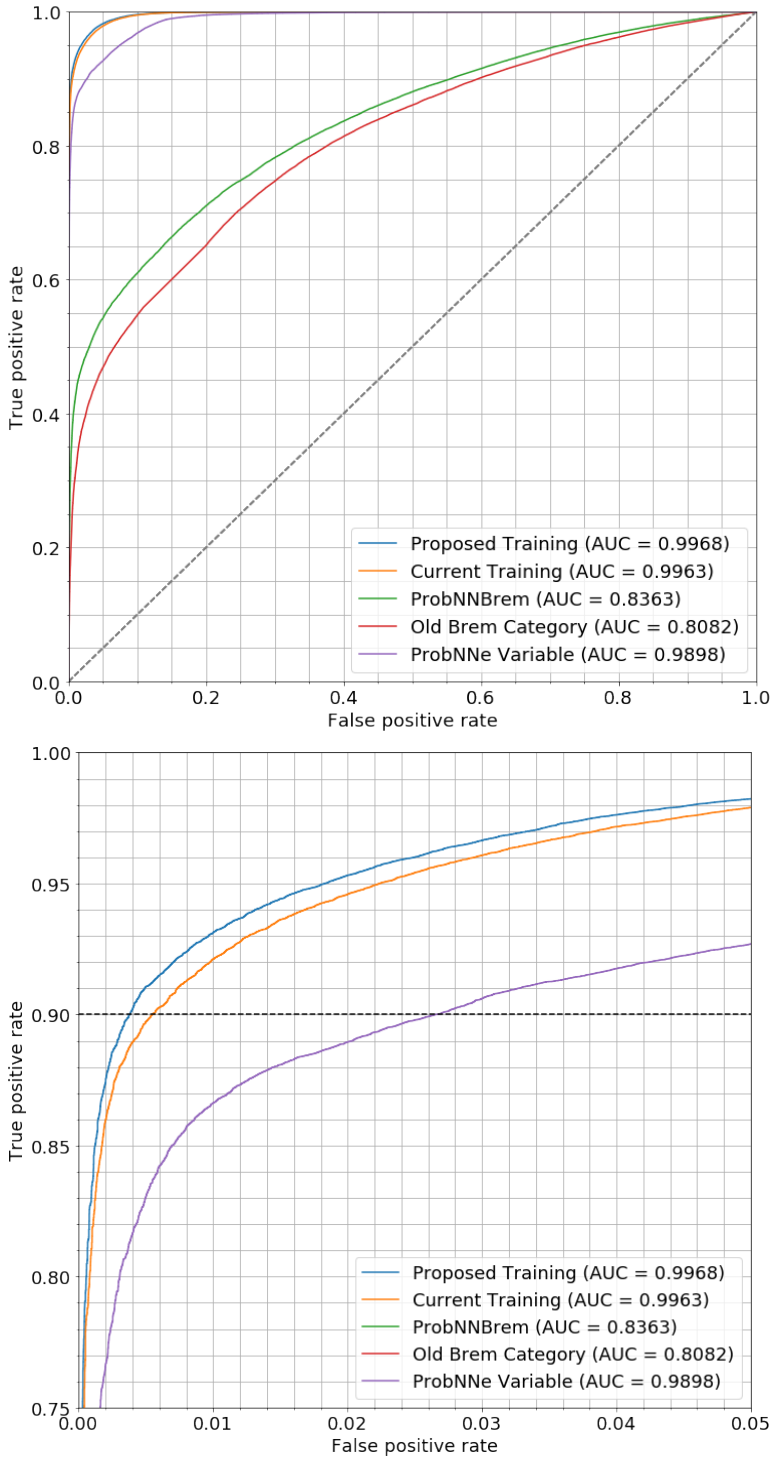

Figure 36: Classifier ROC curves for the full data set of 600 000 events (bottom graph shows zoomed section). Current Training refers to the master list of variables, found in the appendix in section 8.1.1. Proposed Training features the same list, but with the category for bremsstrahlung replaced by the variables added in this work. Old Brem Category and ProbNNBrem refer to these subsets of variables respectively.

From the graphs in Figure 36, one can conclude the following:

1. The classifier trained on the ProbNNBrem set of variables outperforms the classifier trained on the Old Brem Category set.
2. This improvement translates to the overall performance, as can be seen in the difference between the proposed and current training classifiers.

The second point is quantified in Table 5, displaying both the increase in AUC and the decrease in FPR at a fixed TPR.

Table 5: Table summarising the results of Figure 36 for the two main classifiers.

	Current Training	Proposed Training	Improvement
AUC	0.9963	0.9968	0.48‰increase
FPR at TPR = 0.9	5.46×10^{-3}	3.81×10^{-3}	30.2% reduction

As mentioned briefly at the beginning of section 5.1, the ratios of true and false positives vary depending on where the cut is placed on the classifier output. A high threshold value will result in few false positives, but many false negatives too, limiting the signal. The inverse is true for a low threshold. Figure 37 shows signal to noise (True positive classifications as signal and false positive classifications as noise) versus classifier cut threshold for the two main classifiers. The colours are equal to those used in the previous figures, and the optimum cut for the *Proposed Training* classifier is marked on the axes.

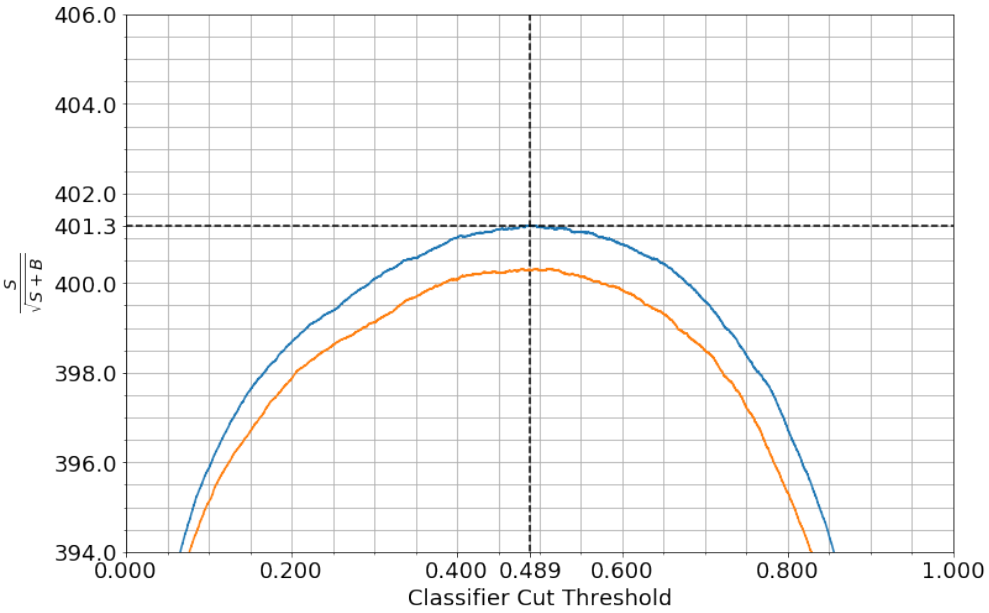

Figure 37: Signal (True positive classifications) to noise (false positive classifications) ratio for various classification cut values. Proposed Training classifier is marked in solid blue, Current Training classifier in solid orange. Optimum cut is marked on the axes.

5.2.1 Per Variable Grouping

The following few figures will show the performances of the newly added, bremsstrahlung-related variables added in this work, appropriately grouped. Detailed descriptions of the

variables can be found earlier in this work, in section 4.3.1. All of the following classifier experiments were performed on the full set of 600 000 events, with the same parameters, outlined in section 4.4.

Starting with the variables relating to the matching of a bremsstrahlung photon to an electron track in Figure 38, a complementary effect can be seen between the two variables. As mentioned before, the BremChi2 variable is generally far more accurate than its CaloBremMatch counterpart, which can be attributed to its more involved way of computation. It is however only available for the subset of events where the BremAdder algorithm has been run. In the graphs this can be seen in the low false positive rate area being dominated by the BremChi2 variable, and higher false positive rates by CaloBremMatch. The initial steep climb in the BremChi2 ROC curve represents its accuracy in classifying true electrons, the transition to a straight line represents the lack of information for most of the particles. It is apparent that combining the two variables leads to a classifier that can reinforce the weaknesses of one variable with the strengths of the other. It should be noted that the HasBremAdded flag in and of itself provides some information as to the identity of a particle, and it is inherently ingrained in the BremChi2 variable. To combat this inequality, all classifiers were trained with the inclusion of the HasBremAdded variable, to allow for a more accurate comparison.

Figure 39 shows ROC curves for the BremOriginZ variable and the HasBremAdded variable, the latter one added for illustrative purposes. Training a classifier on HasBremAdded reveals two things:

1. The HasBremAdded flag contains a sizeable amount of information considering the nature of the particle
2. The HasBremAdded provides no information that is not already present in the BremOriginZ variable

The first point is due to the electrons' higher chance of bremsstrahlung emission: A matched bremsstrahlung photon, and with it the HasBremAdded flag, has a larger chance of belonging to an electron than to another particle. The second point is evident from the construction of the BremOriginZ variable, which obtains a value if the BremAdder algorithm has been run and defaults to a placeholder value if it has not been run, essentially mimicking the HasBremAdded flag.

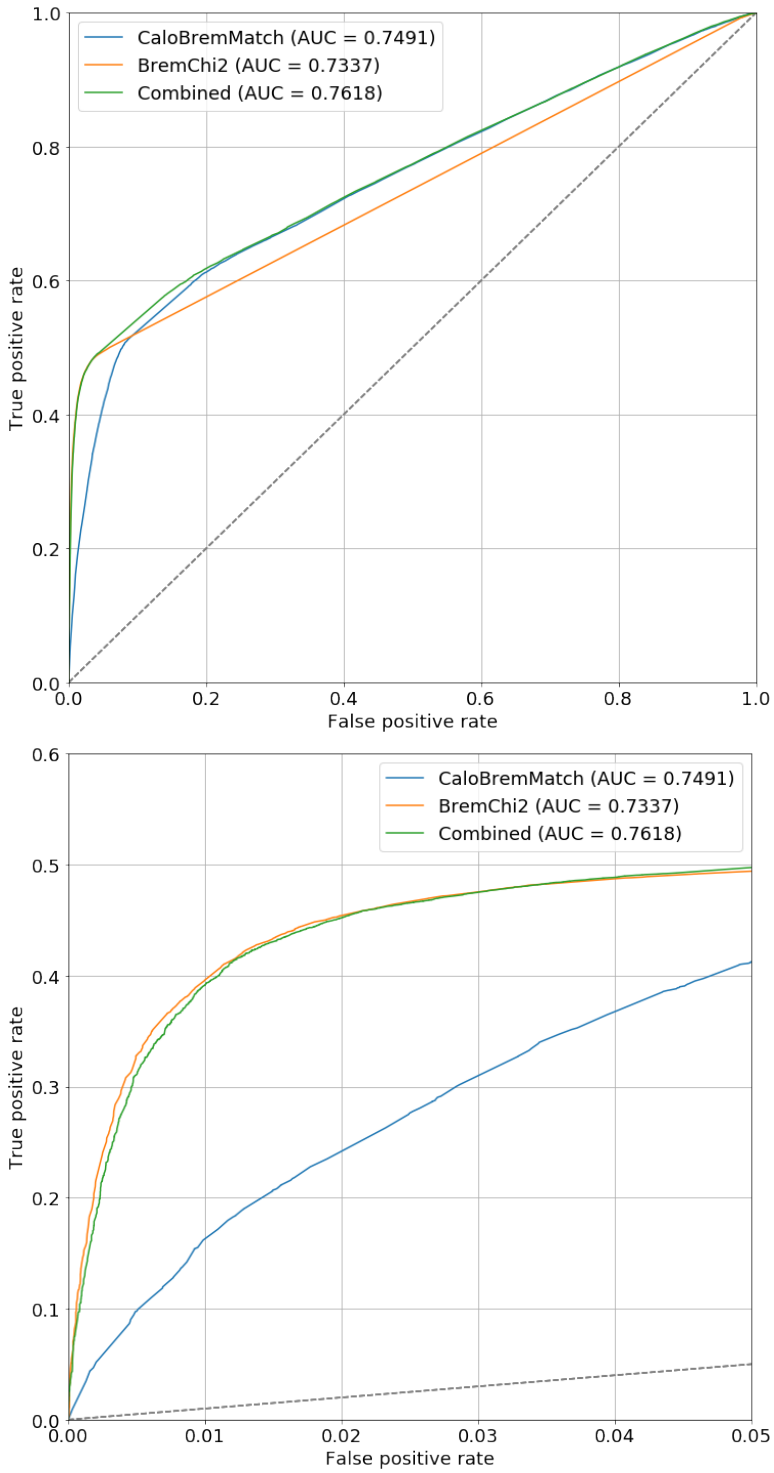

Figure 38: ROC curves for the two variables used to match a photon cluster in the ECAL to an electron track, and their combined curve. The HasBremAdded variable has been added in all cases to allow for a direct comparison.

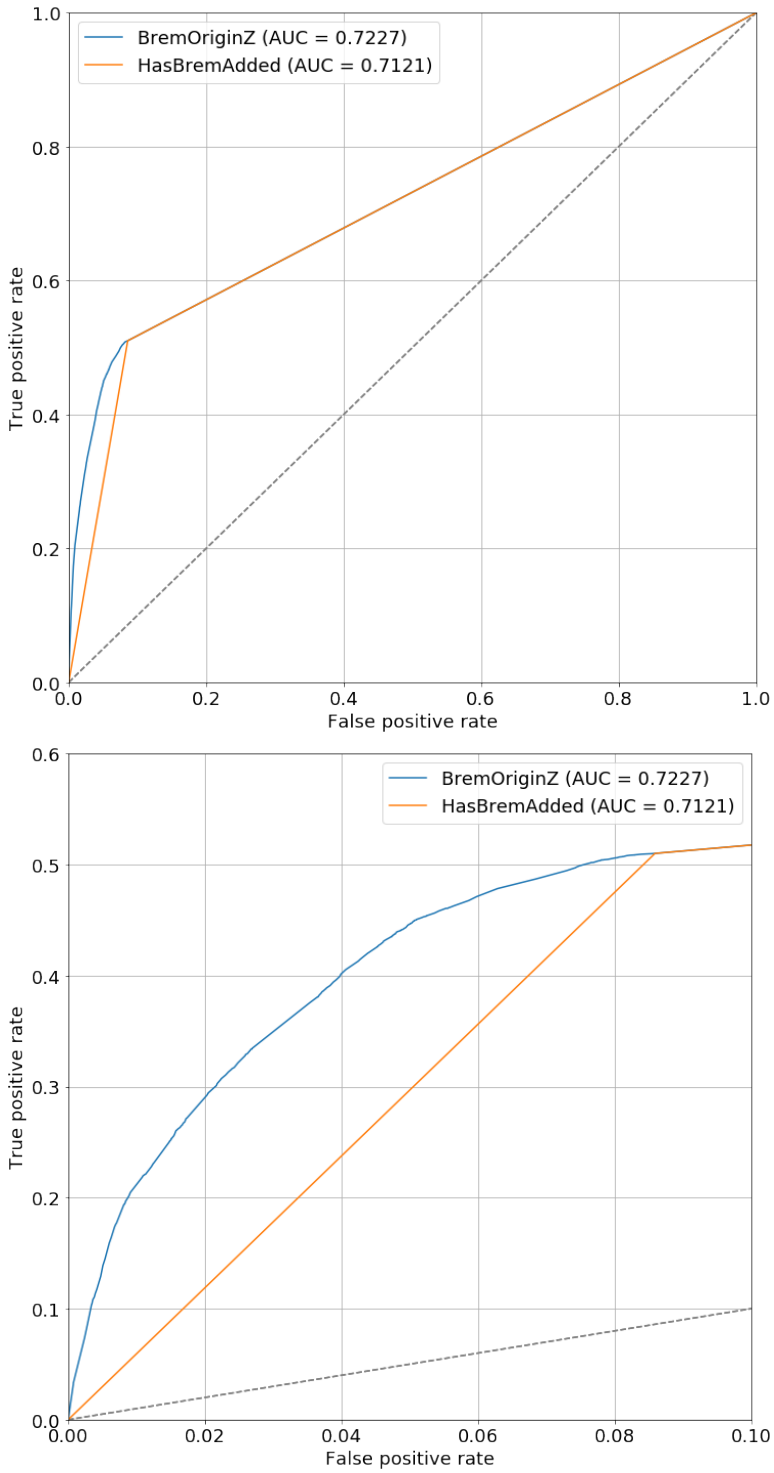

Figure 39: ROC curves for the BremOriginZ variable and the HasBremAdded variable. The two ROC curves overlap at a FPR of $>\tilde{0}.085$, indicating that there is duplicate information in HasBremAdded and BremOriginZ.

The last group of variables are those relating to the calorimeter systems: CaloNeutralPrs and CaloNeutralEcal, relating to the PreShower and ECAL energy deposits of associated bremsstrahlung photons respectively. Their ROC curves are plotted in Figure 40. The similar shapes of the curves hint towards a sizeable portion of the information being shared between the variables, but it is also evident that the combined classifier performs better than the separate ones. In general, the average reduction in FPR is about 0.23 for these variables, indicating their usefulness as discriminating variables.

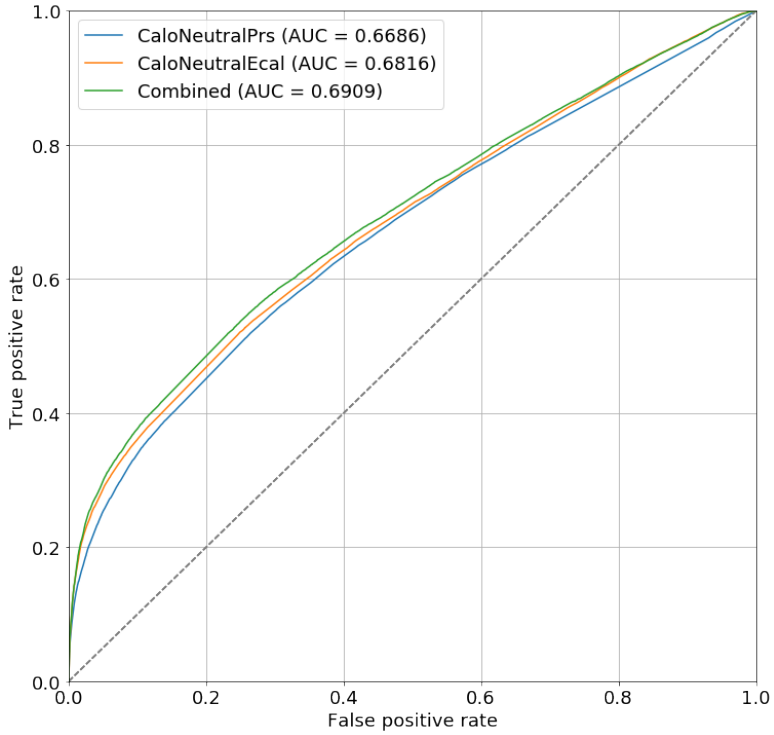

Figure 40: ROC curves for the variables indicating energy deposits of photons in the PreShower and ECAL systems, and their combined ROC curve.

The previous groupings are those that are deemed the most fitting for the selected variables. It is however a daunting task to try to determine the exact intricacies of the correlations and interconnections that are present in the variables. This will be discussed further in section 6.1.

5.2.2 Per Track Type Analysis

The previous analyses make no distinction for the classifier performance for the different track types, long and upstream. It is however expected that most performance increases are to be attained for the upstream tracks. This distinction is shown in Table 6 and Figure 41, which show the results for classifiers trained to identify either long or upstream electrons. The variables used for the *Current Training* classifier can be found in the appendix in sections 8.1.2 and 8.1.3 for long and upstream electrons respectively. The *Proposed Training* classifier features the same lists, supplemented with the variables added in this work. The *ProbNNBrem* classifier has only this supplement as a training set. It is immediately clear that the classification improvements are far larger for upstream electrons, with a decrease of FPR of almost 60% at a TPR of 0.7. The percentual AUC increase is also larger for upstream particles by a factor of 40. Although the effects of the newly added bremsstrahlung related variables is most pronounced for upstream electrons, the increase in performance for long tracks is not negligible.

Table 6: Table summarising the achieved performance increases for the two track types under investigation, upstream and long. The Current Training list of variables differs for long and upstream tracks, and can be found in the appendix in sections 8.1.2 and 8.1.3 respectively. The variables added in this work are appended to this list to obtain the Proposed Training variable list. Note that the TPR comparisons for long and upstream electrons are not equal.

	Current Training	Proposed Training	Improvement
Long			
AUC	0.9977	0.9981	0.38‰increase
FPR at TPR = 0.9	3.10×10^{-3}	2.32×10^{-3}	25.1% reduction

Upstream			
AUC	0.954	0.967	1.45% increase
FPR at TPR = 0.7	37.4×10^{-3}	15.6×10^{-3}	58.4% reduction

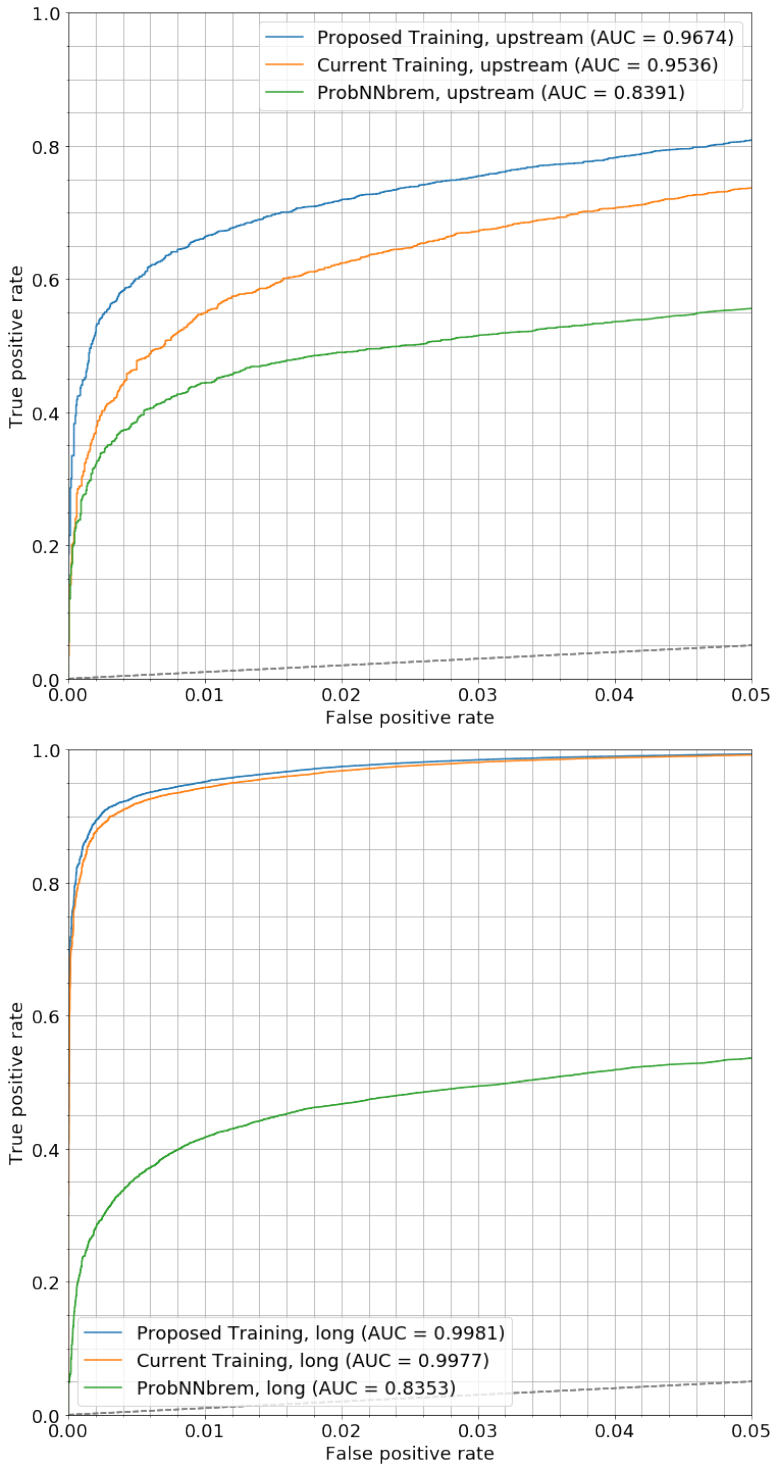

Figure 41: ROC curves for classifiers tasked with identifying either upstream (top) or long (bottom) electrons. Note that these plots do not feature the full length of the FPR axis and instead show the portion of the graph most relevant to analyses.

5.2.3 Prioritisation

With the performance of the ProbNNBrem suite of variables established, the current section will turn to the effectiveness of the variables present in the set currently used for training, and those newly added in this work. To grade this effectiveness, the approach from section 5.1.2 is followed to obtain the 'pseudo-AUC', i.e. the AUC for a pair of variables multiplied by the combinatorial multiplier. The average of this product is calculated for all pairs containing a specific variable, to arrive at a generalised performance, or effectiveness, of that variable. The results are displayed in Figure 42 for the set of upstream classifier variables, and in Figure 43 for the set for long track electrons.

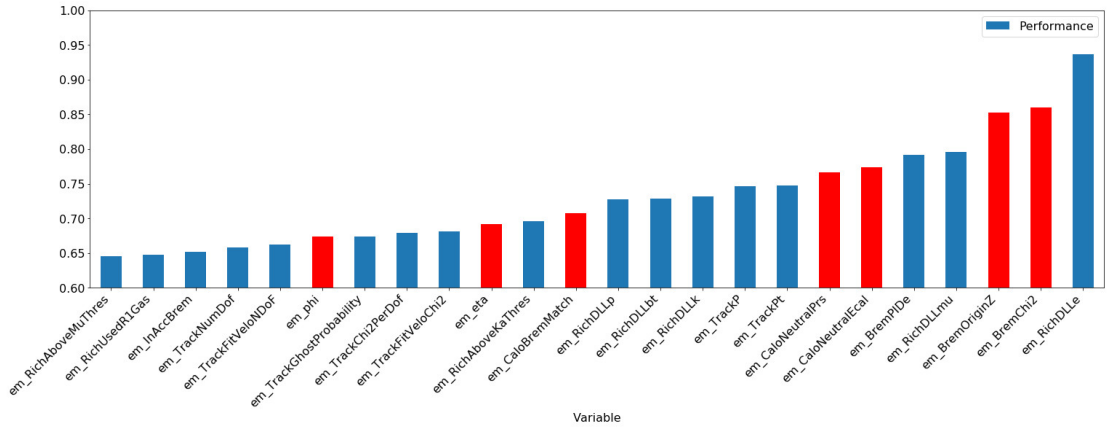

Figure 42: Variables used for upstream electron classification, sorted by effectiveness as calculated in sections 5.1.2 and 5.2.3. Newly added variables part of the ProbNNBrem set are marked in red. An enlarged version of this image can be found in the appendix on page 63.

Figure 42 shows that em_BremOriginZ and em_BremChi2 both grant a sizeable contribution to upstream electron identification efficiency. This effect is still present when only looking at AUC and not accounting for the combinatorial multiplier. The calorimeter-derived variables em_CaloNeutralPrs and em_CaloNeutralECAL also score relatively high on this scale, showing that although they are not normalised to the electron energy, they contain relevant information with regards to electron identification. As expected, the angular variables do not offer much information themselves, and the combinatorial effect they provide is not enough to offset this to a large extent. Their contribution is more apparent by solely looking at their combinatorial multipliers in Figure 34.

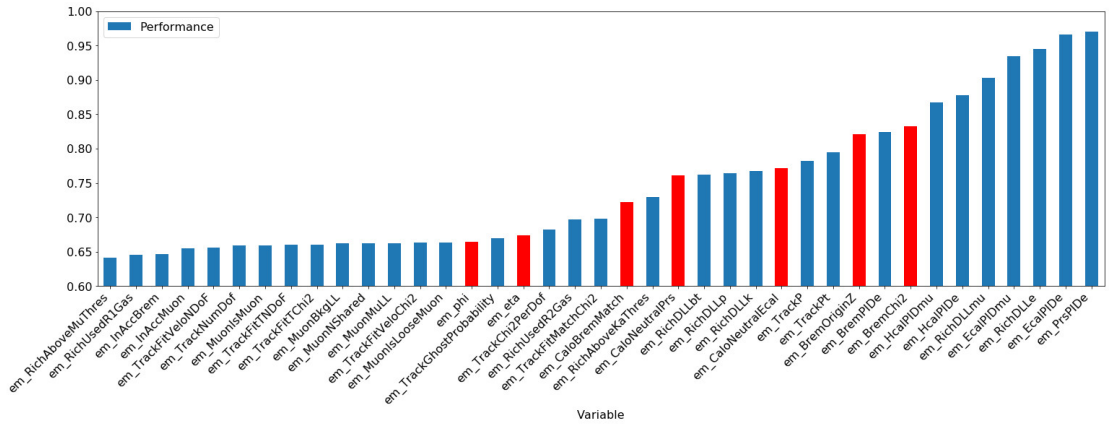

Figure 43: Variables used for long track electron classification, sorted by effectiveness as calculated in sections 5.1.2 and 5.2.3. Newly added variables part of the ProbNNBrem set are marked in red. An enlarged version of this image can be found in the appendix on page 64.

The same hierarchy in performance among the newly added variables is found for the long track analysis in Figure 43. In this case however, there are several variables that perform far better than for example em_BremChi2. These variables are all focused on particle identification in various forms, and due to the greater amount of calorimeter information available for long track particles, their effectiveness is enhanced. The majority of the added ProbNNBrem variables still performs better than the average of the already existing classifier variables, with the exception of the angular variables.

5.2.4 Overtraining

This final result section will cover the relation between sample size and classification accuracy, as well as a note on overtraining. As mentioned before, the Gradient Boosting Classifier is deemed resistant to overtraining, and this is partially reflected in the results obtained in this work. For all of the covered experiments, the split between training sample and testing sample was made at 50%/50%, and at no point has there been a case where an increase in training samples led to a substantial decrease in correctly classified testing samples. Figure 44 does feature a decrease in AUC of about 1 per mille when increasing the training sample count from 10 000 to 300 000, but this is within the statistical variation encountered when performing the same training and classification process multiple times. This variation is primarily encountered for the smaller training sample sizes, the full 300 000 event size has variations below the per mille scale.

With regards to the amount of samples required to train a classifier, the middle graph in Figure 44 reveals that 10 000 events can already train a classifier well enough to perform similarly to one trained on 300 000 events. A training sample size as small as 500 events results in the ROC curves shown in the bottom graph of Figure 44. Here, statistical effects are becoming considerable and AUC can no longer be relied on as much, but classification performance is seen to be still present.

One can conclude that, at least within the data set for this simulated decay, the amount of samples used for training is not as important as it may be for other machine learning applications. The addition of more training samples is shown to be not detrimental to classification performance, and performance is shown to be considerable for even a small amount of training events.

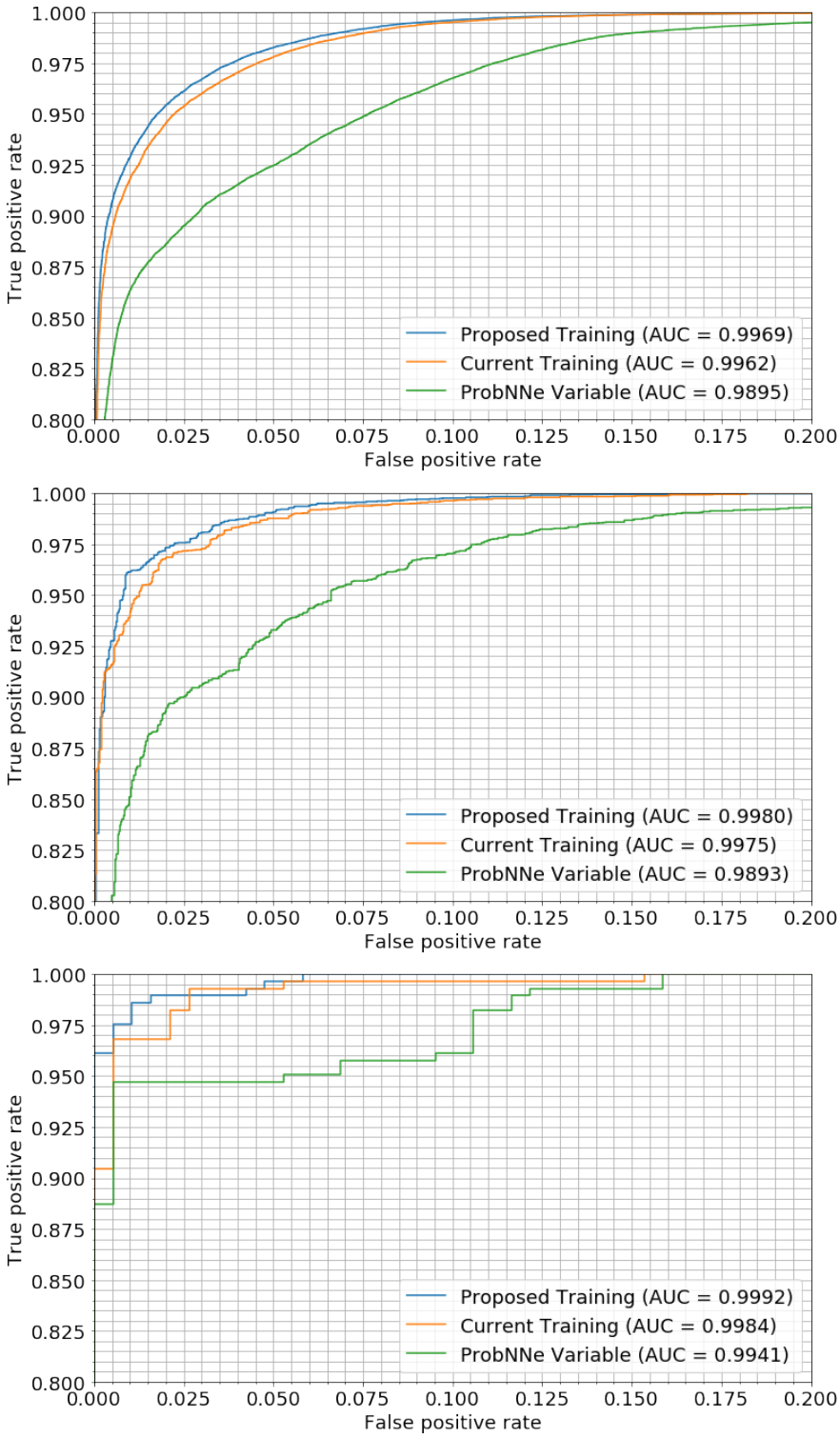

Figure 44: ROC curves for the three classifiers used throughout in this work. Training samples used for each classifier are 300 000, 10 000, and 500, from top to bottom. AUC values are far less reliable for the bottom graph but are left in for completeness.

6 Limitations and Further Research

What follows will be a critical look back on the performed research, as well as a look ahead to see how future endeavours can profit from and improve upon this work. The constituent sections of this chapter are ordered in the same manner.

6.1 Limiting Factors

Perhaps the greatest limitation encountered is the inability to accurately gauge the effect a certain variable has on classifying performance and training bias. Even though sections 5.1.2 and 5.2.1 specifically deal with the isolation of variables in training, higher order effects and correlations can not be ruled out entirely. As has been mentioned in section 4.4, the data set has been resampled so that there is no difference in the distributions of pseudorapidity between the signal electrons and the background. There are however other (kinematic) variables that may have inadvertently introduced a bias in the training sample.

A different issue lies in the quantification of achieved performance increases with respect to the currently used classification variables, which is difficult to solidify. The Gradient Boosting Classifier (GBC) used throughout this work is not the same as the Artificial Neural Network (ANN) that is used in the LHCb reconstruction software. Because of this, any direct comparison between the two is risky, which is made worse by the complex nature of neural networks in general. Relative performance increases for the same GBC are used wherever possible in order to mitigate this issue as well as possible.

Finally, there is the matter of translating the results from this work to real data. In all of the experiments performed, the data has consisted of simulated events. Although the simulation software for the LHCb detector is considered accurate for the most part, there is no guarantee that a GBC or ANN trained on simulated data will perform as well on real data. At the same time, training on real data can prove to be hazardous as there is no way of determining a particles identity with full confidence. The next section will briefly touch on this subject.

6.2 Improving Validation

This section will consist of thoughts on encountered issues and possible avenues of improvement for obtained results.

Although section 5.2.4 reveals that the size of the training sample is not of highest importance, this conclusion is drawn based on the properties of the electrons in this specific decay. By including other decay channels in the training set, one might assume that a more general 'signature' of an electron can be created. This is however generally not the case, as even with a different selection of events within the same channel, differences in classification performance can arise. A calibration will generally be required, either data-driven or extracted from simulations.

In the data set that has been used throughout this work, training has only been done with the variables concerning one of the two leptons. There is however no reason why the other lepton in the event could not be used for training purposes, effectively doubling the sample size. This has not been done in this work due to the sample size already reaching the limit of the available computational resources and added complexity of the algorithm. This issue can be solved by performing the training process on a

centralised supercomputer or by taking steps to limit the memory usage, for example by splitting the training sample and loading each set into memory as it is required. Section 5.2.4 showed that increasing the sample size provided no substantial benefit in the classification performance, and it was thus deemed not necessary to use this extra data.

With regards to the probable discrepancy between simulated results and results obtained from real data, there is the possibility of assigning 'pseudo-truth-flags'. These flags would function the same as those that mark simulated electrons as being true electrons. This information is not available for real data for obvious reasons, but could be approximated by making very tight cuts on the data set. For example, the variable expressing the ratio between energy lost in the calorimeter and total energy, *CaloEoverP*, is one of the key variables that sets electrons apart from other particles. By making stringent cuts on this variable and other variables that are known to have a signature distribution for electrons, a training sample could be constructed for which one can be relatively confident that it consists solely of true electrons. These cuts will however introduce a strong bias, as only 'well behaved' electrons will find their ways into this sample, which will impact classification performance on real data.

7 Conclusion

The Standard Model of Particle Physics provides us with the most accurate description of nature ever achieved, and is a testament to human ingenuity. Predictions of the fundamental constituents of matter and their interactions are possible to an unrivalled degree of precision. However, it is not a complete theory. For example, it does not include the force of gravity, nor does it give an explanation for the presence of dark energy and dark matter, which seem to account for 95% of our universe's energy content. These, among other shortcomings, are the reason for the constant probing of the Standard Model. Attempts to find the limits of the predictive power of the model as well as apparent inconsistencies. One of the predictions of the Standard Model which is under continuous assault from physicists worldwide is that of the universality of leptons. One of the collaborations searching for violations of this universality is that of the LHCb experiment.

The aim of this project has been to aid this search by extracting a larger amount of usable data from previously collected LHCb data. The following question was postulated to serve as a guideline:

Can bremsstrahlung-related information enhance the particle identification of upstream electrons in the LHCb detector?

The results obtained in this work indicate that this question can be answered positively. The various experiments performed with the variables in the ProbNNBrem package show improvements in false positive rejections in the order of 60% for upstream electrons, as well as overall performance increases exceeding 1.4%. The improvements are not limited to upstream tracks, with identification of long track electrons also benefiting from the added bremsstrahlung information.

What remains to be done is the generalisation of the training data and the translation to real data. Using other decay channels may result in classifiers that are more capable of generalisation, and less susceptible to bias, although this may result in a lower per-channel performance. The final step would then be the application of the classifiers to real data, with some form of performance monitoring.

Once completed, the tool can be applied in the many analyses that are performed at LHCb. Together with an ever increasing total amount of data, the extraction of more usable data will result in ever increasing precision, which may one day lead us to a new realm of physics.

References

[1] Y. Amhis et al. "Averages of b-hadron, c-hadron, and τ-lepton properties as of summer 2016". In: *The European Physical Journal C* 77.12 (Dec. 2017). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-017-5058-4. URL: http://dx.doi.org/10.1140/epjc/s10052-017-5058-4.

[2] K. M. and. "Heavy Flavour and Quarkonia production at LHCb". In: *Journal of Physics: Conference Series* 878 (July 2017), p. 012011. DOI: 10.1088/1742-6596/878/1/012011. URL: https://doi.org/10.108%5C%2F1742-6596%5C%2F878%5C%2F1%5C%2F012011.

[3] F. Anghinolfi. *Silicon strip detectors and their readout electronics*. Electronic Systems for Experiments Seminar presentation. 2009. URL: https://indico.cern.ch/event/69666/attachments/1029489/1466033/ESE_0_all.pdf.

[4] S. Bifani et al. "Review of lepton universality tests in B decays". In: *Journal of Physics G: Nuclear and Particle Physics* 46.2 (Dec. 2018), p. 023001. ISSN: 1361-6471. DOI: 10.1088/1361-6471/aaf5de. URL: http://dx.doi.org/10.1088/1361-6471/aaf5de.

[5] M. Bordone et al. "A three-site gauge model for flavor hierarchies and flavor anomalies". In: *Physics Letters B* 779 (Apr. 2018), pp. 317–323. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2018.02.011. URL: http://dx.doi.org/10.1016/j.physletb.2018.02.011.

[6] M. Bordone et al. "Low-energy signatures of the PS3 model: from B-physics anomalies to LFV". In: *Journal of High Energy Physics* 2018.10 (Oct. 2018). ISSN: 1029-8479. DOI: 10.1007/jhep10(2018)148. URL: http://dx.doi.org/10.1007/JHEP10(2018)148.

[7] E. Bos. "Reconstruction of charged particles in the LHCb experiment". PhD thesis. Vrije Universiteit Amsterdam, 2010.

[8] L. Breiman. "Discussion of additive logistic regression: A statistical view of boosting". In: *Annals of Statistics* 28.2 (2000), pp. 374–377.

[9] A. Buja. "Discussion of additive logistic regression: A statistical view of boosting". In: *Annals of Statistics* 28.2 (2000), pp. 387–391.

[10] M. Carena et al. "Z' gauge bosons at the Fermilab Tevatron". In: *Physical Review D* 70.9 (Nov. 2004). ISSN: 1550-2368. DOI: 10.1103/physrevd.70.093009. URL: http://dx.doi.org/10.1103/PhysRevD.70.093009.

[11] A. Ceccucci, Z. Ligeti, and Y. Sakai. *CKM Quark-Mixing Matrix*. Jan. 2018. URL: https://pdg.lbl.gov/2019/reviews/rpp2019-rev-ckm-matrix.pdf.

[12] CERN. *ROOT data analysis framework*. URL: root.cern.ch (visited on 06/14/2020).

[13] C. Collaboration. "Search for $Z' \to e^+e^-$ Using Dielectron Mass and Angular Distribution". In: *Physical Review Letters* 96.21 (May 2006). ISSN: 1079-7114. DOI: 10.1103/physrevlett.96.211801. URL: http://dx.doi.org/10.1103/PhysRevLett.96.211801.

[14] T. A. Collaboration et al. In: *Physics Reports* 427.5-6 (May 2006), pp. 257–454. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2005.12.006. URL: http://dx.doi.org/10.1016/j.physrep.2005.12.006.

[15] T. L. Collaboration et al. "The LHCb Detector at the LHC". In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08005–S08005. DOI: 10.1088/1748-0221/3/08/s08005. URL: https://doi.org/10.1088%2F1748-0221%2F3%2F08%2Fs08005.

[16] G. Corti et al. "Software for the LHCb experiment". In: *IEEE Symposium Conference Record Nuclear Science 2004*. Vol. 4. 2004, pp. 2048–2052.

[17] A. Datta, J. Kumar, and D. London. "The B anomalies and new physics in $b \to se^+e^-$". In: *Physics Letters B* 797 (2019), p. 134858. ISSN: 0370-2693. DOI:

https://doi.org/10.1016/j.physletb.2019.134858. URL: http://www.sciencedirect.com/science/article/pii/S0370269319305726.

[18] M. Dittmar, A.-S. Nicollerat, and A. Djouadi. "Z' studies at the LHC: an update". In: *Physics Letters B* 583.1-2 (Mar. 2004), pp. 111–120. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2003.09.103. URL: http://dx.doi.org/10.1016/j.physletb.2003.09.103.

[19] K.-L. Du and M. Swamy. "Multilayer Perceptrons: Architecture and Error Backpropagation". In: Dec. 2014, pp. 83–126. ISBN: 978-1-4471-5570-6. DOI: 10.1007/978-1-4471-5571-3_4.

[20] Y. Freund and R. E. Schapire. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting". In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 119–139. ISSN: 0022-0000. DOI: https://doi.org/10.1006/jcss.1997.1504. URL: http://www.sciencedirect.com/science/article/pii/S002200009791504X.

[21] E. Haug and W. Nakel. *The Elementary Process of Bremsstrahlung*. World Scientific lecture notes in physics. World Scientific Publishing Company, 2004. ISBN: 9789812795007. URL: https://books.google.nl/books?id=v4FMtIwTri8C.

[22] T. Head. "The LHCb trigger system". In: *Journal of Instrumentation* 9.09 (Sept. 2014), pp. C09015–C09015. DOI: 10.1088/1748-0221/9/09/c09015. URL: https://doi.org/10.1088%2F1748-0221%2F9%2F09%2Fc09015.

[23] W. Heitler. "Über die bei sehr schnellen Stößen emittierte Strahlung". In: *Zeitschrift für Physik* 84 (Mar. 1933), pp. 145–167.

[24] G. James et al. "An Introduction to Statistical Learning: with Applications in R." In: Springer, 2017. Chap. 8, pp. 303–336. ISBN: 978-1-4614-7137-0.

[25] D. Lange. "The EvtGen particle decay simulation package". In: *Nucl. Instrum. Meth. A* 462 (2001). Ed. by S. Erhan, P. Schlein, and Y. Rozen, pp. 152–155. DOI: 10.1016/S0168-9002(01)00089-4.

[26] A. Maevskiy et al. *Fast Data-Driven Simulation of Cherenkov Detectors Using Generative Adversarial Networks*. 2019. arXiv: 1905.11825 [physics.ins-det].

[27] A. Natekin and A. Knoll. "Gradient Boosting Machines, A Tutorial". In: *Frontiers in neurorobotics* 7 (Dec. 2013), p. 21. DOI: 10.3389/fnbot.2013.00021.

[28] M. E. Peskin and D. V. Schroeder. *An Introduction to Quantum Field Theory*. Perseus Books, 1995.

[29] N. Seliya, T. M. Khoshgoftaar, and J. V. Hulse. "A Study on the Relationships of Classifier Performance Metrics". In: *2009 21st IEEE International Conference on Tools with Artificial Intelligence*. 2009, pp. 59–66.

[30] M. Stahl. "Machine learning and parallelism in the reconstruction of LHCb and its upgrade". In: *Journal of Physics: Conference Series* 898 (Oct. 2017). DOI: 10.1088/1742-6596/898/4/042042.

[31] M. Tanabashi et al. "Review of Particle Physics". In: *Phys. Rev. D* 98 (3 Aug. 2018), p. 030001. DOI: 10.1103/PhysRevD.98.030001. URL: https://link.aps.org/doi/10.1103/PhysRevD.98.030001.

[32] J. v. Tilburg. "Track simulation and reconstruction in LHCb". PhD thesis. Vrije Universiteit Amsterdam, 2005.

[33] M. C. van Veghel. "Pursuing forbidden beauty: Search for the lepton-flavour violating decays $B^0 \to e^{\pm}\mu^{\mp}$ and $B_s^0 \to e^{\pm}\mu^{\mp}$ and study of electron-reconstruction performance at LHCb". PhD thesis. Groningen: University of Groningen, July 2020.

[34] E. Wiechert. *Elektrodynamische Elementargesetze*. Jan. 1901. DOI: 10.1002/andp.19013090403. URL: https://doi.org/10.1002/andp.19013090403.

[35] A. Zee. *Quantum field theory in a nutshell*. 2nd ed. Princeton University Press, 2010.

8 Appendices

8.1 Full List of NN variables

8.1.1 Master

Table 7: Masterlist used for the classification of particles.

Event
'NumProtoParticles',
'NumCaloHypos',
'NumLongTracks',
'NumDownstreamTracks',
'NumUpstreamTracks',
'NumVeloTracks',
'NumTTracks',
'NumGhosts',
'NumMuonTracks',
'NumPVs',
'NumRich1Hits',
'NumRich2Hits',
'NumVeloClusters',
'NumITClusters',
'NumTTClusters',
'NumOTClusters',
'NumSPDHits',
'NumMuonCoordsS0',
'NumMuonCoordsS1',
'NumMuonCoordsS2',
'NumMuonCoordsS3',
'NumMuonCoordsS4'

Muon
'InAccMuon',
'MuonIsLooseMuon',
'MuonIsMuon',
'MuonNShared',
'MuonMuLL',
'MuonBkgLL'

Tracking
'TrackP',
'TrackPt',
'TrackChi2PerDof',
'TrackType',
'TrackLikelihood',
'TrackHistory',
'TrackGhostProbability',
'TrackCloneDist',
'TrackFitMatchChi2',
'TrackFitVeloChi2',
'TrackFitVeloNDoF',
'TrackFitTChi2',
'TrackFitTNDoF',
'TrackMatchChi2',
'TrackDOCA',
'TrackNumDof'

ECAL
'InAccEcal',
'CaloChargedSpd',
'CaloChargedPrs',
'CaloChargedEcal',
'CaloElectronMatch',
'CaloTrMatch',
'CaloEcalE',
'CaloEcalChi2',
'CaloClusChi2',
'EcalPIDe',
'EcalPIDmu',
'CaloTrajectoryL'

CombDLL
'CombDLLe',
'CombDLLmu',
'CombDLLpi',
'CombDLLk',
'CombDLLp'

HCAL
'InAccHcal',
'CaloHcalE',
'HcalPIDe',
'HcalPIDmu'

PRS
'InAccPrs',
'CaloPrsE',
'PrsPIDe'

Brem
'InAccBrem',
'CaloNeutralSpd',
'CaloNeutralPrs',
'CaloNeutralEcal',
'CaloBremMatch',
'CaloBremChi2',
'BremPIDe'

RICH
'RichUsedAero',
'RichUsedR1Gas',
'RichUsedR2Gas',
'RichAboveElThres',
'RichAboveMuThres',
'RichAbovePiThres',
'RichAboveKaThres',
'RichAbovePrThres',
'RichAboveDeThres',
'RichDLLe',
'RichDLLmu',
'RichDLLpi',
'RichDLLk',
'RichDLLp',
'RichDLLd',
'RichDLLbt'

VELO
'VeloCharge'

SPD
'InAccSpd',
'CaloSpdE'

8.1.2 Long Electrons

Table 8: Variables used for the classification of long track electrons.

'TrackP',	'RichUsedR1Gas',	'MuonMuLL',	'InAccBrem',
'TrackPt',	'RichUsedR2Gas',	'MuonIsMuon',	'BremPIDe'
'TrackChi2PerDof',	'RichAboveMuThres',	'MuonNShared',	
'TrackNumDof',	'RichAboveKaThres',	'InAccMuon',	
'TrackGhostProbability',	'RichDLLe',	'MuonIsLooseMuon',	
'TrackFitMatchChi2',	'RichDLLmu',	'EcalPIDe',	
'TrackFitVeloChi2',	'RichDLLk',	'EcalPIDmu',	
'TrackFitVeloNDoF',	'RichDLLp',	'HcalPIDe',	
'TrackFitTChi2',	'RichDLLbt',	'HcalPIDmu',	
'TrackFitTNDoF',	'MuonBkgLL',	'PrsPIDe',	

8.1.3 Upstream Electrons

Table 9: Variables used for the classification of upstream track electrons.

'TrackP',	'RichDLLe',
'TrackPt',	'RichDLLmu',
'TrackChi2PerDof',	'RichDLLk',
'TrackNumDof',	'RichDLLp',
'TrackGhostProbability',	'RichDLLbt',
'TrackFitVeloChi2',	'InAccBrem',
'TrackFitVeloNDoF',	'BremPIDe'
'RichUsedR1Gas',	
'RichAboveMuThres',	
'RichAboveKaThres',	

8.2 Additional Images

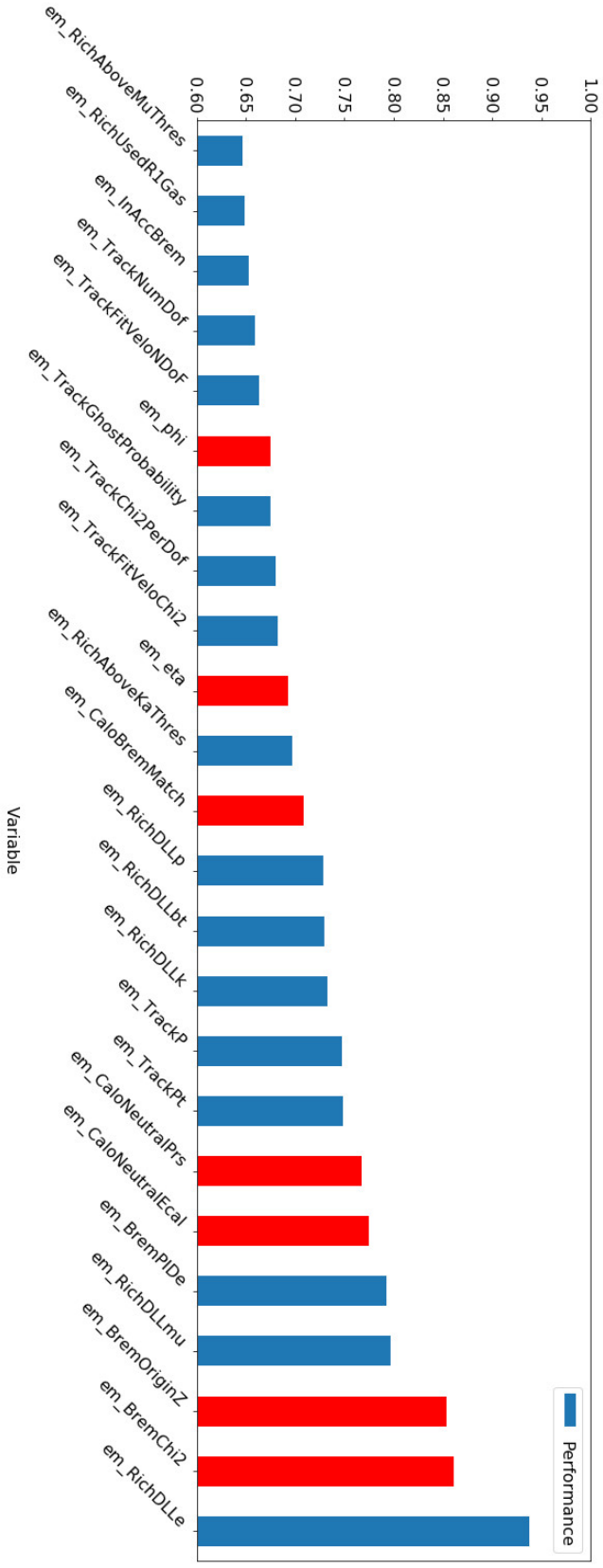

Figure 45: Enlarged version of Figure 42

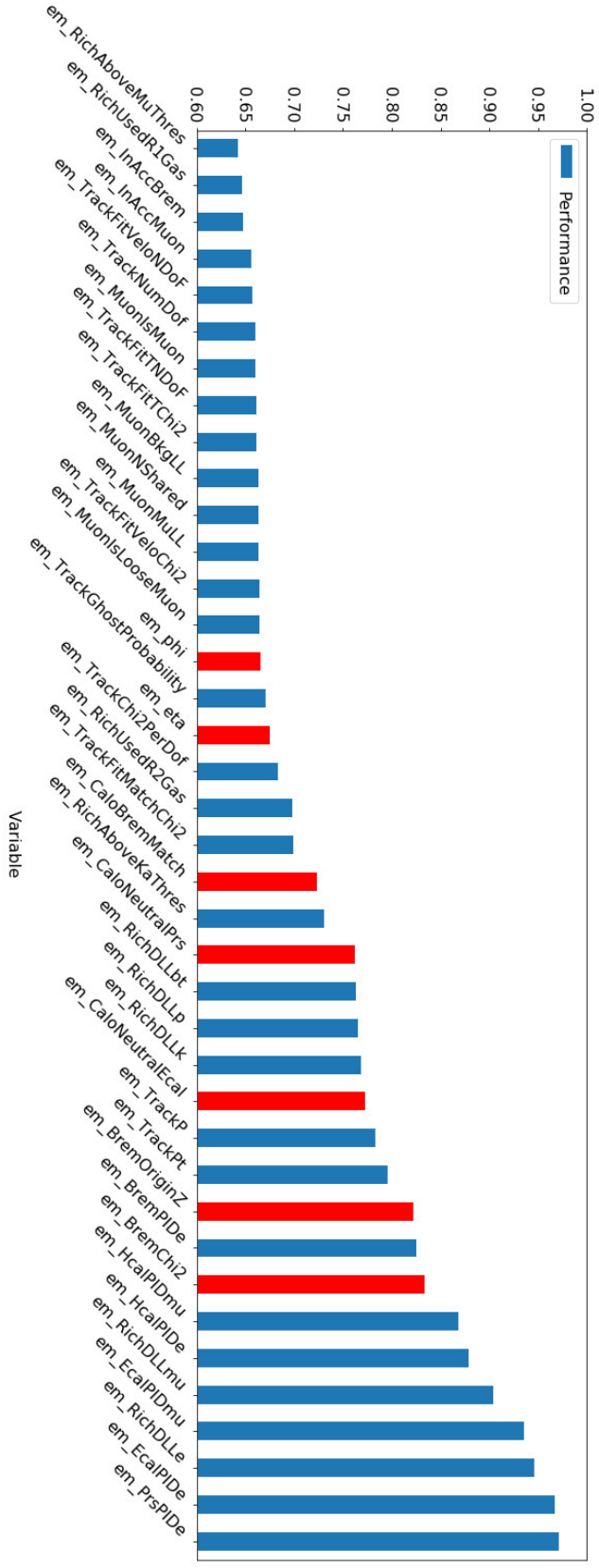

Figure 46: Enlarged version of Figure 43