UNIVERSITÀ DEGLI STUDI DI NAPOLI
"FEDERICO II"

Scuola Politecnica e delle Scienze di Base
Area Didattica di Scienze Matematiche Fisiche e Naturali

Dipartimento di Fisica "Ettore Pancini"

Laurea Magistrale in Fisica

Electron identification in the Emulsion detector of SHiP experiment using machine learning techniques

Relatori:
Prof. Giovanni De Lellis
Dtt.ssa Antonia Di Crescenzo
Correlatore:
Dtt.ssa Autilia Vitiello

Candidato:
Maria De Luca
Matricola *N94/418*

Anno Accademico 2018/2019

Contents

Introduction

The SHiP (Search for Hidden Particles) experiment is a general purpose fixed target facility proposed at CERN SPS with the aim to provide a unique platform to search for New Physics at the Intensity Frontier. The primary physics goal of the SHiP facility consists of exploring hidden portals predicted by a large number of theories beyond the Standard Model and make measurements with tau neutrinos. Using a 400 GeV/c proton beam delivered onto a hybrid Tungsten and Titanium-Zirconium doped Molibdenum (TZM) thick target, the experiment can probe the existence of very weakly interacting long lived particles including Heavy Neutral Leptons - right-handed partners of the active neutrinos, vector, scalar, axion portals to the Hidden Sector. The designed facility is a copious neutrino factory. SHiP experiment allows to perform a wide neutrino physics program: with the unprecedented statistics of ν_τ and $\overline{\nu}_\tau$ it aims to perform the first direct observation of the $\overline{\nu}_\tau$ charged current interactions and measure tau neutrino cross section. It is also suited for the direct search of the Light Dark Matter particles that can be discovered through the observation of their elastic scattering off the electrons of the Scattering and Neutrino Detector (SND). The SND exhibits a modular structure consisting of alternating Emulsion Cloud Chamber and Compact Emulsion Spectrometer walls with electronic detector planes. The ECC is employed as vertex detectors, both primary and short lived particles decay. The CES is designed for the charge and momentum measurement. The Target Trackers are the electronic detectors to provide the time stamp of each event and to connect the emulsion tracks with those reconstructed in the downstream spectrometer. The Emulsion Cloud Chamber can be also considered as an electromagnetic calorimeter with passive plates alternating with nuclear emulsion films acting as a tracking detector. The signature of the interaction of the LDM is the electromagnetic shower produced from the recoil of electron scattered by the LDM particles. On the other hand, also the neutrino interaction can produce an

electron that induces electromagnetic shower in the brick. Since, the most copious source of electromagnetic showers are the photons produced by the neutral pion decays, the electron identification and its energy measurement is crucial to distinguish showers originated by electrons or photons.
This thesis work was performed within the Naples neutrino group that participates to the SHiP experiment. It has been mainly devoted to the study of the reconstruction of the electromagnetic shower in the Emulsion Cloud Chamber with the final task to estimate the electron energy. The thesis is focused on the DESY test beam exposure performed in October 2019 in Hamburg, where the electrons beams are fired on the ECC brick. The challenge we face is to find the shower segments in an environment with high noise density. It is addressed with an innovative algorithm that uses Machine Learning techniques to distinguish the signal segment from the noise one.
The content of this thesis is structured in six chapters:

- **Chapter 1** describes the unsolved problems of the Beyond Standard Model that will be explored with the SHiP facility. A detailed study on the vector portal mediator that is the Dark Photon from which the Light Dark Matter particles can be originates, and an overview of the neutrino physics is provided;
- **Chapter 2** provides a description of the SHiP experiment: the Target complex, the muon shield, the Scattering and Neutrino Detector, the Decay Vessel and Hidden Sector detectors are briefly illustrated in their functionalities and design aspects;
- **Chapter 3** reports a brief overview of the artificial intelligence, in particular of the machine learning techniques and the classifier that we will use in this thesis;
- **Chapter 4** describes the physics of electromagnetic showers and the DESY test beam exposure conducted at Hamburg in October 2019, with the aim to test the capability to reconstruct electromagnetic showers in an Emulsion Cloud Chamber;
- **Chapter 5** describes the procedure for the localization of the showers in a defined space region with the new geometrical parameters for the construction of the rectangles that follow the development of the shower in the brick. New features are introduced to be passed to the classifier;

- **Chapter 6** is dedicated to identification of signal segments through the machine learning techniques- the Random Forest. Different tests are made to study the performances of the classifier changing the datasets. The energy calibration procedure is conducted and a comparison with the classical case is made.

Chapter 1

The SHiP physics case

1.1 The Standard Model of Particle Physics

The Standard Model (SM) of Particle Physics is the most important achievement of high energy physics to date. This highly elegant theory provides a clear description of the subatomic world. It describes all known elementary particles: *fermions*, spin-1/2 particles which are the building blocks of matter and *bosons*, spin-0 particles which consist of the force carrying particles.

The fermions occur in two basic types called *quarks* and *leptons*. Each group consists of six particles organized in three couples or generations. The first generation is made by *up* and *down* quarks followed by the *charm* and *strange* quarks, then the *top* and *bottom* (or *beauty*) quarks. u, c, t are called up-type quarks (q_u) and they share the same electric charge of (2/3) while the down-type quark (q_d), d, s and b, have an identical ($-1/3$) charge. Quarks also are in three different colours and mix in such way to form colourless objects with an entire charge, protons and neutrons. Leptons, also in three generations, are the *electron* and the *electron neutrino*, the *muon* and the *muon neutrino*, and the *tau* and the *tau neutrino*. The electron, the muon and the tau have an electric charge (-1) and a mass whereas neutrinos are electrically neutral and massless. For each of those particles, there is an associated antiparticle which has identical physical properties such as mass but opposite internal quantum numbers which intrinsically define the particle. The bosons mediate three of the four fundamental interactions included in the SM: electromagnetism and the weak force unified in the electro-weak force carried by γ, $W^{\pm}$ and Z^0- they are the result of the

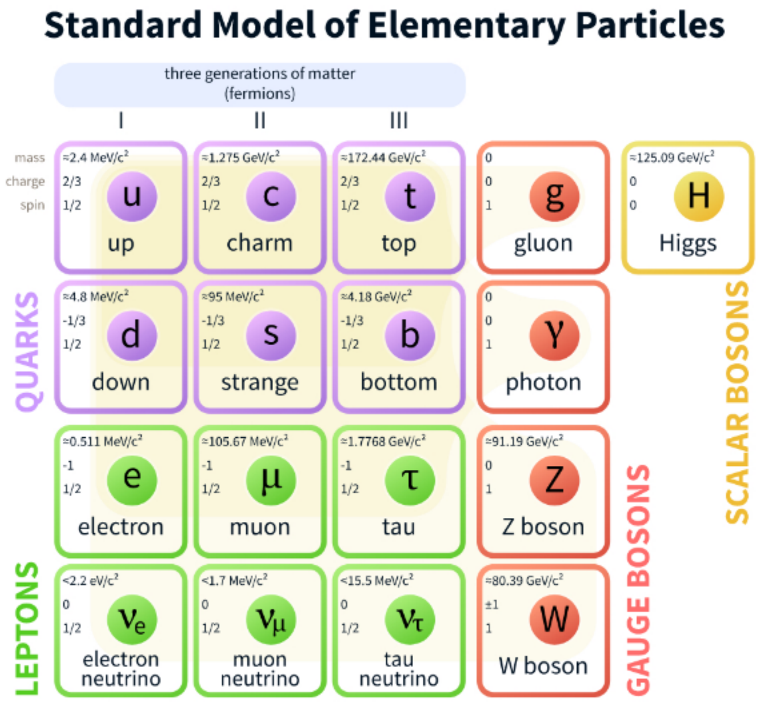


Figure 1.1: Scheme of the Standard Model of elementary particles. In the first three coulums, divided in generations, there are the fermions: leptons and quarks. The forth coulumn shows bosons, mediators of interactions. The Higgs boson is the manifestation of the Higgs mechanism; it is responsible for providing masses to fermions, W and Z bosons.

U(1)$\otimes$ SU(2) symmetry- and the strong force carried by the *gluons*, as a result of SU(3) simmetry. The fourth fundamental interaction, gravity, is missing. Although it is the force we experience most in our daily life, it is negligible in particle physics. The SM was finalized in the 1970s as a joint effort of many prominent physicists and has been studied extensively in the last decades, especially thanks to the contributions given by LEP and Tevatron; its great success is due to the ability to accurately describe the phenomenology of elementary particles observed experimentally, demonstrating an extraordinary agreement between theory and experiment. The latest success was the discovery of the Brout–Englert–Higgs field by the Collaborations of ATLAS [1] and CMS at CERN's Large Hadron Collider in 2012 [2]. Both experiments detected the quantised excitation of the BEH field, the so-called

Higgs boson with a mass of about 126 GeV [3]. With the discovery of the H^0 boson, the picture of all particles predicted by the model is complete, see Figure 1.1. The BEH mechanism explains how elementary particles gain mass [4].
The Standard Model is described by a Lagrangian -as the result of a SU(3)⊗SU(2)⊗U(1) symmetry- and it is found in Equation 1.1:

$$
\begin{aligned}
\mathcal{L}_{SM} = & \overbrace{-\frac{1}{2}TrG_{\mu\nu}G_{\mu\nu} - \frac{1}{2}TrW_{\mu\nu}W_{\mu\nu} - \frac{1}{2}TrB_{\mu\nu}B_{\mu\nu}}^{Bosons\ kinematics\ terms} \\
& + \overbrace{\sum_{\psi=\chi_L,e_r,q_L,u_R,d_r} i\overline{\psi}\gamma^{\mu}D_{\mu}\psi}^{Fermions\ kinematics\ terms} \\
& + \overbrace{\sum_{f,g}[-Y_e^{f,g}\overline{\chi}_L^f\phi_R^g - Y_d^{f,g}\overline{q}_L^f\phi_{d_R}^g - Y_u^{f,g}\overline{q}_L^g\phi_{u_R}^g] + h.c.}^{Fermions\ mass\ term} \\
& + \overbrace{|D_\mu\phi|^2 - V(\phi)}^{Higgs\ term}
\end{aligned}
\tag{1.1}
$$

G,W, and B, are the gauge fields; ϕ is the Higgs field with its potential V(ϕ) $=\lambda(\phi^2 - \frac{v^2}{2})$; D_μ is the covariant derivatives; Y is the Yukawa couplings; and f, g run over the flavours and generations respectively. The fermion fields are χ_L^g for the left-handed leptons and e_R^g for the right-handed charged leptons; and q_L^f for the left-handed quarks, and u_R^g and d_R^g for the right-handed up- and down-type quarks respectively. This separation into left- and right-handed fields is due to the chiral nature of the weak force [5]. As it is formulated, the Standard Model remains mathematically consistent and valid as an effective field theory up to a very high energy scale, possibly all the way to the scale of quantum gravity, the Planck scale [6].

1.2 Beyond the Standard Model

Despite this success, the Standard Model cannot be considered a complete theory, since it fails to explain several well known observed phenomena, both in particle physics, cosmology and astrophysics. These major unsolved challenges are commonly known as "Beyond the Standard Model" (BSM) problems. The most relevant are reported in the following.

- **Neutrino oscillations**

  In the SM the neutrino is a Dirac particle with only one state of helicity: the neutrino left-handed and the antinueutrino right-handed. Neutrino and antineutrino are two different particles with zero-mass. This last assumption is in contrast with the observation of neutrinos oscillation by the SNO [7], the SuperKamiokande [8] and the OPERA Collaborations [9]. The phenomenon, independently proposed in the 1960s by Pontecorvo [10] and Maki Nakagawa and Sakata [11, 12], can only occur if the particles involved are massive. It consists in the mixing of neutrinos of different types (or flavors) and leads, for example, to the transformation of an electron neutrino into a muon neutrino. Direct neutrino mass measurements cannot be determined in oscillation experiments, because these experiments are sensitive only to the difference of the squared masses ($\Delta m_{21(31)}$), therefore they do not even provide information on the ordering of the neutrino masses (neutrino mass hierarchy problem). The experimentally established upper limit for neutrino mass is about 250,000 times smaller than the mass of the electron. At this time, the origin of the neutrino masses and mixings remains an outstanding problem of particle physics. It is trivial to add right-handed neutrinos, which would be sterile and unobservable, to generate Dirac or Yukawa mass terms for neutrinos. Adding these terms would allow for the observed masses, but they would not justify the different magnitude of neutrino masses and those of the others fermions. Their unusually small mass suggest that the neutrino mass generation mechanism is different to that of charged fermions. An interpretation of this behaviour can be found considering the connection with the New Physics: if neutrino was a Majorana particle (i.e. particles that coincide with their own antiparticle) instead of Dirac particles, the way through which this particle acquire mass wold be explained by the See-Saw mechanism [13, 14].

- **The Baryon Asymmetry of the Universe**

  In the SM every particle has a counterpart degenerate in mass and with quantum numbers and charges of opposite sign. Matter and antimatter particles are always produced as a pair and, if they come in contact, annihilate one another, leaving behind pure energy [15]. During the first fractions of a second after the Big Bang, the hot and dense universe was buzzing with particle-antiparticle pairs popping in and out of existence. If matter and antimatter was created and destroyed together, the Universe would contain nothing but energy. However, a tiny portion of matter managed to survive: the universe seems to consist of celestial bodies made of matter only. The relic baryon to photon density ratio today is about $n_b/n_\gamma \approx 10^{-10}$ that is almost ten orders of magnitude bigger than the ratio expected if there was no separation of baryons from antibaryons in the early universe $n_b/n_\gamma \approx 10^{-19}$. So either the universe started with a large quark-antiquark asymmetry or this asymmetry was generated later. The SM does not help in both cases: the initial asymmetry contrasts with the creation of the vacuum (which has baryonic number 0) of $B\overline{B}$ couples and the next generation does not find in the SM a mechanism that justifies it. In the 1967 Sakharov tried to explain the matter-antimatter asymmetry by asserting the need for an interaction that violates the conservation of the baryonic number, C and CP. Nevertheless the amount of CP violation provided by the CKM matrix is not enough to explain current observations [16].

- **Dark matter and dark energy**

  Several observations, such as results obtained by satellites combined with studies on supernovae, have led to the assertion that visible matter, baryonic matter composed of protons and neutrons is only $\sim 4\%$ in the universe. The rest of the universe appears to be made of a mysterious, invisible substance called dark matter $\sim 27\%$ and a force that repels gravity known as dark energy $\sim 68\%$. Unlike normal matter, dark matter does not interact with the electromagnetic force. This means it does not absorb, reflect nor emit light, making it extremely hard to spot. In fact, researchers have been able to infer the existence of dark matter only from the gravitational effect it seems to have on visible matter. The "dark energy" has an unknown origin and nature and it is responsible for the acceleration of the expansion of the universe [17].

These and other phenomena observed in nature which do not find an explanation within the framework of the Standard Model suggest a reason to believe that there should be New Physics; some yet unknown particles or interactions would be needed to explain these puzzles and to answer these questions. At this time, there is no clear guidance on the scale of any New Physics, or on the coupling strength of any

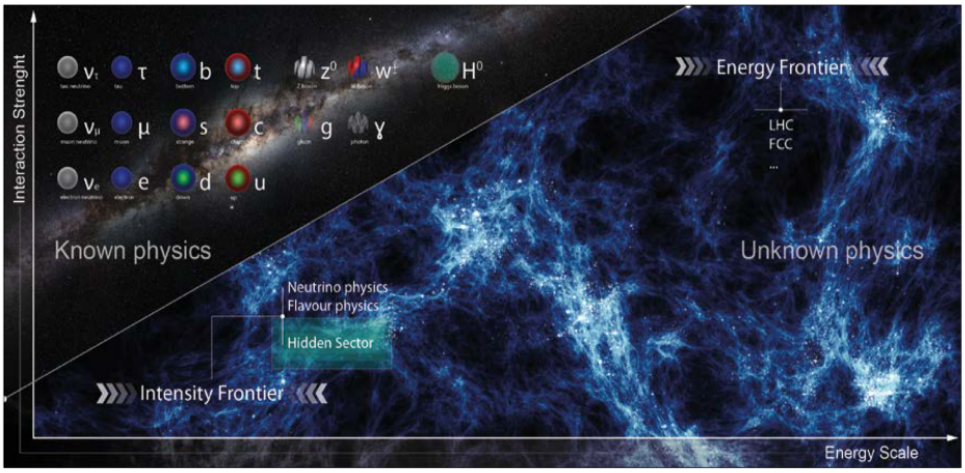


Figure 1.2: The current landscape of known Particle Physics, and the possible extensions: the Energy Frontier, at high energy scale at the Intensity Frontier at low Interaction Strength.

new particles to the SM particles. This lack of theoretical direction necessitates experimental searches at both the *energy* and the *intensity frontiers*, see Figure 1.2. In the first case the interest lies manly in the discovery of new, heavier particles at the TeV energy scale, together with investigation of the quark compositeness. The energy frontier is well investigated by LHC experiments like CMS and ATLAS and by Fermilab, with the future aim of increasing the energy collisions: it is possible that these particles evaded detection so far because they are too heavy to be created by the energy accelerators work today. Alternatively, some of the hypothetical particles can be sufficiently light with a mass below or of the order of the electroweak scale- $\approx$ 100 GeV and feebly interact with particles of the SM. In this case, light new particles can be produced at many high-energy experiments but they were not still observed due to the despite rare events of their production and complexity of their detection.
For the successful production of the hidden particles (to compensate its feeble interaction), this experiment must have the largest possible luminosity. In this respect, beam-dump experiments are ideal as intensity frontier experiments for the search of the GeV-scale hidden particles because their luminosities are several orders of magnitude larger than at colliders. Because of feeble interaction with the SM particles, one can expect its small decay width and long lifetime. So the detector has

to be placed as far as possible from the point of the hidden particle production. In this landscape, the SHiP experiment (Search for Hidden Particles) proposed in 2015, is a new general-purpose experiment at a beam dump facility at CERN Super Proton Synchrotron (SPS). SHiP aims to explore the domain of hidden particles, such as Heavy Neutral Leptons (HNL), dark photons, ligth scalars, supersymmetric particles, axions etc., with masses below O(10) GeV/c^2 and couplings which are orders of magnitude smaller than previously constrained. In addition it will make measurements of tau neutrinos with three orders of magnitude more statistics than available in previous experiments combined. The SHiP facility provides a unique combination of intensity and energy capable of producing the large yields required [18].

1.3 Portals to new physics

There are two different types of Beyond Standard Model theories. In the first case, models assume no new physics between the Fermi and the Planck scales and try to extend the SM using the smallest possible set of fields and renormalizable interactions. For example the *Minimality principle* motivates the νMSM which attempts to explain the pattern of neutrino masses, DM and the observed BAU by introducing three HNLs. The lightest of these, N_1, provides the DM candidate, while N_2, N_3 are responsible for the baryon asymmetry. Through the See-Saw mechanism these HNLs also allow the pattern of neutrino masses and oscillations to be explained [19]. In the second case, there are theories assuming a new energy scale, such as the SuperSimmetry (SUSY), that is a quantum field theory that provides a framework for the unification of particle physics and gravity at the Planck energy scale, $M_p \approx 10^{19}$GeV [20].
A further class of theoretical models with a new energy scale are models with a *dark* or *hidden sector*. The hidden sector is a family of particles that may interact with each other but which do not feel effects of the Standard Model's three forces — strong, electromagnetic and weak [21, 22]. At beam dump experiments, hidden sector particles may be accessible through production:

- in prompt $q\bar{q}$ or qg fusion;
- in decay of mesons produced in the initial proton-on-target collision;

Table 1.1: Portals interactions terms

Portal	Interactions
Scalar (e.g. dark scalar, dark Higgs)	$(\mathcal{H}^{\dagger}\mathcal{H})\phi$
Vector (e.g. dark photon)	$\epsilon\mathcal{F}_{\mu\nu}\mathcal{F}'_{\mu\nu}$
Fermion (e.g. Heavy Neutral Leptons (HNL))	$\mathcal{H}^{\dagger}\overline{\mathcal{N}}\mathcal{L}$
Axion-like particle (ALP)	$\alpha\mathcal{F}^{\mu\nu}\tilde{\mathcal{F}}_{\mu\nu}$

- in quasi elastic scattering of incident protons on nucleons in the target, with a subsequent production of vector states via bremsstrahlung processes.

A dark sector can be connected to the SM sector through "portals", which can be classified by the type of mediator or by the dimension of the interaction operators. Such portals can be renormalizable (mass dimension ≤ 4) or be realized as higher-dimensional operators suppressed by the dimensionful couplings Λ^{-n}, with Λ being the energy scale of the hidden sector. In the latter case the most promising are the dimension 5 operators that are not suppressed too much by the energy scale Λ.
Any extension of the standard model would entail new Lagrangian terms. These generally can be split into those that involve the portal particles ($\mathcal{L}_{portal}$), and those that are of the HS do not ($\mathcal{L}_{HS}$):

$$\mathcal{L} = \mathcal{L}_{SM} + \mathcal{L}_{portal} + \mathcal{L}_{HS} \tag{1.2}$$

Depending on the spin of the mediator, there are three classes of portals mixing with the SM particles: *Scalar Portal* (spin 0, coupling coefficient $sin^2\theta$), *Neutrino Portal* (spin 1/2, coupling coefficient U^2) and *Vector Portal* (spin 1, coupling coefficient ϵ) - in addition there is also the pseudo-scalar Axion-like particle (ALP) portal. The lagragian terms brought by each of the portals are described in Table 1.1. In following, the vector portal will be briefly introduced.

1.3.1 Vector Portal

In the BSM theories, such as GUTs (Grand Unified Theories) the gauge structure of the Standard Model, the celebrated SU(3)⊗SU(2)⊗U(1) combination, could descend from a larger gauge group. In that case, one expects that at least several new vector states are very heavy, e.g. $m_V \approx 10^{16}$, well beyond the direct reach of accelerators.

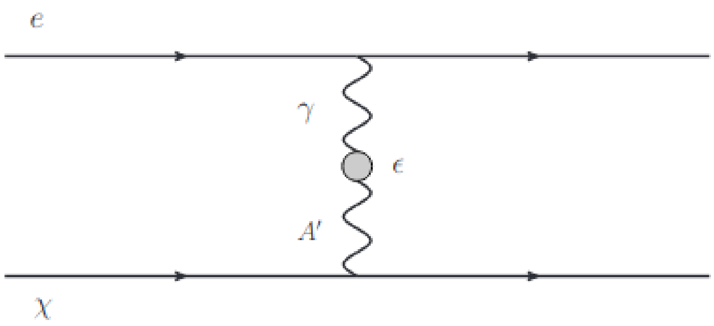


Figure 1.3: Kinetic mixing diagram ($\gamma - A'$) that mediates the interaction between the SM particles and χ dark sector particles charged under $U(1)'$ group.

It is also possible that the SM is accompanied by additional gauge structures that allow for (sub-)TeV gauge bosons, as is the case of multiple $U(1)$'s. An alternative possibility is relatively to the light vector states, in the GeV mass range with small couplings to the SM. While LHC experiments place very strong bounds on the possible existence of new vector states associated with new U(1) gauge groups at high-energy [23, 24], the alternative case is poorly constrained and represents an attractive physics target for many experiments at the intensity frontier [22].
The simplest way to couple a vector particle to the SM is to use the kinetically-mixed portal by introducing an additional U(1)' symmetry, under which hidden sector particles are charged. The new QED-like interaction is mediated by a vector massive boson gauge $\mathcal{A}'$, usually referred as *dark photon*.
The SM lagrangian under the SU(3)⊗SU(2)⊗U(1)⊗U(1)' gauge symmetry group, is extended to:

$$\mathcal{L} = \mathcal{L}_{\psi,A} + \mathcal{L}_{\chi,A'} - \frac{\epsilon}{2}\mathcal{F}_{\mu\nu}\mathcal{F}^{'}_{\mu\nu} + \frac{m^2_{\mathcal{A}}}{2}\mathcal{A}^{'}_{\mu}\mathcal{A}'^{\mu} \tag{1.3}$$

$\mathcal{L}_{\psi,A}$ and $\mathcal{L}_{\chi,A'}$ are the Standard QED-type Lagrangians:

$$\mathcal{L}_{\psi,A} = -\frac{1}{4}\mathcal{F}_{\mu\nu}\mathcal{F}^{\mu\nu} + \overline{\psi}[\gamma_\mu(i\partial_\mu - eA_\mu) - m_\psi]\psi$$
$$\mathcal{L}_{\chi,A'} = -\frac{1}{4}\mathcal{F'}_{\mu\nu}\mathcal{F}'^{\mu\nu} + \overline{\chi}[\gamma_\mu(i\partial_\mu - g'A'_\mu) - m_\chi]\chi$$

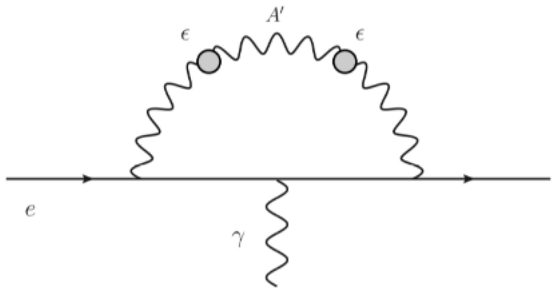


Figure 1.4: One-loop correction to the muon anomalous magnetic moment due to the exchange of the dark photon.

where $\mathcal{F}'_{\mu\nu}$ is the field strenght tensor, that satisfies the Bianchi identity $\mathcal{F}'_{\mu\nu} = \partial_\mu \mathcal{A}'_\nu - \partial_\nu \mathcal{A}'_\mu$, $m_{\mathcal{A}'}$ is the mass of the heavy photon. Using the equations of motion, $\partial_\mu F_{\mu\nu} = eJ^{EM}_\nu$, the interaction term can be rewritten as:

$$\frac{\epsilon}{2} F_{\mu\nu} F'_{\mu\nu} = A'_\mu \times (e\epsilon) J^{EM}_\mu$$

showing that the new vector particle couples to the electromagnetic current with strength, reduced by a small factor ϵ.

The kinetic mixing between the γ (SM photon) and the $\mathcal{A}'$ (dark photon) represented by the term $\mathcal{F}_{\mu\nu}\mathcal{F}'_{\mu\nu}$, could mediate the interaction of the standard Model particles (e.g. electron) and partcles χ charged under new U(1)' group, see Figure 1.3. In the limit of $m_{\mathcal{A}'} \to 0$ the apparent electromagnetic charge of χ is $e\epsilon$.

The vector portals are of great interest because they can provide a remedy for a number of observational anomalies, for example the muon $g - 2$ anomaly [25]. There is a persistent discrepancy between the measured value and the Standard Model prediction at the level of 3σ. The existence of a new light vector particle could explain this issue, through a positive correction from a one loop diagram (see Figure 1.4), that has a coupling equal to $g'/e = \epsilon$. For the choice of $\epsilon \sim few \times 10^{-3}$ at $m_V \sim m_\mu$, the new contribution brings theory and experiment into agreement.

Dark photons can be produced through the production of high-energy (virtual) photons. At a beam dump facility several production channels have been identified for vector states:

- *Meson decay*: the main process is $\pi^0 \to \gamma\mathcal{A}'$ where mesons can be produced

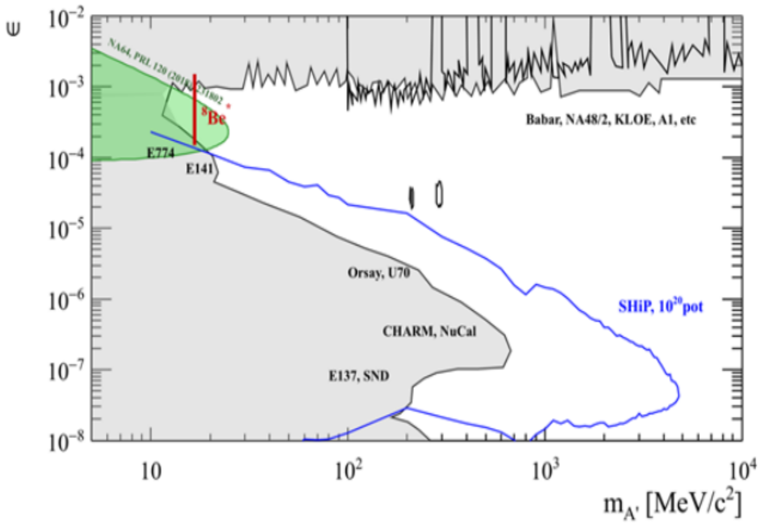


Figure 1.5: Summary of constraints on the dark photon model.

with abundance in proton on target collisions.

- *Proton breemstrahulung*: this process originates from the quasi elastic scattering of incident protons on nucleons in the target, with subsequent production of vector states via bremsstrahlung processes $pp \to pp\mathcal{A}'$
- *Direct perturbative QCD production*: the dominant production mechanism at higher energies available in the center of mass, via the underlying $q\bar{q} \to \mathcal{A}'$ or $qg \to q\mathcal{A}'$

The dark photon decay width is given by:

$$\Gamma_{\mathcal{A}'} = \sum_{l} \Gamma_{\mathcal{A}' \to l^+ l^-} + \sum_{hadrons} \Gamma_{\mathcal{A}' \to hadrons} + \sum_{\chi} \Gamma_{\mathcal{A}' \to \chi\overline{\chi}} \tag{1.4}$$

Depending on the mass of the dark photon, different decay channels predominate. If $m_{\mathcal{A}'} > 2m_\chi$ being m_χ the mass of the Dark Matter candidate, are referred to as dark photon invisible (or dark) decays since consisting in an electrically neutral final state, whereas the case $m_{\mathcal{A}'} < 2m_\chi$ is labelled visible decays mostly favouring the dilepton couple production.
The parameter space $\{m_{\mathcal{A}'},\epsilon^2\}$ places the current constraints on the dark photon model, under the assumption that the dark decays are absent or subdominant. One

can clearly see that the direct constraints can be subdivided into two main categories: those that come from the prompt decays of the dark photons, and those that are derived from the delayed decays of dark photons inside the detectors placed some distance behind beam dumps. The combination of all constraints covers the region of parameter space motivated by the solution to the muon $g-2$ discrepancy. A summary of constraints on the dark photon model is made in Figure 1.5.
Vector portals offer a way to connect the SM to Dark Matter. In the last few years, the direct searches for DM have been intensified, paralleled by broad investigations of theoretical opportunities for dark matter. Dark photons can play the role of dark matter mediators, particularly for light dark matter, for which χ would be an ideal candidate.

1.4 Light Dark Matter search

1.4.1 Dark matter candidate

The experimental evidence that ordinary "luminous" matter is not sufficient to explain some of the gravitational effects observed in the Universe began to appear in the 1930s, but it was only in the 1970s that physicists in the field were forced to introduce an "invisible" matter called *dark matter* in the models. According to the latest estimates, obtained from the Planck satellite and based on the Lambda-CDM (Cold Dark Matter) Model, the "dark matter" content is five times larger than the ordinary matter, it constitutes almost 90% of the mass present in the Universe [26].
Dark matter was initially called "missing mass" because, despite its gravitational effect being clearly visible, the lack of emitted and absorbed electromagnetic radiation made it impossible to observe through spectroscopic analysis.
There are many facts that prove that there is a missing mass in the universe. At first, in 1933 the astronomer Zwicky [27] studied the Coma cluster and derived an estimate of the mass based on its brightness using the virial theorem:

$$<T>=-\frac{1}{2}<U> \tag{1.5}$$

where <T> and <U> are the average kinetic energy and potential energy, respectively. The mass of the cluster resulted to be four-hundred times larger than the observed luminous mass suggesting the presence of a non-luminous material component. The dark matter hypothesis was not further investigated for 40 years, until the observations of Kent Ford and Vera Rubin in 1970 on the rotation curves of galaxies, in particular their observations concerned the Andromeda galaxy [28]. The luminous mass of galaxies, M, generally appears to be more concentrated in their center; for a body of mass but distance r from the center we can easily derive the trend of its rotation speed in relation to the gravitational attraction that it exerts:

$$\frac{GmM}{r^2} = \frac{mv^2}{r} \Rightarrow v \propto \frac{1}{\sqrt{r}} \tag{1.6}$$

The measurement of the velocity curve is possible through the observation of the spectral emission of the hydrogen line at 21cm. None of the measured rotation curves show the expected decrease, not reflecting the detected matter distribution (stars and gas). All the curves instead show an orbital speed which tends to be constant at great distances. This trend would be justified by a spherical halo of dark matter with a density $\rho \propto \frac{1}{r^2}$.

Further proof of the existence of dark matter is the gravitational lensing. According to general relativity, in fact, the presence of mass determines the curvature of space-time, and therefore the shape of geodesics. As a result, the presence of clusters or galaxies between an observer and a celestial body causes particular effects on the detected image, such as distortions, image duplication, or Einstein arcs. From these it is possible to have an estimate of the mass in question; also from these analysis a mass defect between the estimated and the luminous one was found.

If DM was made of particles, they could either be relativistic when decoupling from the early Universe plasma (Hot Dark Matter, HDM) or non relativistic (Cold Dark Matter, CDM). In the Λ_{CDM} model for the early Universe, the dominant component is the CDM. Various possible solutions are plausible but the most promising, especially since it is consistent with all observations, is the hypothesis of the existence of non baryonic-particellar matter, with mass, but interacting only through gravitational and weak interactions with ordinary matter. In the Standard Model there are no particles with all these features and therefore the intrinsic nature of the dark matter is still unknown. Many extensions of the Standard Model, predict the existence of thermal relic Weakly-Interacting Massive Particles (WIMPs). These particles would be stable, neutral, with masses from below GeV$/c^2$ to several TeV$/c^2$ and they interact very weakly with all the other known particles (the cross-section

is $\sigma < 10^{-6} pb$). The identity of the WIMPs reamins a mystery but surely the best motivated and theoretically developed WIMP candidate comes from SuperSymmetry. In particular, in the contest of the Minimal SuperSymmetric Standard Model- an extension of the SM involving SUSY- the WIMP candidate would be the lightest supersymmetric particle (LSP), called neutralino χ, [29].
The existence of DM can be pursued in three scenarios: (i) production of DM at particle accelerators by means of ordinary particles collision, (ii) indirect detection through detection of particles produced by annihilation of dark matter candidates and (iii) direct detection of DM scattering with matter.
Dark matter search at accelerators is currently being performed at the Large Hadron Collider (LHC) at CERN with ATLAS and CMS experiments [30]. They aim is detecting events with missing transferred momentum and energy, thus deducing the presence of DM. So far no deviations form SM expectations have been observed. Possibilities of both light dark matter candidates and light mediator particles have been recently explored, with the motivation of tying various anomalous experimental signatures to the annihilation, scattering or decay of dark matter. In this contest high luminosity fixed target experiments provide a potentially interesting probe of light sub-GeV WIMP dark matter, which is less accessible to underground direct detection. The use of proton-beam fixed target neutrino experiments has been highlighted as a way to efficiently probe the parameter space of LDM [31, 32].

1.4.2 Production of dark matter through dark photon decay

Models of sub-GeV WIMP dark matter generally require light force mediators to ensure efficient annihilation channels that avoid overproduction in the early universe, and these new 'dark' forces provide the primary production mode in fixed target experiments [33]. The three renormalizable "portals" couplings for a SM-gauge neutral field are the natural focal points for model building on effective field theory grounds. The minimal Light Dark Matter (LDM) benchmark model here considered exploits the vector dark portal with a light dark photon mediator. The interaction with the SM is then achieved via kinetic mixing with the hypercharge gauge boson, as described by the following Lagrangian:

$$\mathcal{L} = \mathcal{L}_\chi - \frac{1}{4}\mathcal{V}_{\mu\nu}\mathcal{V}^{\mu\nu} + \frac{1}{2}m_V^2\mathcal{V}_{\mu\nu}\mathcal{V}^{\mu\nu} + \frac{\epsilon}{2}\mathcal{V}_{\mu\nu}\mathcal{F}^{\mu\nu} + q_B g^{'}\mathcal{V}_{\mu\nu}\mathcal{J}_B^{\mu} \tag{1.7}$$

with

$$\mathcal{L}_\chi = \begin{cases} i\overline{\chi}\not{D}\chi - m_\chi\overline{\chi}\chi & \text{Dirac fermion DM,} \\ |D_\mu\chi|^2 - m_\chi^2|\chi|^2 & \text{Complex scalar DM,} \end{cases} \tag{1.8}$$

where the $U(1)^{'}$ covariant derivative takes the form $D_\mu = \partial_\mu$- $iq^{'}g^{'}V_\mu$, with $g^{'}$ is the $U(1)^{'}$ gauge coupling (and throughout we set the charge $q^{'} = 1$ for the DM field χ). The variable q_B represents the charge belonging to the extra gauge dark symmetry, here identified with the baryonic number. As consequence, two possible mediation channels are accepted:

- *Vector portal*: the pure vector portal is obtained by setting $q_B = 0$ and $q^{'} = 1$ in the Lagrangian above. This mediation channel is UV complete;
- *Baryonic vector portal*: the limit of a gauged baryon number current is obtained by setting $\epsilon = 0$, $g^{'} \to g_B$ with $q_B = 1$ and $q^{'} = 1$ as before. This mediation channel requires an additional UV completion to ensure anomaly cancellation.

The minimal vector portal model provides a natural thermal relic dark matter candidate, and thus the s-channel annihilation rate in the early universe determines the freeze-out abundance. This imposes a bound on the parameter space to ensure that the relic abundance is not too large, and a one-dimensional constraint on the parameters (ϵ, $\alpha^{'} = g^{'}/(4\pi)$, $m_\mathcal{V}, m_\chi$) when the relic abundance matches the observed DM abundance. In this contest, three main LDM production modes from dark photon decay are considered:

- π^0, η, ω in flight decays relevant for low m_V

  $$\pi^0/\eta \to \gamma + \mathcal{V}^* \to \gamma + \chi^\dagger + \chi$$

  if kinematically allowed, on-shell $\mathcal{V}$ -production is dominant, but it is also includes off-shell $\mathcal{V}^* \to \chi^\dagger + \chi$ decays.
- emission of $\mathcal{V}$ particles through proton bremsstrahlung:

  $$p + N \to p + N + \mathcal{V}$$

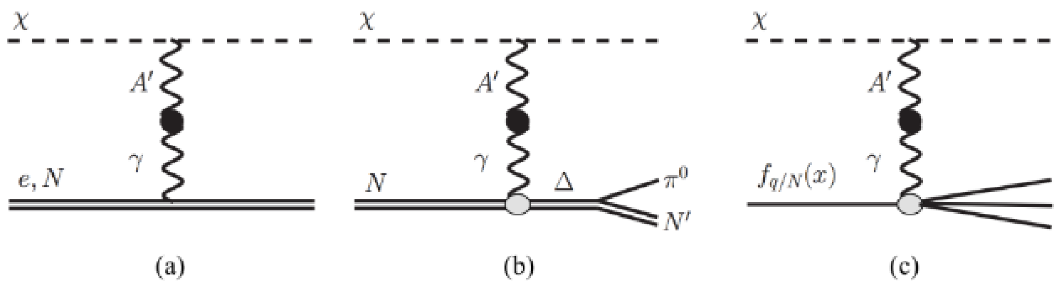


Figure 1.6: Feynman diagrams for LDM detection in the scattering on nuclei or electrons (a), incoherent single pion (resonant) mode (b) and the deep inelastic scattering (c)

with N = p or n, could provide a significant dark matter source with very little angular spread

- Direct production corresponds to parton-level processes such as:

$$p + N \rightarrow \mathcal{V}^* \rightarrow \chi^\dagger + \chi$$

it is rilevant for the mass range $m_{\mathcal{V}} > 1$ GeV

LDM particles can be detected by observing one of the following scattering signatures off SM particles:

- electron scattering (a);
- elastic Neutral Current-like scattering of nucleons (a);
- resonant scattering with nucleons (b);
- deep inelastic scattering off nucleons (c).

The Feynman diagrams of all detection modes are represented in Figure 1.6.

A rather systematic study of the detection of light dark matter produced via the dark photon portal has been performed both in e^+e^- colliders and also in fixed target experiment. In the first case, the missing energy constraints on dark photons were analyzed.

The invisible decays of $\mathcal{A}'$ are usually harder to detect, except $K^+ \to \pi + \mathcal{A}' \to \pi$+missing energy, where the competing SM process, $K^+ \to \pi + \nu\overline{\nu}$ is extremely suppressed [34]. A typical detection scheme in the proton beam dump experiments is built on the following chain of events:

$$pp \to \pi^0 + X, \quad \pi^0 \to \mathcal{V}\gamma, \quad \mathcal{V} \to \chi\overline{\chi} \tag{1.9}$$

with χ scattering on electrons or nuclei.
Currently, the MiniBooNE collaboration is conducting a dedicated search for such states in a beam dump mode run [35]. The SHiP facility will be able to explore a new parameters space $\{m_{\mathcal{A}'},\epsilon\}$, up to $m_{\mathcal{A}'} \sim$ 3GeV. If the values of $\epsilon \sim 10^{-7}$ are thought to be generated by loops of charged particles, this region corresponds to the two- or three-loop mediation mechanism [36].

1.5 Neutrino Physics

1.5.1 Tau neutrino cross section

In a beam dump experiment, with a high intensity and high energy proton beam impinging on a target, a high flux of neutrinos of all flavours are expected. It is ideally suited to perform studies on neutrino and anti-neutrino physics.
In fact, despite the progresses done in the recent years in the studies of neutrino physics, there are still lacking information about the third neutrino and anti-neutrino generations with only 9 ν_τ events reported by the DONUT [37] and 10 by the OPERA collaborations [9]. Both experiments measured the tau neutrino cross section. The errors, however, are high due to the low statistics collected.
In SHiP, several thousands neutrino and anti-neutrino interactions are expected, there will be the possibility to measure separately ν_τ and $\overline{\nu}_\tau$ cross sections in the same experiment and with higher precision, identifying for the first time $\overline{\nu}_\tau$ interactions. They are produced in the decays of charmed mesons $D_S \to \tau\nu_\tau$ and $\tau \to \nu_\tau$. The Feynmann diagramm is shown in Figure 1.7. These processes will allow to determine the structure functions in charged current neutrino cross section.

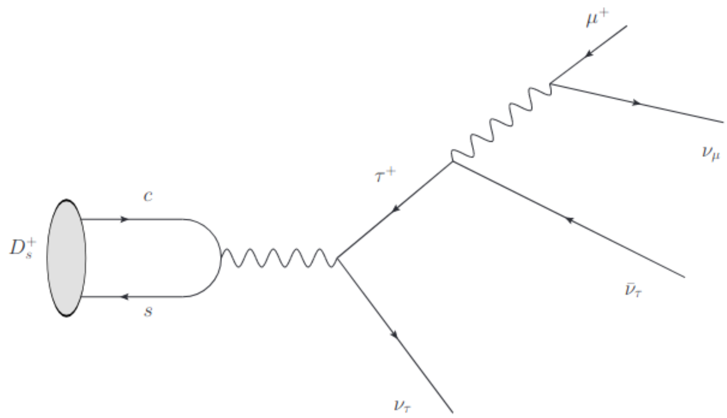


Figure 1.7: Feynman diagram showing the D_S decay chain for ν_τ and $\overline{\nu}_\tau$ production

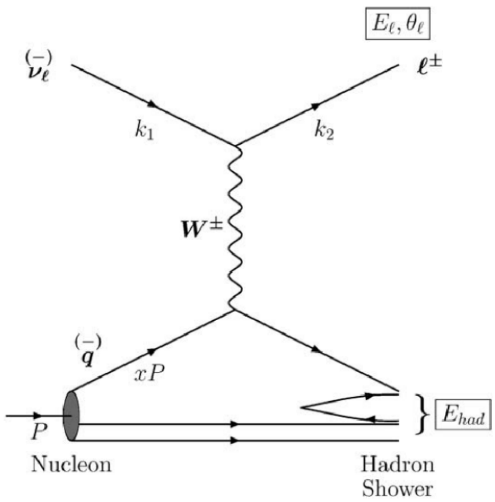


Figure 1.8: Feynman diagram showing the deep inelastic charged current interaction

Using the Deep Inelastic Scattering (DIS) variables $x = \frac{q^2}{2p \cdot q}$ and $y = \frac{p \cdot q}{p \cdot k}$ where the momentum assignments are:

$$\nu_\tau/\overline{\nu}_\tau(k) + N \to \tau^-/\tau^+(k') + X$$

$$q^2 \equiv (k - k')^2 = -Q$$

where k and k' are the the four-momenta of the incoming neutrino and outgoing lepton, see Figure 1.8. The ν_τ and $\overline{\nu}_\tau$ charged current cross section in terms of the

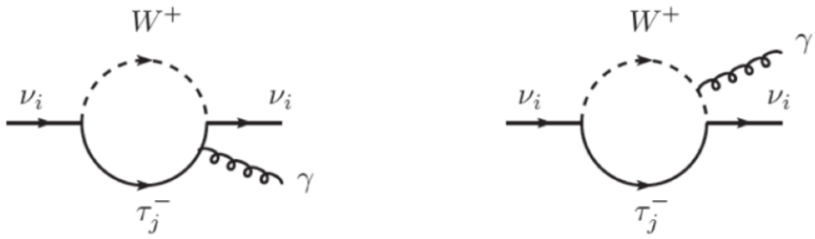


Figure 1.9: Feynman diagrams showing the contribution of W-loop to the neutrino magnetic moment.

structure functions is:

$$\begin{aligned}\frac{\mathrm{d}^2\sigma}{\mathrm{d}x\mathrm{d}y} =&\frac{G_F^2 M E_\nu}{\pi(1+Q^2/M_W^2)^2}((y^2x+\frac{m_\tau^2 y}{2E_\nu M})F_1+[(1-\frac{m_\tau^2}{4E_\nu^2})-(1+\frac{Mx}{2E_\nu})y]F_2+\\ &\pm[xy(1-\frac{y}{2})-\frac{m_\tau^2 y}{4E_\nu M}]F_3+\frac{m_\tau^2(m_\tau^2+Q^2)}{4E_\nu^2 M^2 x}F_4-\frac{m_\tau^2}{E_\nu M}F_5),\end{aligned} \tag{1.10}$$

where the plus sign applies to neutrino scattering and the minus one to antineutrino scattering. M and m_τ are the nucleon and τ lepton masses, respectively, M_W is the W boson mass, E_ν is the initial neutrino energy and G_F is the Fermi constant. A precise estimation of the tau neutrino and anti-neutrino cross section would allow a measurement of the structure functions F_4 and F_5 which are negligible in other flavoured neutrino interactions. The SHiP experiment offers the exclusive chance to perform the first measurement of F_4 and F_5 structure functions.

1.5.2 Tau neutrino magnetic moment

In the minimal extension of the SM, neutrinos can acquire non zero magnetic moment μ_ν and give rise to electromagnetic interactions.
This is possible due to the contribution of loop Feynman diagrams with a W boson and a lepton, which can couple to an external magnetic field, see Figure 1.9. The magnetic moment created in this way is extremely small and proportional to the mass of the neutrino m_ν:

$$\mu_\nu = \frac{3eG_F m_\nu}{8\pi^2\sqrt{2}} \tag{1.11}$$

The presence of a non-zero magnetic moment adds an extra-component to the elastic cross-section for the process $\nu + e^- \rightarrow \nu + e^-$ that in the SM involves only the neutral current. Therefore, an anomalous increment of the measured value of this cross-section can prove the hypothesis of anomalous magnetic moment, predicted by many New Physics (BSM) theories. The DONUT experiment set the upper limit to $3.9 \times 10^{-7} \mu_B$ [38], where μ_B is the Bohr magneton. With SHiP, if the systematic uncertainty, mostly coming from the knowledge of the neutrino flux, could be reduced down to 5%, the evidence for the tau neutrino anomalous magnetic moment with a significance of 3σ requires the observation of an excess of about 540 events over the background. Hence, a region down to a magnetic moment $\mu_\nu = 1.5 \cdot 10^{-7}$ could be explored.

Chapter 2

The SHiP experiment

In April 2015 the SHiP Collaboration submitted its Technical Proposal (TP) [18] to the Cern Scientific Committe: a new general purpose fixed target facility was proposed at Cern Super Proton Syncrotron (SPS) accelerator which could provide a unique platform to search for New Physics at the Intensity Frontier. Since then, during the Comprehensive Design Study (CDS) [39], the entire experiment has undergone several rounds of re-optimisation. While the overall concept remain the same, the design of the geometry and the sub-systems have been significantly improved and have matured as result of simulation studies, test beam measurements and prototyping resulting in a new layout, reported in a Progress Report (PR) [40]. This chapter is intended to provide a brief review of all the designed components of the SHiP experiment.

2.1 The SHiP facility at SPS

The primary physics goals of the SHiP facility consists of exploring hidden portals predicted by a large number of models beyond the SM and make measurements with tau neutrinos. The experiment is aimed at searching for very weakly interacting long lived particles including Heavy Neutral Leptons - right-handed partners of the active neutrinos, vector, scalar, axion portals to the Hidden Sector, and it is also optimised to search for Light Dark Matter particles through scattering signatures. Since many hidden particles are expected to be accessible through decays of heavy hadrons, the facility must maximize their production and the detector acceptance in the

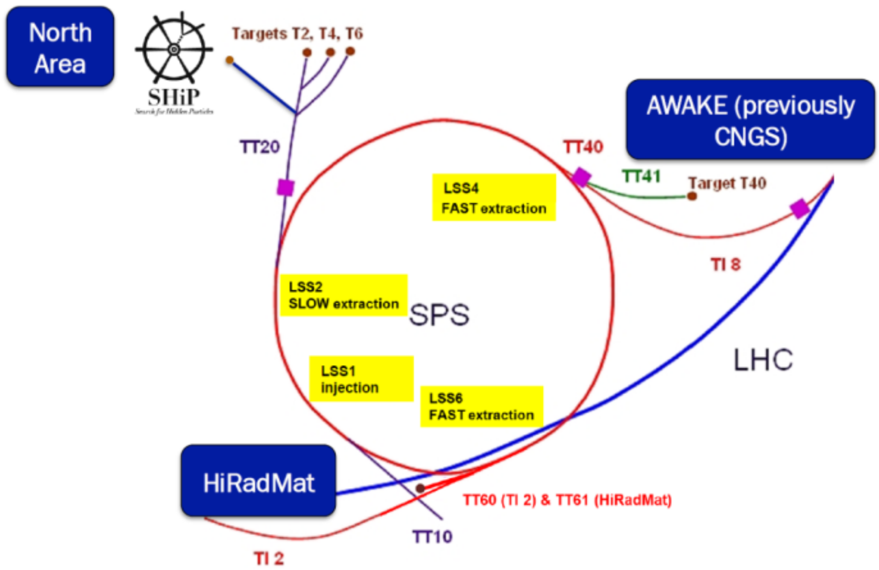


Figure 2.1: Overview of the SPS accelerator complex. The TT20 transfer line will be shared by the SHiP facility alongside other fixed target programmes.

cleanest possible environment. Hadrons, electrons, and photons must be stopped, the production of neutrinos from light hadron decays must be minimized, and the detector must be shielded from the residual muon flux. In this contest, SHiP will be able to reconstruct the decay vertex, measure the invariant mass of the decayed particle, and identify its final state using particle identification (PID). The building of the SHiP dump facility will take place in the Prevessin site, belonging to the North Area of the CERN SPS accelerating complex, see Figure 2.1.
An overview of the experimental layout is shown in Figure 2.2.
A 400 GeV/c proton beam will be delivered onto a thick hybrid target, specifically designed to maximise heavy hadron production yield and thus hidden sector particles and tau neutrino yields. A hadron stopper of approximatively 5 m of iron, follows the beam-dump target, with the goal to absorb SM particles produced in the beam interaction but it still remains a large flux of muons and neutrinos, produced by the decays of pions and kaons. A 48 m long muon shield will deflect the residual muons that constitute the main background for the Hidden Sector particles search further cleaning the flux of particles from leftover backgrounds to hidden sector particles

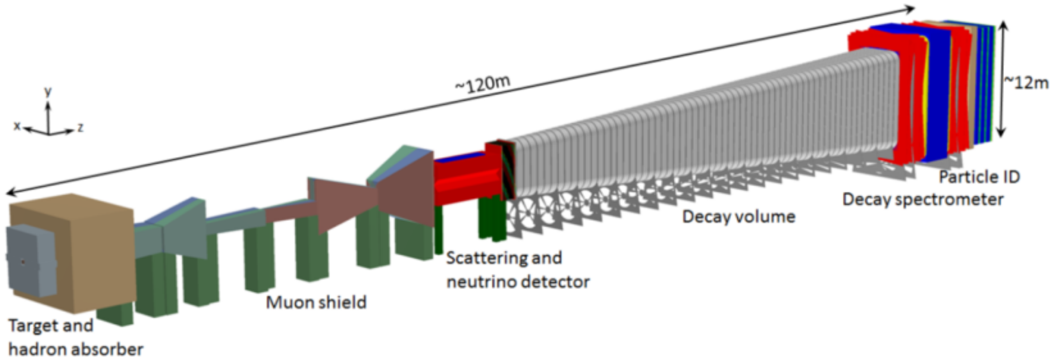


Figure 2.2: Overview of the SHiP experiment.

and neutrino searches.
The SHiP detector consists of two complementary apparatus, the Scattering and Neutrino Detector (SND), and the Hidden Sector (HSS) spectrometer. The SND is located in a magnetised region of 1.2 T downstream of the muon sweeper. It consists of a $(90 \times 75 \times 321)$ cm^3 high-granularity tracking device which exploits the Emulsion Cloud Chamber (ECC) technique developed by the OPERA experiment [41]. The SND perfectly tailored for neutrino physics using all three flavours, as well as detection of light dark matter particles scattering off of electrons of the SND. The HSS is an approximately 50 m long decay vessel positioned downstream of the SND detector. In order to eliminate the background from neutrinos interacting in the decay volume, it is maintained at a pressure of $O(10^{-3})$ bar. To maximise the sensitivity to HS particles the muon shield was shortened in combination with changing the shape of the decay volume from an elliptic cylinder to a pyramidal frustum. The decay volume is followed by a large spectrometer with a rectangular acceptance of 5 m in width and 10 m in height.
Downstream of the HSS, there are a set of calorimeters and muon detectors which provide particle identification, thus allowing for discriminating between the very wide range of HS model. The electromagnetic calorimeter (SplitCal) is optimised to provide electron identification and π^0 reconstruction and moreover it provides a reconstruction of axion-like particles (ALP) decaying to the two photon final state. The muon shield and the SHiP detector systems are housed in an ~ 120 m long underground experimental hall at a depth of ~ 15 m. To minimise the background

induced by the flux of muons and neutrinos interacting with material in the vicinity of the detector, no infrastructure systems are located on the sides of the detector, and the hall is 20 m wide along its entire length.

2.1.1 The proton beam

The SHiP operational scenario implies returning to full exploitation of the capacity of the SPS. The *Super Proton Syncrotron* is a 7 km circumference accelerating machine, located in an underground tunnel and used as final injector for LHC collisions and also as proton beam supplier for several fixed target experiments, such as NA61/SHINE, NA62 and COMPASS. The location of the SHiP facility is ideally suited since it allows a full integration on CERN land with almost no impact on the existing facilities. It allows reuse of about 600 m of the present TT20 transfer line, which has a sufficient aperture for the slow-extracted beam at 400 GeV. In the baseline scenario for SHiP, the beam sharing delivers an annual yield of 4×10^{19} protons to the Beam Dump Facility and a total of 10^{19} to the other physics programs at the CERN North Area, while respecting the beam delivery required by the HL-LHC. The physics sensitivities are based on acquiring a total of 2×10^{20} protons on target, which may thus be achieved in five years of nominal operation. Each proton spill, with a nominal intensity of 4×10^{13} p.o.t., will be delivered to the TT20 transfer line every 7.2 seconds with a ~ 1 s long flat top, in order to minimize the muon pile-up induced background. Those requirements can be achieved by means of the challenging technique known as slow beam extraction.

2.1.2 The target

The physics scope of the SHiP experiment requires a proton target which maximises the production of heavy hadrons- charms and beauty and photons. At the same time, the proton interactions give rise to copious direct production of short-lived meson resonances, as well as pions and kaons. Considering the SPS operational parameters leading to a high pulse intensity, a cycle averaged beam power on target of 320 kW is expected. Integrating over the spill duration ($\Delta t_{spill} \sim 1s$), the expected beam power is of 2.56 MW. So, while a hadron stopper of a few meters of iron is sufficient to absorb the hadrons and the electromagnetic radiation emerging from the target, the decays of the pions, kaons and short-lived meson resonances result

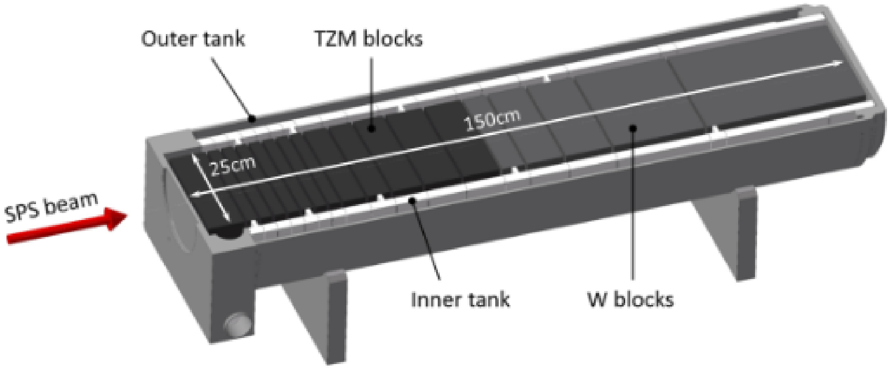


Figure 2.3: Layout of the target assembly, showing the TZM and the tungsten blocks.

in a large flux of muons and neutrinos. In order to reduce the flux of neutrinos, in particular the flux of muon neutrinos and the associated muons, pions and kaons should be stopped as efficiently as possible before they decay. The target should thus be long and be made of a material with the highest possible atomic mass and atomic charge. It should be sufficiently long to intercept virtually all of the proton intensity and to contain the majority of the hadronic shower with minimum leakage. This performance is achieved with a longitudinally segmented hybrid target consisting of of several collinear cylinders, with a diameter of 250 mm. made of titanium-zirconium doped molybdenum- TZM (0.08% titanium-0.05% zirconium-molybdenum) alloy and pure tungsten. In particular, while the downstream section uses pure tungsten, since it satisfies all the physics requirements (high atomic number and short interaction length) this material can not be adopted in the first section due to the high stress, for this purpose the TZM alloy has been used. In order to preserve the target from the very high energy intensity, each block is interleaved with sixteen 5 mm slits for water cooling. To respect the material limits derived from thermo-mechanical stresses, the thickness of each block has been optimised to provide a relatively uniform energy deposition and a sufficient heat flux for an efficient energy extraction. The infrastructure complex to house the proton target with the associated services and remote handling, which is fully compatible with the radiation protection and environmental considerations, consists of a compact multi-compartment complex with a floor size of 7.8×10 m^2 at a depth of 11 m.

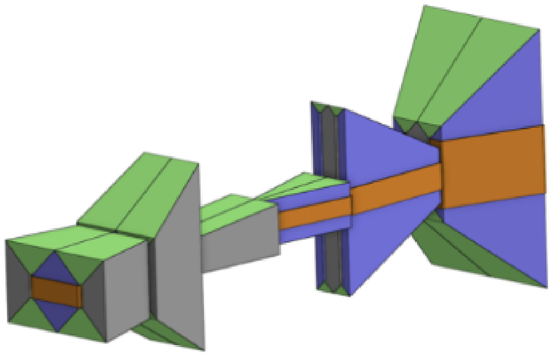


Figure 2.4: Three dimensional view of the active shield.

2.1.3 The muon shield

The SHiP experiment requires to operate in an environment of low background. The muons, that are more penetrating, can not be stopped in the iron placed immediately after the target as happens for the heavier particles, pions or kaons. Thus, muon represent the most source of background, a shield is designed to reduce the flux of muons by six orders of magnitude in the detector acceptance, in this way it is possible to make negligible the background from random combinations and deep inelastic scattering. The design of the muon shield is shown in Figure 2.4.
At first, active and passive shields were considered. Several studies have demonstrated that passive shielding of muons would require a long distance between the target and the detectors, which would drastically decrease acceptance, and large amounts of dense material, such as tungsten or lead, so it was clear, that the baseline decision is focusing on an active shield where the reduction of the muon background must be achieved in the shortest possible distance from the target. A first set of magnets provides the separation of μ^+ and μ^- and tries to sweep out the muons as far as possible to create empty space for a second set of magnets with a flipped magnet polarity. This allows the second set to catch any muons swept back in by the return field of the first magnet set. While it is not practically possible to magnetise the target in the SHiP set-up, the design includes a magnetisation of the hadron stopper over a length of 4 m with the help of a magnetic coil integrated into the shielding. The resulting system of magnets is now 35 m long and has 1300

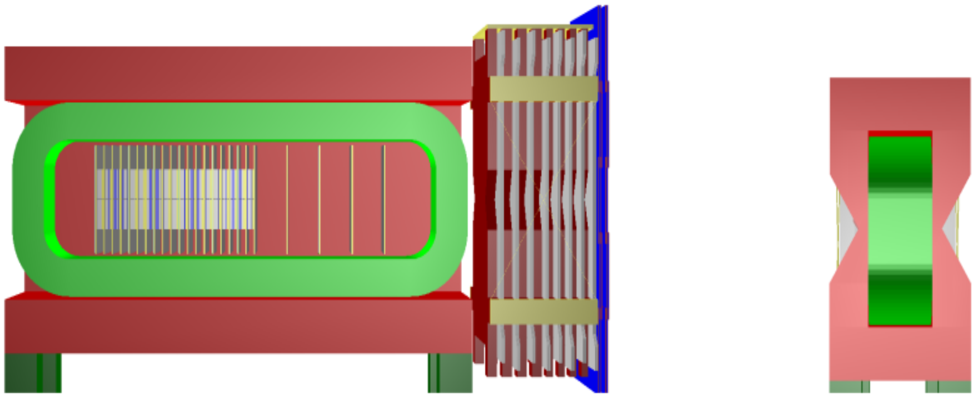


Figure 2.5: Layout of the scattering and neutrino detector.

tons of magnets. Since it has been possible to maintain the spectrometer aperture of $5 \times 10\text{m}^2$, the reduced length of the shield gives the detector a larger geometric acceptance for signal decays.
The muon shield optimisation is based on the use of Grain-Oriented (GO) steel allowing a high magnetic flux density at a very limited current. Grain-oriented steel sheets comes in 0.3 - 0.5 mm thickness. An average field of 1.7 T could be achieved throughout the magnets. The study made using the model on Figure 2.4 to produce a realistic magnetic field simulation shiws that the magnetic field simulation reveals a good field homogeneity and strength in the regions critical for deflecting the high momentum muons and minor degradation in the less important outer regions.

2.2 Scattering and Neutrino Detector

The Scattering and Neutrino Detector of the SHiP experiment is designed to perform the first direct observation of the $\overline{\nu}_\tau$ and, thanks to an unprecedented statistics of tau neutrinos and anti-neutrinos it will be able to study their properties and cross-sections. On the other hand, the same detector is suitable to detect LDM through scattering off electrons.
The muon flux immediately downstream of the last magnet of the Muon Shield is shown in Figure 2.6. The hourglass-shaped region delineates the area in the transverse plane which is cleared from the bulk of the muon flux, and, which in turn

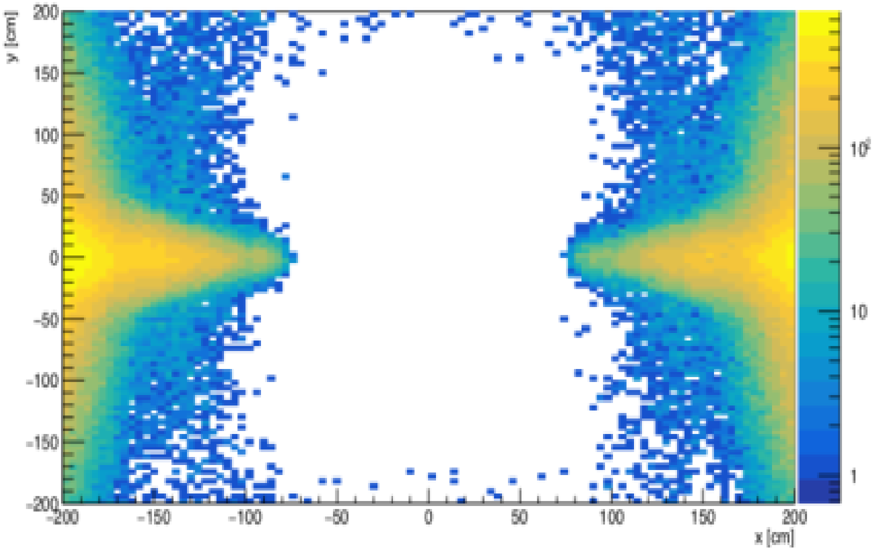


Figure 2.6: Muon flux distribution in the transverse plane, immediately down-stream of the last sweeper magnet.

defines the allowed region for the detector. The geometry of the SND is shown in Figure 2.5. It consists of a $\sim$ 7 m long magnet providing a 1.2 T horizontal magnetic field. The magnet hosts the Emulsion Target, interleaved with the Target Tracker planes and a Downstream Tracker. The magnetised region measure the momenta of both hadrons and muons in a single spectrometer, followed by a simplified muon identification system.
While the Emulsion Target with the Compact Emulsion Spectrometer (CES) is efficient at lower momenta, the Downstream Tracker is tailored to measure the charge and momentum of muons above 10 GeV/c.

2.2.1 SND spectrometer magnet

A significant effort has been put on designing the SND spectrometer magnet adapted to host the SND and at the same time maximises the aperture available to the detector [42]. The layout of the SND spectrometer magnet, showing the opening

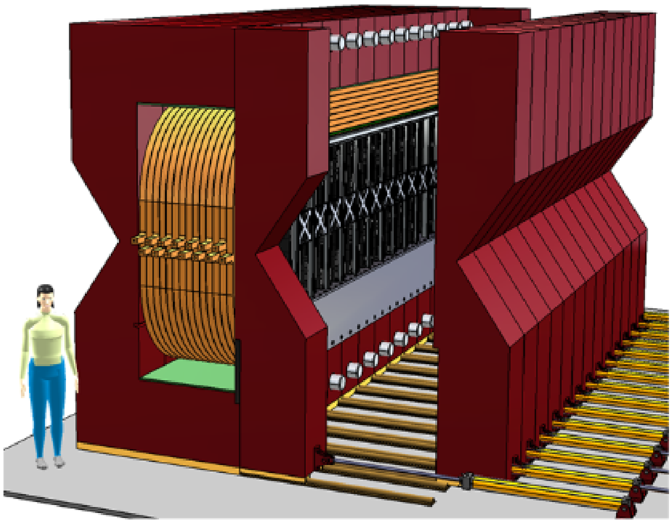


Figure 2.7: Schematic layout of the SND magnet.

mechanism for the replacement of the emulsion films is shown in Figure 2.7.
A field higher than 1.2 T is able to magnetized a large volume $\sim 1\times1.6\times6.4$ m^3 which accommodating the Emulsion Target, the Target Trackers and the Downstream Tracker. At the same time, the yoke was designed to produce a low stray field to avoid interfering with the muon background flux on the sides of the magnet. The two different options for the coil have been considered- warm aluminum and copper, and high and low temperature superconducting. Requirements on the access and maintenance as well as constraints from the associated detectors suggest that the better option is a warm magnet. The coils have been designed with an active cooling system, in fact the emulsion system requires that the thermal management of the detector volume respects the operating limit at a temperature of 18° C. The magnet can be opened regularly for the emulsion film replacements. This procedure should take no more than a day. The designed coil (copper/aluminum option) has a nominal cross-section of 0.41 m^2 and a fill factor of 0.68/0.73, respectively, and produces a uniform magnetic field of 1.25 T with $\Delta B/B < 1\%$ in the detector volume. An additional 22 mm thick thermal shield of aluminium inside the magnet is required to guarantee the required temperature. It will be cooled with an average coolant temperature at 18° C and a maximum variation of 0.4° C, requiring a refrigerator

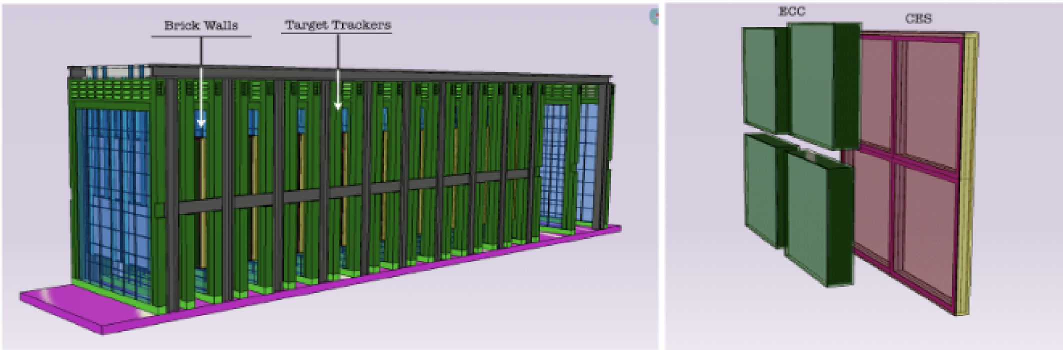


Figure 2.8: Schematic representation of the Emulsion Target.

power of about 24 kW.

2.2.2 The emulsion target

The Emulsion Target exhibits a modular structure and its fundamental unit is made of two parts: a *Emulsion Cloud Chamber* (*ECC*) and a *Compact Emulsion Spectrometer* (CES). The target is complemented by planes of electronic detectors (Target Trackers) to provide time stamp of the event and to identify the target unit where the neutrino interaction occurred. The current design of the Emulsion Target foresees a target made of 19 emulsion brick walls and 19 Target Tracker planes. . The walls are divided in 2×2 cells, each with a transverse size of 40×40 cm^2 , containing an ECC and a CES. A schematic representation of the Emulsion Target is shown in Figure 2.8.

Emulsion Cloud Chamber

The ECC technology consists of a sequence of passive material plates interleaved with nuclear emulsion films. A unit ECC cell is made of 36 emulsion films with a transverse size of 40×40 cm^2 , interleaved with 1 mm thick plates of tungsten alloys. The resulting brick has a total thickness of ~ 5 cm, corresponding to ~ 10 X_0 , and a total weight of ~ 100 kg.
It acts as a tracking device with sub-micrometric position and milli-radiant angular resolution. Thanks these features, it is capable of detecting τ leptons and charmed

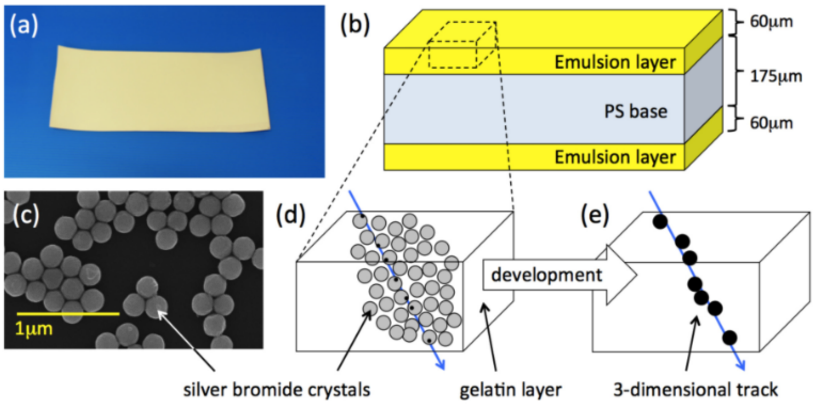


Figure 2.9: (a)Nuclear emulsion film picture with its internal structure (b). The silver bromide crystals themselves are seen in an electron microscope image in (c). (d) and (e) illustrate the principle of detection of a charged particle in the nuclear emulsion. Black small dots on silver bromide crystals show latent images in (d) [45].

hadrons by disentangling their production and decay vertexes. Given the short lifetime of the tau lepton, both the neutrino interaction and the tau lepton decay vertices are in the same ECC and a full event reconstruction can be achieved exploiting the features of the detector. It is also suited for Light Dark Matter detection through the direct observation of the scattering off electrons in the absorber planes. The high spatial resolution of the nuclear emulsion allows measuring the momentum of charged particles through the detection of multiple Coulomb scattering in the passive material [43]. ECC is also a fine sampling calorimeter where the electrons are identified by observing electromagnetic showers in the brick, allowing the electron energy and incident angle to be measured accurately [44].

Nuclear Emulsion Nuclear emulsion films are produced by Nagoya University in collaboration with the Fuji Film Company and by the Slavich Company in Russia. They consist of AgBr crystals, with a diameter of 0.2 μm, dispersed in a gelatine binder. The AgBr crystal is a semiconductor with a band gap of 2.6 eV. The passage of a charged particle creates electron-hole pairs. The electrons are grabbed by reticular flaws on the crystal surface, leading to the production of Ag metal atoms (the so called grains) acting as latent image centers. Through a process called devel-

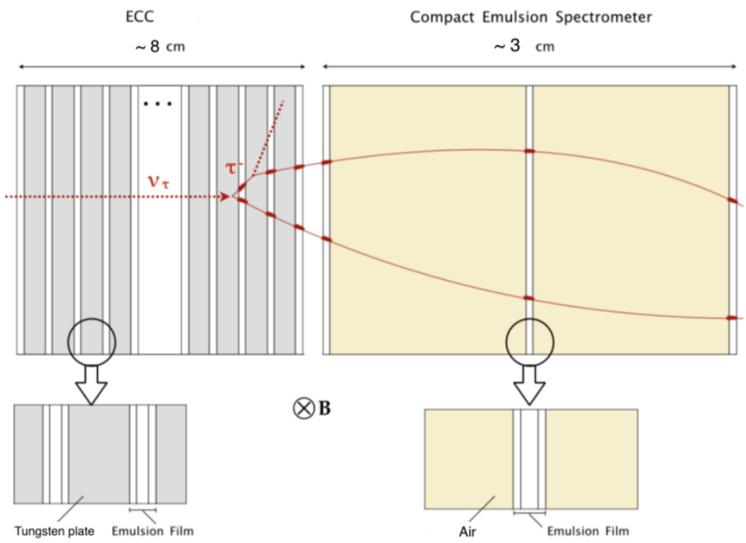


Figure 2.10: Schematic representation of the SND unitary cell.

opment, the number of silver atoms multiplies and the size of silver grains increases from 0.2 μm to 0.6 μm, becoming visible with an optical microscope. For a Minimum Ionizing Particle (MIP) the emulsion sensitivity is about 36 grains/100 μm. A nuclear emulsion film has two sensitive layers on both sides of a transparent plastic base, see Figure 2.9. By connecting the two hits left by a charged particle on both sides of the base, the slope ofthe track can be measured with milli-radian accuracy. Nuclear emulsions integrate all tracks including cosmic rays and natural radiations from their production to development. To prevent thermal deformations, the films are required to be at a temperature of around 20° C or lower, and stable within one degree. The data taking and processing in the emulsion films is accomplished by of automated microscopes. Afterwards, data analyses techniques are applied to reconstruct the trajectory of the particle in the whole ECC.

The Compact Emulsion Spectrometer

The Compact Emulsion Spectrometer modules aim at measuring the electric charge and the momentum of hadrons produced in τ lepton decays, thus providing the unique feature of disentangling ν_τ and $\overline{\nu}_\tau$ charged current interactions also in their hadronic decay channels. The CES also measures the momentum of soft muons

which are emitted at large angles and which do not reach the Muon Tracker. The basic structure of the CES is made of three emulsion films interleaved by two layers of low density material, attached immediately downstream of each ECC brick as shown in Figure 2.10. The momentum is measured from the curvature in a magnetic field, via the sagitta method. It is used to discriminate between positive and negative charge. The two dimensional sagitta $s = (s_x, s_y)$ is defined as the distance between the track position in the middle sheet and the straight line joining the track positions in two external sheets. The sagitta s_x is very nearly equal to s and reflects the curvature due to the magnetic field, while the s_y is centered at zero and its width reflects the deviations due to multiple scattering.
The layout of the detector is optimized in order to have good performances in the charge measurement for hadrons produced in the τ decay. Efficiencies as high as 90% for hadronic τ daughters reaching the end of ECC brick can be achieved with a magnetic field of 1 T.

The Target Tracker System and Downstream Tracker

The target trackers are electronic detectors which provide the time stamp to the events reconstructed in the emulsion bricks and connect the emulsion tracks to those reconstructed in the downstream detector. A set of 19 Target Tracker planes is placed within the target, complementing each brick wall with a plane. The first plane precedes the first brick, thus acting as a veto for charged particles entering the Emulsion Target. After the target, more planes of electronic detectors are located downstream to measure momentum and charge, working as a second spectrometer to complement the Compact Emulsion Spectrometer for long charged tracks, especially muons. The distance between the target tracker plane and the emulsion bricks must be a few mm and kept uniform over the whole surface, to assure constant tracking capabilities. The thickness of a TT plane must be contained within about 5 cm, requiring a flat surface to get a good planarity and uniform tracking performances. The SND TT system consists of a scintillating fibre tracker, namely SciFi- the same technology developed for the LHCb experiment [46]. SciFi tracking detectors show high spatial resolution, of about $30 - 50$ μm and a very high efficiency to minimum ionising particles ($> 99.6\%$) and are totally unaffected by magnetic fields. Furthermore the detector is extremely uniform and has no dead zones due to its

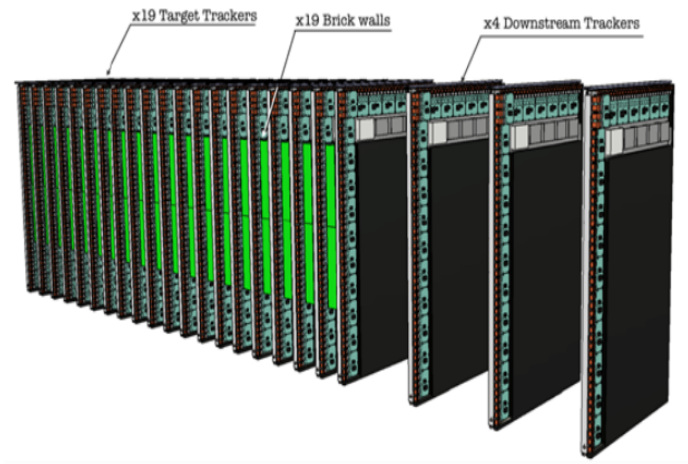


Figure 2.11: Schematic view of the Target Tracker layers interleaved with the bricks and CESs.

thin scintillating fibres 250μm of diameter. The same technology used for the TT system is employed in the Downstream Trackers (DT). The main purpose of the detector is to measure the momentum and the charge for long tracks, in particular muons, thanks to the presence of the magnetic field. Moreover it provides additional informations on the identification of muons to the downstream detector.

2.2.3 Muon identification system

The muon identification system is designed to identify with high efficiency the muons produced in neutrino interactions and in τ decays occurring in the Emulsion Target. The current design involves four iron layers of 15 cm thickness, followed by smaller four iron layers of 10 cm thickness. The layout is shown in Figure 2.12.
The transverse size of the magnet is driven by the requirement to measure muons emitted with angles up to $\pi/4$ with respect to the incoming neutrino direction. In the horizontal direction, the size is limited to 4 m to avoid the external high muon rate. The muon absorber will accomplish not only the task of muons identification, but will also act as a veto system for the hidden sector detectors. Each RPC plane is composed of three gaps with an active area of 1900×1200 mm^2 each. The planes are read via orthogonal strip panel with ~ 1 cm pitch and they are staggered by

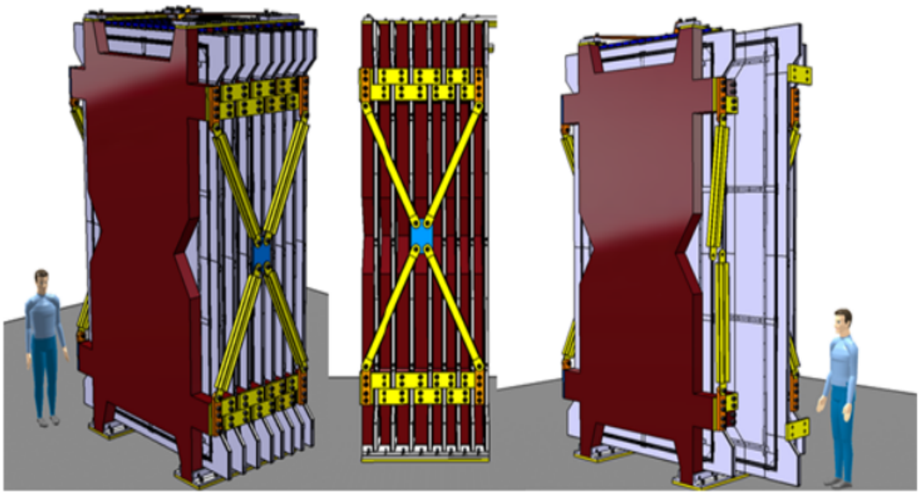


Figure 2.12: Layout of the SND muon identification system.

± 10 cm, to compensate from the acceptance losses due to the dead areas between adjacent gaps in the same plane. The overall transverse dimension of one plane is 4290×2844 mm^2 . The RPC were operated in avalanche mode using the standard CMS/ATLAS gas mixture, 95.2% C2H2F4/4.5% Iso-C4H10/0.3% SF6 with a flux of about $2l/h$ at $HV = 9.6$ kV.

2.3 Hidden Sector Detector

The SHiP experiment will be able to detect the particles of the Hidden Sector which come from the decay of the charmed hadrons. They exhibit a relatively large opening angle to the beam direction and since they couple very weakly with SM particles, they also have a long lifetime ($c\tau \sim O$(km)). This implies that their detection requires a long decay volume equipped with detectors at the far end in order to be able to reconstruct their decay products.

2.3.1 Decay volume

The search for the Hidden Particles depends critically on the background rejection capability and one of the main challenges of the SHiP experiment is to have a zero

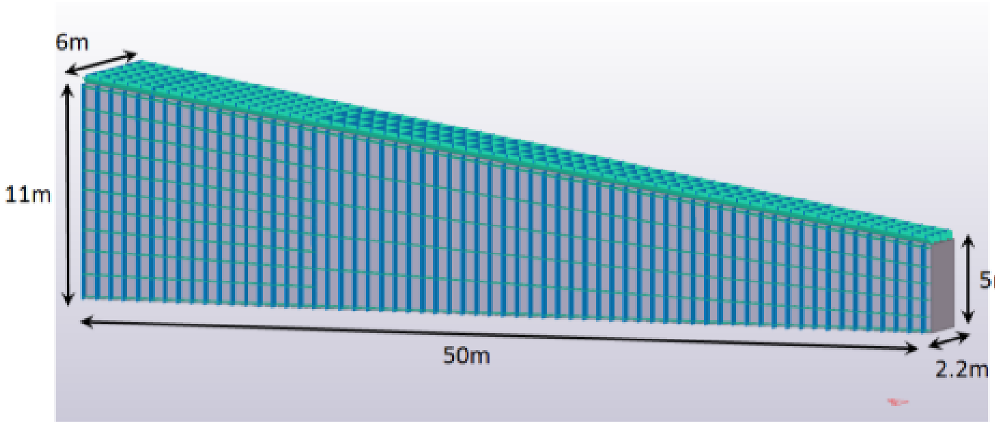


Figure 2.13: Overview of the structure of the decay volume.

background. For this reason the decay volume, designed to host hidden particles decays has to be evacuated at the level of 10^{-2} bar, to suppress the neutrino and muon interactions with the residual air, and be instrumented with a highly efficient background event tagger.
The decay vessel, whose layout is shown in Figure 2.13, is $\sim$ 50 m long with upstream outer/inner dimensions of 2.2×5.0 m^2/1.5×4.3 m^2, and downstream dimensions of 5.9×11.9 m^2/5×11 m^2. Its pyramidal frustum shape is mainly motivated by the requirement of proximity to the proton target together with the shape of the region which is cleared from the muon flux.
The total vacuum volume is of $\sim$ 2040 m^3. The vacuum volume end-cap is located just behind the last tracker station upstream of the timing and the particle identification detectors. To further ensure that signal candidates are not produced by neutrino or muon interactions in the upstream Scattering and Neutrino Detector or the walls of the decay volume, the decay volume section is surrounded by a high efficiency Surround Background Tagger (SBT) system which is capable of detecting the charged particles produced in the interactions.
The SBT system is based on a liquid scintillator detector to provide a large detection efficiency and good sensitivity. It is organized in compartments (referred to as "SBT cells") with dimensions of $80cm \times 120cm \times 30cm$. Each cell is readout by a ring of 24 SiPMs of 3×3 mm^2 area equipped with two wavelength-shifting optical modules.

2.3.2 Decay Spectrometer

The purpose of the HS spectrometer is to reconstruct with high efficiency the tracks of charged particles from the decay of hidden particles, while rejecting background events. Additionally, the spectrometer must provide an accurate determination of the track momentum and of the flight direction within the fiducial decay volume, measure the momenta, and provide PID information.

The SHiP Decay Spectrometer consists of the Spectrometer Straw Tracker (SST), the Timing Detector (TD), the Electromagnetic Calorimeter and the downstream Muon System.

The SST spectrometer is able to measure track parameters and momentum of charged particles with high efficiency and accurate enough to reconstruct decays of hidden particles and to reject background events. The invariant mass, the vertex quality, the timing, the matching with background veto taggers and the pointing to the production target are crucial tools for the back-ground suppression.

The main task of the Timing Detector (TD) is to provide unambiguous start times to the SST for the drift time measurement and to the coincidence criterion on the decay vertex candidates. In order to reduce combinatorial di-muon background to an acceptable level, a timing resolution of better than 100 ps is necessary.

The SplitCal is an electromagnetic calorimeter located in the region immediately downstream of the decay vessel with the aim to identify photons, electrons and neutral pions and to provide measurements of their energy and positions. The SpliCal is currently foreseen to use 40 absorber/scintillator planes of 6×12 m^2 cross-section and 20 X_0 depth. The layers of the calorimeter whith the high precision is made of MicroMegas chambers, similar to those developed for the upgrade of the ATLAS muon system but of smaller area of 80×80 cm^2. This option guaranteed the mrad resolution and high spatial segmentation of ~ 200 μm.

The muon detector is located downstream of the electromagnetic calorimeter. The main goal of the muon system is to identify muons with high efficiency ($> 95\%$) in the momentum range of $\sim 5 - 100$ GeV/c, in addition it can contribute to the rejection of combinatorial background requiring a time coincidence of candidate muon tracks.

2.4 Performance of the SHiP detector

2.4.1 Light Dark Matter detection

The χ particles might be detected through its elastic scattering process off of the electrons of the SND target. From the event $\chi e^- \to \chi e^-$, the scattered electron can be sufficiently energetic to generate an electromagnetic shower within the brick.
The Emulsion Cloud Chamber (ECC) bricks interleaved with the Target Tracker planes, act as sampling calorimeter allowing to reconstruct a sufficient portion of the shower produced by the recoil electron and to determine the particle angle and energy. The background events which are mostly due to the neutral current ν_μ and ν_e scattering on electrons, and charged current elastic, resonant and deep inelastic ν_e scattering off nuclei are generated by GENIE MonteCarlo [47]. The main variables to separate signal from background are the electron energy and the angle with respect to the neutrino direction and the number of detectable particles at the neutrino interaction vertex, which need to be accurately measured from the electromagnetic shower reconstruction.
Assuming LDM pair-production ($\chi\overline{\chi}$) in the prompt decays of Dark Photon (DP). In the considered DP mass range of $M_{A'} \sim O(1\text{GeV/c}^{\ 2})$, only contributions from the decay of light mesons (π, η, ω) and proton bremsstrahlung have been included. Prompt-QCD and heavier Drell Yan-like production mechanisms have been proven to be negligible. Assuming a benchmark scenario with relevant theory parameters corresponding to a dark coupling constant of $\alpha_D = 0.1$ and an LDM mass of M $M_\chi = M_{A'}/3$, and a total of ~ 800 neutrino background events, the SHiP 90% CL sensitivity is computed in the plane $(M_\chi, Y = \epsilon^2 \alpha_D (M_\chi/M_{A'})^4\ )$, where ϵ is the dark photon coupling, see Figure 2.14 and Table 2.1.

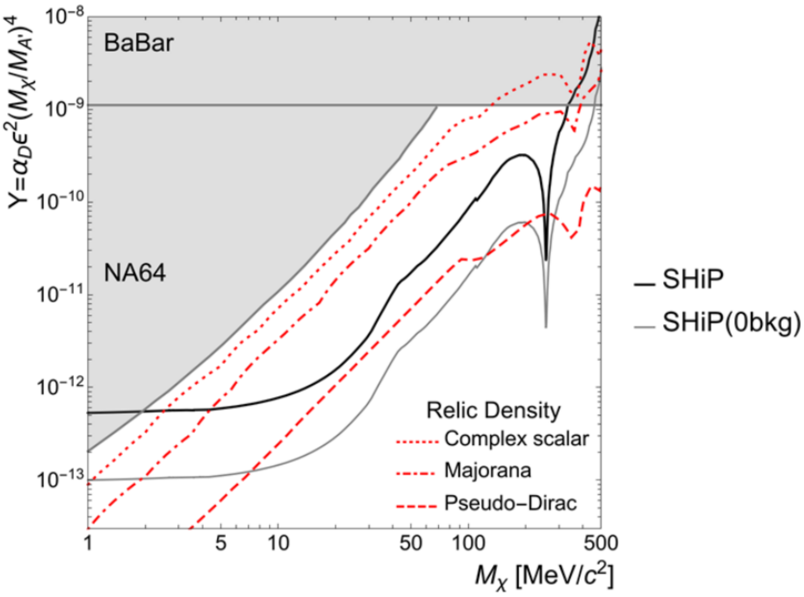


Figure 2.14: SHiP exclusion limits at 90% CL as a function of the LDM mass M_χ, compared to the current experimental limits by NA64 [48] and BaBar [49] (grey shaded area) and the predicted thermal relic abundances. The coupling is provided as $Y = 2\epsilon\alpha_D((M_\chi/M_{A'}))^4$.

Table 2.1: Number of background events to the dark matter search after cuts, from neutrino interactions with $2 \cdot 10^{20}$ p.o.t.

	ν_e	$\overline{\nu}_e$	ν_μ	$\overline{\nu}_\mu$	all
Elastic scattering on e^-	81	45	56	35	217
Quasi-elastic scattering	245	236			481
Resonant scattering	8	126			134
Deep inhelastic scattering		14			14
Total	334	421	56	35	846

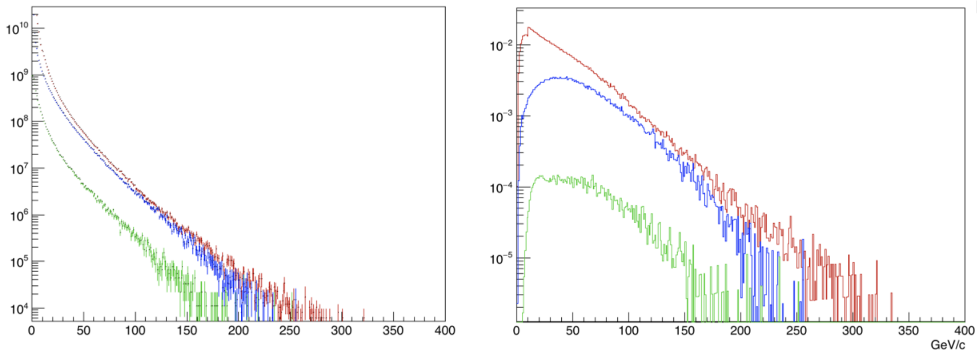


Figure 2.15: Left: energy spectra of muon (red), electron (blue) and tau (green) neutrinos at the beam dump. Right: energy spectra of neutrinos after their deep inelastic interactions with the Scattering Spectrometer. Units are arbitrary. [18].

2.4.2 Neutrino detection

The Scattering and Neutrino detector has the unique capability of detecting all three neutrino flavours and of distinguishing neutrinos from anti-neutrinos. The neutrino fluxes produced at the beam dump were estimated with FairShip, that is the tool to describe the SHiP experiment, including the contribution from cascade production in the target. The energy spectra of different neutrino flavors are shown in Figure 2.15(left), and the integrated yield at the beam dump for 2×10^{20} protons on target is shown in the left column of Table 2.2. The number of charged-current (CC) deep-inelastic interactions in the neutrino target is evaluated by convoluting the generated neutrino spectrum with the cross section provided by GENIE. The expected number of CC DIS in the target of the SND detector is reported in the right column of Table 2.2, and the corresponding energy spectra in Figure 2.15 (right).

The neutrino flavour is determined through the detection of the primary charged lepton produced in neutrino charged-current interactions. The lepton identification is also used to classify the tracks produced in the τ decay and, therefore, to identify the τ decay channel. In addition, the target magnetisation will allow for the first time to distinguish between ν_τ from $\overline{\nu}_\tau$. Two procedures, the event location and decay search, are applied. The location of a neutrino interaction consists of reconstructing the neutrino interaction vertex and its three-dimensional position with micrometric

Table 2.2: Expected neutrino flux for different neutrino flavors at the beam dump (left) and charged-current deep-inelastic interactions in the Scattering Spectrometer (right). 2×10^{20} protons on target were assumed.

	<E>[GeV]	Beam dump	<E>[GeV]	CC DIS interactions
N_{ν_e}	4.1	2.8×10^{17}	59	1.1×10^{6}
N_{ν_μ}	1.5	4.2×10^{17}	42	2.7×10^{6}
N_{ν_τ}	7.4	1.4×10^{18}	52	3.2×10^{4}
$N_{\overline{\nu}_e}$	4.7	2.3×10^{16}	46	3.6×10^{5}
$N_{\overline{\nu}_\mu}$	1.6	2.7×10^{18}	36	6.0×10^{5}
$N_{\overline{\nu}_\tau}$	8.1	1.4×10^{17}	70	2.1×10^{4}

Table 2.3: Expected number of ν_τ and $\overline{\nu}_\tau$ signal events observed in the different τ decay channels, except for the $\tau \to e$, where the lepton number cannot be determined.

Decay channel	ν_τ	$\overline{\nu}_\tau$
$\tau \to e$	1200	1000
$\tau \to h$	4000	3000
$\tau \to 3h$	1000	700
Total	6200	4700

accuracy. The decay search, consisting of the detection of the τ decay vertex, where the electron, muon and hadron channels are investigated. The identification of the neutrino flavor is performed through the identification of the charged lepton produced in the charged current interaction with the passive material of the brick. The lepton flavor identification is also important to classify the daughter tracks produced by τ decay, thus identifying the decay channel, as well as to identify charmed hadrons induced by neutrino interactions. The number of charged-current (CC) DIS interactions in the neutrino target is evaluated by convoluting the generated neutrino spectrum with the cross-section provided by GENIE. The unprecedented statistics of about 6000 ν_τ and 5000 $\overline{\nu}_\tau$ detected interactions is expected for 2×10^{20} protons on target, as reported in Table 2.3.

Chapter 3

Machine Learning Techniques

This chapter will be dedicated to a description of artificial intelligence, principally the machine learning field and the Random Forest classifier that will be used in our work. Reference used for this chapter is [50].

3.1 Artificial Intelligence and Machine Learning

Artificial intelligence (AI) was born in the 1950s, when a handful of pioneers began to ask whether computers could be made to "think". This issue is still being explored today. It is a wide-ranging branch of computer science concerned with building smart machines capable of performing tasks that typically require human intelligence. AI is an interdisciplinary science with multiple approaches, it covers machine learning and deep learning, but it also includes many methods that do not involve any learning. The machine learning (ML) arises from the question: could a computer go beyond "what we know how to order it to perform" and learn on its own how to perform a specified task? Could a computer automatically learn the processing rules by looking at data - instead of be passed by a programmers? This question opens the door to a new programming paradigm. In classical ones, human beings input rules (a program) and data to be processed according to these rules, and out come answers. With machine learning, men input data as well as the answers expected from the data and out come the rules. It finds statistical structure in these examples that eventually allows the system to come up with rules for automating the task. These rules can then be applied to new data to produce original answers.

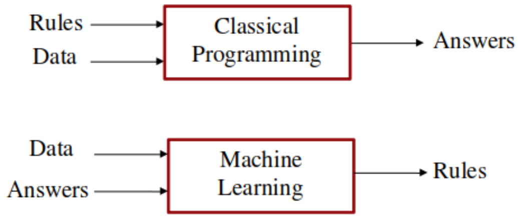


Figure 3.1: Classical programming vs Machine learning.

Not only ML is becoming increasingly important in computer science research but it also plays an ever greater part in our everyday life. It is also playing a significant role in advancing the current state of the art in particles physics, in which the reach of the experiments can be limited by the performance of algorithms and computational resources. Using ML, it is possible to develop self-learning algorithms capable of acquiring knowledge from data with the aim of making predictions.

3.2 The three types of machine learning

There are three different types of machine learning algorithms: *supervised, reinforcement* and *unsupervised learning* .

Supervised machine learning

Supervised learning is the most common subbranch of the machine learning. The goal of supervised ML is to learn a model from labeled training data that allows us to make predictions about unseen or future data. The term supervised refers to a set of samples where the desired output signals (labels) are already known.

Using a set of inputs x_i and the correct answers y_i the algorithm has to infer the mapping function $y = f(x)$ building a model for $f(x)$. Later this model is applied over a new data-set (the test set) and the predictions $\hat{y}_i$ are compared to the correct answers y_i.

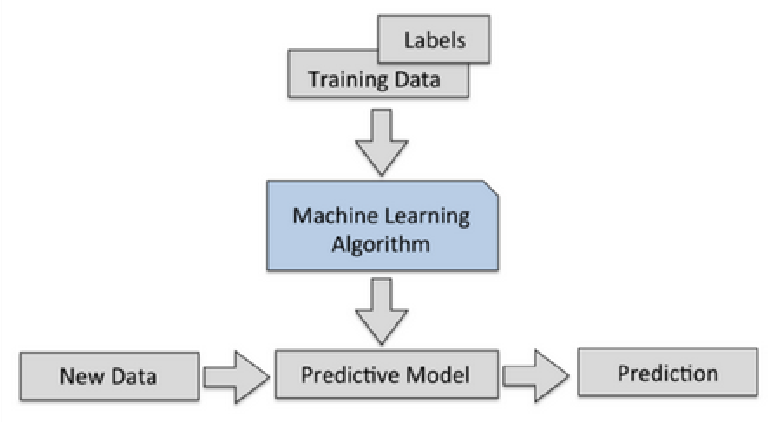


Figure 3.2: Supervised machine learning process.

Supervised learning problems can be further grouped into *classification* and *regression*.

The *classification* is a process of categorizing a given set of data into classes, that are often referred to as target, label or categories. It consists of approximating the mapping function from input variables to discrete output variables with the main goal to identify in which class/category the new data will fall into. The set of class labels can have a binary nature or not. In the first case the problem is a binary classification, in the second case instead, the predictive model can assign any class label that was presented in the training dataset to a new, unlabeled instance, in this case we have a multi-class classification. The Figure 3.3a illustrates the concept of a binary classification task given 30 training samples: 15 training samples are labeled as negative class (circles) and 15 training samples are labeled as positive class (plus signs). In this scenario, the supervised machine learning algorithm can be used to learn a rule - the decision boundary represented as a black dashed line - that can separate those two classes and classify new data into each of those two categories given its values.

A second type of supervised learning is the prediction of continuous outcomes, which is also called *regression* analysis. Given a number of predictor (explanatory) variables and a continuous response variable (outcome), it tries to find a relationship between those variables that allows us to predict an outcome. The Figure3.3b illustrates the concept of linear regression. Given a predictor variable x and a response

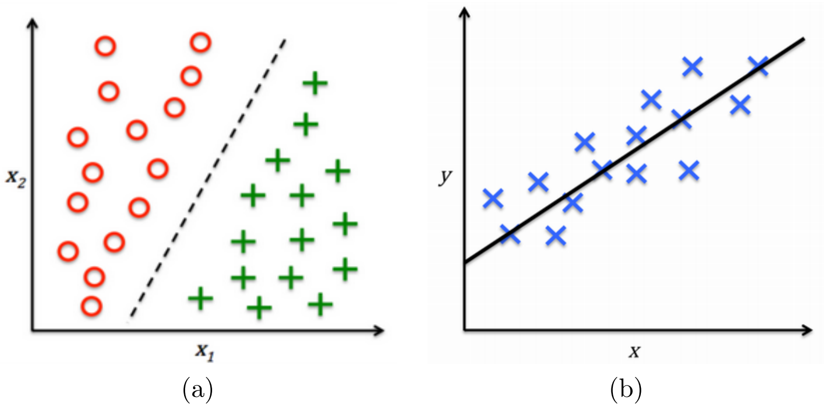


Figure 3.3: (a) Schematic representation of binary classification and (d) linear regression.

variable y, we fit a straight line to this data that minimizes the distance between the sample points and the fitted line. The intercept and slope learned from this data can be used to predict the outcome variable of new data. Linear regression is just one of the simplest cases of regression.

Reinforcement learning

The reinforcement learning allows an agent to discover in total autonomy the correct behavior to perform, see Figure 3.4. The learning of the agent is given by the continuous interaction with the environment, defined as everything that surrounds the agent and with which it can interact. From the environment it obtains a state and a reward. The agent is not told which action to take, he must instead find out which action will lead to obtain a greater reward simply trying and exploring the set of possible actions. In the most interesting problems, actions can condition not only the immediate reward but also the subsequent situation and state that will be created, and therefore also affect subsequent rewards. We can think of reinforcement learning as a field related to supervised learning. However, in reinforcement learning this feedback is not the correct ground truth label or value, but a measure of how well the action was measured by a reward function. Through the interaction with the environment, an agent can then use reinforcement learning to learn a series of actions that maximizes this reward via an exploratory trial-and-error approach or deliberative planning.

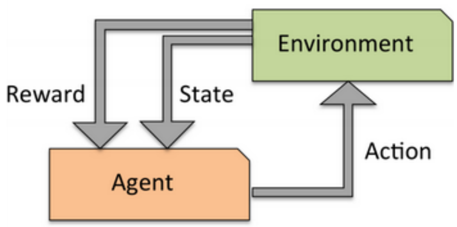


Figure 3.4: Reinforcement learning process.

Unsupervised learning

The unsupervised ML is used with unlabeled data or data of unknown structure, in order to extract meaningful information without the guidance of a known outcome variable or reward function. There is no quantity to compare the algorithm's scores with: the program is supposed to discover underlying structures in data by itself. *Clustering* is an exploratory data analysis technique that allows us to organize a pile of information into meaningful subgroups (clusters) without having any prior knowledge of their group memberships. Each cluster that may arise during the analysis defines a group of objects that share a certain degree of similarity but are more dissimilar to objects in other clusters, which is why clustering is also sometimes called "unsupervised classification". K-means is one of the most popular algorithms for clustering problems.

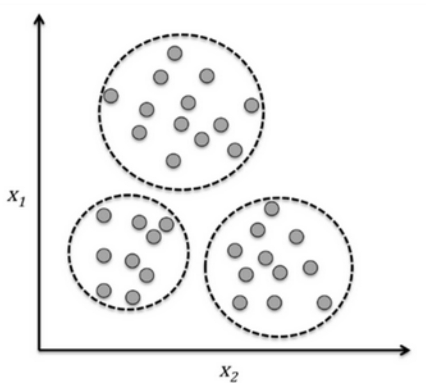


Figure 3.5: Clustering process.

3.3 Workflow diagram

The diagram below shows a typical workflow scheme for using machine learning in predictive modeling:

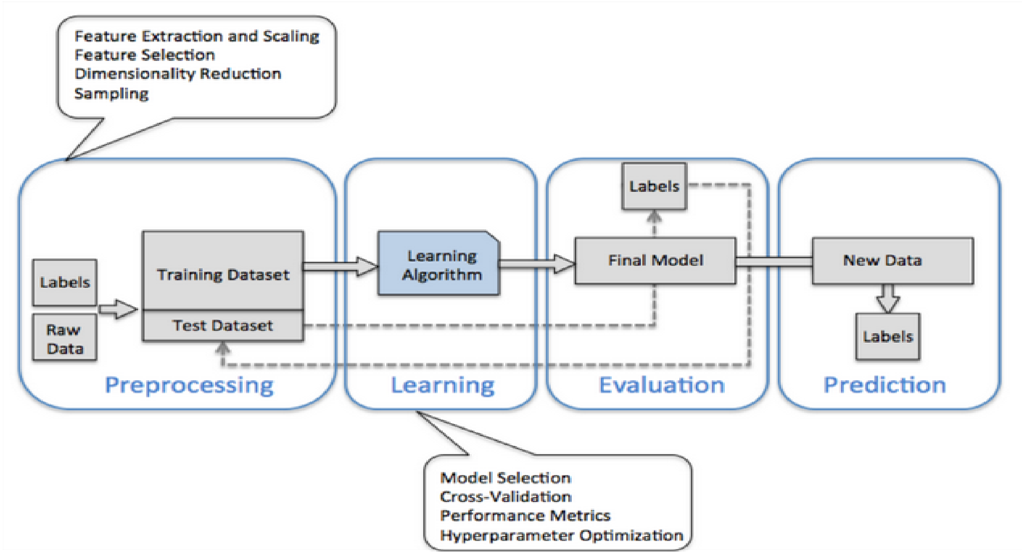


Figure 3.6: Typical workflow diagram for using ML.

The raw data comes from the measurements and labels that are the results of the measure constitutes the initial dataset. It is divided into two parts, a training and a test sets. A training dataset is used to better understand the data and to prepare them for modeling. It is also used to evaluate a set of different algorithms and to tune the hyperparameters of the selected model. On the other hand, the test set is used until the very end to evaluate the final model.

3.3.1 Data pre-processing

The pre-processing of the data is one of the most crucial steps in any machine learning application, since the raw data rarely comes in the better form and shape for the learning algorithm. From the raw data we want to extract meaningful features. Many machine learning algorithms also require that the selected features are on the same scale for optimal performances, it is name a *feature scaling* procedure. There are two approaches to bringing different characteristics on the same scale: the normalization and standardization. In general, normalization refers to the change in scale of the character in a range [0,1], which is a special case of min-max scaling. The standardization centers the characteristic column at mean 0 with standard deviation 1, so that the characteristic column assumes the form of a normal distribution, from which it is easier to derive the weights. In addition, standardization retains useful information about anomalies and makes the algorithm less sensitive to this problem than scaling min-max, which returns the data to a limiting range of values. The standardization procedure can be expressed by the following equation:

$$x_{std}^{(i)} = \frac{x^{(i)} - \mu_x}{\sigma_x} \tag{3.1}$$

μ_x is the sample mean of a given column of characteristics and σ_x is the corresponding standard deviation.
When the dataset has many variables, one risks considering them all at the same level of importance. But often this is not the case. Indeed, it can happen that some features may be highly relevant and therefore redundant to a certain degree or irrelevent. To consider only the most important features we could proceed through the features selection or dimensionality reduction.
The feature selection is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. The dimensionality reduction techniques are used to compress features into lower-dimensional sub-spaces. The dimension of the feature space has the following advantages: less storage space and the learning algorithm can run faster.

3.3.2 Learning phase

Each classification algorithm has its inherent biases, and no single classification model enjoys superiority if we do not make any assumptions about the task. In practice, it is therefore essential to compare at least a handful of different algorithms in order to train and select the best performing model. But before we can compare different models, we have to decide upon a metric to measure performance. The metrics that we know, are:

- accuracy: it is a synthetic indicator that summarizes the model's ability to respond correctly- it is the proportion of correct predictions (both true positives and true negatives) among the total number of cases examined:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{3.2}$$

- precision: it is the ability of a classifier not to label as positive instance that is actually negative. It is also called Positive Predictive Value being the number of true positives divided by all positive predictions. It is a measure of a classifier's exactness: a low precision indicates a high number of false positives

$$Precision = \frac{TP}{TP + FP} \tag{3.3}$$

- recall: it is the ability of a classifier to find all the positive instances. Recall is also called Sensitivity or the True Positive Rate being the number of true positives divided by the number of positive values in the test data. It is a measure of a classifier's completeness and a low recall indicates a high number of false negatives

$$Recall = \frac{TP}{TP + FN} \tag{3.4}$$

- f-score: it can be interpreted as a weighted average of the precision and the recall, where an F-score reaches its best value at 1 and worst score at 0:

$$F - score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{3.5}$$

- confusion matrix: a table showing correct predictions and types of incorrect predictions. The confusion matrix is a $N \times N$ table where N is the number of the classes - in binary classification they are only two classes. Each row of the matrix represents the instances in a actual class while each column represents the instances in a predicted class. The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another).

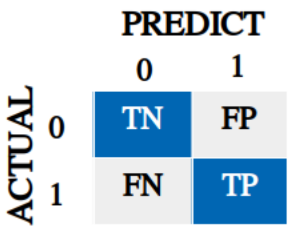


Figure 3.7: Confusion metric as a table that shows true negative and true positive on the main diagonal and false negative and false positive on the other.

All the metrics, can be computed by the confusion metrics. TP and TN are the true positive and the true negative instances, they are correctly classified for each class that are on the main diagonal, FP and FN are the false positive and false negative defined as the number of mis-classified instances of each class, shown on the secondary diagonal, see Figure 3.7.

3.3.2.1 Metrics estimation for balanced and imbalanced dataset

In the most of the classification problems, accuracy is a common indicator to evaluate the performance of a classifier. The reason it is widely used is because it is easy to calculate, easy to interpret, and is a single number used to summarize the function of the model. It must be said that this works well when the classes in the data set are in the same proportions, but when the data set is not balanced, the accuracy becomes meaningless.
An imbalanced classification problem is an example of a classification problem where the distribution of examples across the known classes is biased or skewed. The distribution can range from slight deviation to severe imbalance, where there is one example in the minority class and hundreds, thousands or millions of examples in the majority class. In the first case, accuracy can still be used as a useful measure, but when the skew in the class distribution is severe, accuracy may become an unreliable measure of model performance.
Typically in imbalanced classification problem, the minority class is of the most interest - in particle physics it can be represented by a rare or extremely rare events that we are interested in. This means that a model's skill in correctly predicting

the class label or probability for the minority class is more important than the majority class. Since the minority class contains only a few examples in the data set, it is difficult to predict. For a model, it may be more challenging to learn the characteristics of examples from this class and to differentiate them from the majority class.
In this scenario, the confusion metrics help better to understand how many instances are correctly classified and mis-classified for each class and the precision and recall are more useful metrics than the accuracy. They provide more information because they are more focused on the positive class than negative one. Precision and recall can be combined into a single score that seeks to balance both concerns, called the F-score or the F-measure. It is also a popular metric for imbalanced classification.

3.3.2.2 Overfitting and underfitting in the Machine Learning

The goal of a good machine learning is to achieve models that generalize well to never-before-seen data. The two biggest causes for poor performance of machine learning algorithms are overfitting and underfitting. Overfitting is by far the most common problem in applied ML. It happens when a model performs well on training data but does not generalize well to unseen data (test data). If a model suffers from overfitting, we also say that it has a high variance, which can be caused by having too many parameters that lead to a model that is too complex given the underlying data. Non-parametric and nonlinear models have greater flexibility in learning the target function, so the possibility of overfitting is greater. For this reason, those algorithms also include parameters or techniques to limit and constrain the amount of detail of model learning. For example, for a decision tree that is a very flexible non-parametric machine learning algorithm sensitive to overfit the training data, the problem can be solved by pruning it after the tree has learned it to delete some of the details it has picked up. Our model can also suffer from underfitting (high bias), which means that our model is not complex enough to capture the pattern in the training data well and therefore also suffers from low performance on unseen data. Figure 3.8 shows the three different situations, the underfit and overfit in the first and last plot. The second plot shows the good compromise between the two cases.
Cross-validation is a powerful preventative measure against overfitting. The idea is

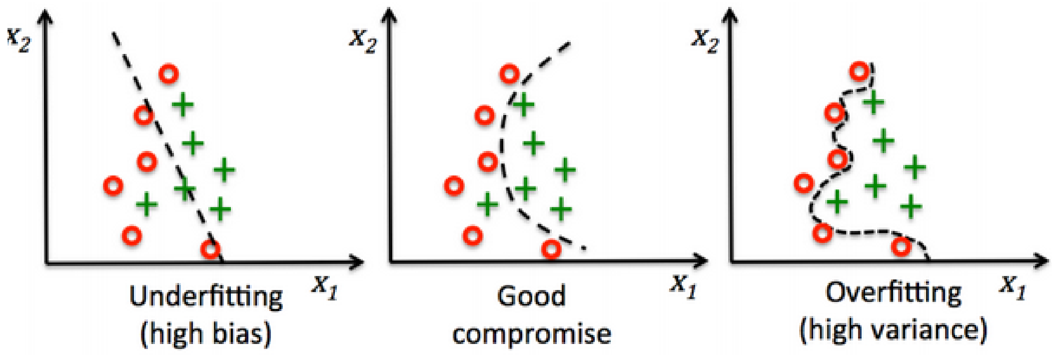


Figure 3.8: The first plot shows the underfitting, it happens when the model not capture the pattern in the training data. The last plot shows the overfitting, when the model learn too better the relationship of the training features. In both cases, the model does not give a good result on the test set. The second plot shoes the good compromise between the overfit and underfit.

to use the initial training data to generate multiple train-test splits to tune your model. Cross-validation is also use to tune hyperparameters with only your original training set. This allows to keep the test set as a truly unseen dataset for selecting the final model.

3.3.2.3 Grid search cross-validation

In typical machine learning applications, we are also interested in tuning and comparing different parameter settings to further improve the performance for making predictions on unseen data. This process is called model selection, where the term model selection refers to a given classification problem for which we want to select the optimal values of tuning parameters (also called hyperparameters).
We know that the models have two types of parameters: those that are learned from the training data, for example, the weights in logistic regression, and the parameters of a learning algorithm that are optimized separately. They are the tuning parameters, also called hyper-parameters of a model, that are for example, the regularization parameter in logistic regression or the depth parameter of a decision tree. We cannot expect that the default parameters of the different learning algorithms provided by software libraries are optimal for our specific problem task. Therefore, we will make frequent use of hyperparameter optimization techniques that help us

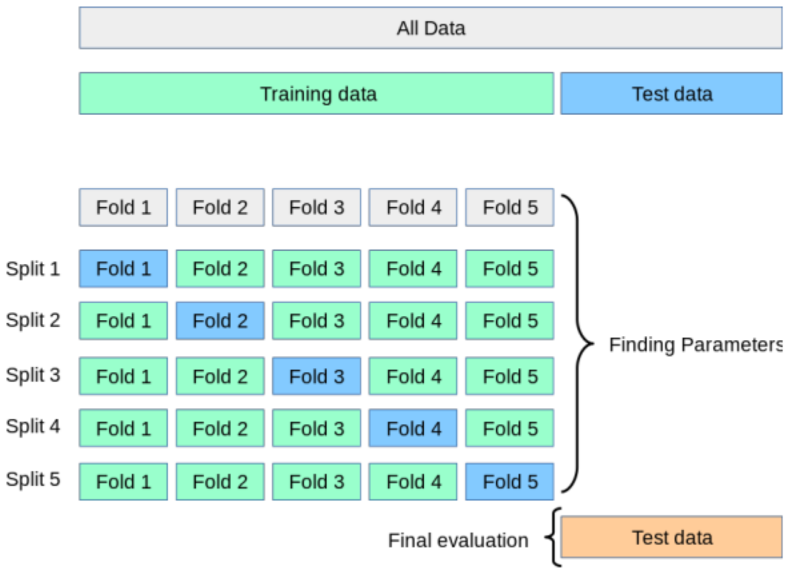


Figure 3.9: K-fold cross validation method implemented in *GridSearchCV* to estimate the best combination of hyper-parameters for each model. The complete dataset is divided into two parts: training and test sets. The first part is splitted into k folds (in this case k=5) where k -1 folds are used for the model training and one fold is used for testing. Found the best combination of hyper-parameters, the model performances are evaluated on the test set.

to fine-tune the performance. Intuitively, we can think of those hyperparameters as parameters that are not learned from the data but represent the knobs of a model that we can turn to improve its performance.
One of the best way to perform hyper-parameter optimization is grid search that is a brute-force exhaustive search paradigm where we specify a list of values for different hyper-parameters, and the computer evaluates the model performance for each combination of those to obtain the optimal set. The grid search also implements the k-fold cross-validation to estimate this generalization performance, see Figure 3.9: the initial dataset is divided into two parts, a training and a test sets. The first part is randomly splitted into k folds without replacement, where k *-1* folds are used for the model training and one fold is used for testing. This procedure is repeated k times so that we obtain k models and performance estimates. We then

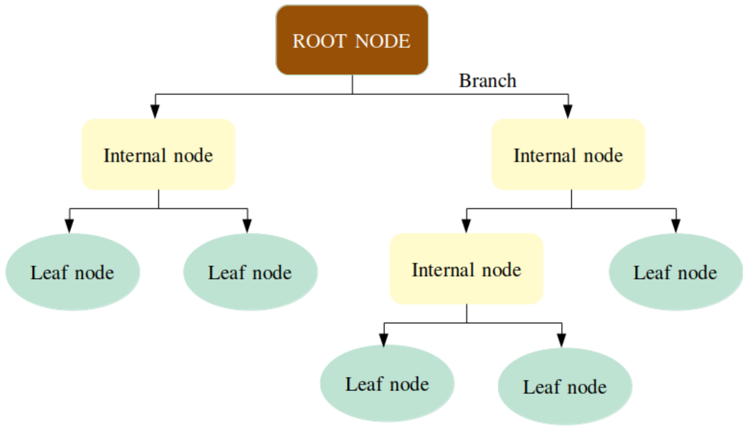


Figure 3.10: Visualization of the decision tree structure.

calculate the average performance of the models based on the different, independent folds to obtain a performance estimate that is less sensitive to the sub-partitioning of the training data. Once, k-fold cross-validation finds the optimal hyper-parameter values that yield a satisfying generalization performance. We can retrain the model on the complete training set and obtain a final performance estimate using the independent test set.

After we have selected a model that has been fitted on the training dataset, we can use the test dataset to estimate how well it performs on this unseen data to estimate the generalization error. If we are satisfied with its performance, we can now use this model to predict new, future data.

3.4 Decision tree-based classifiers

A class of very powerful machine learning models are based on the decision trees. The reason it is widely used is because it is not expensive to build, easy to interpret and to integrate with databases and it gives good metric response in many applications, even compared to other classification methods. The most disadvantage of decision trees include the overfitting that can be avoided with the pruning mechanism, setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree.

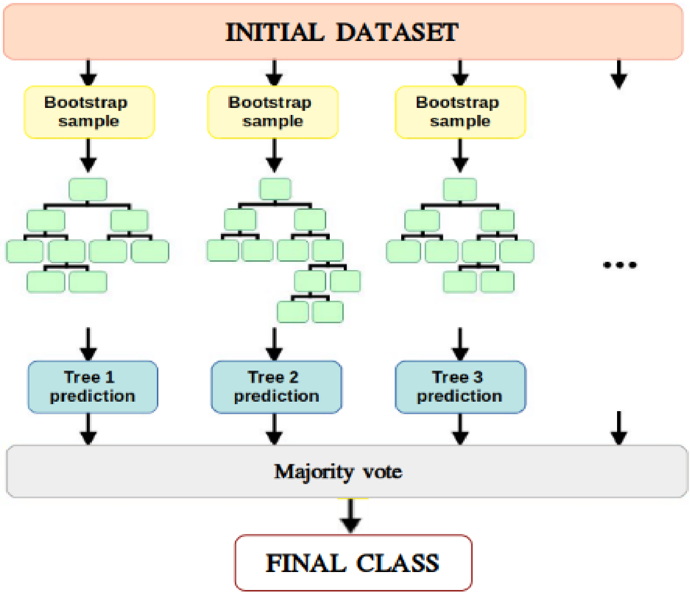


Figure 3.11: Visualization of the Random Forest classifier structure.

A decision tree is a structure similar to a flowchart, consisting of a nodes, edge-branch and leaf node. The topmost node in the decision tree is called the root node. It learns to partition based on attribute values. The internal nodes represent the subsets of the initial dataset divided according to the attributes, the branches represent the decision rules according to which the division takes place, the child nodes represent the final subsets into which the initial dataset has been divided.
A decision tree can be visualized in Figure 3.10. In order to improve classification performance and improve the stability of statistical fluctuations in training samples, multiple trees grow into a forest.

3.4.1 Random Forest Classifier

Random Forest can be considered as an ensemble of decision trees. The idea is to combine multiple "weak" learning systems to build a "strong" learning model more robust and powerful with better generalization errors and smaller errors easy to overfit. Figure 3.11 shows the structure of the Random Forest classifier. The algorithm can be summarized in the following four steps:

- Draw a random bootstrap sample of size n (randomly choose n samples from the training set with replacement).
- Grow a decision tree from the bootstrap sample. At each node:
  - randomly select d features without replacement.
  - split the node using the feature that provides the best split according to the objective function, for instance, by maximizing the information gain.
- Repeat the steps 1 to 2 k times.
- the forecasts of each tree are aggregated to assign the label on the based on a majority vote.

The information gain is defined as:

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j) \tag{3.6}$$

f is the features on which the classification is made, D_p and D_j are the dataset of the parent and the j-th child node, N_p and N_j are the size of the dataset of parent and *j-th* child respectively and I is the measure of "impurity". The information gain is therefore defined as the difference between the impurity of the node parent and the sum of the impurities of the child nodes: the lower the impurity of the child nodes, the greater the information gain. There are several estimates for the impurity of a node: Gini impurity (I_G), entropy (I_H) and mis-classification (I_E).
The entropy is calculated using the following formula:

$$I_H(t) = -\sum_{i=i}^{c} p(i|t) log_2 p(i|t) \tag{3.7}$$

p(i/t) is the proportion of the samples that belongs to class i for particular node. The entropy is maximal if the classes are perfectly mixed.
The Gini impurity is a criterion to minimize the probability of misclassification:

$$I_H(t) = \sum_{i=i}^{c} p(i|t)(1 - p(i|t)) = 1 - \sum_{i=1}^{c} p(i|t)^2 \tag{3.8}$$

At the end, there is the classification error:

$$I_H(t) = 1 - max p(i|t) \tag{3.9}$$

The “majority vote” consists in choosing as a label for a sample that of the class that was predicted by most classifiers.
Random forests do not offer the same degree of interoperability as decision trees, a big advantage of random forests is that we don’t have to worry so much about choosing good hyper-parameter values. We typically do not need to prune the random forest since the ensemble model is quite robust to noise from the individual decision trees. The only parameter that we really need to care about in practice is the number of trees k (step 3) that we choose for the random forest. Typically, the larger the number of trees, the better the performance of the random forest classifier at the expense of an increased computational cost.

Chapter 4

Electromagnetic showers

The SHiP experiment can probe the existence of Light Dark Matter through the observation of its scattering off electrons of the Emulsion Target. The Emulsion Could Chamber can be considered as an electromagnetic calorimeter where nuclear emulsion films inside the brick act as an active detector interleaved with passive plates. In the case of the SHiP experiment, it shows more than 5 active layers every X_0 over a total thickness of 10 X_0. The signature of Light Dark Matter interaction is the electromagnetic shower produced from the recoil of electron and that propagates throughout the ECC. In the SHiP target other sources of electromagnetic showers are expected, first of all photons. The micrometric tracking accuracy, provided by emulsion films interleaved by passive plates employed by the Emulsion Cloud Chamber allows to separate the showers produced by photons from those produced by electrons. For most of the events, the electron path in the brick is long enough for the electromagnetic shower to develop, thus allowing its detection and energy measurement.

The present work is focused on the development of an algorithm for the identification and the energy measurement of electromagnetic showers induced by electrons in SHiP-like Emulsion Cloud Chamber. A machine learning approach was used in order to fulfill the requirement of an high signal-to-noise ratio in an environment where the instrumental background is orders of magnitude larger than the signal.

Detailed Monte Carlo simulations have been performed in order to study the propagation and identification of the shower in an Emulsion Cloud Chamber. The newly developed algorithm was tested on data samples from a dedicated test beam con-

ducted at DESY laboratories in October 2019.
In this chapter, I will describe the physics principle of electromagnetic shower development, the DESY test beam and the performance study with Monte Carlo simulation in the ECC brick.

4.1 Interaction of particles with matter and e.m. showers

When charged particles penetrate a material, they interact with its constituents, i.e. the atoms it is made of. The interaction takes place between electromagnetic fields: if the incident particle has low energy it will interact with the electromagnetic field generated by the atomic shell's electrons while, at higher energies, it can interact with the one generated by the atomic nucleus (i.e. protons). The energy and the nature of incident particles also determine different ways of interacting with matter and different interaction processes are associated with different energy deposits inside the material. Moderately relativistic charged particles other than electrons lose energy in matter primarily by ionization and atomic excitation. The mean rate of energy loss (or stopping power) is given by the Bethe-Bloch equation:

$$-\frac{dE(x)}{dx} = 2\pi\frac{N_A Z\rho}{A}r_e^2 m_e c^2\frac{z^2}{\beta^2}\Big[ln\Big(\frac{2m_e\gamma^2 v^2 W_{max}}{I^2}\Big) - 2\beta^2 - \delta(\gamma)\Big] \tag{4.1}$$

where N_A is Avogrado's number 6.022×10^{23} mol^{-1}; Z and A are the atomic number and the atomic mass of the absorber respectively; r_e is the classical electron radius that is 2.817×10^{-13} cm; z is the charge of the incident particle; W_{max} is the maximum kinetic energy which can be imparted to a free electron in a single collision, I is the mean excitation energy measured in eV and $\delta(\gamma)$ is the density correction which limits the energy loss for high values of γ [51]. The dependence on the material traversed appears in terms of A, Z, ρ, I as well as inside the additional term δ while the quantities dependent on the incident particle are z and, indirectly, the mass M, which appears in β and γ.
The units for the energy loss in Eq. 4.1 are MeV/cm; dividing by the density the energy loss is measured in MeVcm^2/g with the advantage that the value of dE/dx does not depend much on the crossed material.

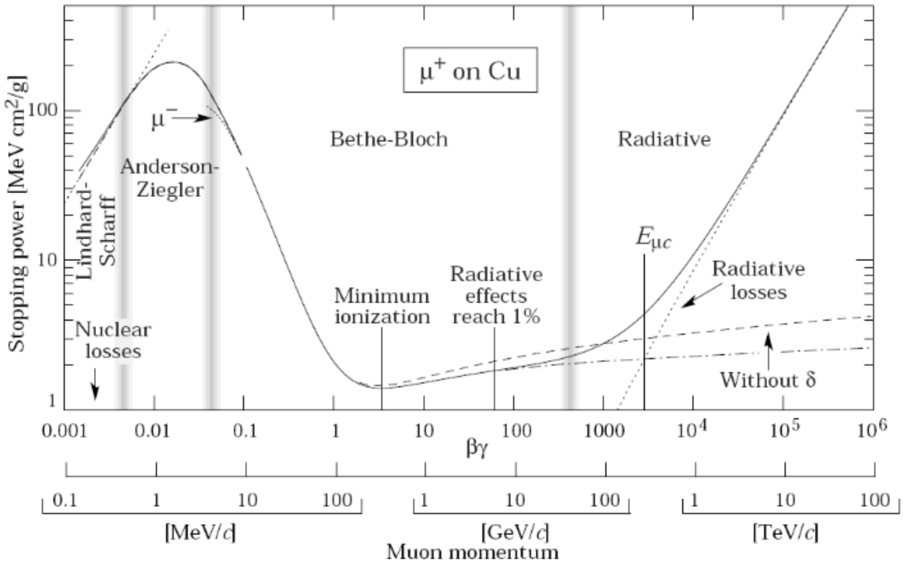


Figure 4.1: Stopping power ($=< -dE/dx >$) for positive muons in copper as a function of $\beta\gamma = p/Mc$, [52].

Figure 4.1 shows the trend of the Bethe Bloch equation (divided by the density of the material) for muons in copper, as a function of the product $\beta\gamma = p/Mc$. The curve has a minimum for $\beta\gamma \sim 3$: particles with energy corresponding to this value are commonly called at the "minimum ionising particles" (MIPs).

Multiple scattering through small angles

A charged particle traversing a medium is deflected by many small-angle scatters. Most of this deflection is due to Coulomb scattering from nuclei, and hence the effect is called multiple Coulomb scattering. The Coulomb scattering distribution is roughly Gaussian for small deflection angles, but at larger angles (greater than a few θ_0, defined in equations 4.2 and 4.3) it behaves like Rutherford scattering, tails larger than those expected for a gaussian distribution. If we define:

$$\theta_0 = \theta_{plane}^{rms} = \frac{1}{\sqrt{2}}\theta_{space}^{rms} \tag{4.2}$$

then it is sufficient for many applications to use a Gaussian approximation for the central 98% of the projected angular distribution, with a width given by:

$$\theta_0 = \frac{13.6MeV}{\beta cp} z\sqrt{x/X_0}[1 + 0.038ln(x/X_0)] \tag{4.3}$$

Here p, c, and z are the momentum, velocity, and charge number of the incident particle,and $x = X_0$ is the thickness of the scattering medium in radiation lengths (defined below). The larger the momentum, the less the particles is deflected.

Model of electromagnetic shower

At high energy, the loss of energy for the electrons incident on a thick absorber, occurs through the "avalanche" creation of secondary particles, i.e. an *electromagnetic shower*. The particles that compose the shower are electrons, positrons and photons. The simplified description of the development of electromagnetic showers was first developed by Heitel and Rossi [53]. There are two high-energy (E $>$ 100 MeV) electromagnetic energy loss mechanisms through which the electromagnetic cascade is propagated, the bremsstrahulung for e^+ and e^- particles and pair production for γ particles.
The bremsstrahulung effect consists of the radiation of electromagnetic waves (photons) by charged particles when they are subjected to a central force field. The acceleration that an incident electron undergoes due to the electrostatic attraction of the atomic nuclei of the crossed material generates photons in the direction tangent to the motion. This effect, known as synchrotron radiation, is at the basis of accelerator machines like synchrotrons.
The energy loss by radiation is described by the equation 4.4:

$$-\frac{dE(x)}{dx} = 4\alpha N_A\Big(\frac{e^2}{mc^2}\Big)^2 ln\Big(\frac{183}{Z^{1/3}}\Big)\frac{Z(Z+1)}{A}Q^2E \tag{4.4}$$

where α is the fine structure constant, N_A the Avogadro number, m and Q are the mass and charge of incoming particle, while A and Z are mass and atomic number of the material. So, in the presence of a nucleus, for energy-momentum conservation an electron emits a photon. In this interaction the γ is emitted at a given angle and the electron will be diverted, therefore there will be an enlargement of the shower with respect to the line of flight of the initial particle. Photons, being particles without mass, lose energy inside a material in a different way than particles with mass and charge.

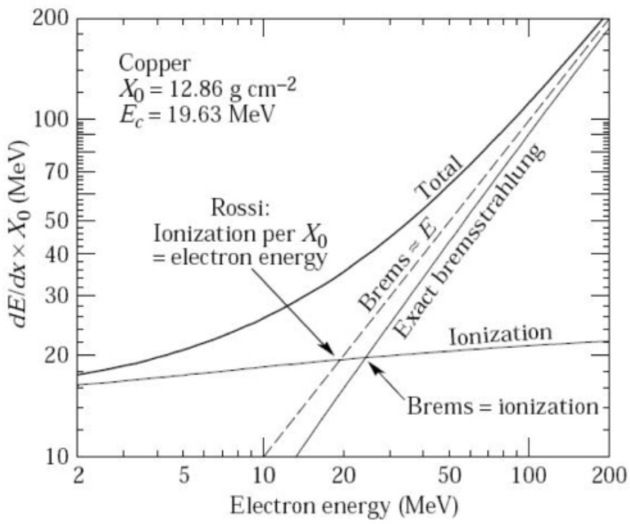


Figure 4.2: Energy loss mechanism for electrons and positrons in copper. Above the "critical energy" the main mechanism is the emission of bremsstrahlung radiation.

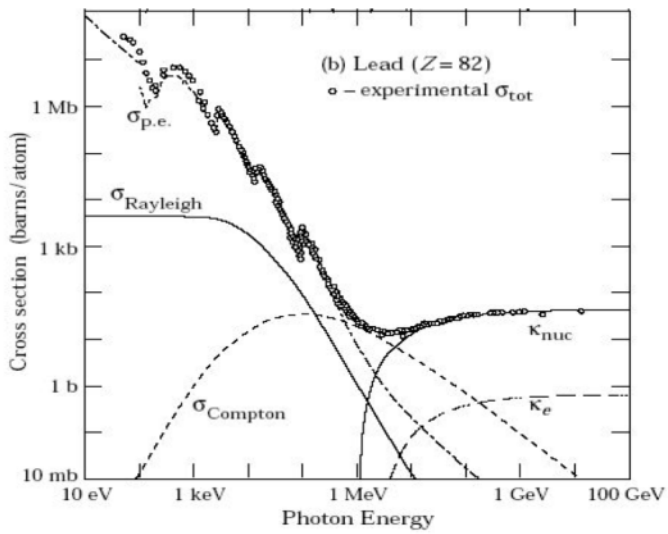


Figure 4.3: Cross section for the interactions of photons in lead. At high energies (> 10 MeV) the main contribution is pair production.

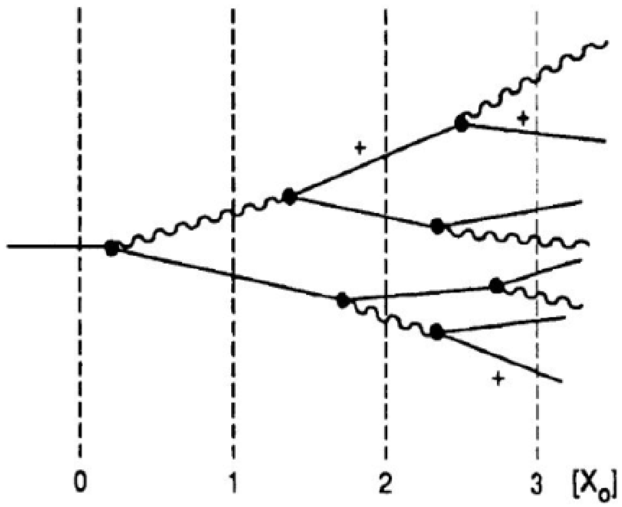


Figure 4.4: Two-dimensional model of an electromagnetic shower. Two processes, bremsstrahlung and pair production are responsible for the energy degradation and the shower development.

The main phenomena they encounter are: the photoelectric effect, most likely to occur at low energies, Compton scattering, happening mostly in an intermediate region, and pair-production, which dominates at high energies and consist of the conversion of the photon into an e^- and e^+ pair. The threshold condition for the latter process is $E_\gamma > 2m_ec^2$.

Since the energy loss per unit of path increases with energy and, for high energy, it is approximately proportional to energy of the particle, the equation 4.4 can be rewritten as:

$$-\frac{dE(x)}{dx} = \frac{E(x)}{X_0} \tag{4.5}$$

where X_0 is the radiation length, expressed in gcm^{-2}. It represents the distance at which the energy of an electron or positron is reduced to a factor of 1/e of the initial energy.

Figure 4.2 shows the energy loss mechanism for electrons and positrons in copper.

Figure 4.3 shows photon total cross sections as a function of energy in lead, showing the contributions of different processes: $\sigma_{p.e.}$ is the atomic photoelectric effect (electron ejection, photon absorption); $\sigma_{Rayleigh}$ shows the Rayleigh (coherent) scattering-atom neither ionized nor excited; $\sigma_{Compton}$ is the Compton scattering of an electron;

K_{nuc} shows the pair production for nuclear field; k_e shows the pair production for electron field.
An electromagnetic shower extends inside the matter both longitudinally and laterally. In the Rossi and Heitel description it is assumed that at each radiation length there is an energy loss, i.e. if at the beginning there is an electron after X_0 it has irradiated a photon and after a further X_0 the photon will have converted into an e^+ and e^- pair. Indeed, the law that governs the conversion of photons into electrons is:

$$N(x) = N_0 e^{-x/\lambda} \tag{4.6}$$

where $\lambda = 9/7X_0$. Since $9/7 \approx 1$, the model of multiplying the number of particles by a factor 2 at each step of X_0 is reasonable.
With this model, after n times X_0 there are 2^n particles where n is the number of X_0 crossed, see Figure 4.4. The bremsstrahulung mechanism remains dominant until the energy of the particles are above the *critical energy* E_C, which represents the value where the average energy loss per ionization equals the radiation loss. Once the critical energy is reached the shower begins to extinguish and disappear.
The maximum propagation of the electromagnetic shower inside the matter, is given by:

$$X_{max} = \frac{log(E/E_c)}{log_2} X_0 \tag{4.7}$$

In addition to the longitudinal propagation of the electromagnetic shower, there is a lateral development, which is not explained in Rossi's formalism. The creation of a shower is strongly linked to the process of couples production and to the bremsstrahlung process, the latter is characterized by a maximum angle of photon emission which, at first approximation, has the following dependence on the relativistic parameter $\gamma_{rel} = \frac{1}{\sqrt{1-\beta^2}}$, that is:

$$\theta_{emission} = \frac{1}{\gamma_{rel}} \tag{4.8}$$

for $\gamma_{rel} \to 0$, i.e. at low energy, photons tend to be emitted at large angle. This, together with the multiple diffusion phenomenon, produces a progressive widening of the shower as its energy decreases and this widening is well parameterized by an exponential trend. We can define a characteristic magnitude to describe the lateral damping, the Moliere radius, i.e. the radius of the cylinder in which 95% of the

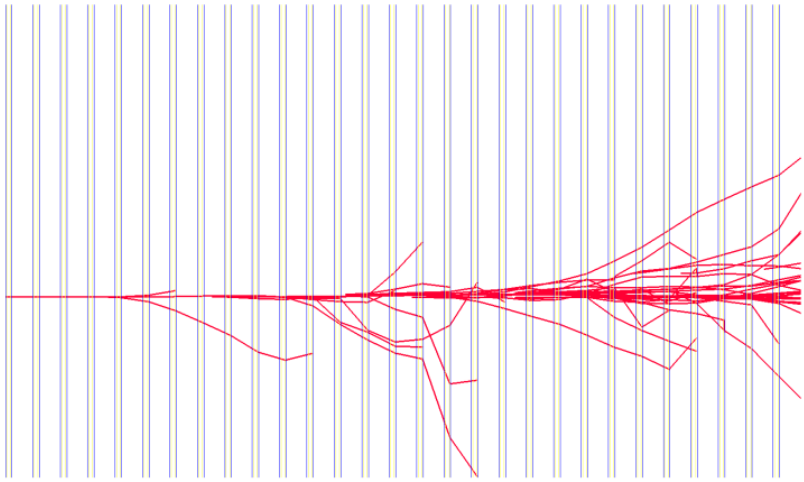


Figure 4.5: ZX display of the simulation of electromagnetic shower in the Emulsion Cloud Chamber produced by an electron of 6 GeV. The dimension of x and z are in μm.

shower is contained in an attenuation length X_0. The Molier radius depends on the X_0 trough this equation:

$$R_M = \frac{21MeV}{E_c} X_0 \tag{4.9}$$

Figure 4.5 shows the development of electromagnetic shower in an Emulsion Cloud Chamber that acts as a calorimeter where nuclear emulsion films are the active detector interleaved with passive plates.

4.2 DESY Test Beam

In October 2019, the Emulsion Cloud Chamber was set up in a test beam campaign in the area 22 of DESY II synchrotron and data was taken for one week.
The aim of the test beam is the identification of low energy electromagnetic showers in a target like the one expected for SHiP. The Emulsion Cloud Chamber together with the SciFi plane (Target Tracker) have been exposed to a pure electron beam, whose parameters have been optimized to have a number of electromagnetic showers inside the brick equal to that expected for the SHiP experiment. The reconstruction

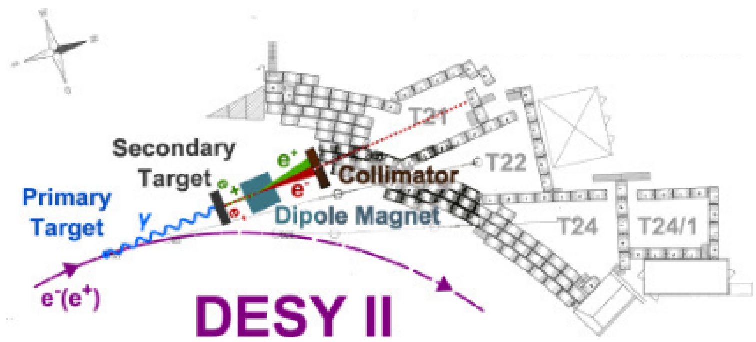


Figure 4.6: Schematic view of the DESY II beam generation for test beam areas.

of electromagnetic shower will be performed integrating the information about the two detectors.

4.2.1 The beam line

In DESY II, electrons or positrons are accelerated up to an energy of 7 GeV. A bremsstrahlung beam is generated by a carbon fiber put in the circulating beam of the electron/positron synchrotron DESY II. The photons are converted to electron/-positron pairs with a metal plate. Electrons with energies ranging from 1 to 6 GeV/c are then selected by a dipole magnet, spreading the beam into a horizontal fan and a collimator (formation of the extracted beam). Finally they are conducted to the test beam area. In this energy range, electrons can indeed be considered minimum ionising particles (MIPs). A schematic layout of a test beam line at DESY is shown in Figure 4.6.

4.2.2 Experimental setup

The conceptual setup of the experiment is shown in Figure 4.7: an ECC target is followed by two SciFi stations, with a passive 50-mm-thick lead target in the middle. Nuclear emulsion films used in the ECC target have a transverse size of (125×100) mm^2. Figure 4.8 shows a open brick with all the emulsion films. The used passive material is lead for almost all the bricks, except one brick built in tungsten, that

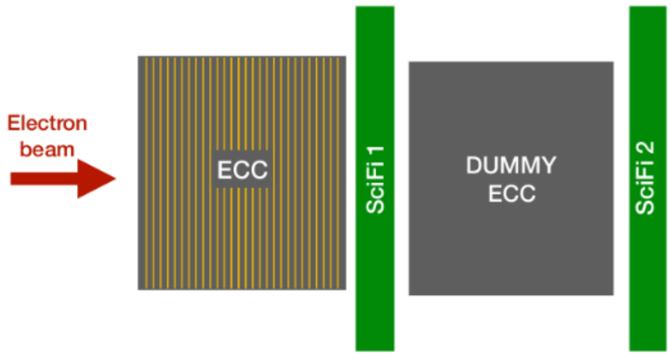


Figure 4.7: Conceptual layout of the apparatus used in the DESY test beam.

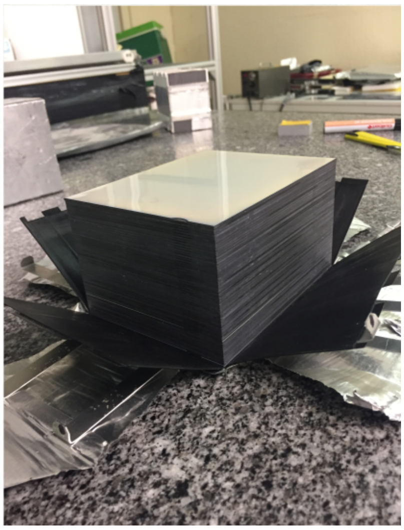
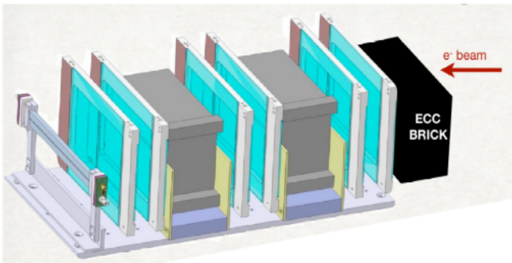


Figure 4.8: A brick of nuclear emulsions.

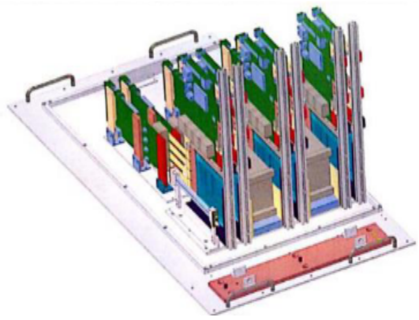
thanks to its higher density and its shorter radiation length and smaller Molière radius, improves the electromagnetic shower containment.

The SciFi tracker consists of six fibre mat planes spread along the beam direction, the z-axis of the detector was aligned with the beam axis and the detector measures 350mm along this axis.

Each plane consists of two layers for measurements in the x and y directions. The 12 detection layers are divided into six X layers ($X_1 ... X_6$) and six Y layers ($Y_1 ... Y_6$). The total active area of the detector is 135 mm^2 in the transversal direction allow-



(a)



(b)

Figure 4.9: (a)SciFi and ECC brick. (b) Overview of the stations composing the SciFi detector.

ing precise spatial information and track reconstruction. The scintillation light is transported and read out by multichannel arrays of silicon photomultipliers coupled to one end of the fibre mats. The SiPMs convert the light signal into an electronic one. A metallic shield covering SiPMs and fibre mats was set up, to ensure that the detector remains "light-tight" (light shielded). Their layout can be seen in Figure 4.9. Each single plane provides a time resolution better than 500 ps, allowing to acquire at a trigger rate of 50 kHz for single track and better than 1 kHz for shower readout. The detectors were placed on a movable table (x-ytable) that could be controlled in the control hut allowing to spread the beam on the whole surface of the brick. Eight runs have been performed. Each one with different combination of thickness of passive materials ranging from 14 mm to 56 mm and energy of the beam; in this way the performances can be studied as a function of the depth of the shower in the ECC and with the corresponding different SciFi sampling. The electron beam was at fixed energies, with exposure runs of 2, 4 and 6 GeV, and spread on the whole emulsion surface thanks to the movable table on which the emulsion and SciFi are positioned. The electron density was between 2 and 3 e^-/cm^2 , while the event rate was 1.5 Hz. A summary of all 8 RUNs is reported in Table 4.1. RUN8 is the only one with two ECC detectors, the latter replacing the passive lead block between SciFi stations.

Table 4.1: Summary of exposures at DESY19.

Configuration	Energy	ECC Passive Plates	N emulsion films
RUN1	6 GeV	56×1 mm Pb	57
RUN2	6 GeV	42×1 mm Pb	43
RUN3	6 GeV	28×1 mm Pb	29
RUN4	6 GeV	14×1 mm Pb	15
RUN5	2 GeV	28×1 mm Pb	29
RUN6	4 GeV	28×1 mm PB	29
RUN7	6 GeV	18×0.9 mm W	19
RUN8A	6 GeV	14×1 mm Pb	15
RUN8B	6 GeV	28×1 mm Pb	29
Total	-	283 mm Pb 18 mm W	265 ~ 3.3 m^2

4.3 Simulation of electromagnetic showers

Monte Carlo simulation is used to estimate the detector performance in the identification of electromagnetic shower and measuring the energy of electrons.
Electron propagation and interactions in the apparatus are simulated through Geant4 framework [54]. The software allows to build a simulated experiment taking into account the geometry, tracking and response of detectors. The properties of detector material, of the properties of the particles and of all the physical processes involved in their interaction are provided to the software.
The following reference system is used:

- the z coordinate follows the electron beam line;
- the y coordinate represents the height.
- Finally, the x coordinate represents the width.

The origin is set downstream of the Emulsion Cloud Chamber brick.
To study the propagation of the showers in an Emulsion Cloud Chamber 10.000 electrons have been simulated in the configuration of the RUN1 in which the detector is made of 57 emulsion films.

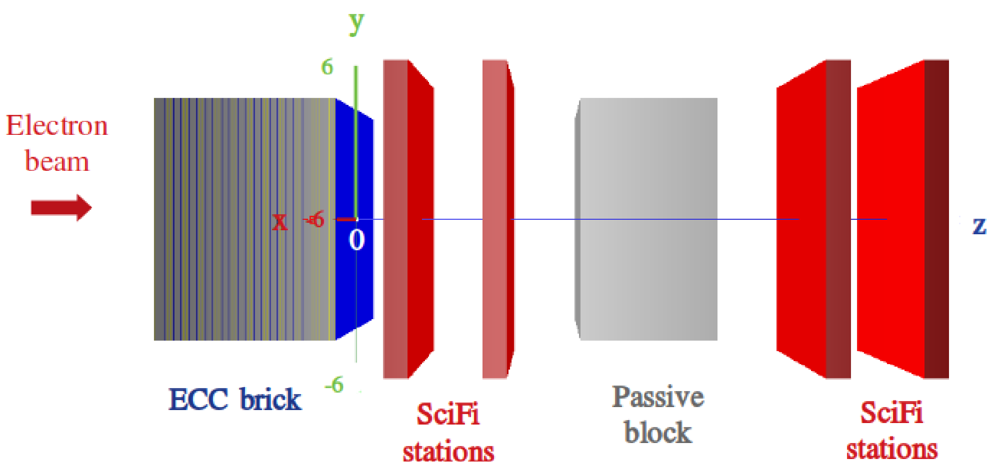


Figure 4.10: Simulation display with brick and SciFi stations. The axes and direction of the beam are shown.

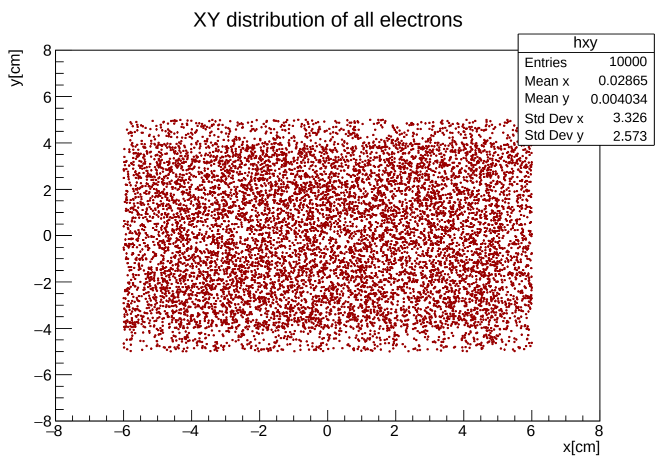


Figure 4.11: XY distribution of all electrons on the first emulsion film of the target surface.

All the incident electrons are orthogonal to the target surface with the energy of 6 GeV, and they are uniformly distributed in the xy plane within half a centimeter from the edge of the emulsion, see Figure 4.11.

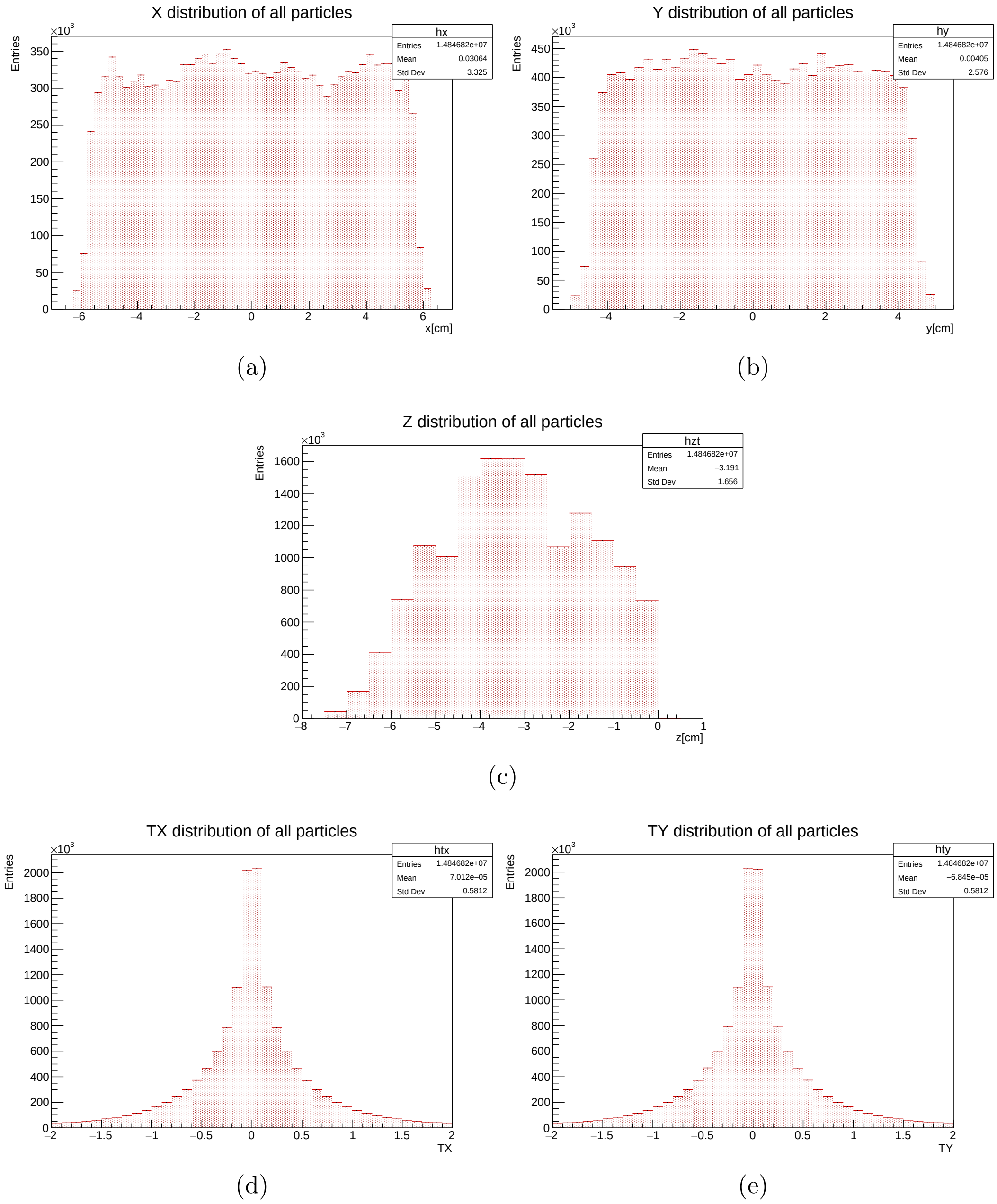


Figure 4.12: The figures show the x, y, z and TX, TY distribution of all showers particles.

Starting from the first electron with high energy, an avalanche of secondary particles

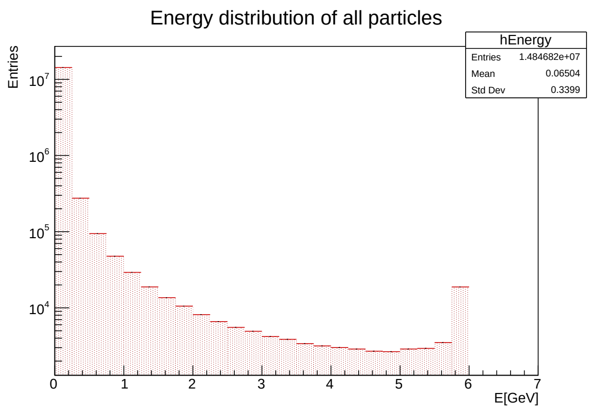


Figure 4.13: Energy distribution of all showers particles.

is generated. Except the photon which is neutral and therefore not revealed, the other charged particles such as electrons and positrons leave traces in the sensitive layers of the detector.

Figure 4.12 shows the distribution of particles' position, angle and energy. In particular, Figures 4.12d and 4.12e shows the TX and TY distributions, that are the tangent of the impact angles computed as $TX = p_x/p_z$ and $TY = p_y/p_z$ where $\vec{p} = (p_x, p_y, p_z)$ is the particle momentum.

The number of particles increases with the depth of the detector reaching the maximum halfway detector, see Figure 4.12c. After the peak, the number of revealed particles decreases, this means that the majority of the electrons no longer have the energy required to make breemsstrahulung, their energy becomes less than the critical energy which for lead is 6.2 MeV.

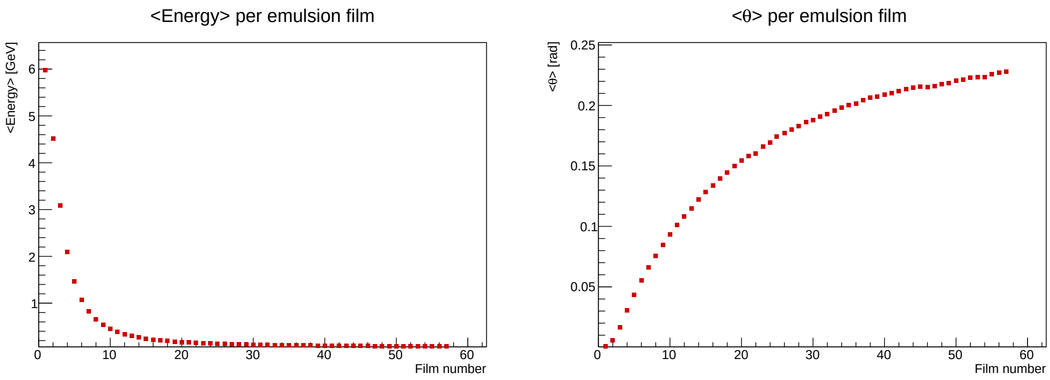
The energy distribution of all the particles in the electromagnetic shower is shown in Figure 4.13, the most of e^+ and e^- have an energy smaller than 1 GeV.

Not all hits stored in the emulsion layers are accepted but only those satisfying the so called visibility conditions:

$$\begin{cases} E > 30MeV \\ \theta < 1rad \end{cases} \tag{4.10}$$

The lower energy threshold is required to accurately track a real particle traversing the emulsion film, while the upper angular threshold corresponds to the maximum

angle acquired during the scanning of the emulsions.
The mean energy and mean angle of the particles that satisfy the visibility conditions are shown as function of the film in Figure 4.14.



(a) (b)

Figure 4.14: (a) Mean energy and (b) mean angle computed for each emulsion film.

The first plot shows that already after the 7th film the energy is below 1 GeV.
As the depth inside the brick is larger the energy of electrons and positrons decreases, resulting in larger scattering angles.
With the depth of the detector the particles decrease their energy and scatter more, resulting in greater angular dispersion.

In this preliminary analysis, a simulation of 10.000 electrons with the same energy of 6 GeV in the thickest brick, with the largest number of plates equal to 57 has been analyzed. The development of the showers in the detector was studied through the reconstruction of the tracks left by charged particles in the sensitive layers.
However, in the real case of the DESY TB, there are RUNs with 360 electrons and the thickness of the brick is variable, from 19 to 57. Since the quality of the reconstruction of the showers depends on the number of plates and emulsions that make up the detector, in the next chapter we will study the RUN3 in which the detector is made of 29 plates.

Chapter 5

Signal and background characterization

The identification of electromagnetic showers in an Emulsion Cloud Chamber is based on the selection of segments produced by electron and positron pairs during the shower development in the target material. The energy of the shower is estimated via calorimetric measurement, by counting the number of selected segments and applying a calibration procedure. The main challenge of a shower identification algorithm is the selection of signal segments out of a huge background (that is a few order of magnitudes larger than the signal).

The algorithm currently used by the Collaboration is based on a topological selection of segments that are selected as signal if they fall within a cone, having the vertex in the primary electron position and with an opening angle of 20 mrad, followed by a cylinder, having a radius of 800 μm. This stringent selection is due to the presence of a large background and the absence of any visual inspection of selected segments. This method guarantees an energy resolution limited to 30%.

The work done in this thesis is focused to develop a new algorithm implementing the machine learning techniques for the identification of electromagnetic showers and the background rejection. This algorithm is divided into two parts. In the first part, a detailed study on the new topological variables that define a space region in which we expect to have a shower is performed. In the second part these features are used as input for the classification algorithm. This chapter is focused on the characterization of signal and background by using the Monte Carlo simulation described in section 4.3.

5.1 Segment and tracks reconstruction

The tracks left by charged particles in emulsion are recorded by a series of sensitized AgBr grains, which grow to a maximum diameter of 0.6μm during the development process of the emulsion.

The emulsion films are analysed by means of an automated system which allow for fast extraction of physical information from emulsion films. The Super Ultra Track Selector (S-UTS) [55] and the European Scanning System (ESS) are the scanning systems used for the study of neutrino interactions in OPERA [56]. The ESS is a software-based system which makes use of commercial hardware components. The working principle is that it can reconstruct all the trajectories in each field of view by measuring the spatial position of the developed particles along their trajectories, regardless of their slope with track finding efficiency of above 90% for particles with tg$\theta \leq 0.7$, where θ is the angle with respect to the perpendicular direction to the film, and the track positioning error is about 0.3μm.

The upgraded version of ESS developed by the Emulsion Group in Naples will be used for SHiP data reconstruction. It implements the new scanning software LASSO [57], a faster camera and a small pixels for an scanning speed of 190 cm^2/h improved ten times with respect to the previous one [58]. The emulsion scanning takes place on a glass-made mechanical stage, movable along both X and Y axes and in the Z direction. It is equipped with a computer-controlled monitored stage, a dedicate optical system and a computer with a frame grabber, a vision processor board and a motor control board. A microscope with a vertical resolution of a few microns is used in such a way raising or lowering the optimal focus plane of the objective to cover the entire emulsion thickness, the three-dimensional structure of the track can be reconstructed. To guarantee a homogeneous optical path inside the layers, the immersion lens and oil used have the same refraction index of the emulsion films, see Figure 5.1. The automated optical microscope analyses the whole thickness of the emulsion, acquiring various topographic images. These raw images are processed using a high-pass digital filter to enhance the contrast between the pixels that are the focused particle image and the pixels that are darkened by non-focused particles or electronic noise. The reconstruction of a sequence of at least 20 aligned grains constitutes a single $micro - track$ inside one emulsion layer. The possibility of obtaining sub-micron tracking accuracy is closely related to the stability of the relative position of the particles within the emulsion volume and after development. Two main deformation effects must be considered: shrinkage and distortion. The shrinkage is defined as the thickness of the emulsion at the time of exposure divided by the thickness after development. The tracking algorithm takes this factor into account, micro-tracks are indeed multiplied by this factor to obtain the actual slope. The base-track reconstruction is performed by projecting micro-track pairs across the plastic base, and looking for consistency within given slope and position tolerances, see Figure 5.2. They are selected on the basis of a χ^2 cut, a quality estimator of the angular agreement between the micro-tracks and the corresponding base-track, defined as:

$$\chi^2 = \frac{1}{4}\sqrt{\left[\left(\frac{\Delta S_{lon}}{\sigma_{lon}}\right)^2_{top} + \left(\frac{\Delta S_{tra}}{\sigma_{tra}}\right)^2_{bot} + \left(\frac{\Delta S_{tra}}{\sigma_{tra}}\right)^2_{top} + \left(\frac{\Delta S_{lon}}{\sigma_{lon}}\right)^2_{bot}\right]} \tag{5.1}$$

where ΔS_{tra} and ΔS_{lon} are the angular differences between the slope of the top (bottom) micro-track and the corresponding base-track calculated in the transverse and longitudinal planes, respectively. The σ_{tra} is the transverse resolution and has a value of 10 mrad while the longitudinal resolution, σ_{lon}, is given as:

$$\sigma_{lon} = (1 + 5tan(\theta))\sigma_{tra} \tag{5.2}$$

Low momentum particles are subjected to higher scattering and are then reconstructed with a worse χ^2.

The angle and position of the base-tracks are initially in the local reference of the single emulsion film. The different layers have to be aligned in order to have all the segments in the reference system of the exposure. The inter-calibration between two emulsion films is achieved by dividing the latter in several cells and carrying out relative shifts between them. The correct translations are those that maximize the number of coincidences between the base-tracks. The alignment between two plates is described by an affine transformation defined as:

$$\begin{pmatrix} x_M \\ y_M \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_f \\ y_f \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \tag{5.3}$$

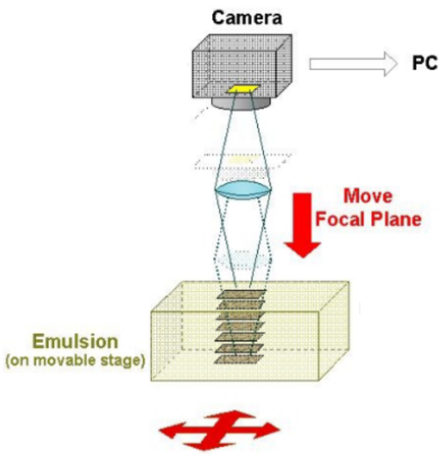


Figure 5.1: The readout: several images are taken at different focal depths.

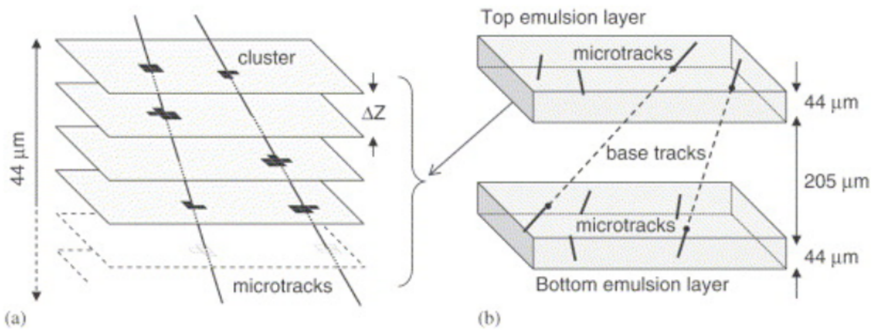


Figure 5.2: Right: reconstruction of the micro-tracks in a layer of emulsion by combining clusters belonging to different levels. Left: the connection of the micro-tracks through the plastic base forms the base-tracks.

where $x(y)_M$ are the aligned coordinates, i.e. in the global reference system and $x(y)_t$ are the tracks coordinates in the single emulsion film. The other parameters are determined from the data of cosmic rays and penetrating particles.

Once all the films have been aligned, the base tracks are merged to form the volume tracks with a process named *tracking*. The first step is the identification of sequences of consecutive base-tracks. They are the input to a Kalman filtering algorithm [59] which performs the linear fit and follows the propagation inside the brick, setting the maximum number of consecutive missing segments accepted.

5.2 Simulation of RUN3 of the DESY test beam

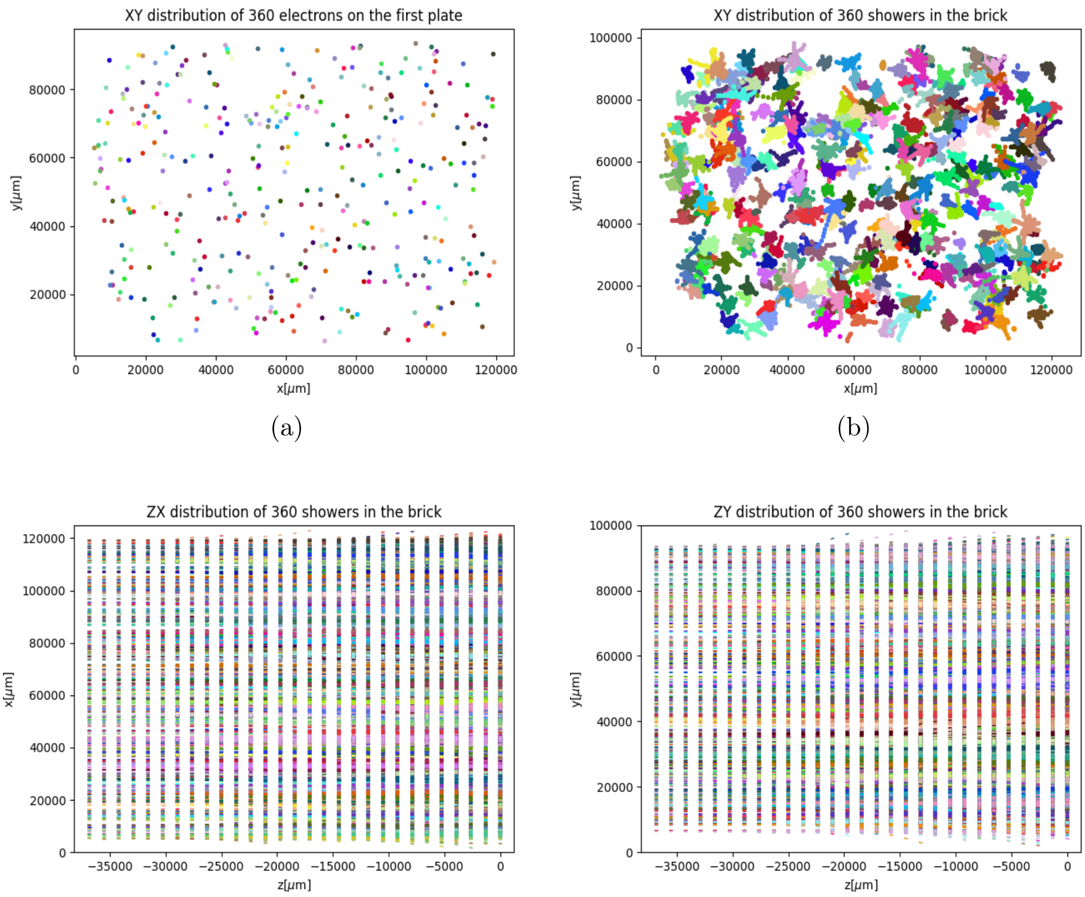
In order to characterize signal and background a full simulation of one of the runs performed at the DESY test beam was performed. RUN3 was selected, being made by 28 lead plates and 29 emulsion films and exposed to 6 GeV electrons.

5.2.1 Signal

The Geant4 software has been used to simulate the electrons propagation and interactions in the apparatus. For the RUN3, 360 electrons with the energy of 6 GeV are fired orthogonally to the target surface and evenly distributed on the emulsion surface, see Figure 5.3a. The number of the electrons is set to 360 to reproduce the same density expected for the SHiP experiment. Each electron produces an electromagnetic shower which propagates in the detector, as shown in Figure 5.4. Hits released in the emulsion films are accepted if they satisfy the visibility cuts, see equation 4.10.

All the information of the segments are collected in a data-frame characterised by the following variables:

- ID: identification number associated to each segment;
- PID: number of the layer where the segment is. PID= 28 is the most upstream film and PID= 0 the most downstream;
- x, y, z: coordinates of the hit. z= 0 correspondes to the most downstream film;
- TX, TY: tangents of the segments's slopes;
- P: a momentum of the particles
- MCEvent: identification number of the simulated event.



(a) (b) (c) (d)

Figure 5.3: (a)XY distribution of all 360 electrons on the most upstream layer. Each electron is identified with a different color. The propagation of the showers in the ECC seen in the transverse plane - xy (b), in the zx (c) and in the zy plane (d).

The number and position of simulated events generates a rather uniform distribution of segments in ECC brick. The 360 showers produce a great number of the base-tracks that occupy the entire volume of the detector with segments of the shower overlapping each other, see Figure 5.3b, 5.3c, 5.3d. The number of hits per shower varies from a minimum of 20 segments to a maximum of 368 segments, see Figure 5.5. The average of segments is 224±3 and the number of these segments in emulsion films increases with the depth in the detector reaching the maximum in film 21, see Figure 5.6.

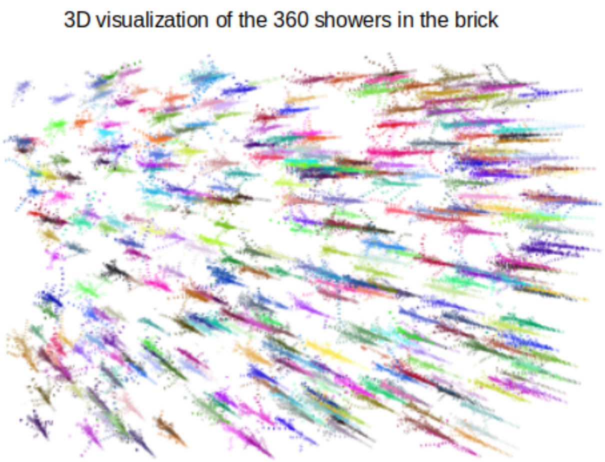


Figure 5.4: 3D visualization of electromagnetic showers simulated in the ECC.

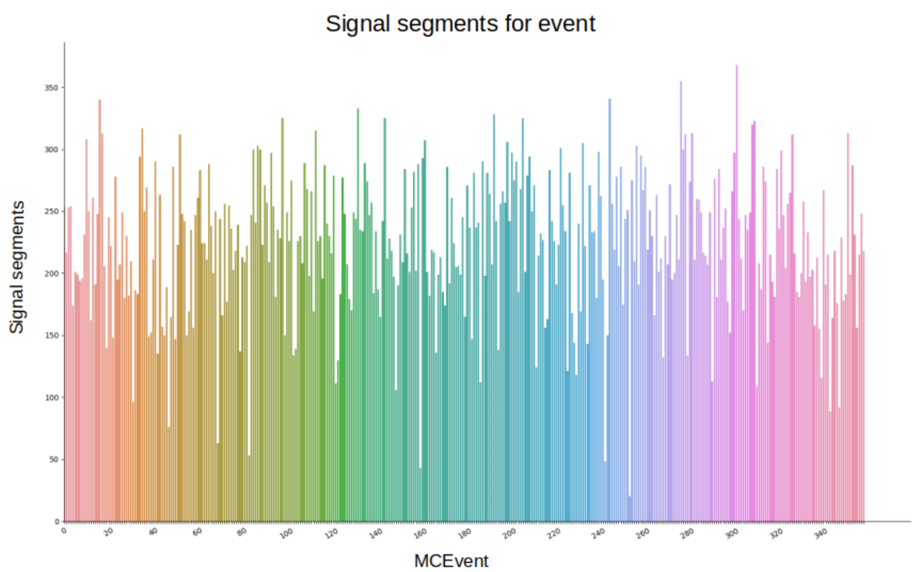
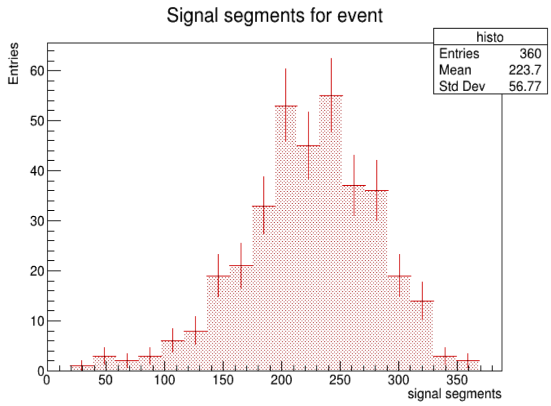
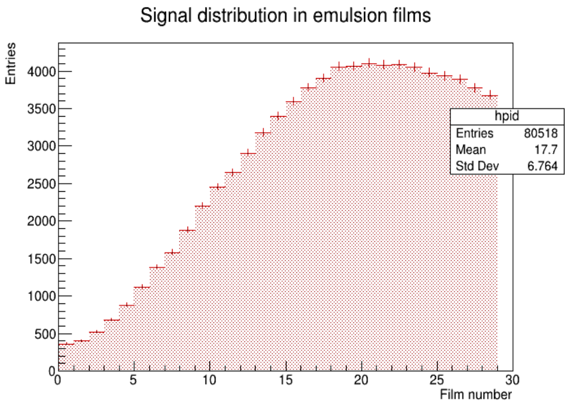


Figure 5.5: Number of segments associated to each simulated event.



(a)



(b)

Figure 5.6: (a)Signal segments per event, (b)Number of segments in the emulsion films.

Shower opening angle

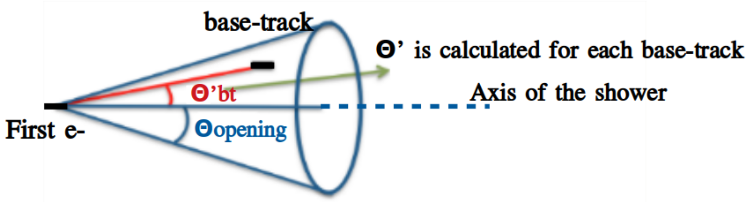
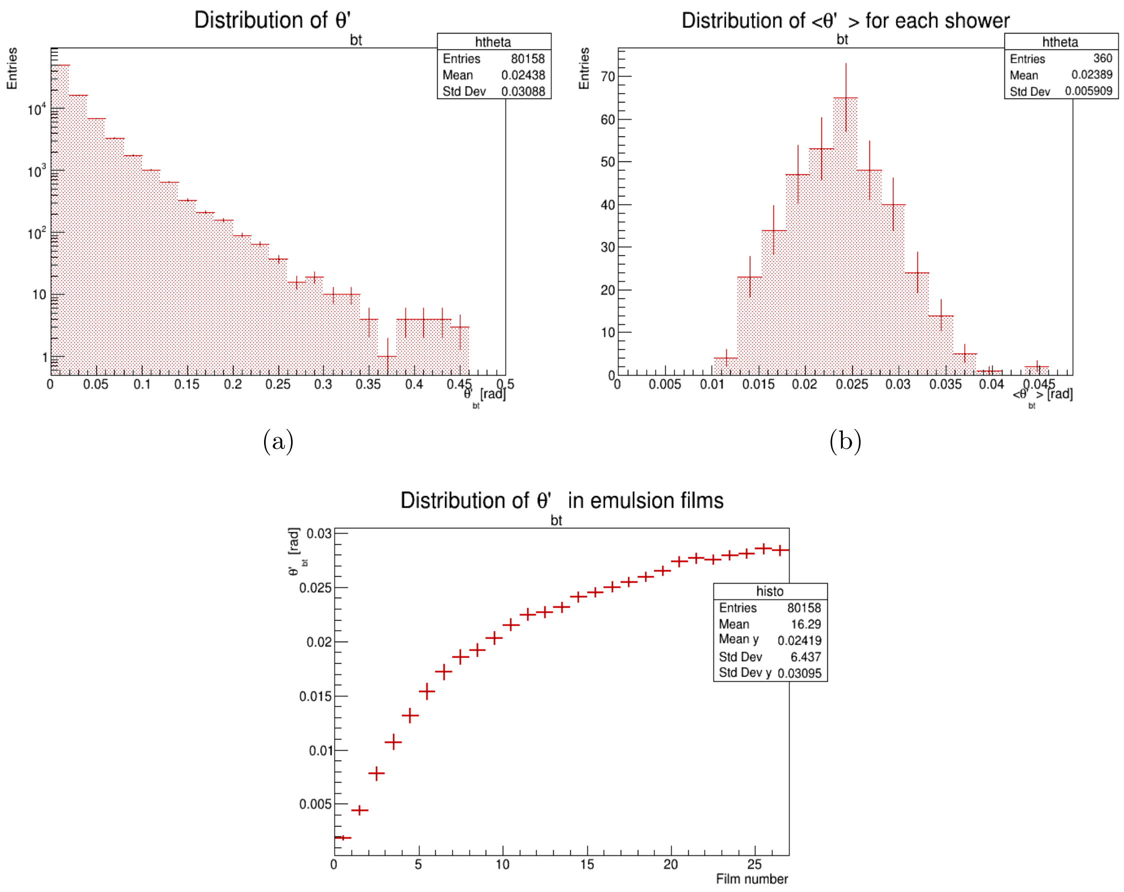


Figure 5.7: The first electron defines the axis of the shower. θ'_{bt} (in red) is calculated for each base-track inside the cone. $\theta_{opening}$ (in blue) is the opening angle of the shower.

A cone can be opened from the first segments of the shower. Defined the axis of the cone for a given event, each base-track inside the cone makes an angle (θ'_{bt}) with respect to the axis of the cone by connecting the base-track with the vertex, see Figure 5.7. The distribution of θ'_{bt}, is shown in Figure 5.8a. The mean value of θ'_{bt} for each shower is plotted in Figure 5.8b, obtaining an average value of about (23.9 ± 0.3)mrad.

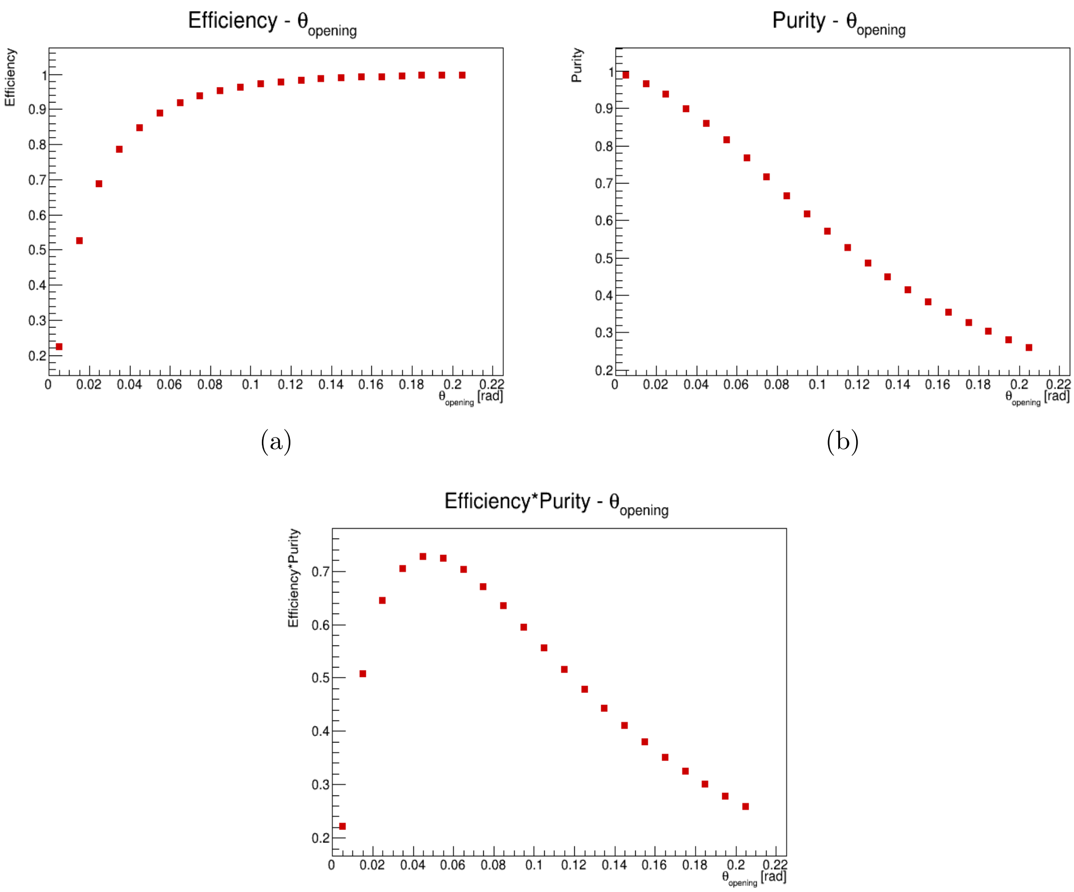
The 360 electrons impinge orthogonally on target surface with an angle close to zero (TX=TY=0). Segments in the most upstream films have a small angle, while



(a) (b) (c)

Figure 5.8: (a) θ'_{bt} distribution for all events. (b)The distribution of $< \theta'_{bt} >$ for all showers. (c) Profile histogram of the θ'_{bt} as a function of the film number.

when particles lose energy, they scatter the most and the segments are detected at a large angle. θ'_{bt} is therefore, expected to increase with the depth in the detector, see Figure 5.8c.
It is obvious that the number of base-tracks inside the cone is dependent on the angle of the cone, see Figure 5.7 where $\theta_{opening}$ is shown in blue. For a given $\theta_{opening}$, all the segments with $\theta'_{bt} \leq \theta_{opening}$ can be collected. An estimation of the efficiency and the purity has been obtained as shown in Figure 5.9. The efficiency is the ratio between segments of the same event satisfying the selection and the total number of segments of the event, i.e. the fraction of shower segments correctly associated to the shower. The purity is defined as the fraction of selected segments that come



(a) (b) (c)

Figure 5.9: (a)Efficiency as a function of the shower opening angle. (b)Purity as a function of the shower opening angle. (c) Product of efficiency and purity as a function of the shower opening angle.

from the electron inducing the shower.
The two metrics have opposite behaviours: when $\theta_{opening}$ is small, the efficiency is low while the purity is high. In other words, in a small cone the number of selected segments of the event is small but we can be confident that the contamination from other events is small. The best value for the opening angle is the one that maximises the product $Efficiency * Purity$, Figure 5.9c. For $\theta_{opening} = 50$ mrad, $Efficiency * Purity = (73.6 \pm 0.2)\%$.

The starting point of the shower is defined as the most downstream film where only

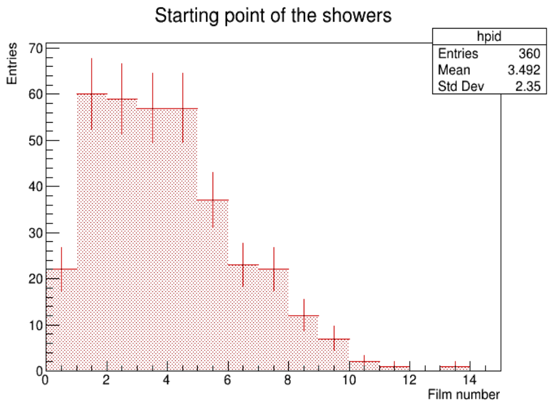


Figure 5.10: Distribution of the most downstream plate where only one particle is seen. Most of the electrons convert between the 3rd and 4th plate.

one particle (base-track in the emulsion) passes through, it is shown in Figure 5.10.

5.2.2 Background

Background segments are largely dominant in data with respect to the signal. There are two main background sources:

- the physical background is due to cosmic rays collected by the detector from the moment they were produced until development. This background affects the base-tracks but not the volume tracks because the order of the emulsions during transport is different from that assumed in the exposure configuration;
- the instrumental background is mainly related to uncorrelated grains that are produced randomly inside emulsion films. These grains create fake segments during the base-tracks reconstruction process.

To simulate the noise, two most upstream films of the RUN1 and RUN3 data taken during the DESY test beam have been used. The number of signal tracks in the most upstream plates, indeed, is totally negligible with respect to background tracks, see Figure 5.6b. The number of films available are duplicated from the initial 4 to 29. At least 8 distinct transformations are required $(7 + 7 + 7 + 8)$, using a one-dimensional reflections:

$$x' = 125000 - x, \quad y' = 110000 - y, \quad TX' = -TX, \quad TY' = -TY \tag{5.4}$$

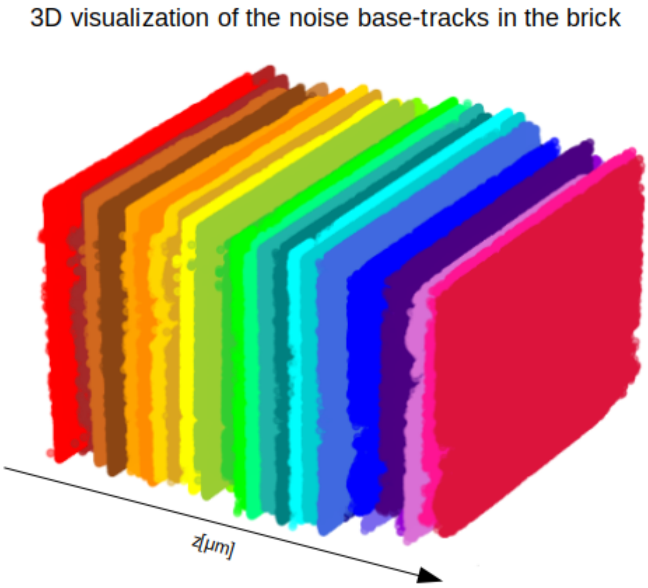
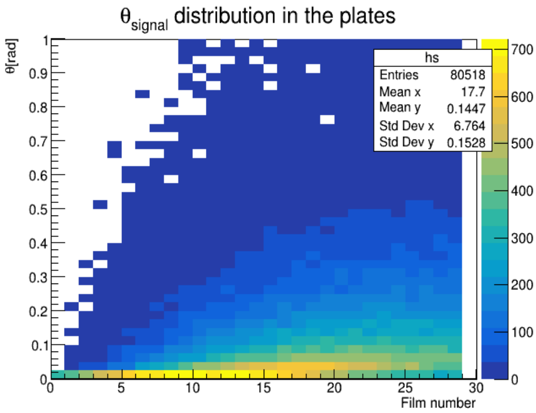


Figure 5.11: 3D visualization of the noise base-tracks in the brick. Each plate has a different color.

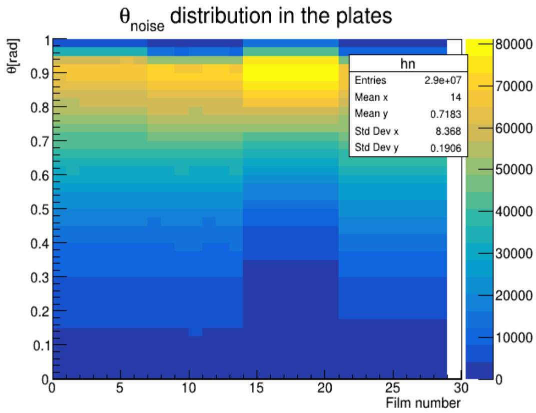
The final noise dataset is obtained in such a way that in each plate there is the expected number of background segments, amounting to 10^6, as measured in data. Figure 5.11 is the 3D visualization of the noise base-tracks in the brick. To distinguish one film to the another, different colours are used.

5.2.3 Signal to noise discrimination

When an electromagnetic shower propagates inside the ECC brick, the energy estimation of its shower will be done by counting the number of produced base-tracks inside the electromagnetic shower. Therefore, the rejection of background is fundamental in this analysis.
The slope distribution of signal and noise segments can help in the discrimination but only in the most upstream plates, where the base-tracks of the two classes are rather different, Figure 5.12. Indeed, signal tracks impinge orthogonally to emulsion films and have rather small angles while cosmic-rays are mostly parallel to the surface given their inclination during the transportation.



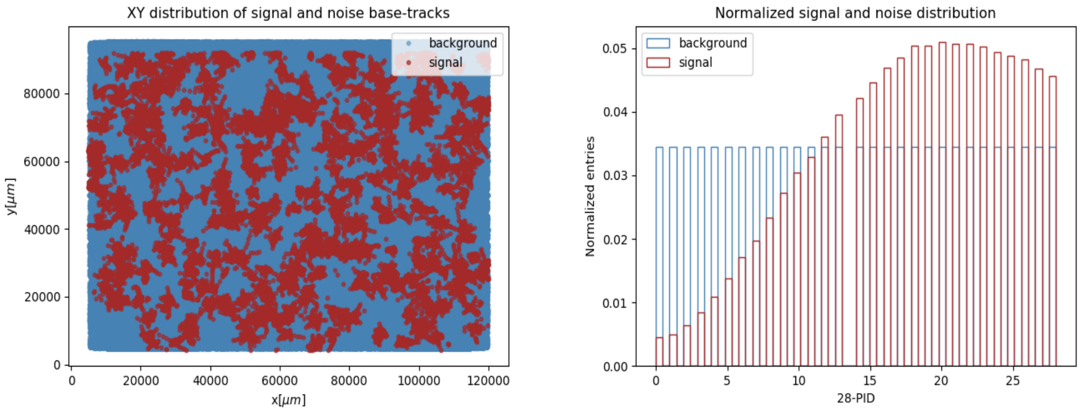
(a)



(b)

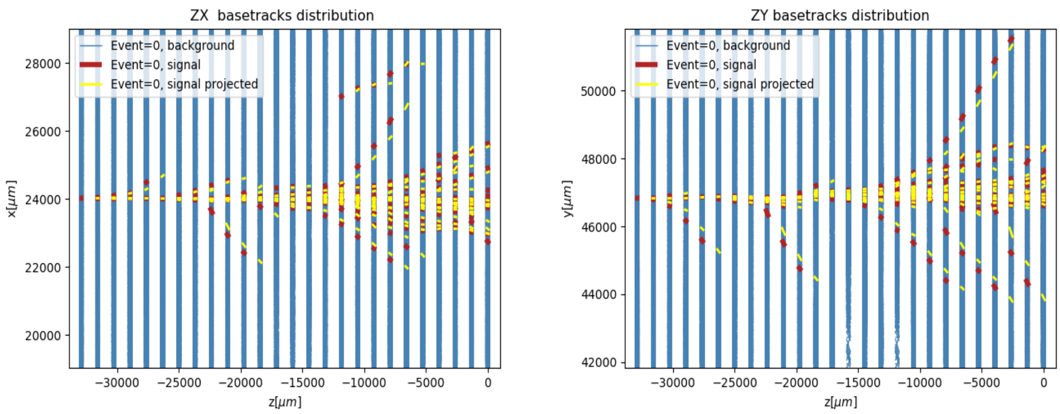
Figure 5.12: Slope distributions of the signal(a) and background(b) base-tracks in the emulsion films. The differences in slope between the two classes in the first 10 films can be used to select the injectors of the showers.

Therefore, in the 10 most upstream plates this difference can be used to discriminate between signal and background. In the downstream plates, instead, the data-set is highly imbalanced with a signal to noise ratio of about 1 to 483 - Figure 5.13a emphasizes that the signal segments are overwhelmed by the noise, Figure 5.13b shows the distribution of the two categories in each film, normalized to unity.
The noise is constant while the signal increases with the depth in the detector.
From the analysis performed, we can conclude that the discrimination between signal and background cannot be based on the single segment, i.e. the angle and position are not enough to discriminate one class from the another. There is a need to introduce new variables based on the physical knowledge we have of the problem.
We know that the noise has no geometric structure while the signal forms a shower.
We expect that from one plate to the others the correlation between signal segments is more powerful than for noise segments. Our idea is to use the projections of segments from one film to the downstream one. The difference in x, y, TX and TY between the projections of the signal segments and those found in the next film could significantly discriminate the signal from background segments, see Figure 5.14. The problem of an imbalanced data-set and the introduction of new variables is addressed in four steps where geometrical parameters are introduced to reconstruct the electromagnetic shower in the brick.



(a) (b)

Figure 5.13: (a)XY distribution of all signal and noise segments in the brick. Signal is in red while noise in blue. (b)Signal and noise distribution in the brick normalized at the unity. The noise is uniform - there is 1million of base-track per film while the signal increases with the depth in the detector.



(a) (b)

Figure 5.14: zx (a) and zy (b) profiles of a single shower on the ECC. Signal segments are reported in red while background in blue. The projection of each signal segment in the downstream film is represented as yellow line.

Step1

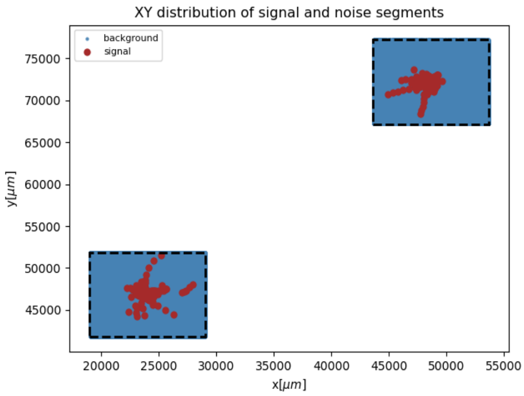
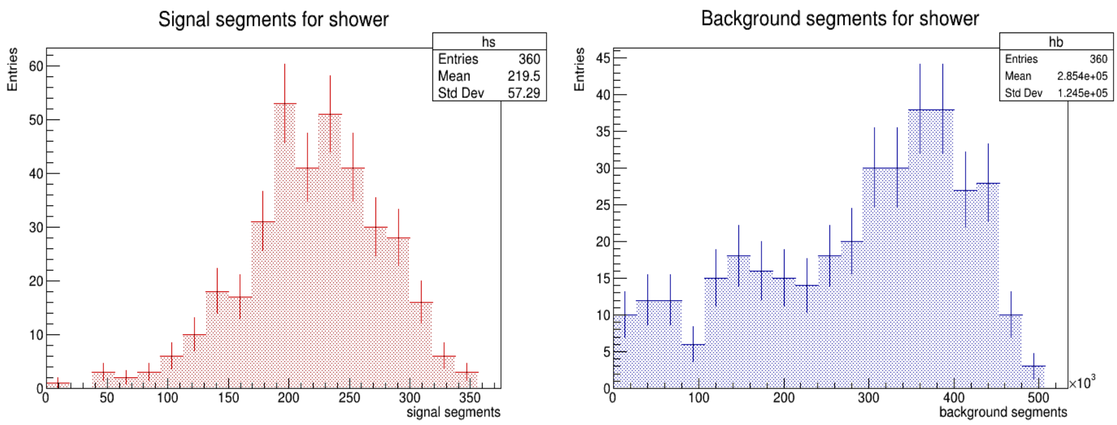


Figure 5.15: XY distribution of the noise and signal segments collected in a square with a side of 1 cm made for all emulsion films. It is shown for two different events.

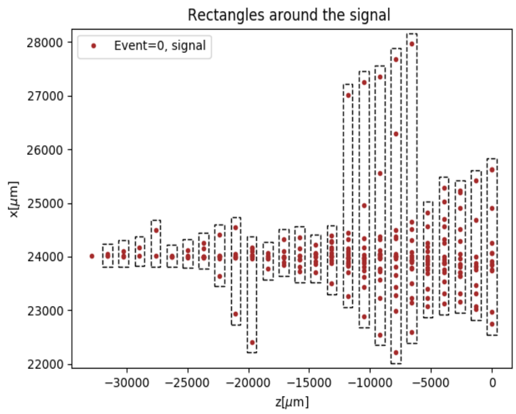


(a) (b)

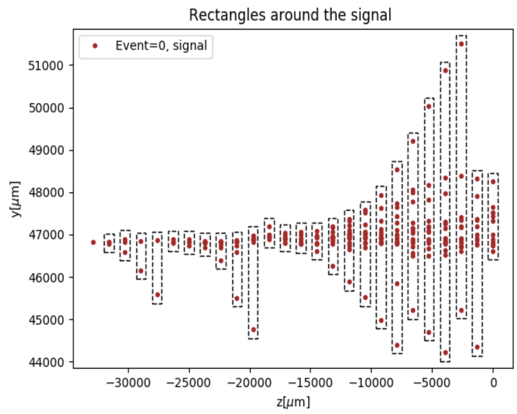
Figure 5.16: For each shower, distribution of signal (a) and noise (b) segments collected for each film in a square of a side of 1 cm.

A square with a side of 1 cm is constructed around the center of each shower for all the emulsion films and all the noise and signal segments in this volume are collected, see Figure 5.15. The square is large enough to contain the signal with an efficiency of $(99.70 \pm 0.02)\%$.

Step 2



(a)



(b)

Figure 5.17: (a)ZX and (b)ZY distribution of the signal segments for a single event with rectangles constructed around the signal with $x = (x_{max} - x_{min} + 500\mu\text{m})$ and $y = (y_{max} - y_{min} + 500\mu\text{m})$.

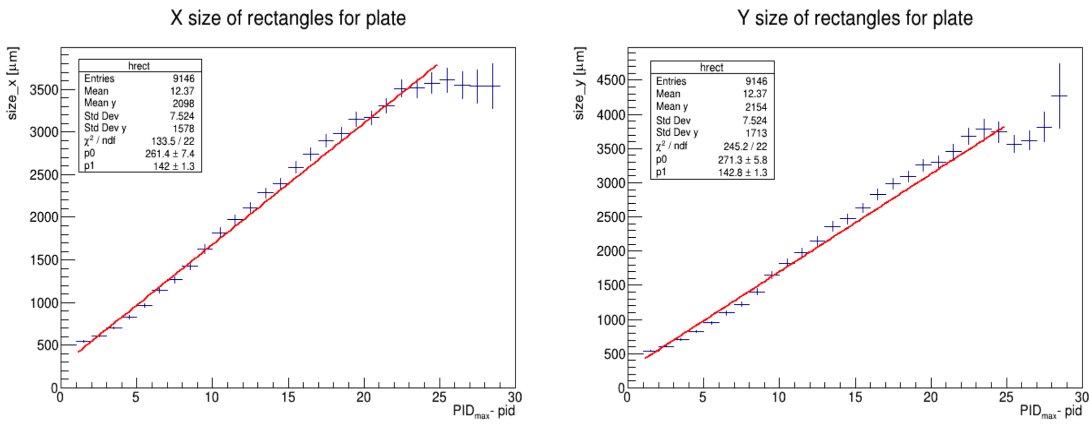
For this step, we start to consider only the signal. From the initial point of the shower defined above, a pyramid with a rectangular section is built to contain the shower development. The size of the rectangle is defined as:

$$\begin{aligned} x &= (x_{max} - x_{min}) + par \\ y &= (y_{max} - y_{min}) + par \end{aligned} \tag{5.5}$$

where x_{max}, y_{max} and x_{min}, y_{min} are the maximum and minimum position of signal segments in that rectangle and par is initially equal to 500μm, as the minimum value when only one segment is recorded, see Figure 5.17. The size of the rectangles depend on the film because the number of the signal segments increases with the depth in the brick. The relationship between the two variables are:

$$\begin{aligned} x &= p1_x \times (PID_{max} - pid) + p0_x \\ y &= p1_y \times (PID_{max} - pid) + p0_y \end{aligned} \tag{5.6}$$

PID_{max} is the film where the shower starts and pid varies between PID_{max} and PID_{min} that is the last film where there are the segments. They are different for each shower.



(a) (b)

Figure 5.18: Profile histogram of x and y dimensions (a-b) of the rectangles created for each plate fixing a shower. A linear fit is conducted to obtain a values of p0 and p1 parameters.

To find the parameters $p0$ and $p1$ a linear fit is performed, see Figure 5.18. Since not all the events reach the bottom of the detector, the fit is made in a linear region with $PID_{max} - pid = [1, 25]$. For x dimension the fit returns: $p0_x = (261.5 \pm 7.4)[\mu m]$ and $p1_x = (141.9 \pm 1.3)[\mu m]$ and for y dmension we have $p0_y = (271.5 \pm 5.8)[\mu m]$ and $p1_y = (142.8 \pm 1.4)[\mu m]$.
These are the first values of the rectangle size. Since $p0_x$ and $p0_y$, $p1_x$ and $p1_y$ are consistent within the errors we can put $p1_x = p1_y$ and $p0_x = p0_y$ and assume a square section of the pyramid.

From the start point of the shower, the squares with increasing dimensions are built. These values are varied in a range for $p0_{[x,y]}$ from $20\mu m$ to $200\mu m$ and for $p1_{[x,y]}$ from $0\mu m$ to $1000\mu m$, in order to obtain the combination maximising the product of efficiency and purity.
The efficiency and the purity for differernt p0 and p1 values are reported in 5.19a and 5.19b. The product $Efficiency * Purity$ is used to choose the values of p0 and p1 that maximize both metrics, see Figure 5.19c. For each p0 and p1, the number of signal and noise inside the rectangles is shown in Figures 5.19d and 5.19e.

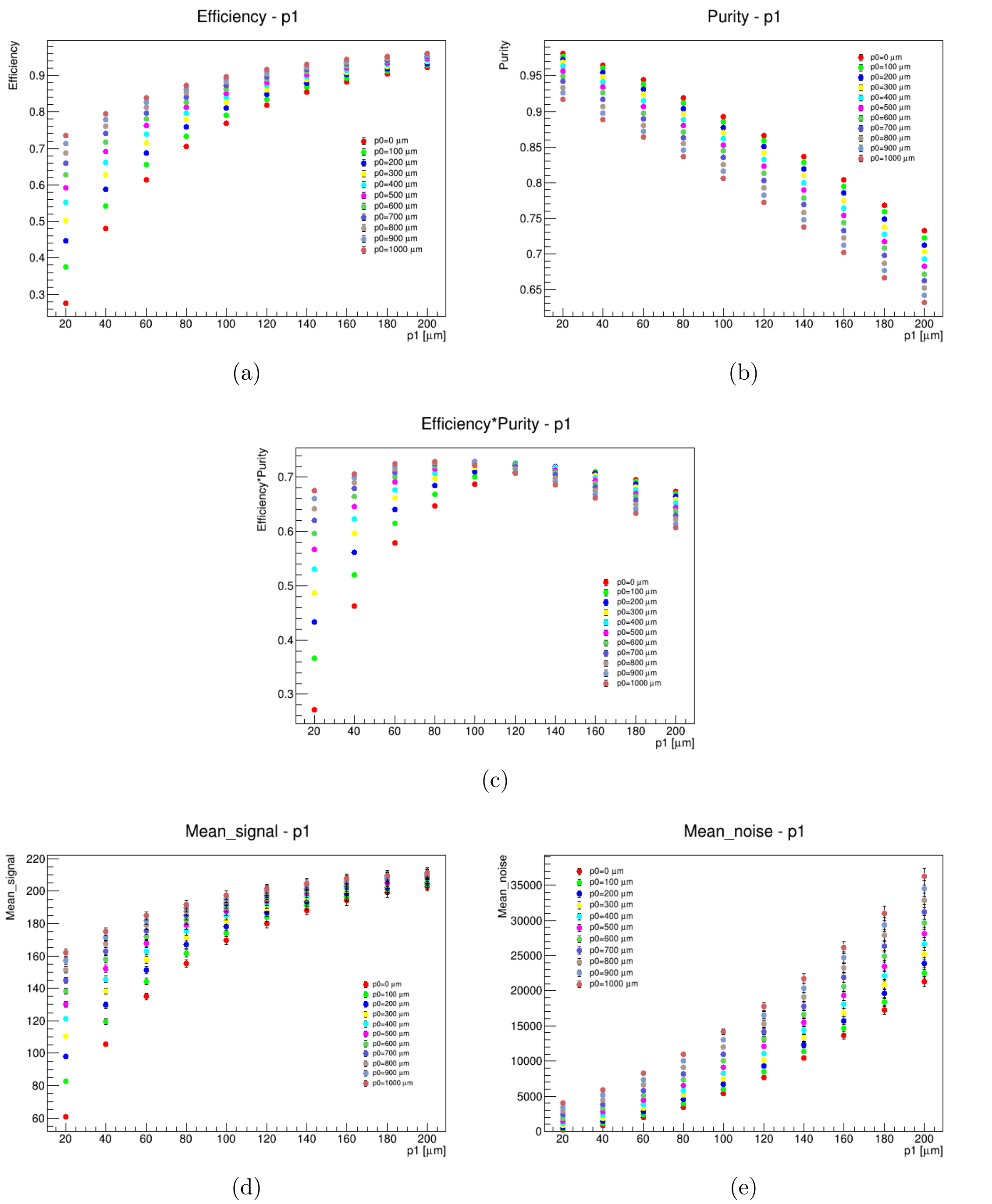
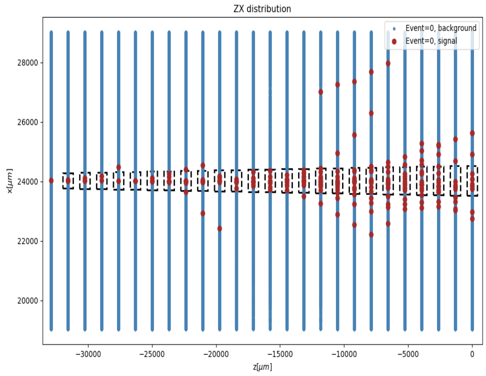
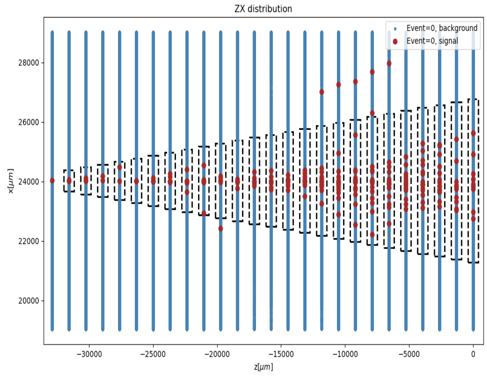


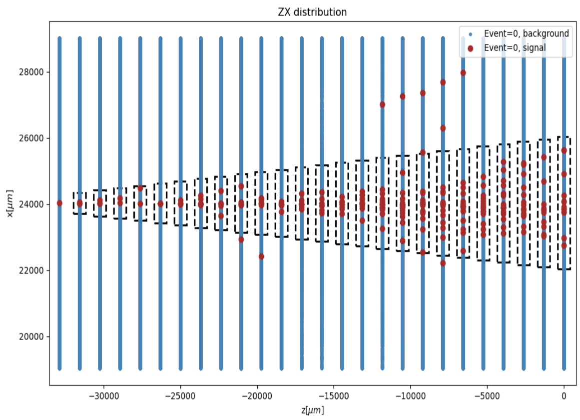
Figure 5.19: P0 varies from 20μm to 200μm and p1 from 0μm to 1000μm. For each combination of these two values, the efficiency, the purity and the product of the two metrics is shown (a-b-c) and the average number of signal and noise segments collected (c-d).



(a)



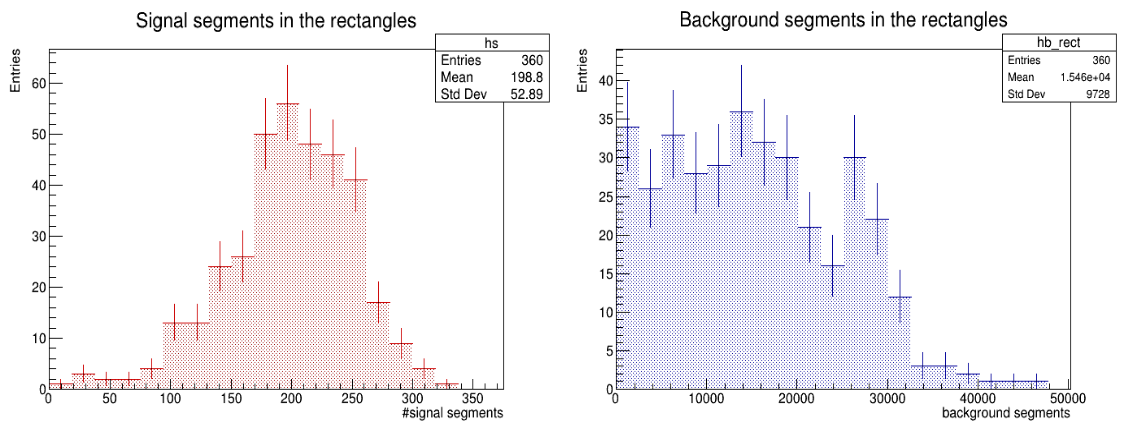
(b)



(c)

Figure 5.20: ZX distribution of the signal (in red) and noise (in blue) for one event in the increasing squares, with p0 = 20μm (a), 200μm (b) and 140μm (c) and p1 = 500μm in each case.

A display of the zx distribution of signal and noise base-tracks- for a single event- with the minimum and maximum size of rectangles is shown in the first two plots of Figure 5.20. The last display shows the zx distribution of the base-tracks for a given value of p0 = 140μm. For each plot p1 is fixed to 500μm. Choosing p0 = 140μm and p1 = 500μm the $Efficiency$ is $(90.2 \pm 0.2)\%$, $Purity$ is $(78.9 \pm 0.1)\%$ and $Efficiency * Purity$ is $(71.2 \pm 0.2)\%$. The average number of selected signal and noise segments is 198.9 ± 2.8 and $(1.54 \pm 0.05) \times 10^4$ respectively, see Figure 5.21.



(a) (b)

Figure 5.21: The distribution of the signal (a) and noise (b) segments in the square-based pyramid.

Step 3

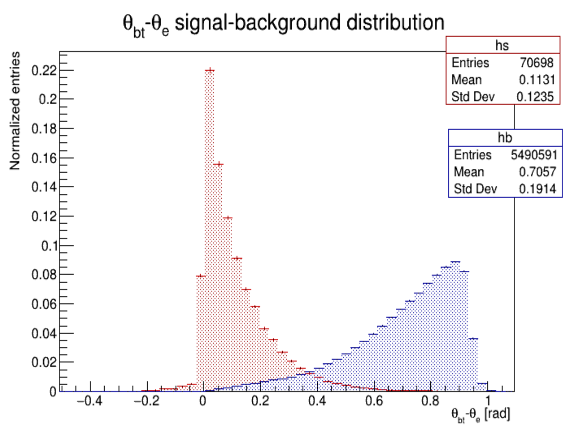
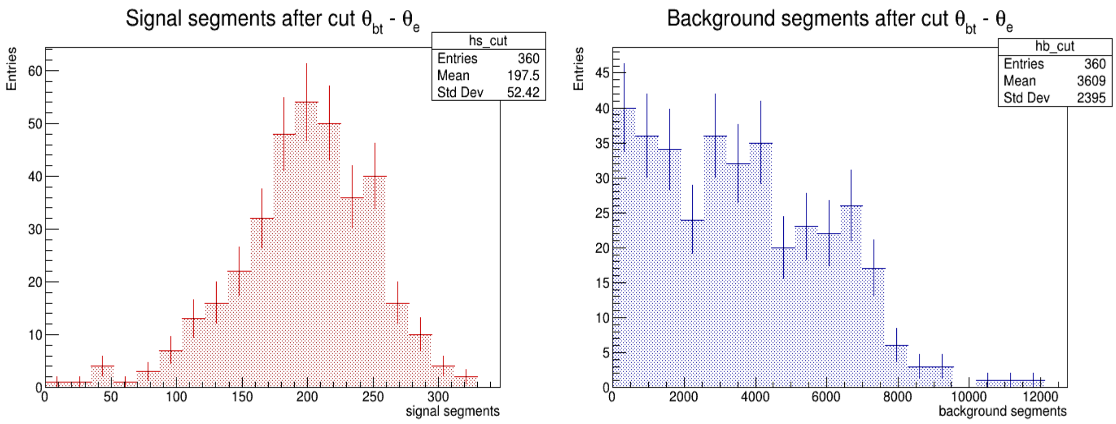


Figure 5.22: $\theta_{bt} - \theta_e$ distribution normalized to unity. A cut on $(\theta_{bt} - \theta_e) \leq 0.6$ rad has been made to reject most of the noise.

A study on the introduction of a new variable is performed. It is the difference between the angle of each base-track and the angle of the axis of the shower. This new variable, $\theta_{bt} - \theta_e$ is very different for signal and noise segments, see Figure 5.22. A cut on $(\theta_{bt} - \theta_e) \leq 0.6$ rad allows to eliminate a lot of noise, see Figure 5.23 to evaluate the number of signal and noise segments collected after the cut.



(a) (b)

Figure 5.23: Distribution of signal (a) and noise (b) segments after the cut in $\theta_{bt} - \theta_e$.

Step 4

In this last step, we use the information of segments projected from one plate to the next. Our starting idea is based on the analyses that the differences between the component - x, y, TX an TY as well as the distance dR and dθ - of the segments projected and those found in the next plate would be considerably different for signal and noise segments, see Figure 5.24.
The new analysis is performed to search the k segments "nearest" to the projections. It is made into three steps:

- starting from the injector of the shower we make the projection of this segment from one film to the downstream one;
- the differences in x, y, TX and TY and the distance dR and dθ between this projection and all segments found in the downstream film are performed;
- a condition on the d$\theta \leq 0.4$ rad is imposed to perform a selection of the base-tracks.

The difference between the signal and noise segments is especially evident in Figure 5.24f, where the dθ is different for noise and signal. For this purpose, a cut on d$\theta \leq 0.4$ rad is applied discarding much of the noise.

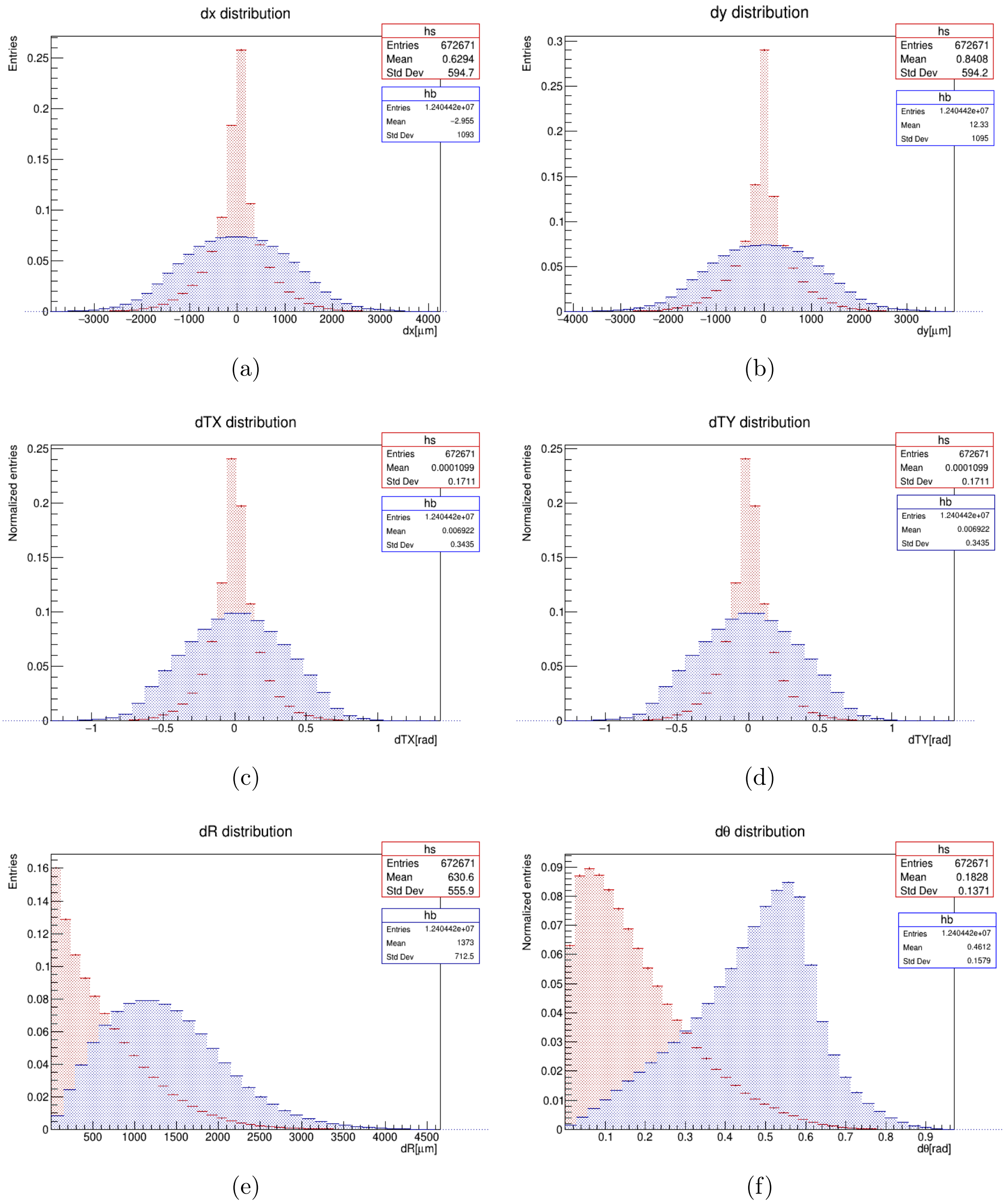
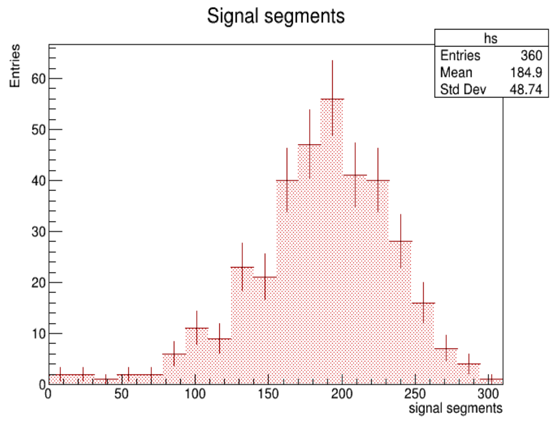
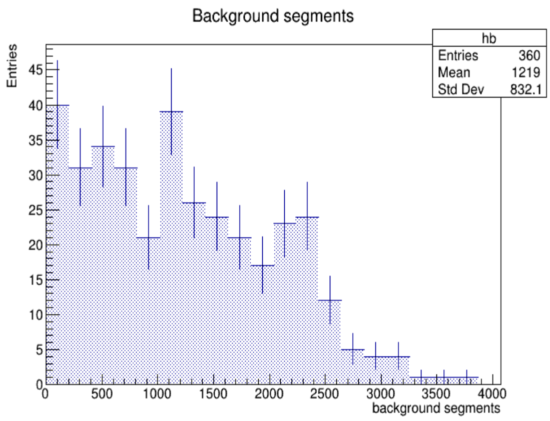


Figure 5.24: The distributions of the differences - dx, dy, dTX, dTY and the distances dR and dθ made between the signal projected segments in the downstream film with the signal (in red) and noise (in blue) segments in that film.

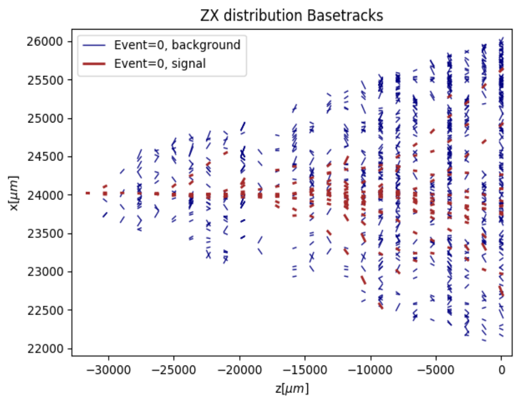


(a)

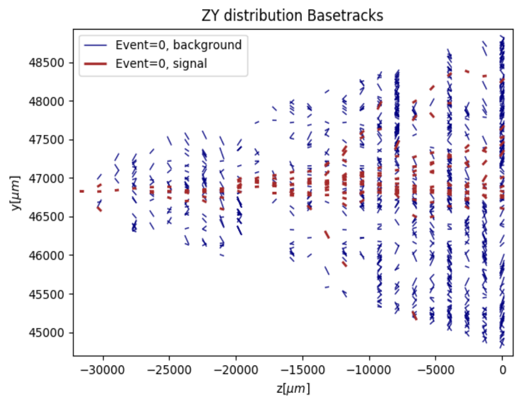


(b)

Figure 5.25: Signal (a) and noise (b) segments for each event.



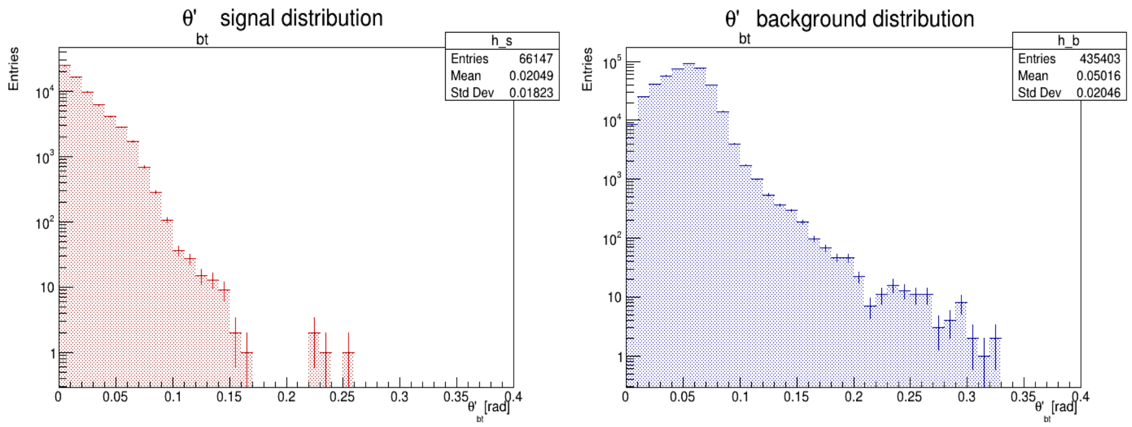
(a)



(b)

Figure 5.26: ZX(a) and ZY(b) distribution of the signal and background (in red and in blue) after the four-steps procedure.

Once the segments on the downstream film are selected with this procedure, they are projected into the next film and the four steps are applied again. The algorithm for selection of the nearest segments is an iterative process that ends when it arrives at the most downstream film where there are shower segments. The distribution of all signal and noise segments collected for each shower is shown in Figure 5.25 while Figure 5.26 shows a display of all signal and noise segments selected through this procedure for a single event.



(a) (b)

Figure 5.27: The distribution of the θ'_{bt} for the signal (a) and background (b).

After the four-steps procedure the signal efficiency is of about $(83.70 \pm 0.02)\%$. The differences in x, y, TX, and TY are the new variables added to the initial data-set and that will be used in the binary classification. Other features calculated for signal and background base-tracks are:

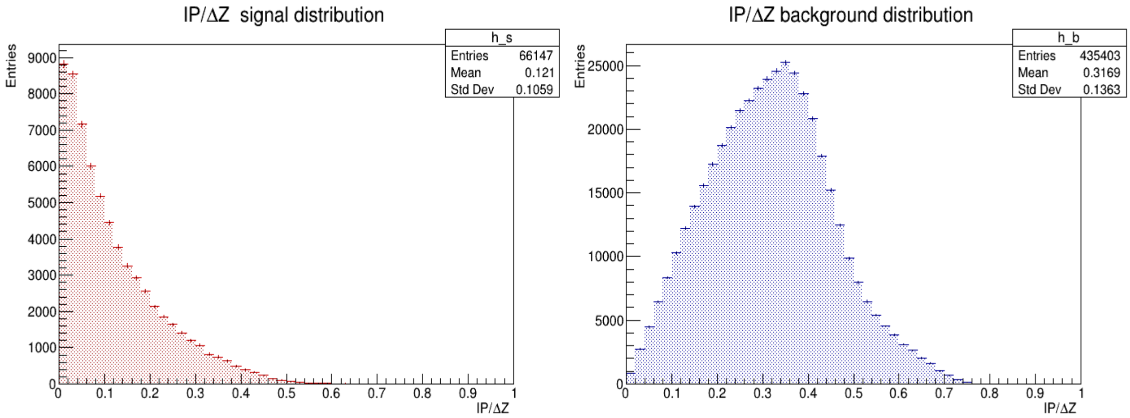
- θ'_{bt}: already defined, see Figure 5.7, as the angle that each base-track forms with the pyramid vertex. The θ'_{bt} distribution for signal and noise is shown in Figure 5.27. The base-tracks produced by electron have higher momentum and therefore they propagate mostly around the direction of the electron until, they lose their energy.

- **Impact parameter divided by ΔZ**: IP is the distance between the vertex position and the projection of each base-track on the transverse plane of vertex, and it is defined as:

$$IP = \sqrt{(X_{ver} - X_{proj})^2 + (Y_{ver} - Y_{proj})^2} \tag{5.7}$$

where X_{ver} (Y_{ver}) is the position of the pyramid vertex and X_{proj} (Y_{proj}) is the extrapolated position of each base-track on the vertex plane, where the shower originates. X_{proj} and Y_{proj} are defined as:

$$X_{proj} = X - (\Delta Z \times S_x), \quad Y_{proj} = Y - (\Delta Z \times S_y) \tag{5.8}$$



(a) (b)

Figure 5.28: The distribution of the impact parameter computed for each base-tracks divided by its distance from the vertex position - ΔZ- for the signal (a) and background (b) class.

where X(Y) is the position of each base-track, ΔZ is the longitudinal distance between the position of related base-track and the cone vertex and S_x (S_y) is the slope of each base-track in the X (Y) projection.
IP is computed for each base-track and it is divided by its distance from the vertex position, ΔZ . In case of signal, base-tracks closer to the vertex have smaller IP. However, the IP for the background base-tracks is independent on the distance from the vertex. Therefore, dividing the IP by ΔZ improves the signal/background separation. The distribution of $IP/\Delta Z$ for the signal and noise class is shown in Figure 5.28.

This chapter is focused on the introduction of new features that highlights the differences of the signal segments from the noise ones. The strategy consists of four steps, through which the noise is localized around each shower, new geometrical parameters are defined to follow the shower development in the brick, a cut on a new variable is made to reject the most of background and the search for signal segments is conducted.
With all the steps, a final dataset shows a signal-to-noise ratio of about 1/7 as shown in Figure 5.28, therefore reducing the noise by a factor of 70 and the introduction of new variables - dx, dy, dTX, dTY, $\theta_{bt} - \theta_e$, θ'_{bt}, $IP/\Delta Z$. These new features will be

added to the initial dataset and will be used to make a binary classification between the signal and noise class. The application of machine learning techniques and in particular of the Random Forest as classifier will be described in detail in the next chapter.

Chapter 6

Electron identification with machine learning techniques

The challenge of electron identification in ECC lies in the selection of segments related to the electromagnetic shower out of a large background, overwhelming the signal by a few order of magnitudes.
The selections on the variables defined in the previous chapter, reduced the background leading signal/noise to about 1/7. By applying these selections, the initial sample with new features has been defined for the analysis with machine learning techniques. In this chapter, the application of the Random Forest to make the binary classification between signal and noise segments will be conducted in order to identify the electromagnetic shower in the brick.

6.1 Choice of the Machine Learning algorithm

Different supervised machine learning models are compared in order to train and select the best performing ones. They are: *support vector machine*, *decision tree*, *random forest*, *adaboost* and the *k-nearest neighbor*, [60] [61]. The initial dataset is divided into two parts: a training and a test. Since we want that the classifier is trained on the whole showers, the samples are made of showers, with 70% in the training and 30% in the test. For each model the *Grid Search CV*[1] is implemented

[1]See Chapter 3 for an exposition of the Machine learning techniques and 3.3.2.3 for the Grid Search CV treatment.

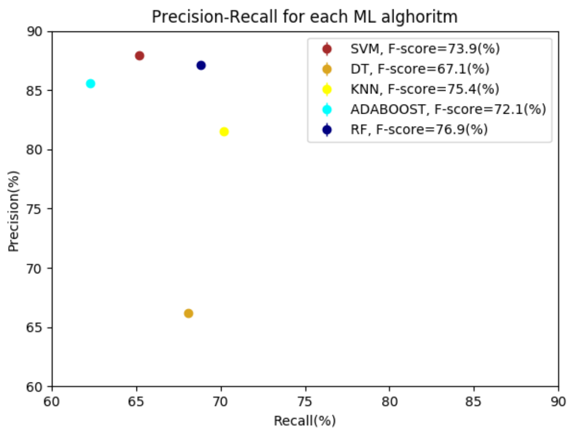


Figure 6.1: Precision-Recall metrics for each machine learning algorithm. In the legend, the f-score is reported for the signal class.

to find the best combination of the hyper-parameters. It is the one that returns the best value of *f-score* [2] metric. Once the best combination of hyper-parameters is found, the choice of the best model is made by comparing the value of the f-score for the signal class on the test data-set for each machine learning algorithm. As we see in Figure 6.1, the Random Forest is the one that returns the highest f-score value amounting to 76.9% with the following hyper-parameters optimized with the Grid Search CV application: *number of estimators = 500*, *maximum depth = 200* and *criterion = entropy*.

6.2 Random Forest performances

Once the Random Forest has been selected as the best model, we have tested its performances in the identification of the electromagnetic shower. The confusion matrix and the classification report [2] are used on purpose. The confusion matrix shows the number of true negative and true positive (TN and TP) on the main diagonal, i.e. the number of the background and signal instances correctly classified and the false positive and false negative, that are the number of the instances mis-classified. The classification report shows the precision, the recall and the f-score

[2] See Section 3.3.2 for metrics's explanation used in this work.

metrics. The precision and the recall corresponds to the purity and the efficiency of the selection respectively. The f-score can be interpreted as a weighted average of the precision and the recall.

RUN3 - dataset 1

At first, we consider only one data-set that is divided into two parts: a training and a test sets with 250 showers in the first set and 110 in the second set. The results of the model's application are shown in Tables 6.1 and 6.2.

Table 6.1: Confusion matrix: 0:0 and 1:1 are background and signal correctly classified while 0:1 and 1:0 are the background and signal mis-classified.

	0	1
0	130238	2034
1	6356	14021

Table 6.2: Classification report: the precision, the recall and the f-score are shown.

	Precision (%)	Recall (%)	F-score (%)
0	95.44	98.46	96.88
1	87.33	68.88	76.91

In the test set, the number of signal segments is 20377. The Random Forest classifies correctly 14021 signal segments while 6356 are the mis-classified ones. They are the true positive (1:1) and false negative (1:0) shown in the confusion matrix, see Table 6.1. The number of background segments, in the test set, is 132272: 130238 are correctly classified and 2034 are the noise mis-classified. They are the true negative (0:0) and false positive (0:1) respectively shown in the Table 6.1.
The confusion matrix shows that the Random Forest recognizes better the noise segments than signal ones - more than a 30% of the signal base-tracks (6356 on a total signal entries of 20377) is lose while only a 2% of the noise is confused as signal. An explanation of this behavior can be found considering that the dataset

is imbalanced, for each signal base-tracks there are 7 noise segments. It means that the classifier can train on a wider noise class and so it tends to recognize it better. As a result, the f-score of the noise class is very high amounting to 96.8%. For the signal class, the f-score is 76.9% with a high purity of the selection 87.3% as shown in Table 6.2.

RUN3 - datasets 2, 3, 4, 5

We created other datasets in RUN3 configuration that differ from the first one only for the simulated signal while the background remains unchanged. The datasets differ for the position of the first electrons on the target surface and the starting point of the showers. For each of dataset the four-steps procedure is conducted to create the set of data with new features. At first, we use one of these dataset to train the model and the others to test it. Then, some of these datasets are combined to make a training or a test set and see how the Random Forest behaves by changing the test datasets or increasing the statistic.

Training RUN3 - dataset 1 RUN3 - dataset 1 is used to train the model and other datasets to test it. The confusion matrix and the classification report are shown in Tables 6.3 and 6.4.

Table 6.3: Confusion matrix: 0:0 and 1:1 are background and signal correctly classified while 0:1 and 1:0 are the background and signal mis-classified.

Test set	Class	0 (Predict)	1 (Predict)
RUN3 - dataset 2 (495915)	0	424245	6599
	1	20202	44869
RUN3 - dataset 3 (485712)	0	414760	6481
	1	20087	44384
RUN3 - dataset 4 (506160)	0	435976	6889
	1	19606	43689
RUN3 - dataset 5 (497784)	0	427059	6300
	1	19867	44558

Table 6.4: Classification report.

Test set	Class	Precision (%)	Recall (%)	F-score (%)	Support
RUN3 - dataset 2	0	95.44	98.48	96.93	430844
	1	87.26	68.88	76.99	65071
RUN3 - dataset 3	0	95.37	98.47	96.90	421241
	1	87.37	68.82	76.99	64471
RUN3 - dataset 4	0	95.69	98.44	97.05	442865
	1	86.37	69.02	76.73	63295
RUN3 - dataset 5	0	95.55	98.54	97.02	433359
	1	87.61	69.16	77.30	64425

The first column of the confusion matrix reports the datasets used to test the Random Forest classifier with the number of the instances for each one. The second column is the class that is 0 for background and 1 for signal. The confusion matrix has to be read separately for each test sets and shows the number of true negative (0:0) and true positive (1:1) instances and the number of signal and noise segments mis-classifed. They are the false negative (1:0) and false positive (0:1). In the same way, the classification report in Table 6.4 shows for each test sets the values of metrics. The last column, *support*, reports the number of noise and signal segments in each set of data.

Training RUN3 - dataset 1+2 RUN3 - dataset 1+2 is used to train the model and the others are used to test it. The training set contains 1003844 instances. The confusion matrix and classification report are shown in Tables 6.5 and 6.6.
For each test sets, shown in the first column of the Tables with their instances and for each class the confusion matrix reports the number of the instances correctly classified (0:0 for background and 1:1 for signal) and number of background and signal instances mis-classified (0:1 and 1:0). The classification report shows the metrics values and the total number of background and signal in test sets.

Table 6.5: Confusion matrix.

Test set	Class	0 (Predict)	1 (Predict)
RUN3 3 (485712)	0	414826	6415
	1	19864	44607
RUN3 4 (506160)	0	435030	7835
	1	18979	44316
RUN3 5 (497784)	0	426212	7147
	1	19137	45288

Table 6.6: Classification report.

Test set	Class	Precision (%)	Recall (%)	F-score (%)	Support
RUN3 3 (485712)	0	95.54	98.29	96.89	421241
	1	86.25	70.03	77.30	64471
RUN3 4 (506160)	0	95.82	98.23	97.01	442865
	1	84.97	69.01	76.77	63295
RUN3 5 (497784)	0	95.70	98.35	97.01	433359
	1	86.36	70.29	77.50	64425

Training RUN3 - dataset 1+2+3 RUN3 - dataset 1+2+3 is used to train the model and others to test it. The training dataset contains 1483177 instances.

Table 6.7: Confusion matrix.

Test set	Class	0 (Predict)	1 (Predict)
RUN3 4 (506160)	0	434347	8518
	1	18507	44788
RUN3 5 (497784)	0	425550	7809
	1	18688	45737
RUN3 4+5 (1003944)	0	859897	16327
	1	37195	90525

Table 6.8: Classification report.

Test set	Class	Precision (%)	Recall (%)	F-score (%)	Support
RUN3 4 (506160)	0	95.91	98.07	96.98	442865
	1	84.02	70.76	76.82	63295
RUN3 5 (497784)	0	95.79	98.19	96.98	433359
	1	85.42	71.00	77.53	64425
RUN3 4+5 (1003944)	0	95.85	98.13	96.98	876224
	1	84.72	70.87	77.18	127720

The confusion matrix and classification report are shown in Tables 6.7 and 6.8. In this last case, the datasets used to test the model are considered separately and also combined, see the first column of the tables. For each test set and for each class, the confusion matrix shows the number of true negative (0:0), true positive (1:1), false positive (0:1) and false negative (1:0) and the classification report shows the metrics values.

All the tests on the performances of the Random Forest classifier show that the values metrics do not change significantly varying the training or the test set, even if the datasets are used separately or combined as training or test. It highlights important properties of the classifier: the *robustness* and the *stability*.

6.2.1 Features analysis

It is interesting to observe the distribution of the features used to make the classification for the four categories shown in the confusion matrix. The RUN3 - dataset 3 with 360 showers is used as test set.
Figures 6.2 and 6.3a show the distribution normalized to unity for the four categories of Table 6.5 related to the RUN3 - dataset 3.
The true positive cases corresponding to signal correctly classified (in red) are concentrated around zero while the true negative, i.e. the noise correctly identifies (in blue) is much wider. The distributions of the mis-classified classes, signal classified as noise (in orange) and noise classified as signal (in azure) behave similarly to the correctly classified of noise and signal respectively.

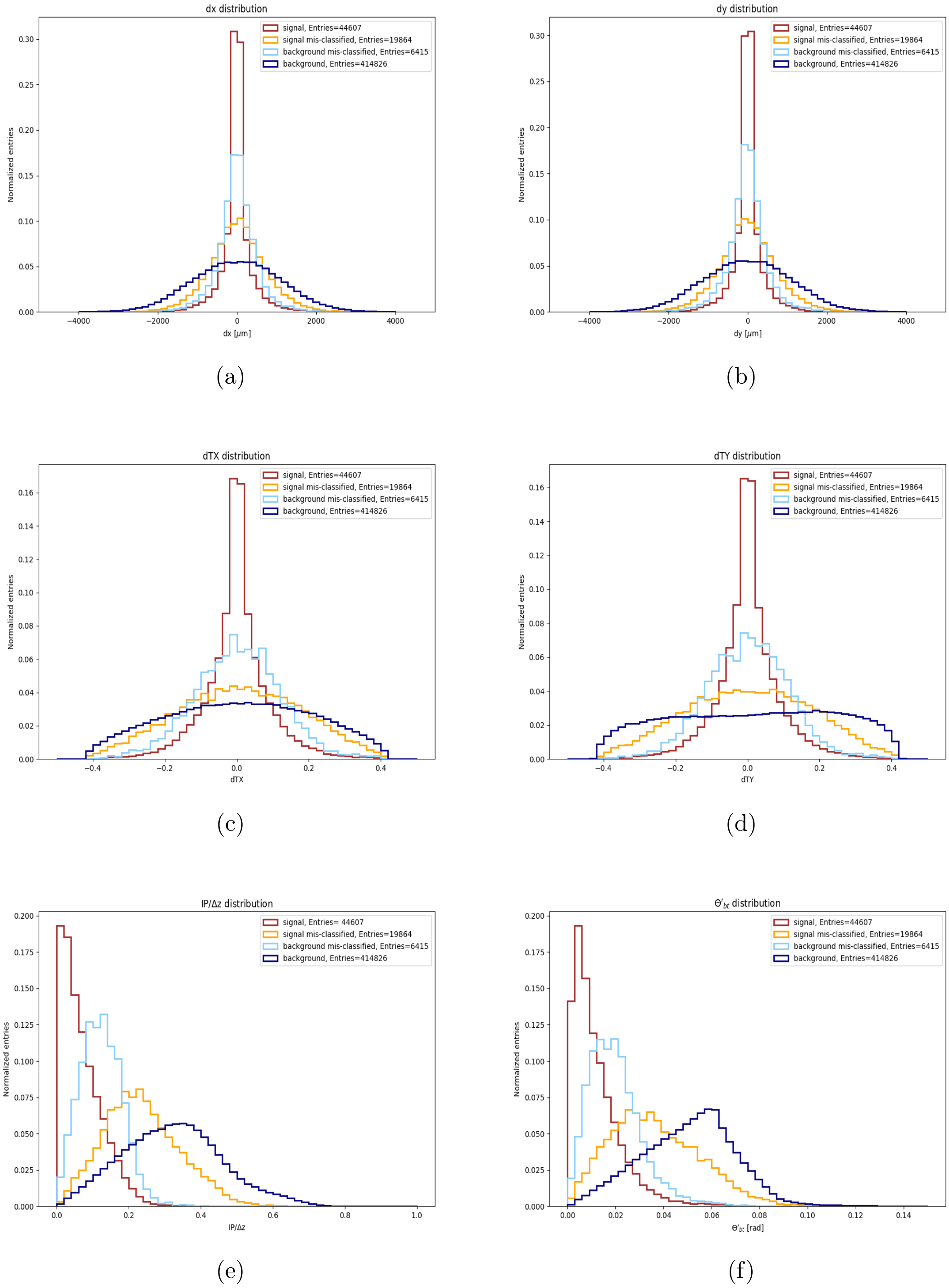
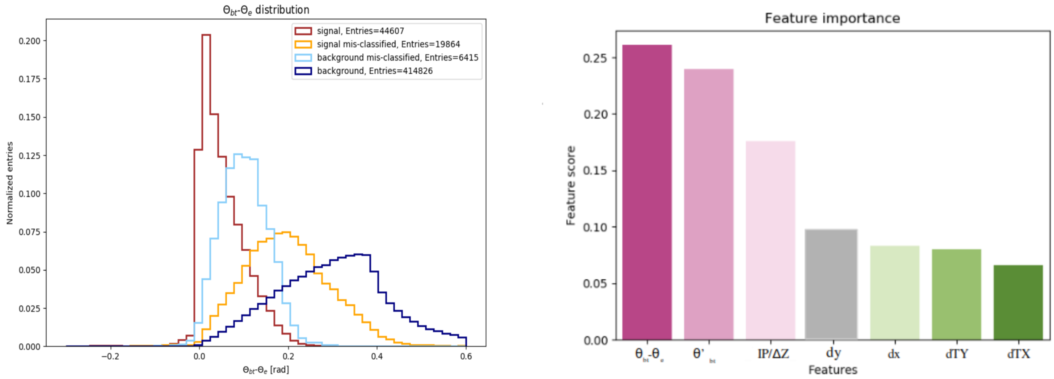


Figure 6.2: dx, dy, dTX, dTY, $IP/\Delta Z$ and θ'_{bt} distributions for signal and background correctly classified (red and blue) and mis-classified (orange and azure).



(a) (b)

Figure 6.3: (a) The $\theta_{bt} - \theta_e$ distribution- normalized to unity for the four categories. (b) The scores importance for each feature: $\theta_{bt} - \theta_e$ is the one with the highest score.

The right panel of Figure 6.3b shows the features importance: for each feature it is reported how much it weighs in the classification. The $\theta_{bt} - \theta_e$ is the most important one.

For each of 360 showers in the RUN3 - dataset 3 used as test set, the number of signal and background segments correctly classified and mis-classified are shown in the distributions in Figure 6.4, each one with different color. We expect to have similarly populated events. Out of the 184.9±2.6 segments per shower in the dataset before the application of the Random Forest - see Figure 5.25a, about 123.9±1.8 are on average the number of correctly classified signal segments while about a third of the signal is lost - 55.2 ± 1.2 are the base-tracks confused as noise segments. On the other hand, only a small part of the noise is classified as a signal, it is about 17.8±0.8 per event, since the most of the background segments are classified correctly.
For each of these showers, the precision, the recall and the f-score metrics are estimated starting from the number of true positive (TP), false positive (FP) and false negative (FN) after the Random Forest application. They are shown in Figures 6.5a, 6.5b, 6.5c. The efficiency is computed as the ratio of signal segments identified by the Random Forest and the total number of signal segments belonging to the shower before the reconstruction and identification procedure. About the $(57.4 \pm 0.5)\%$ of the total segments survive, see Figure 6.5d.

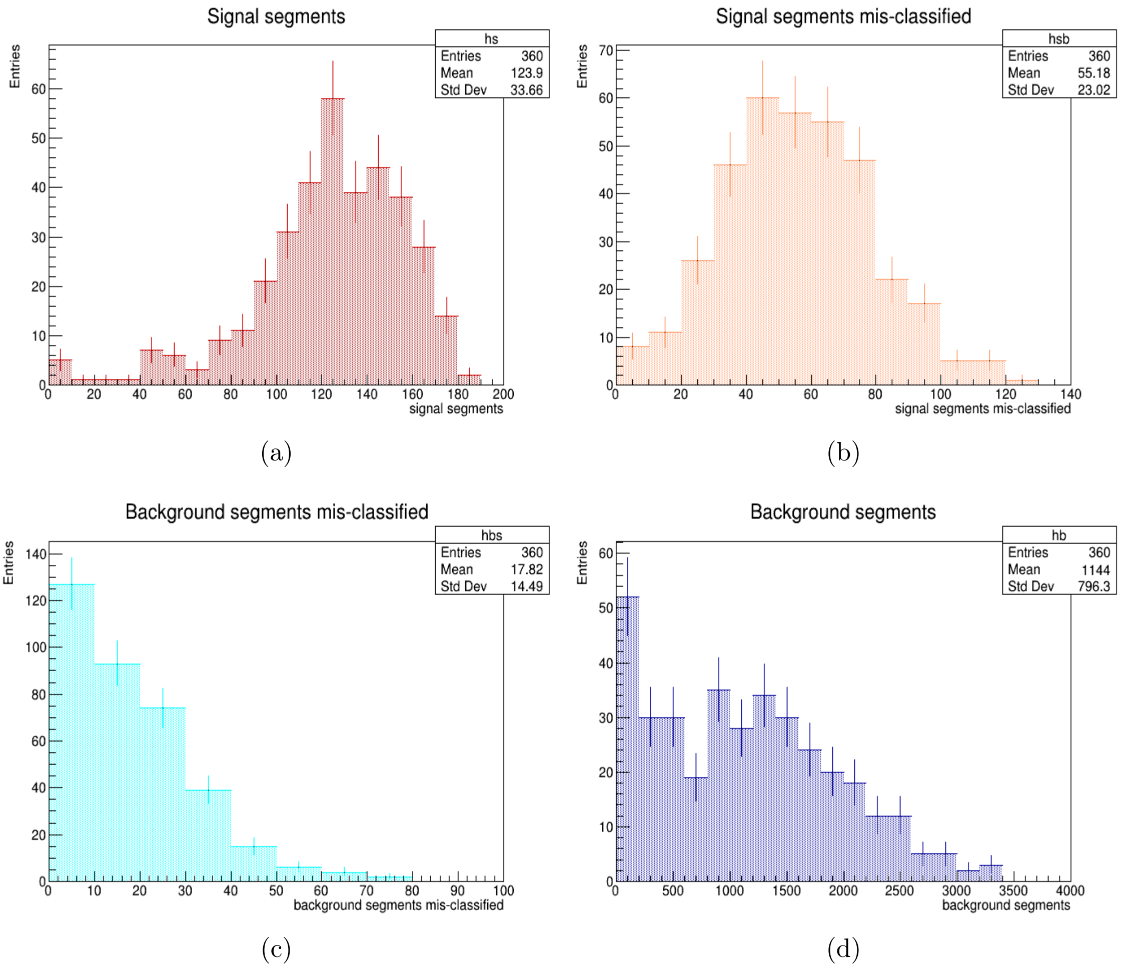
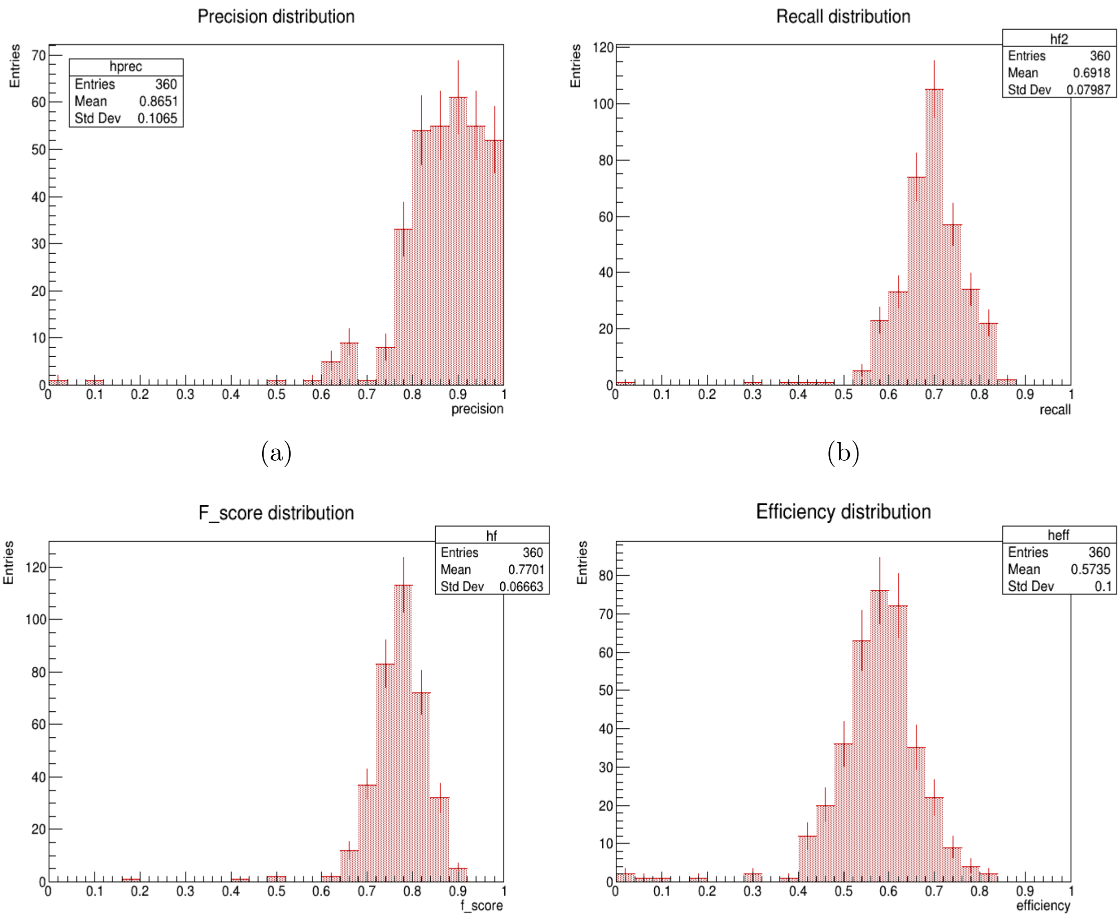


Figure 6.4: The distribution of the signal and background correctly classified (a and d) and mis-classified (b and c) for the RUN3 - 3 dataset.



(a) (b) (c) (d)

Figure 6.5: Precision, recall, f-score and efficiency for each shower.

6.2.2 Precision-Recall curve

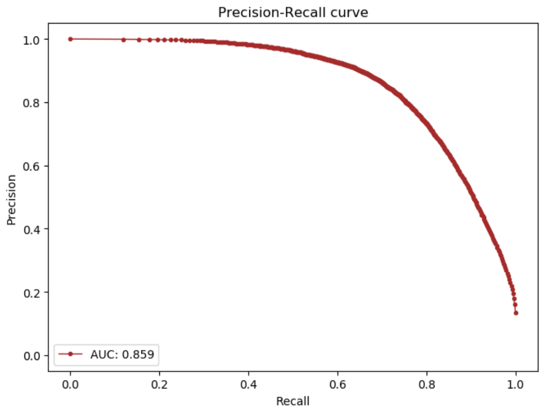


Figure 6.6: Precision and recall curve. The area under the curve is 85.9%.

At the end of this study, we would to introduce the Precision-Recall (PR) curves. They have been cited as an alternative to ROC curves to study the output of a classifier in binary classification, when the classes are very imbalanced [62]. The difference exists because in this domain the number of negative examples largely exceeds the number of positive ones. Consequently, a large change in the number of false positives can lead to a small change in the false positive rate used in ROC analysis. Precision, on the other hand, by comparing false positives to true positives rather than true negatives, captures the effect of the large number of negative examples on the algorithm's performance.

The precision-recall curve shows the trade-off between precision and recall for different thresholds. A model with perfect skill is depicted as a point at a coordinate of (1,1). A no-skill classifier will be a horizontal line on the plot with a precision that is proportional to the number of positive examples in the dataset. For a balanced dataset this will be 0.5.

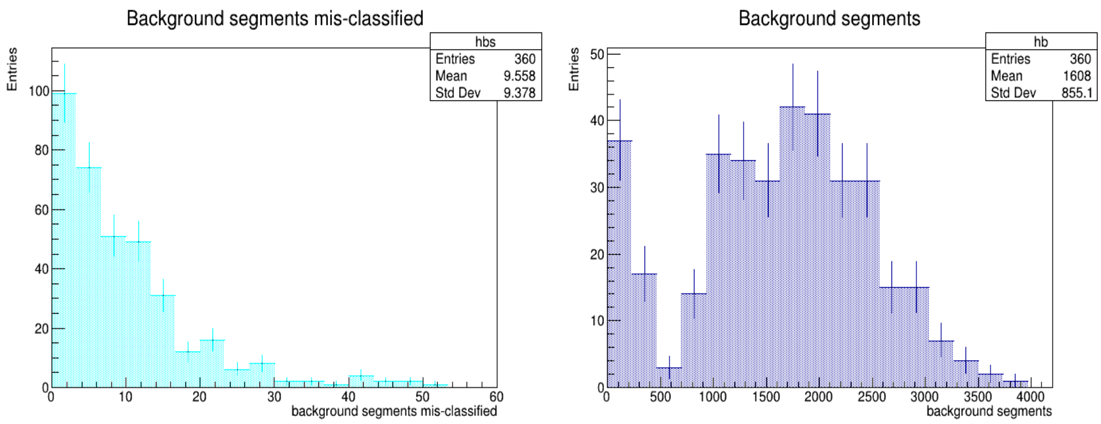
Often, the area under the curve (AUC) is used as a simple metric to define how an algorithm performs over the whole space. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of positive results (high recall). Figure 6.6 shows the precision-recall, with AUC = 85.9%.

6.3 Electron identification

A track reconstructed in an ECC is identified as an electron or a positron if it produces an electromagnetic shower. In turn, the shower is recognised if the number of segments classified as signal by the Random Forest exceeds a given threshold. The value of the threshold is optimised by maximizing the f-score[3] of the selection.

Table 6.9: Confusion matrix: 0:0 and 0:1 are background correctly classified and background mis-classified.

	0	1
0	578723	3447

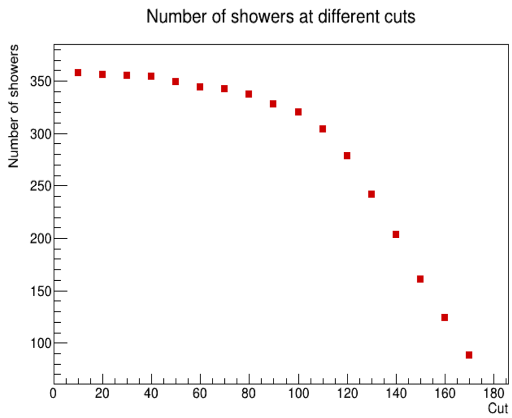


(a) (b)

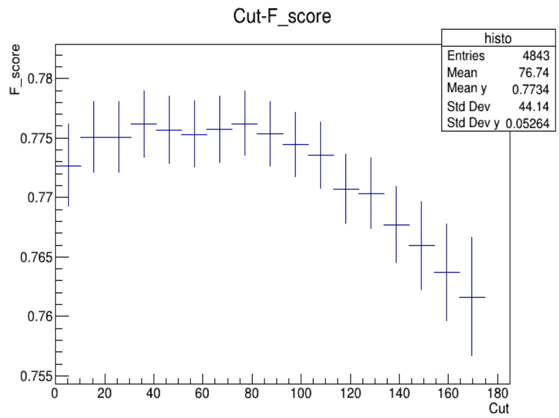
Figure 6.7: The distributions of the noise segments mis-classified (a) and correctly classified (b).

The evaluation of the f-score metric has been performed considering only the noise dataset. 360 injectors are selected from the first four plates with small angles ($\theta <$

[3] F-score is defined as the geometric mean of the recall and precision. Maximize the f-score means maximize both metrics.



(a)



(b)

Figure 6.8: (a)Number of showers exceeding the cut as a function of the cut. (b) F-score value computed for each shower satysfying the cut.

0.05 rad) and they are considered as fake electrons. Also for this dataset, the four-steps procedure is conducted to create the new features. To train the model, it is used one of the previous datasets and as a test set the dataset generated by the 360 fake electrons. Table 6.9 shows the number of the background segments correctly classified (0:0) and mis-classified (0:1). Figure 6.7 shows the distribution of the two categories computed for each shower produced by fake electrons. The left panel shows that the mean value of the noise segments mis-classified is of about 10 and the maximum number is 53.

The cut-analysis has been conducted using one dataset with signal and noise class. Changing the cut, i.e. the requirement on the number of signal segments classified by the Random Forest, the number of showers satisfying the condition and the f-score trend are reported in Figure 6.8.
The f-score metric computed for each of the events decreases when the threshold is 80. At higher thresholds, the showers are mostly populated by mis-classified noise segments. If we set a cut ≥ 50, the number of the selected showers are 349, with 100 they are 320, 150 there are 160 showers.

6.4 The energy calibration procedure

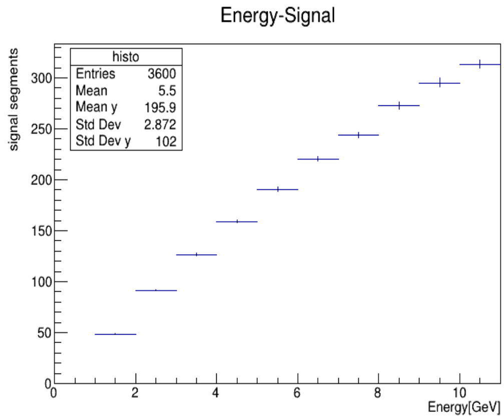
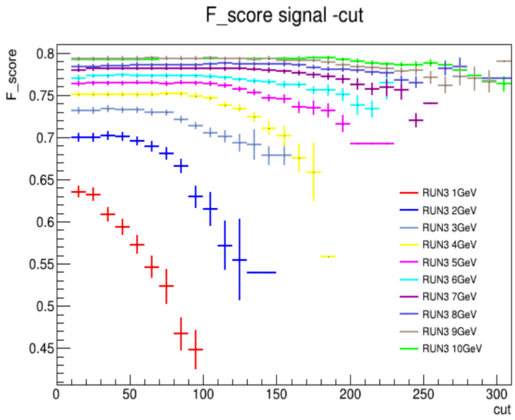
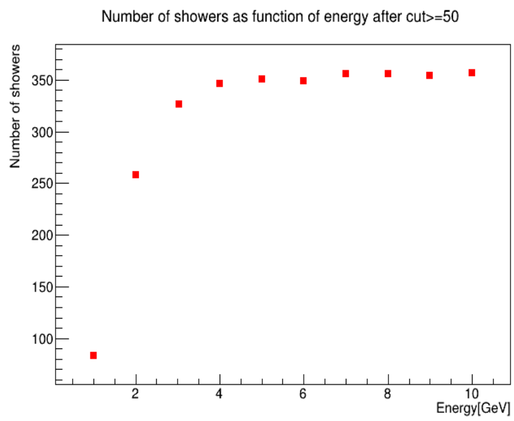


Figure 6.9: Signal segments collected for each shower at different energies. This number increases with the energy.

The final task of this work is to estimate the energy of the particle that produces the electromagnetic shower in the brick. The energy is proportional to the number of signal segments associated to the showers after the reconstruction. To make the energy calibration procedure, we start to study the RUN3 geometry - 360 electrons in a brick made of 29 emulsion films - at different energies, from 1 to 10 GeV.
For each dataset, signal is simulated while noise is the same produced for the previous sets of data. Different datasets differ for the number of signal segments that increases with the energy, see Figure 6.9. The identification of the segments is made with the application of the classifier. One of the previous dataset at 6 GeV is used to train the model and the others at different energies as test sets.
The Random Forest classifier returns for each shower the number of signal and background segments correctly classified and mis-classified, from which we can compute the recall, the precision and the f-score metrics.
Only showers satisfying the cut on the number of segments (Nseg) are selected as signal. Figure 6.10a shows the f-score for signal class and for each dataset at different energies changing the cut. Except the dataset at 1 GeV, for which the f-score already decreases after Nseg= 20, for the others sets Nseg = 50 is the best.
Figure 6.10b shows the number of the showers that exceeds the cut Nseg=50 as a function of the electron energy.

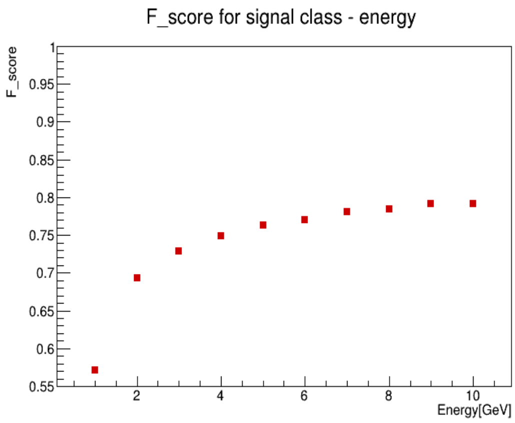


(a)

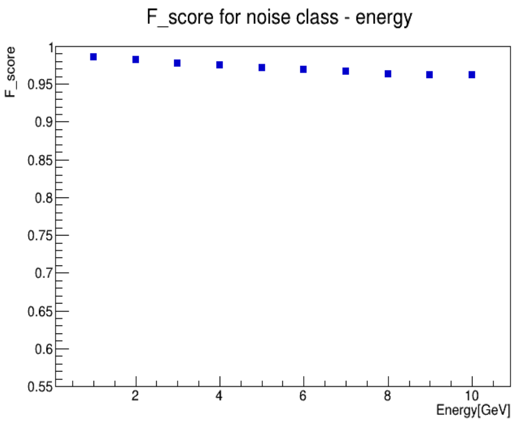


(b)

Figure 6.10: (a) The trend of f-score metric computed for each shower of datasets at different energies - each one is shown with different color. (b) Number of showers as a function of the energy above the threshold 50 on the number of selected segments.



(a)



(b)

Figure 6.11: F-score for signal (a) and noise (b) class for the RUN3 at different energies, evaluated after the cut of at least 50 segments per showers.

The trend of the f-score metric, shown in Figure 6.11, can be explained considering that the introduction of the new features is based on the differences made between the projected segments from one film with those found in the downstream one. Not all segments are selected but only those satisfying a condition on the angular displacement. When the energy increases, there are much more signal segments

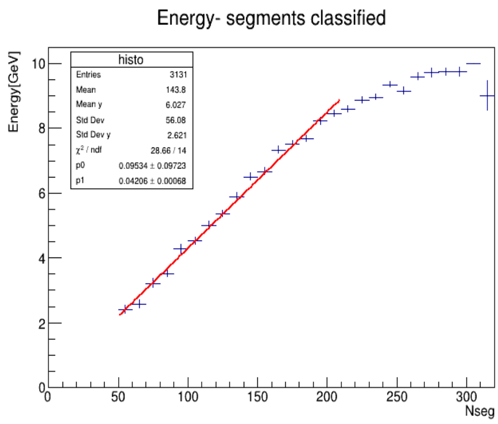


Figure 6.12: Energy as a function of the segments classified as signal. A linear fit is made in a linear region [50, 210] classified as signal.

in the films that can satisfy that condition, as a consequence the Random Forest identifies more signal segments. While the f-score for the signal class goes from 57.1% at 1 GeV to 79.2% at 10 GeV, the f-score for the noise class remains almost constant, see Figure 6.11b.

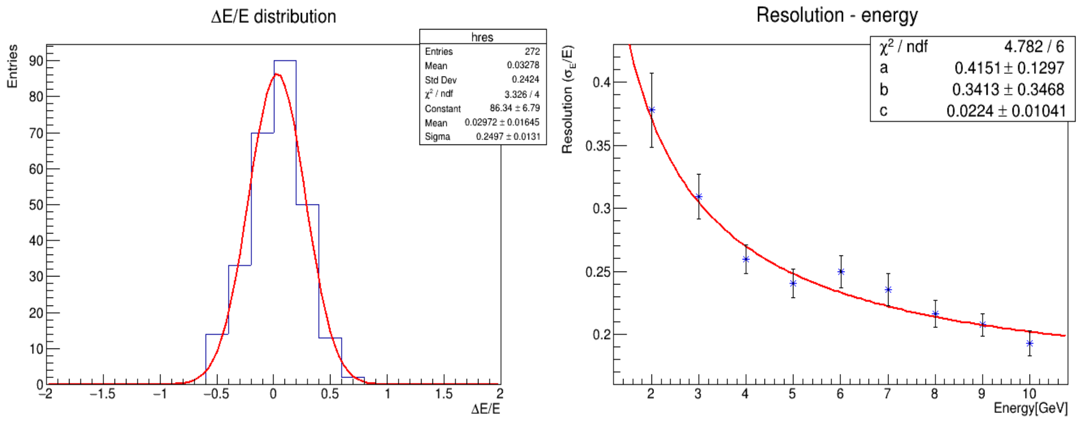
Once an electromagnetic shower is reconstructed, the electron energy can be correlated to the number of segments associated to the shower, as it is shown in Figure 6.12. There is indeed a linear correlation between the two quantities:

$$E_{rec} = p0 + p1 * (Nseg) \tag{6.1}$$

To find the p0 and p1 parameters, a linear fit is made in a region between [50, 210] segments, that returns:

$$p0 = (0.095 \pm 0.097)GeV, \quad p1 = (0.0421 \pm 0.0007)GeV \tag{6.2}$$

We use another set of 10 datasets with different energies to estimate the energy resolutions. Only the showers that exceed the threshold of 50 segments classified as signal are considered.
The energy resolution is given by σ_E/E and it is the sigma of the Gaussian fit of the relative difference.
It is made separatly for each dataset - Figure 6.13a shows the $(E_{true} - E_{rec})/E_{true}$ distribution with Gaussian fit for the dataset at 6 GeV.



(a) (b)

Figure 6.13: (a) ΔE/E computed for the dataset at energy of 6 GeV. (b) σ_E/E as function of electron energy.

Figure 6.13b shows that the fractional energy resolution improves with the energy. On the other hand, the Emulsion Cloud Chamber acts as a sampling calorimeter that satisfy the realtionship:

$$\frac{\sigma_E}{E} = \sqrt{\left(\frac{a}{\sqrt{E}}\right)^2 + \left(\frac{b}{E}\right)^2 + c^2} \tag{6.3}$$

Using this formula, a fit has been performed given the parameter a, b and c. In particular, the first term is stochastic fluctuation, it is $a = 0.41 \pm 0.13$ GeV$^{1/2}$ and the third is the calibration term $c = 0.02 \pm 0.01$. The noise term b, is compatible with zero.

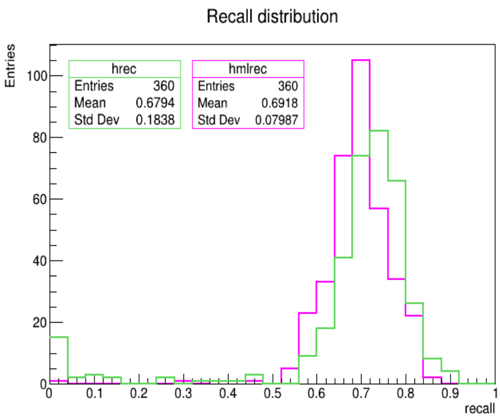
6.5 Machine learning vs classical approach

The reconstruction of the electromagnetic shower is conducted in this work with a new algorithm divided in two parts. At first, the description of the space region where we expect to find the shower is made with the definition of the geometrical parameters to built the rectangles that follow the development of the showers in the brick. This analysis leads to the introduction of new features. These variables are used in the second step, by the Random Forest classifier in order to recognize the signal segments and discard the noise ones. Different tests on the performances of

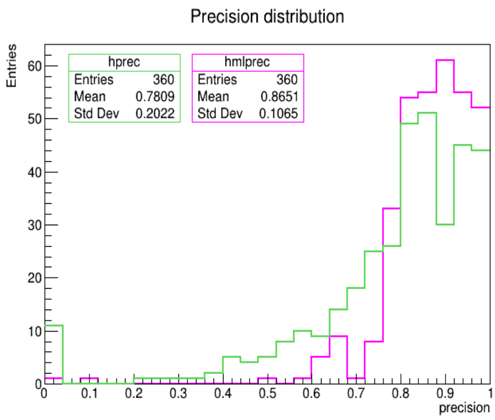
the Random Forets model are conducted and show that it is a robust and stable classifier: the metrics values do not change significantly varying the training or the test sets. In this last section, we are interested to compare the machine learning with the classical results.
The classical procedural algorithm used by the OPERA collaboration selects the tracks of the shower if their positions are inside a structure made of a cone followed by a cylinder. Starting from the injector segment, a cone with an opening angle α is defined to collect all the base-tracks within it. However, hits must be in a cylinder of radius ΔR_{max}. From the first base-track, each other collected segment is connected to the shower, if they meet specific requirements about angular and position displacements (dθ and dR) with respect to one track already added to the shower - i.e. if they are within the setted values.
In order to compare the two approaches, we have used the same dataset: RUN3 - 3 dataset.
It contains 485712 segments in total, 13.3% being signal segments and 86.7% noise. It is passed to the classical algorithm after the definition of the rectangles but before the application of the Random Forest Classifier.
For the classical procedure, the parameters are setted to:

- Cone aperture: 40 mrad;
- Cylinder radius: 1000μm;
- dR: 250μm;
- dθ: 150 mrad;

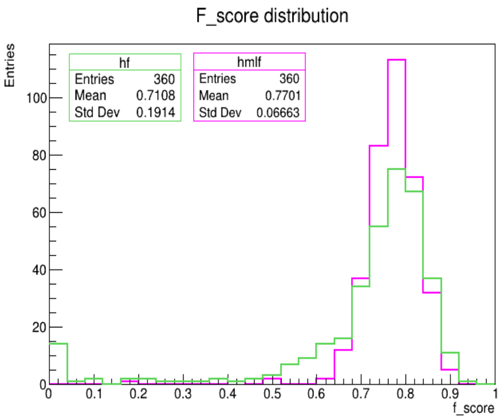
Each metric is computed separately for each shower both for the machine learning and classical approach. Table 6.10 shows the average of the precision, recall and f-score.
Figure 6.14 shows the overlapping distribution of each metric for the two approaches. The Machine Learning shower classification presents larger precision than the classical cone reconstruction, without significant loss in efficiency (recall).



(a)



(b)



(c)

Figure 6.14: The distributions of recall, precision and f-score for the machine learning(in purple) and classical approach (in green).

Table 6.10: Comparison between the classical algorithm and Random Forest model.

Algorithm	Precision (%)	Recall (%)	F-score (%)
Classical	78.1 ± 1.0	67.9 ± 0.9	71.1 ± 1.0
Random Forest	86.5 ± 0.5	69.2 ± 0.4	77.0 ± 0.3

Conclusions

The Search for Hidden Particles (SHiP) experiment is a new beam dump facility proposed at Cern SPS accelerator with the aim of exploring the Beyond Standard Model theories in the intensity frontier. The experiment will search for weakly interacting particles in the mass range below 10 GeV/c^2, lighter than the ones searched at LHC experiments with a complementary program of the energy-frontier one. The discovery of these particles can shed light on the open questions in particle physics about the nature of the dark matter, the baryon asymmetry in the universe and neutrino masses. The designed facility is also a copious neutrino factory where tau neutrino physics will be studied with three orders of magnitude more statistics than previous experiments. SHiP aims to perform the first observation of the $\overline{\nu}_\tau$ charged current interactions and measure tau neutrino cross section.
The micrometric tracking accuracy, provided by emulsion films interleaved with tungsten plates according to the Emulsion Cloud Chamber technology in the SND Emulsion target, will allow the identification of both the tau lepton production and decay vertices. The experiment is also suited for the direct search of Light Dark Matter (LDM) particles. Those weakly interacting particles, produced in the sub-GeV mass range via Dark Photon decays, can be discovered through the observation of their elastic scattering off the electrons of the ECC target. The signature of a LDM interaction is the electromagnetic shower produced from the recoil of electron scattered by LDM particles. On the other hand, also the neutrino interactions can produce an electromagnetic shower in the final state. The separation of electrons and photons and the measurement of the shower energy are key tasks to distinguish LDM from neutrinos.

The present thesis has been devoted to the study of the identification of the electromagnetic showers in an Emulsion Cloud Chamber, including the estimate of the

electron energy.
The work has been focused on the DESY test beam, conducted at Hamburg in October 2019, to identify low energy electromagnetic showers in a SHiP-like target. An apparatus consisting of an Emulsion Cloud Chamber and SciFi detectors - a prototype of the Target Trackers - was exposed to a pure electron beam. The beam parameters were adjusted in such a way to have the same number of electromagnetic showers as those expected during the SHiP experiment run. The study has been conducted on the RUN3 configuration of the DESY test beam, where electrons with an energy of 6 GeV have been fired in the ECC brick made of 29 emulsion films. The main purpose is the identification of segments of the showers in an environment with high noise density that is a few orders of magnitude larger than the signal. This is made by implementing machine learning techniques to identify signal and background segments.

The analysis of the reconstructed algorithm has been divided in two parts. The first part has been dedicated to a detailed study on the topological variables defining the space where most of the shower tracks were included. The size of this space is such that the efficiency is $(90.2 \pm 0.2)\%$, and the purity is $(78.9 \pm 0.1)\%$. The introduction of new features has been conducted to enhance the differences between signal and noise segments. With this study, the signal to noise ratio is 1/7 with a noise reduction of a factor of 70 and a signal efficiency of $(83.70 \pm 0.02)\%$.
In the second part, the new variables have been used as input for the Random Forest to make a binary classification between signal and noise segments.
To estimate the performances of the classifier, the precision, recall and f-score metrics have been considered.
A first result using only one dataset divided into training and test sets, shows a F-score value for the signal class of about 76.9%, with an high precision of about 87.3% and recall of 68.8%. Other tests have been performed with new datasets. Changing the training and test sets or combined them, the F-score values do not change significantly, meaning that the classifier is *robust* and *stable*. The Precision-Recall curve has been used to estimate the area under the curve, AUC that is 85.9%. The signal efficiency after the Random Forest classification is $(57.3 \pm 0.5)\%$.

The energy of showering electrons is determined by its correlation with the number of identified segments in the classification. As a result, the fractional resolution, σ_E/E, goes from $(37.8 \pm 3.0)\%$ at 1 GeV down to $(19.3 \pm 1.0)\%$ at higher energies.

A comparison between machine learning and classical approaches has been conducted. It showed that the ML has a higher F-score than the classical algorithm, 77.0% compared to 71.1% of the classical case.

Bibliography

[1] G. e. a. Aad, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Physics Letters B **716**, 1–29 (2012).

[2] S. e. a. Chatrchyan, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Physics Letters B **716**, 30–61 (2012).

[3] A. Maas, *Brout–Englert–Higgs physics: From foundations to phenomenology*, Progress in Particle and Nuclear Physics **106**, 132–209 (2019).

[4] V. A. Bednyakov, N. D. Giokaris, and A. V. Bednyakov, *On the Higgs mass generation mechanism in the Standard Model*, Physics of Particles and Nuclei **39**, 13–36 (2008).

[5] S. M. Bilenky and J. Hošek, *Glashow-Weinberg-Salam theory of electroweak interactions and the neutral currents*, Physics Reports **90**, 73 (1982).

[6] M. Holthausen, K. S. Lim, and M. Lindner, *Planck scale boundary conditions and the Higgs mass*, Journal of High Energy Physics **2012** (2012).

[7] Q. R. Ahmad, R. C. Allen, T. C. Andersen, J. D. Anglin, G. Bühler, J. C. Barton, E. W. Beier, M. Bercovitch, J. Bigu, S. Biller, and et al., *Measurement of the Rate of* $\nu_e + d \to p + p + e$ *Interactions Produced by B8 Solar Neutrinos at the Sudbury Neutrino Observatory*, Physical Review Letters **87** (2001).

[8] J. Y. Kim, *Discovery of Neutrino Oscillations in the Super-Kamiokande Experiment*, Phys. High Technol. **24**, 8 (2015).

[9] N. Agafonova et al., *Final results of the search for ν $\mu\rightarrow$ ν e oscillations with the OPERA detector in the CNGS beam*, Journal of High Energy Physics **2018**, 151 (2018).

[10] B. Pontecorvo, *Neutrino experiments and the problem of conservation of leptonic charge*, Sov. Phys. JETP **26**, 165 (1968).

[11] Y. Fukuda, T. Hayakawa, E. Ichihara, K. Inoue, K. Ishihara, H. Ishino, Y. Itow, T. Kajita, J. Kameda, S. Kasuga, and et al., *Evidence for Oscillation of Atmospheric Neutrinos*, Physical Review Letters **81**, 1562–1567 (1998).

[12] C. Giganti, S. Lavignac, and M. Zito, *Neutrino oscillations: The rise of the PMNS paradigm*, Progress in Particle and Nuclear Physics **98**, 1–54 (2018).

[13] F. d. Aguila, J. Syska, and M. Zrałek, *Neutrino oscillations beyond the standard model*, Journal of Physics: Conference Series **136**, 042027 (2008).

[14] M. Lindner, T. Ohlsson, and G. Seidl, *Seesaw mechanisms for Dirac and Majorana neutrino masses*, Physical Review D **65** (2002).

[15] L. Canetti, M. Drewes, and M. Shaposhnikov, *Matter and antimatter in the universe*, New Journal of Physics **14**, 095012 (2012).

[16] A. D. Sakharov, *Violation ofCPin variance,Casymmetry, and baryon asymmetry of the universe*, Soviet Physics Uspekhi **34**, 392 (1991).

[17] D. Edmonds, D. Minic, and T. Takeuchi, *Dark Matter, Dark Energy and Fundamental Acceleration*, 2020.

[18] S. Collaboration, *A facility to Search for Hidden Particles (SHiP) at the CERN SPS*, 2015.

[19] T. Asaka, S. Blanchet, and M. Shaposhnikov, *The nuMSM, dark matter and neutrino masses*, Phys. Lett. B **631**, 151 (2005).

[20] M. E. Peskin, *Supersymmetry in Elementary Particle Physics*, 2008.

[21] S. Alekhin and et al., *A facility to search for hidden particles at the CERN SPS: the SHiP physics case*, Reports on Progress in Physics **79**, 124201 (2016).

[22] B. Batell, M. Pospelov, and A. Ritz, *Exploring portals to a hidden sector through fixed targets*, Physical Review D **80** (2009).

[23] G. Aad, B. Abbott, J. Abdallah, S. Abdel Khalek, O. Abdinov, R. Aben, B. Abi, M. Abolins, O. AbouZeid, H. Abramowicz, and et al., *Search for high-mass dilepton resonances in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, Physical Review D **90** (2014).

[24] V. Khachatryan, A. M. Sirunyan, A. Tumasyan, W. Adam, T. Bergauer, M. Dragicevic, J. Erö, C. Fabjan, M. Friedl, and et al., *Search for physics beyond the standard model in dilepton mass spectra in proton-proton collisions at $\sqrt{s} = 8$ TeV*, Journal of High Energy Physics **2015** (2015).

[25] S. Gninenko and N. Krasnikov, *The muon anomalous magnetic moment and a new light gauge boson*, Physics Letters B **513**, 119–122 (2001).

[26] N. Arkani-Hamed, D. P. Finkbeiner, T. R. Slatyer, and N. Weiner, *A theory of dark matter*, Physical Review D **79** (2009).

[27] F. Zwicky, *The redshift of extragalactic nebulae*, Helv. Phys. Acta **6**, 138 (1933).

[28] V. Rubin, J. Ford, W.K., and N. Thonnard, *Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605 (R=4kpc) to UGC 2885 (R=122kpc).*, ApJ **238**, 471 (1980).

[29] S. P. MARTIN, *A SUPERSYMMETRY PRIMER*, Advanced Series on Directions in High Energy Physics , 1–98 (1998).

[30] J. M. Butler, *Searches for Dark Matter at the LHC*, PoS **ALPS2018**, 030 (2018).

[31] P. deNiverville, D. McKeen, and A. Ritz, *Signatures of sub-GeV dark matter beams at neutrino experiments*, Physical Review D **86** (2012).

[32] P. deNiverville, M. Pospelov, and A. Ritz, *Observing a light dark matter beam with neutrino experiments*, Physical Review D **84** (2011).

[33] R. E. et al., *Dark Sectors and New, Light, Weakly-Coupled Particles*, 2013.

[34] M. Pospelov, *Secluded U(1) below the weak scale*, Physical Review D **80** (2009).

[35] A. A. Aguilar-Arevalo et al., *Low Mass WIMP Searches with a Neutrino Experiment: A Proposal for Further MiniBooNE Running*, 2012.

[36] S. Collaboration, *Sensitivity of the SHiP experiment to light dark matter*, 2020.

[37] K. Kodama et al., *Final tau-neutrino results from the DONuT experiment*, Phys. Rev. D **78**, 052002 (2008).

[38] R. Schwienhorst et al., *A New upper limit for the tau - neutrino magnetic moment*, Phys. Lett. B **513**, 23 (2001).

[39] C. Ahdida and et all, *SHiP Experiment - Comprehensive Design Study report*, Technical Report CERN-SPSC-2019-049. SPSC-SR-263, CERN, Geneva, 2019.

[40] C. a. Ahdida, *SHiP Experiment - Progress Report*, Technical Report CERN-SPSC-2019-010. SPSC-SR-248, CERN, Geneva, 2019.

[41] T. Nakamura et al., *The OPERA film: New nuclear emulsion for large-scale, high-precision experiments*, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **556**, 80 (2006).

[42] C. Ahdida, R. Albanese, A. Alexandrov, A. Anokhina, S. Aoki, G. Arduini, E. Atkin, N. Azorskiy, J. Back, A. Bagulya, and et al., *The magnet of the scattering and neutrino detector for the SHiP experiment at CERN*, Journal of Instrumentation **15**, P01027–P01027 (2020).

[43] N. Agafonova et al., *Momentum measurement by the multiple Coulomb scattering method in the OPERA lead-emulsion target*, New journal of physics **14**, 013026 (2012).

[44] F. Juget, *Electromagnetic shower reconstruction with emulsion films in the OPERA experiment*, in *J. Phys. Conf. Ser.*, volume 160, page 012033, 2009.

[45] K. Morishima, A. Nishio, M. Moto, T. Nakano, and M. Nakamura, *Development of nuclear emulsion for muography*, Annals of Geophysics **60**, 0112 (2017).

[46] L. Collaboration et al., *LHCb tracker upgrade technical design report*, Technical report, 2014.

[47] C. Andreopoulos, C. Barry, S. Dytman, H. Gallagher, T. Golan, R. Hatcher, G. Perdue, and J. Yarba, *The GENIE Neutrino Monte Carlo Generator: Physics and User Manual*, 2015.

[48] D. Banerjee et al., *Dark matter search in missing energy events with NA64*, Physical review letters **123**, 121801 (2019).

[49] J. Lees et al., *Search for invisible decays of a dark photon produced in e+ e- collisions at B a B ar*, Physical review letters **119**, 131804 (2017).

[50] S. Raschka, *Python machine learning*, Packt publishing ltd, 2015.

[51] W. Leo, *Techniques for nuclear and particle physics experiments*, Nucl Instrum Methods Phys Res **834**, 290 (1988).

[52] N. M. D.E. Groom and S. Striganov, *Muon stopping- power and range tables*, Atomic Data and Nuclear Data Tables, 2000.

[53] B. B. Rossi, *High-energy particles*, (1952).

[54] S. Agostinelli et al., *GEANT4—a simulation toolkit*, Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment **506**, 250 (2003).

[55] K. Morishima and T. Nakano, *Development of a new automatic nuclear emulsion scanning system, S-UTS, with continuous 3D tomographic image read-out*, Journal of Instrumentation **5**, P04011 (2010).

[56] L. Arrabito and et al., *Track reconstruction in the emulsion-lead target of the OPERA experiment using the ESS microscope*, Journal of Instrumentation **2**, P05004–P05004 (2007).

[57] A. Alexandrov et al., *The continuous motion technique for a new generation of scanning systems*, Scientific reports **7**, 1 (2017).

[58] M. De Serio et al., *High precision measurements with nuclear emulsions using fast automated microscopes*, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **554**, 247 (2005).

[59] R. K. Bock, H. Grote, and D. Notz, *Data analysis techniques for high-energy physics*, volume 11, Cambridge University Press, 2000.

[60] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research **12**, 2825 (2011).

[61] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, *API design for machine learning software: experiences from the scikit-learn project*, in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[62] J. Davis and M. Goadrich, *The relationship between Precision-Recall and ROC curves*, in *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

Acknowledgements

I would like to thank all the people who contributed to this thesis work.
First of all, I would like to express my gratitude to my supervisor, Professor Giovanni De Lellis, for his advice and precious teachings.
My most sincere thanks go to Dtt.ssa Antonia Di Crescenzo, for the kindness and professionalism with which she has guided me throughout this work, helping me to improve my skills and knowledge.
A heartfelt thank to Dr. Antonio Iuliano for his constant help and support for this thesis.
I sincerely thank my co-supervisor Dtt.ssa Autilia Vitiello for her important support for this work.
Finally, I would like to thank the whole Emulsion Group in Naples for their wonderful welcome.
Thank you all.