UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULDADE DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Luan Gomes Mattosinhos de Carvalho

Advancing Background Rejection Capabilities in the CYGNO Experiment via Machine Learning

Juiz de Fora
2026

Luan Gomes Mattosinhos de Carvalho

Advancing Background Rejection Capabilities in the CYGNO Experiment via Machine Learning

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Engenharia Elétrica. Área de concentração: Sistemas Eletrônicos

Orientador: Prof. Dr. Rafael Antunes Nóbrega

Juiz de Fora

2026

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Carvalho, Luan Gomes Mattosinhos de.
Advancing Background Rejection Capabilities in the CYGNO Experiment via Machine Learning / Luan Gomes Mattosinhos de Carvalho. – 2026.
124 f. : il.

Orientador: Rafael Antunes Nóbrega
Dissertação (Mestrado) – Universidade Federal de Juiz de Fora, Faculdade de Engenharia. Programa de Pós-Graduação em Engenharia Elétrica, 2026.

1. Dark matter. 2. CYGNO experiment. 3. Machine learning. 4. Deep learning. 5. Background rejection. I. Nóbrega, Rafael Antunes, adv. Título.

Luan Gomes Mattosinhos de Carvalho

Advancing Background Rejection Capabilities in the CYGNO Experiment via Machine Learning

Dissertation submitted to the Graduate Program in Electrical Engineering of the Federal University of Juiz de Fora as a partial requirement for obtaining a Master's degree in Electrical Engineering. Concentration area: Electronic Systems

Approved on 05 of August of 2026.

EXAMINING BOARD

Prof. Dr. Rafael Antunes Nóbrega – Academic Advisor
Federal University of Juiz de Fora

Prof. Dr. Augusto Santiago Cerqueira
Federal University of Juiz de Fora

Prof. Dr. João Victor da Fonseca Pinto
Federal University of Rio de Janeiro

Juiz de Fora, 08/05/2026.

Documento assinado eletronicamente por **Augusto Santiago Cerqueira**, **Professor(a)**, em 05/08/2026, às 18:03, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **JOAO VICTOR DA FONSECA PINTO**, **Usuário Externo**, em 05/08/2026, às 18:26, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Rafael Antunes Nobrega**, **Professor(a)**, em 05/08/2026, às 18:31, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Dedico este trabalho aos meus pais, irmã, familiares e amigos.

ACKNOWLEDGEMENT

Agradeço primeiramente a meu professor orientador, Rafael, pela paciência, conselhos valiosos e apoio ao longo do desenvolvimento deste trabalho, além de todas as oportunidades que me foram concedidas.

Agradeço profundamente aos meus pais, Lucielena e Rone, pelo amor, apoio, suporte e dedicação ao longo de toda a minha jornada. À minha irmã, Luene, pelo carinho e compreensão constantes.

Aos meus professores, pela orientação e contribuições essenciais à minha formação.

Aos meus amigos e companheiros da faculdade, Augusto, Bernardo, Elian, Luís e Otton, pela amizade e companheirismo durante todo o processo.

Ao Igor Pains e demais colegas do NIPS, por terem sido essenciais para o desenvolvimento deste trabalho.

Aos membros da colaboração CYGNO que tornaram este trabalho possível, em especial: Giulia, Giorgio, Pietro e Stefano.

Por último, um agradecimento especial à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), à Universidade Federal de Juiz de Fora, à Faculdade de Engenharia, ao Programa de Pós-Graduação em Engenharia Elétrica e ao Instituto Nazionale di Fisica Nucleare (INFN), por todo suporte e apoio para o desenvolvimento deste trabalho.

RESUMO

A detecção direta de Matéria Escura requer a identificação de raros eventos de Recuo Nuclear (NR) na presença de um fundo dominante de Recuos Eletrônicos (ER). O experimento CYGNO aborda esse desafio por meio de uma Câmara de Projeção Temporal gasosa com leitura óptica, capaz de reconstruir a topologia das trilhas de ionização produzidas por interações de partículas. Este trabalho apresenta o desenvolvimento e a avaliação de métodos de aprendizado de máquina para a discriminação ER/NR utilizando dados simulados do protótipo LIME. Para isso, foi estabelecida uma cadeia completa de análise, incluindo a reconstrução das imagens, a extração de variáveis físicas, a caracterização da resposta do detector, o ajuste de simulações de Monte Carlo, e a otimização dos classificadores. Um conjunto de 30 variáveis relacionadas à topologia e energia foi extraído das trilhas reconstruídas e utilizado no treinamento de algoritmos clássicos de aprendizado de máquina, enquanto modelos de aprendizado profundo baseados em imagens foram treinados diretamente sobre as imagens capturadas pelo detector. Os classificadores otimizados foram avaliados por validação cruzada e por meio de um conjunto de teste independente, alcançando uma ROC AUC máxima de 0,902. Uma análise detalhada em função da energia mostrou que o desempenho da discriminação é fortemente limitado em baixas energias, aproximando-se de uma separação perfeita em energias mais elevadas. Na região intermediária de energia, onde a tarefa de classificação é mais desafiadora, os métodos baseados em variáveis extraídas apresentaram fatores de rejeição de fundo comparáveis ou superiores aos obtidos pelos modelos baseados em imagens. Entre os métodos avaliados, o XGBoost e o Perceptron Multicamadas apresentaram a melhor combinação entre desempenho de classificação e eficiência computacional, atingindo taxas de inferência superiores a 10^5 eventos por segundo enquanto mantinham elevado poder de rejeição. Os resultados demonstram que as variáveis desenvolvidas são capazes de capturar as informações físicas relevantes contidas nas trilhas de ionização, permitindo que modelos clássicos de aprendizado de máquina igualem e, em algumas condições de operação, superem o desempenho de redes neurais convolucionais. De forma geral, o estudo indica que abordagens baseadas em variáveis físicas constituem uma solução eficiente para a discriminação ER/NR no detector LIME, preservando a possibilidade de interpretar os observáveis responsáveis pela classificação.

Palavras-chave: matéria escura; experimento CYGNO; aprendizado de máquina; aprendizado profundo; rejeição de fundo.

ABSTRACT

The direct detection of Dark Matter requires the identification of rare Nuclear Recoil (NR) interactions in the presence of dominant Electron Recoil (ER) backgrounds. The CYGNO experiment addresses this challenge through a gaseous Time Projection Chamber with optical readout, capable of reconstructing the topology of ionization tracks produced by particle interactions. This work presents the development and evaluation of machine learning frameworks for ER/NR discrimination using simulated data from the LIME prototype. A complete analysis pipeline was established, including image reconstruction, feature extraction, detector response characterization, Monte Carlo simulation tuning, and classifier optimization. A set of 30 topological and energy-related observables was extracted from reconstructed tracks and used to train classical machine learning algorithms, while image-based deep learning models were trained directly on the detector images. The optimized classifiers were evaluated using cross-validation and an independent test set, achieving a maximum ROC AUC of 0.902. A detailed energy-dependent analysis showed that discrimination performance is strongly limited at low energies and approaches perfect separation at higher energies. In the intermediate energy region, where the classification task is most challenging, feature-based approaches achieved background rejection factors comparable to or exceeding those of the image-based models. Among the evaluated methods, XGBoost and a Multilayer Perceptron provided the best combination of discrimination performance and computational efficiency, reaching inference rates above 10^5 events per second while maintaining competitive rejection power. The results demonstrate that the engineered observables successfully capture the relevant physical information contained in the ionization tracks, allowing classical machine learning models to match and, in some operating regimes, exceed the performance of convolutional neural networks. Overall, the study indicates that feature-based machine learning provides an efficient solution for ER/NR discrimination in the LIME detector while preserving the possibility of interpreting the physical observables responsible for the classification.

Keywords: dark matter; CYGNO experiment; machine learning; deep learning; background rejection.

LIST OF ILLUSTRATIONS

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ADC	Analog-to-Digital Converter
AI	Artificial Intelligence
AUC	Area Under the Curve
BDT	Boosted Decision Tree
CCD	Charge-Coupled Device
CEvNS	Coherent elastic neutrino-nucleus scattering
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CSV	Comma-Separated Values
CYGNO	CYGnus module with Optical readout
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DM	Dark Matter
ER	Electron Recoil
FLOPS	Floating-point Operations Per Second
FN	False Negative
FP	False Positive
FPR	False Positive Rate
FT	Fine-Tuning
GEM	Gas Electron Multiplier
GPU	Graphics Processing Unit
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
INFN	Istituto Nazionale di Fisica Nucleare
LArTPC	Liquid Argon Time Projection Chamber
LHC	Large Hadron Collider
LIME	Long Imaging ModulE
LNGS	Laboratori Nazionali del Gran Sasso
LY	Light Yield
MC	Monte Carlo
ML	Machine Learning
MLP	Multilayer Perceptron
MPGD	Micro Pattern Gas Detector
MSE	Mean Squared Error
NaI	Sodium Iodide
NaI(Tl)	Thallium-doped Sodium Iodide
NR	Nuclear Recoil
PCA	Principal Component Analysis
PMT	Photomultiplier Tube
QF	Quenching Factor
RANSAC	Random Sample Consensus
RFE	Recursive Feature Elimination

RGB	Red, Green, Blue
RMS	Root Mean Square
ROC	Receiver Operating Characteristic
ROI	Region of Interest
sCMOS	scientific Complementary Metal-Oxide-Semiconductor
SFS	Sequential Feature Selection
SMP	Standard Model Particle
SRIM	Stopping and Range of Ions in Matter
TL	Transfer Learning
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPC	Time Projection Chamber
TPE	Tree-structured Parzen Estimator
TPR	True Positive Rate
VRAM	Video Random Access Memory
WIMP	Weakly Interacting Massive Particle
XGBoost	Extreme Gradient Boosting

LIST OF SYMBOLS

A	Amplitude of the Gaussian fit (for charge profiles)
A_u	Amplitude of the Gaussian fit for the longitudinal profile
A_v	Amplitude of the Gaussian fit for the transverse profile
α	Learning rate (MLP/CNN) or regularization parameter
β	Gain saturation parameter of the GEM
C	Regularization parameter (SVM)
C_{conv}	Photon-to-ADC conversion factor
D_L	Longitudinal diffusion coefficient
D_T	Transverse diffusion coefficient
ΔE	Energy deposition (Geant4)
Δt	Time interval between interaction and detection
ΔV	Voltage drop across the GEM
ϵ_{extr}	Electron extraction efficiency from the GEM
ϵ^B	False Positive Rate (background acceptance fraction)
ϵ^S	Signal Efficiency (True Positive Rate)
E_{cal}	Calibrated energy (electron-equivalent energy, $\mathrm{keV_{ee}}$)
E_{source}	Energy of the calibration source
η	Charge density metric (MaxDen/Skeleton_Length)
G	GEM gain
I	Integrated track charge (ADC counts)
I_{ij}	Adjusted pixel intensity at (i, j)
λ	Electron attenuation (absorption) length
λ_1, λ_2	Eigenvalues of the spatial covariance matrix (PCA)
μ, σ	Mean and standard deviation parameters (Gaussian or Z-score)
Ω	Solid angle acceptance
ϕ	Activation function (Neural Networks)
ρ	Charge density metric normalized by dispersion
σ_L, σ_T	Longitudinal and transverse diffusion standard deviations
σ_u, σ_v	Charge distribution standard deviation along PCA axes
Σ	Spatial covariance matrix
χ^2	Goodness-of-fit statistic
v_d	Electron drift velocity

CONTENTS

1 INTRODUCTION

Over the past decades, multiple astrophysical observations have consistently indicated that most of the matter content of the Universe is composed of an unknown component known as Dark Matter (DM). Among the various hypotheses proposed to explain DM, Weakly Interacting Massive Particles (WIMPs) remain one of the most investigated candidates (1). WIMPs theoretically could interact with atomic nuclei through rare scattering processes, producing observable nuclear recoils (NR) in detectors.

Direct detection experiments are designed to search for these interactions. According to the standard galactic halo model, the motion of the Solar System through the galaxy produces an apparent “WIMP wind” originating from the direction of the Cygnus constellation. As a result, recoil signals detected on Earth are expected to exhibit directional signatures that could help distinguish a potential DM signal from conventional backgrounds.

The CYGNO experiment (2), part of the CYGNUS collaboration (3), investigates this possibility using gaseous Time Projection Chambers (TPCs) with optical readout. By combining Micro Pattern Gas Detectors (MPGDs), scientific Complementary Metal-Oxide-Semiconductor (sCMOS) cameras, and photomultiplier tubes (PMTs), the detector records detailed information about the topology and energy deposition of ionization tracks produced by particle interactions. This approach provides access to observables that can be explored to distinguish nuclear recoils, expected from WIMP-induced interactions, from the dominant electron recoil (ER) background.

However, the discrimination between ER and NR events represents one of the main challenges for rare-event searches. Beyond the detector itself, the increasing complexity and dimensionality of the acquired data motivate the use of advanced computational techniques for event classification. In recent years, Machine Learning (ML) methods have become increasingly important in experimental particle physics, providing a way to combine multiple observables and explore correlations between them that are difficult to capture through traditional univariate analyses.

Within this context, this work investigates the application of ML techniques for ER/NR discrimination in the LIME prototype, one of the CYGNO detectors. The study combines detector simulation, image reconstruction, feature extraction, and supervised classification to evaluate both feature-based and image-based approaches. The goal is to compare these strategies in terms of background rejection capabilities, computational cost, and model interpretability within the low-energy region relevant for DM searches.

1.1 MAIN CONTRIBUTIONS

The first contribution of this work consists of the development of a complete feature extraction pipeline for reconstructed ionization tracks. This includes the implementation and the evaluation of a set of 30 morphological, geometrical, and energy-related observables used to characterize the events.

A second contribution is the calibration and optimization of the detector simulation parameters to reproduce the experimental response of the LIME prototype. Starting from particle interactions simulated with the collaboration's Geant4 framework, the digitization stage was tuned using experimental data, allowing the generation of simulated datasets consistent with the detector response observed experimentally.

The main methodological contribution is the implementation and evaluation of ML techniques for ER/NR discrimination. This includes feature selection, hyperparameter optimization, transfer learning strategies, architecture search for a custom Convolutional Neural Network (CNN), and a comparison between feature-based and image-based approaches in terms of background rejection capability, computational cost, and model interpretability.

1.2 DECLARATION OF GENERATIVE AI USE

During the preparation of this dissertation, the author used generative Artificial Intelligence (AI) tools as supplementary tools in the writing and development process. Their use was limited to English language revision, translation, grammar correction, and structural refinement of the text. AI tools were also used as auxiliary resources during the development of Python code employed in the data analysis and machine learning tasks. All scientific content, methodological decisions, analyses, results, and interpretations presented in this dissertation were developed and evaluated by the author.

1.3 DOCUMENT STRUCTURE

To provide a comprehensive understanding of the study, this document is structured as follows:

- **Chapter 2** reviews the foundations of dark matter searches, discussing the astrophysical evidence for DM, the WIMP hypothesis, and the principles of direct detection experiments.
- **Chapter 3** presents the CYGNO experiment and the LIME prototype, describing the detector architecture, the optical readout system, and the physical processes involved in ionization track formation.

- **Chapter 4** details the event reconstruction pipeline, including image preprocessing, cluster identification, and feature extraction from reconstructed tracks.
- **Chapter 5** describes the simulation framework and dataset generation procedure, covering detector simulation, calibration using experimental data, and dataset preparation for supervised learning.
- **Chapter 6** presents the Machine Learning methodology, including feature-based models, feature selection, hyperparameter optimization, cross-validation procedures, and image-based deep learning approaches.
- **Chapter 7** discusses the obtained results, comparing the performance of the evaluated models through optimization studies, energy-dependent analyses, and background rejection measurements.
- **Chapter 8** summarizes the main conclusions of this work and discusses possible directions for future research.

2 LITERATURE REVIEW

The search for dark matter has motivated the development of increasingly sensitive detectors and analysis techniques capable of identifying extremely rare particle interactions. Although a wide range of astrophysical observations indicate the existence of DM, its nature remains unknown. Direct detection experiments seek to observe the interaction of DM particles with ordinary matter, but this task is complicated by the presence of background events produced by natural radioactivity sources.

This chapter reviews the physical and experimental concepts that motivate the present work. Section 2.1 introduces the principles of DM direct detection, discusses the expected signatures produced by WIMPs, and presents the main experimental approaches currently employed in the field. Section 2.2 then focuses on the challenge of discriminating nuclear recoils from electron recoil backgrounds, reviewing the evolution of background rejection techniques from traditional cut-based approaches to modern machine learning and deep learning methods applied to particle detectors.

2.1 DARK MATTER DIRECT DETECTION

Over the last century, the development of particle physics has led to the establishment of the Standard Model, which describes the fundamental particles and their interactions. The Standard Model has been remarkably successful in explaining a wide range of experimental observations in high-energy physics and astronomy, but some exceptions remain unexplained (4).

One of these observations is the evidence that most of the matter in the Universe is not composed of the particles described by the Standard Model. As a result, astronomers proposed the existence of a form of matter known as dark matter. Since it does not emit, absorb, or reflect electromagnetic radiation, DM cannot be observed directly through conventional astronomical techniques (5). Current cosmological models indicate that DM is five times more abundant than ordinary baryonic matter (6).

Several theoretical models have been proposed to explain DM (7, 8, 9, 10, 11). Among the most studied candidates are the so-called Weakly Interacting Massive Particles, hypothetical particles that interact only weakly with ordinary matter.

The search for these particles relies on three main approaches:

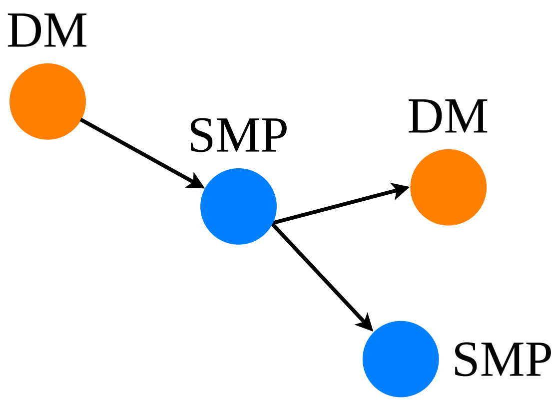
- Identifying DM particles produced in hadron colliders, such as the Large Hadron Collider (LHC) (12).

- Observing the products of DM annihilation or decay, particularly in regions with a high DM density (13).

- Direct detection of WIMPs through scattering processes in highly sensitive experiments, such as CYGNO, which operate with low background noise (14).

The possibility of directly detecting DM particles through WIMP interactions was first proposed in 1985 (15). In this approach, the interaction is studied through elastic collisions between WIMPs and atomic nuclei. These collisions, illustrated in Figure 1, result in a scattered WIMP and a recoiling Standard Model Particle (SMP).

Figure 1 – Simplified diagram of direct detection. DM particles interact with a SMP, resulting in a scattered DM particle and a recoiling SMP.



Source: Adapted from (16).

Since these particles carry no electric charge, interactions with atomic electrons are expected to be strongly improbable. However, elastic scattering with atomic nuclei may occur, producing a detectable nuclear recoil through momentum transfer. The total energy loss associated with this recoil in a WIMP detector can be described by the following equation:

$$\left(\frac{dE}{dx}\right)_{tot} = \left(\frac{dE}{dx}\right)_{elec} + \left(\frac{dE}{dx}\right)_{nucl} \tag{2.1}$$

During a WIMP-nucleus interaction, the recoiling nucleus dissipates energy through in two ways. A fraction converts into heat, causing atomic vibrations. The remaining fraction transfers through electronic losses, which ionize or excite surrounding atoms. As the ionized material returns to its ground state, it emits scintillation light. Photosensors detect these emitted photons, an approach utilized by experiments like XENON (17). However, due to the low intensity of the generated signal, the number of photons available for detection is extremely limited.

2.1.1 Event signatures and backgrounds

Interactions occurring inside DM detectors are commonly classified into two categories: nuclear recoils and electron recoils. Electron recoils are produced when radiation deposits energy through interactions with atomic electrons, whereas nuclear recoils result from momentum transfer directly to the atomic nucleus. Since WIMPs are expected to interact through elastic scattering with nuclei, their signatures are expected to resemble NR events.

The identification of a potential dark matter signal is complicated by the presence of background radiation, which can also produce detectable interactions inside the detector. One important source of background originates from cosmic rays. These high-energy particles, mainly protons and light nuclei, interact with the Earth's atmosphere and generate secondary particles such as muons, neutrons, pions, and photons, some of which can reach underground experiments and contribute to the recorded event rate.

Background events can also occur from the detector materials themselves. The materials used in the detector construction can have unstable isotopes, such as uranium, thorium, and potassium, that are radioactive and decay, emitting radiation.

For this reason, direct detection experiments rely on a combination of shielding, material selection, underground operation, and event discrimination techniques to suppress background contributions and improve sensitivity to rare NR signals.

2.1.2 WIMPs detection techniques

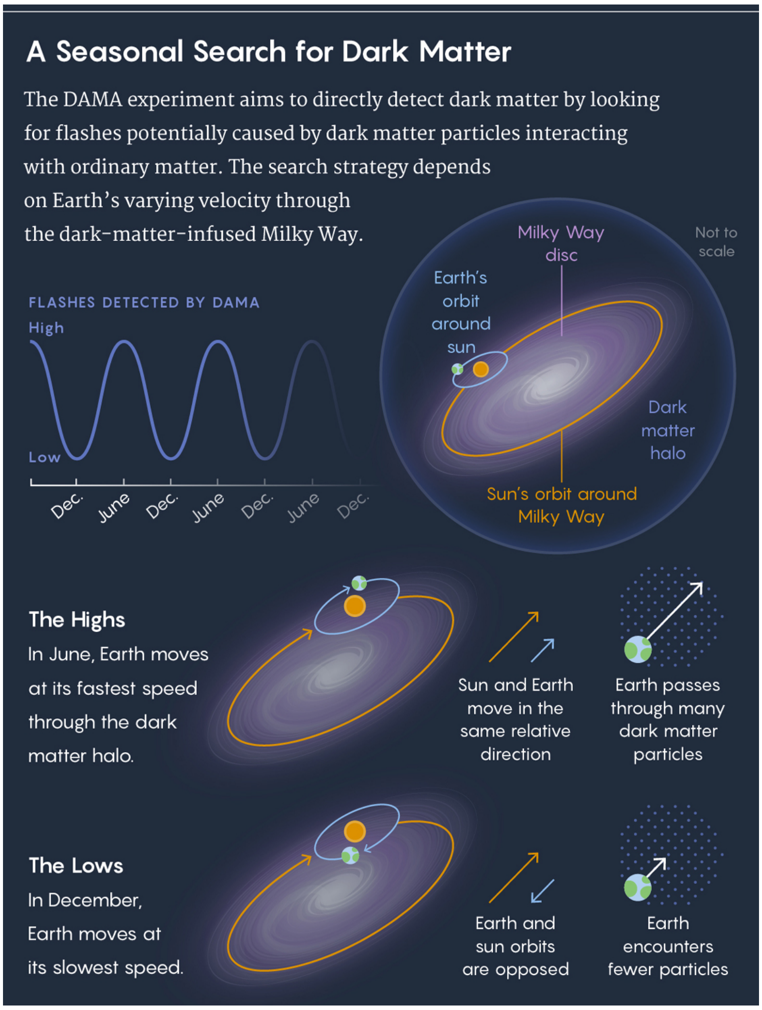
Direct DM detection faces the challenge of discriminating rare nuclear recoils from dominant backgrounds. To overcome this, experimental strategies leverage two effects for signal identification: seasonal modulation and directional dependence.

Because the Solar System moves through the galactic DM halo, the Earth's orbit around the Sun causes a seasonal variation in the relative velocity of the WIMP wind, expected to peak in June and reach a minimum in December, as illustrated in Figure 2. The DAMA/LIBRA experiment (18) explores this annual modulation using sodium iodide (NaI) crystals. DAMA has reported a modulation signal over decades, but no other experiment has observed similar results, rendering these claims controversial.

For this reason, several other detector technologies and approaches have been developed worldwide to investigate these findings and capture WIMP signals. These experimental techniques can be categorized as follows (20):

- **Inorganic Crystal Detectors:** These detectors use high-purity crystals to capture DM-induced charge signals. Early experiments used a 0.72kg Germanium crystal (21), later evolving to Silicon-based detectors like DAMIC (22) and SENSEI (23),

Figure 2 – Expected annual modulation of the WIMP interaction rate due to the Earth's orbit around the Sun.



Source: Extracted from (19)

improving sensitivity at lower masses. DAMA/LIBRA (18) and COSINE-100 (24) use NaI(Tl) scintillators.

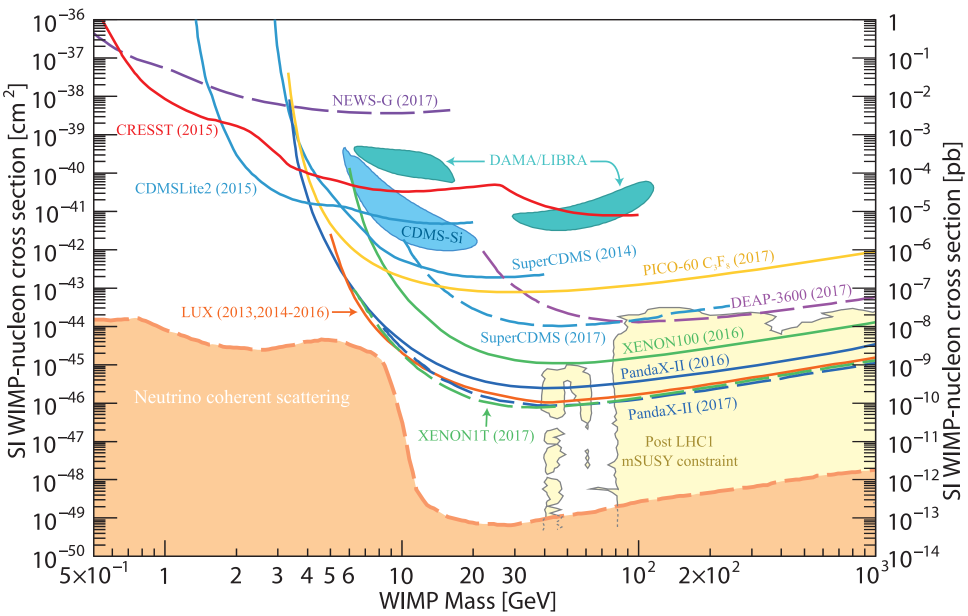
- **Cryogenic Detectors:** Operating at temperatures of a few tens of millikelvin, these detectors measure the phonons produced by particle interactions in the target material (25). Representative experiments include EDELWEISS (26) and CDMS (27).
- **Noble Liquid Detectors:** These detectors measure scintillation and ionization signals in single or dual-phase setups with PMTs, using liquid noble gases like argon or xenon. Examples: DarkSide (28, 29), DEAP (30, 31), XMASS (32), LUX (33), and XENON (17, 34).
- **Bubble Chambers:** In these detectors, particle interactions inside a superheated liquid (e.g., CF3I, C3F8, C4F10) induce bubble formation. The resulting bubbles are

recorded optically and can also be characterized through acoustic information. This technique is employed by experiments such as PICASSO (35, 36), PICO (37, 38), and COUPP (39).

The continuous development of these distinct technologies has pushed the boundaries of WIMP sensitivity, defining a rigid parameter space. Figure 3 illustrates the current exclusion limits for spin-independent WIMP-nucleon interactions as a function of the hypothetical WIMP mass. In this representation, each curve corresponds to an experimental upper bound on the interaction cross-section: parameter combinations above a given curve are excluded, as they would have produced detectable signals in that experiment, while regions below the curves remain experimentally allowed. The horizontal axis represents the WIMP mass (expressed in GeV/c^2), and the vertical axis corresponds to the interaction cross-section (in cm^2, shown in logarithmic scale), which quantifies the probability of WIMP scattering off nucleons.

However, as detector capacities scale, experiments will eventually reach the "Neutrino Floor". Below this physical threshold, coherent elastic neutrino-nucleus scattering (CEvNS) from solar and atmospheric neutrinos generates recoil signatures identical to those expected from WIMP interactions, making statistical separation extremely challenging.

Figure 3 – Exclusion limits for spin-independent WIMP-nucleon scattering as a function of WIMP mass. Each curve represents an upper bound from a given experiment; parameter space above the curves is excluded, while the region below remains allowed. The shaded region at low cross-sections indicates the irreducible "Neutrino Floor".



Source: Extracted from (40).

To provide an unambiguous DM signature, researchers also exploit directional dependence. The Earth's rotation leads to a diurnal modulation of the DM flux direction. Because the expected peak flux aligns with the Cygnus constellation, measuring the recoil direction provides a unique topological signature that background events cannot reproduce. This principle is the basis of **Directional Detectors**, such as Gas-based TPCs, which reconstruct ionization tracks to identify the recoil direction. Along with CYGNO, experiments such as MIMAC (41), DRIFT-II (42) and TREX-DM (43) fall into this category.

2.2 ER/NR DISCRIMINATION

The identification of DM signals in direct detection experiments relies on the capacity to reduce dominant background contributions. In gaseous detectors, this is achieved by analyzing the morphology and energy deposition of ionization tracks to distinguish electron recoils from nuclear recoils. Over the years, background rejection techniques have evolved from simple selection criteria based on track properties to Machine Learning and Deep Learning methods that combine multiple observables for event classification.

2.2.1 Discrimination techniques in Particle Detectors

Traditionally, the discrimination between ER and NR events in TPCs relied on algorithms extracting hand-crafted observables. Studies like those by Ghrear et al. (44) mapped these features, demonstrating how geometric and energy-based variables, like track length, transverse width, and charge distribution, characterize ionization tracks in gas TPCs. In optical detectors, such as the GEM-based TPC with Charge-Coupled Device (CCD) imaging described by Phan et al. (45), standard strategies classify events by applying fiducial cuts on these 2D space parameters. However, classical cut-based methods show limitations near the low-energy threshold, where diffusion and low signal-to-noise ratios blur the topological differences between recoils, causing significant signal efficiency loss.

To overcome the limitations of classical methods, experimental collaborations shifted to Multivariate Analysis and Machine Learning. In directional DM detection, Billard et al. (46) demonstrated this approach by applying Boosted Decision Trees to reject low-energy electron recoils in the MIMAC gas TPC. As ML techniques became more established, researchers also began exploring Deep Learning approaches. Baldi et al. (47) demonstrated that Deep Neural Networks (DNNs) can outperform traditional ML models in collider searches for exotic particles. The authors showed that deep architectures can learn relevant patterns directly from reconstructed event variables, reducing the need for additional feature engineering. Still in this approach, Sadowski et al. (48) adapted neural architectures to reconstruct tracks and detect antihydrogen annihilation events in antimatter physics experiments.

As optical readouts and segmented detectors capture spatial representations of energy depositions, the detector outputs started to be treated as digital images. Baldi et al. (49) implemented this computer vision approach by applying CNNs to "jet images" for substructure classification in colliders. In the context of TPCs, the MicroBooNE collaboration (50) implemented this image-based methodology using CNNs to classify neutrino interaction topologies in a Liquid Argon TPC (LArTPC). Expanding this concept to gaseous detectors, Renner et al. (51) introduced transfer learning for background rejection in the NEXT high-pressure xenon TPC. By adapting the pre-trained GoogLeNet architecture (52), they demonstrated that established computer vision models can classify ionization tracks based on their topological signatures.

Recent efforts applied these computer vision models to the specific requirements of directional DM searches. Schueler et al. (53) developed a 3D Convolutional Neural Network (3D-CNN) for high-resolution gas TPCs. The authors compared the 3D architecture against classical cut-based methods relying on track length, as well as feature-based algorithms including a Boosted Decision Tree (BDT) and a fully connected Multilayer Perceptron (MLP). In this specific case, the 3D-CNN outperformed all baseline models.

3 THE CYGNO EXPERIMENT

The CYGNO collaboration (a CYGNus module with Optical readout), as part of the CYGNUS proto-collaboration, aims to investigate the directional signature of the DM wind expected to originate from the direction of the Cygnus constellation. To achieve this objective, the collaboration seeks to establish a network of underground observatories, with the construction and operation of a detector with an active volume of $\mathcal{O}(30)$ m^3 as its ultimate goal.

The experimental approach adopted by CYGNO combines optical readout technology with multi-GEM structures in large-volume TPCs. The detector operates with a He:CF_4 (60:40) gas mixture at atmospheric pressure and room temperature, a configuration specifically designed to study rare events with energy releases ranging from hundreds of eV up to tens of keV. For charge multiplication, a MPGD amplification stage is used, employing a Triple-GEM stack. The readout system combines high-resolution sCMOS cameras with PMTs, which record the light produced on the Gas Electron Multipliers (GEMs). While the camera provides the two-dimensional (x, y) projection of the ionization tracks, the PMT signals recover the third dimension (z) through longitudinal temporal evolution, making 3D reconstruction of the events possible.

Throughout this chapter, an overview of the CYGNO experimental configuration and detector principles is provided in Section 3.1. Section 3.2 outlines the project development roadmap, and Section 3.3 details the architecture of the LIME prototype.

3.1 DETECTOR OVERVIEW

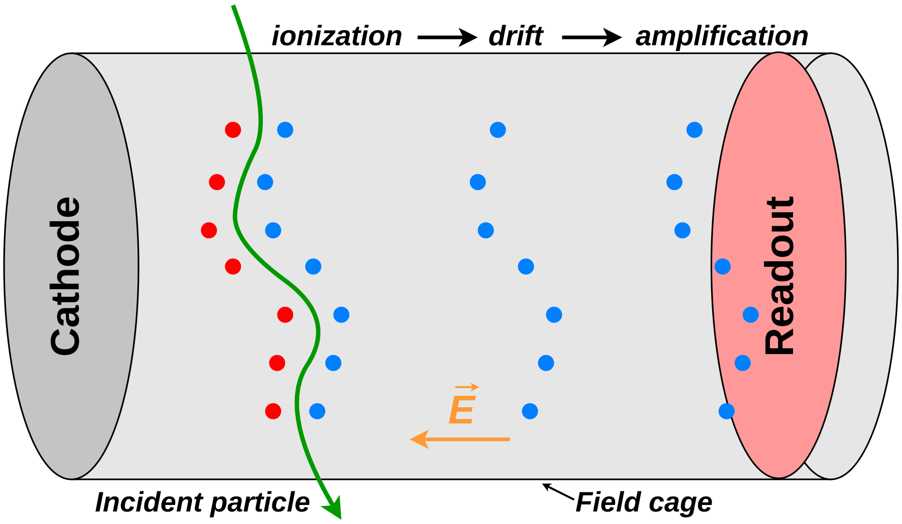
The operation of the CYGNO detector can be divided into four main stages: primary ionization and electron diffusion in the TPC, charge amplification in the GEM stack, optical signal detection, and three-dimensional event reconstruction. The operating principles of each stage are described in the following sections.

3.1.1 Time Projection Chamber

A Time Projection Chamber (54) is the active volume of the detector, where particle interactions produce ionization tracks in the gas. As illustrated in Figure 4, when a high-energy incident particle traverses the gas, it transfers kinetic energy through collisions with gas atoms and molecules, creating electron-ion pairs along its trajectory.

A uniform external electric field $(\vec{E})$, established by the field cage, separates positive ions toward the cathode and ionization electrons toward the amplification plane at the anode. Since the number of produced electrons is proportional to the energy deposited in the gas, the collected charge preserves information about the particle energy loss (dE/dx).

Figure 4 – Schematic and working principle of a TPC



Source: Prepared by the author (2026).

During the drift process, the electrons undergo diffusion due to collisions with gas molecules (55). This effect causes the electron cloud to spread both along the longitudinal drift direction (z) and in the transverse (x, y) plane. The resulting charge distribution is approximated by a Gaussian distribution, whose width increases with drift distance.

Assuming a drift distance z, the standard deviations of the transverse (σ_T) and longitudinal (σ_L) distributions can be parameterized as:

$$\sigma_\mathrm{T} = \sqrt{\sigma_{\mathrm{T}0}^2 + D_\mathrm{T}^2 z} \tag{3.1}$$

$$\sigma_\mathrm{L} = \sqrt{\sigma_{\mathrm{L}0}^2 + D_\mathrm{L}^2 z} \tag{3.2}$$

where D_T and D_L are the transverse and longitudinal diffusion coefficients, while $\sigma_{\mathrm{T}0}$ and $\sigma_{\mathrm{L}0}$ represent constant contributions associated with the amplification stage (56).

Diffusion is one of the main factors limiting the reconstruction of low-energy events, since it progressively blurs the topology of the original ionization track as the drift distance increases.

3.1.2 Gas mixture

The choice of gas mixture directly affects the detector performance, influencing the ionization process, electron drift and diffusion, charge amplification, and light production. When charged particles traverse the gas, part of their deposited energy is converted into ionization and molecular excitation. The subsequent de-excitation of these molecules produces scintillation light, whose intensity and spectral distribution depend on the gas composition. The CYGNO detector operates with a He:CF_4 mixture in a 60:40 volume ratio at atmospheric pressure and room temperature.

Carbon tetrafluoride (CF_4) is a well-established scintillation gas (57) and produces a significant fraction of its emission in the visible range. It also exhibits low electron diffusion during drift, helping preserve the topology of ionization tracks. The addition of helium provides a light target for particle interactions and allows operation at atmospheric pressure while maintaining recoil tracks long enough to be reconstructed.

The He:CF_4 (60:40) composition was adopted following studies comparing different gas proportions, particularly the 60:40 and 70:30 mixtures (56). These studies showed that the 60:40 configuration provides a higher light yield (LY), defined as the number of detected photons per keV of deposited energy, and improved long-term detector stability while maintaining comparable energy resolution. The emission spectrum of the He:CF_4 (60:40) mixture, together with the spectral response of the optical sensors used in CYGNO, is further illustrated in Figure 6.

Recent studies have also investigated alternative gas mixtures containing additives such as isobutane (C_4H_{10}) (58) and methane (CH_4) (59). These additives modify the electroluminescence and charge properties of the He:CF_4 base mixture. Isobutane increases charge multiplication while reducing light production per avalanche electron, with limited impact on energy resolution at low concentrations. Methane improves detector stability, enabling operation at higher GEM voltages, leading to improved overall performance.

3.1.3 Gas Electron Multiplier

The amplification stage is responsible for increasing the charge signal produced in the drift region, compensating for the limited photon collection efficiency of the optical readout system. This can be achieved using Gas Electron Multipliers (60, 61), a type of MPGD widely used in gaseous detector systems due to their operational stability over long periods of time. A schematic representation of the GEM structure is shown in Figure 5.

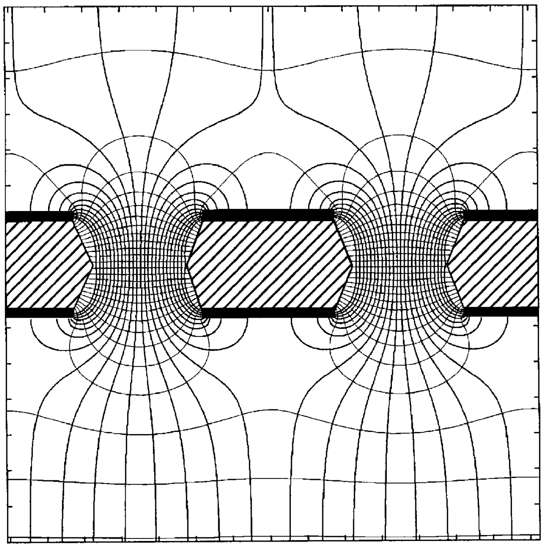
A GEM consists of a thin insulating polymer foil, metal-coated on both sides, and perforated by a regular array of microscopic holes. When a voltage difference is applied across the electrodes, the structure produces localized high-field regions that enable charge amplification.

To achieve higher and more stable gain, multiple GEM foils can be cascaded. In CYGNO, a Triple-GEM stack is employed, following preceding studies (63, 64). In this configuration, the total amplification is distributed across stages, improving operational stability and reducing the probability of discharges. This setup allows amplification gains on the order of 10^4 to 10^6 (65, 66).

3.1.4 Optical readout

The optical readout approach differs from conventional segmented charge readout systems by recording the light produced during charge amplification instead of collecting

Figure 5 – Schematics and fields of the gas electron multiplier.



Source: Extracted from (62).

the amplified charge directly. In gaseous detectors, this light can be classified as primary scintillation, produced directly by the initial particle interaction in the medium, and secondary scintillation, generated during the charge amplification process. In CYGNO, only the secondary scintillation produced in the GEM stages is recorded due to light intensity and geometrical constraints of the detector.

3.1.4.1 CMOS Camera

The light emitted from the GEMs is recorded using a sCMOS camera placed outside the active volume. An optical lens projects the emitted photons onto the sensor plane, where they are converted into electrical signals at pixel level. sCMOS devices combine high sensitivity, low readout noise, and large dynamic range, making them suitable for low-light applications such as gaseous detector readout.

CYGNO has evaluated multiple generations of Hamamatsu sCMOS cameras for optical readout, including the ORCA-Flash (67), ORCA-Fusion (68), and ORCA-Quest (69) models, ordered from oldest to most recent release date. These devices show progressive improvements in sensor size, pixel density, and readout noise, with the most recent models enabling single-photon sensitivity under low-noise operating conditions.

In this work, the Hamamatsu ORCA-Fusion camera is employed, as all results were obtained using the LIME prototype, which will be described in Section 3.3. The choice of this sensor was motivated by previous studies that identified the ORCA-Fusion as the most suitable option among the available models at the time (70). The ORCA-Quest was not included in these comparisons, as it became available only during the later stages of

LIME operation.

3.1.4.2 Photomultiplier Tubes

While the sCMOS camera records the (x, y) projection of the emitted light, Photomultiplier Tubes provide temporal information with nanosecond-scale resolution. PMTs are highly sensitive photon detectors capable of converting low-intensity light signals into measurable electrical pulses through an internal amplification process. In CYGNO, Hamamatsu R7378A PMTs (71) are employed due to their spectral response, which, together with the ORCA-Fusion camera, are well matched to the emission spectrum of the He:CF_4 gas mixture (72), as shown in Figure 6.

The signal generated by the PMTs is proportional to the amount of detected light and therefore to the charge amplified in the GEM stage. Since the arrival time of drifting electrons depends on their distance from the amplification region, the PMT waveform contains information about the longitudinal drift of the event.

The PMT signals are digitized using two acquisition systems with different sampling frequencies: a CAEN v1742 digitizer operating at 750 MS/s and a CAEN v1720 digitizer operating at 250 MS/s (73, 74). The faster digitizer provides the timing resolution required to capture short signals, while the slower digitizer offers a longer acquisition window for extended events. The specifications of both devices are summarized in Table 1. In addition to waveform acquisition, the PMTs are used as the primary trigger system, initiating data acquisition whenever the detected signals satisfy predefined conditions.

Table 1 – Fast and Slow digitizers specifications

	Resolution	**Sampling Rate**	**Window length**	**Time window**
CAEN v1742	12 bit	750 MS/s	1024 Samples	1.365 μs
CAEN v1720	12 bit	250 MS/s	4000 Samples	16 μs

Source: Prepared by the author (2024).

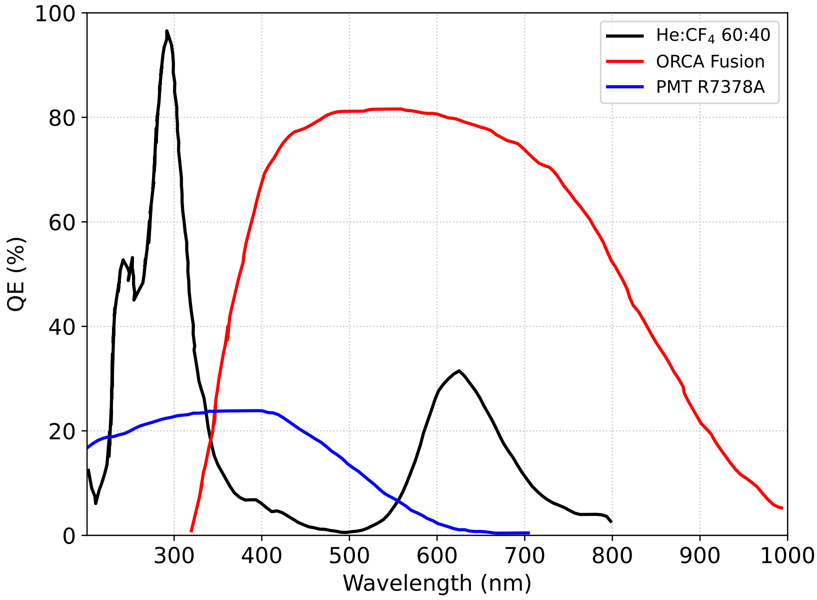
3.1.5 3D Reconstruction

The combination of the optical information provided by the sCMOS camera and the PMTs allows the reconstruction of particle trajectories in three dimensions. The camera records the transverse projection of the event on the (x, y) plane, while the PMT waveforms provide information along the drift direction through the arrival time distribution of the electrons at the amplification stage.

Assuming a known electron drift velocity v_d, the depth coordinate can be obtained from the drift time according to

$$z = v_d \cdot \Delta t, \tag{3.3}$$

Figure 6 – Normalized emission spectrum of He:CF_4 (60:40) and the quantum efficiencies of the optical sensors used in CYGNO.



Source: Prepared by the author (2026).

where Δt is the time interval between the interaction and the detection of the corresponding light signal. Likewise, the temporal extension of the PMT waveform can be related to the longitudinal extent of the event.

Figure 7 illustrates a recent reconstruction approach developed within the CYGNO collaboration (75). In this study, the PMT waveforms are used to estimate the position of the light emission along the drift direction through a Bayesian inference framework, while the camera image provides the transverse coordinates.

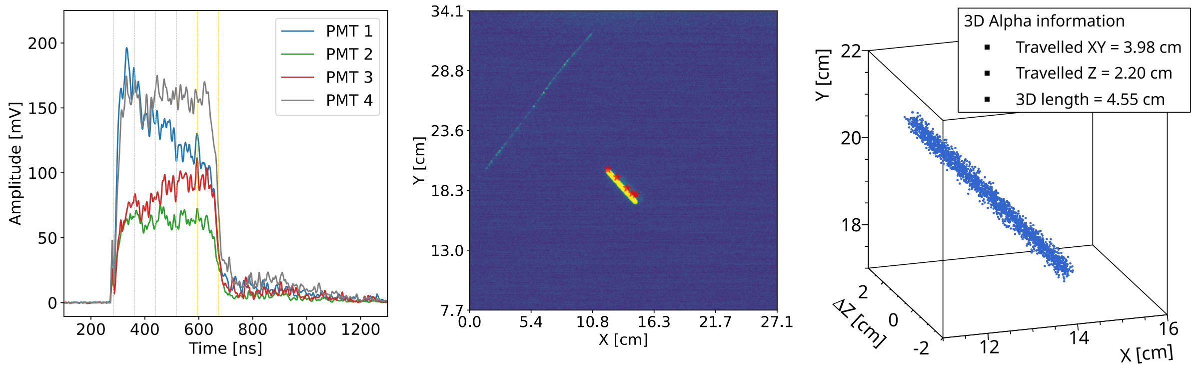
3.2 CYGNO ROADMAP

The technologies described in the previous sections form the basis of the CYGNO experiment. To validate the detector concept and guide its development toward larger-scale implementations, the collaboration has adopted a phased strategy based on successive detector generations.

The roadmap is organized into three main phases:

- **PHASE_0:** deployment of a 50-liter prototype at the INFN - Laboratori Nazionali del Gran Sasso (LNGS). This phase has been completed, and the present work is based on its results.

Figure 7 – Example of a 3D reconstruction of a particle track. (a) PMT signals; (b) overlay of reconstructed positions (red stars) on the camera image; (c) final 3D track combining PMT and camera data.



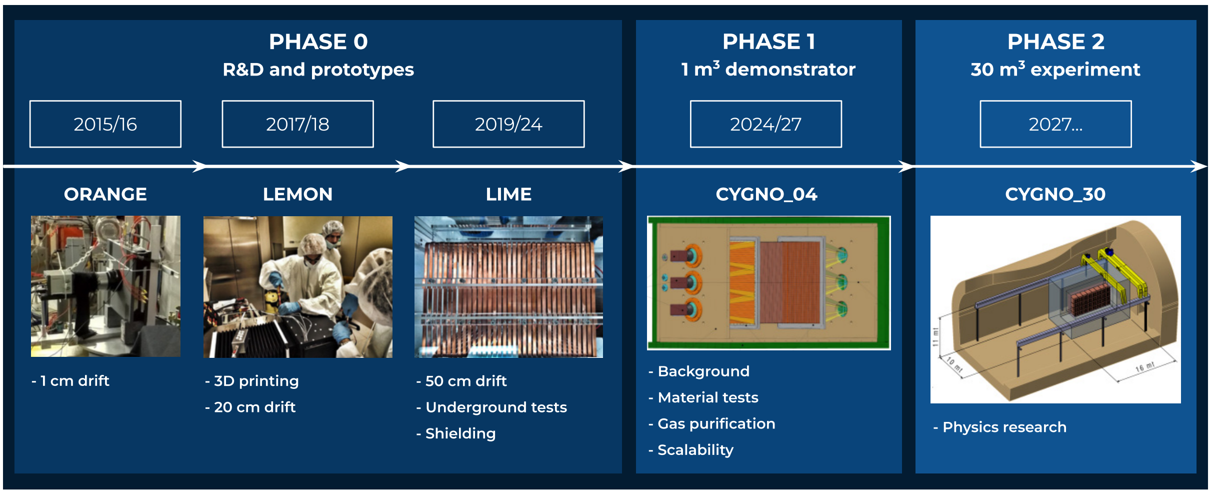
(a) (b) (c)

Source: Extracted from (75).

- **PHASE_1:** deployment of a detector with an active volume of $\mathcal{O}(1)$ m^3, aimed at validating the scalability of the experimental approach.
- **PHASE_2:** development of a large-scale detector with a target volume between 30 and 100 m^3.

The complete development plan is summarized in Figure 8.

Figure 8 – CYGNO experiment roadmap



Source: Adapted from CYGNO Collaboration (2026).

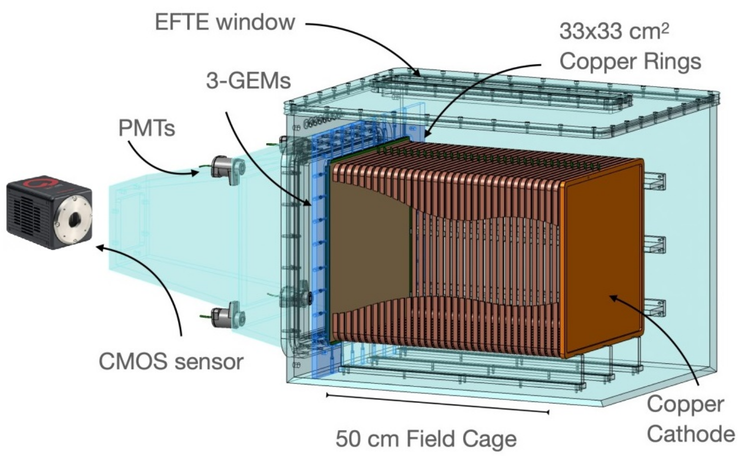
3.3 LIME PROTOTYPE

The Long Imaging ModulE (LIME) detector is the larger prototype foreseen to conclude the PHASE_0 of the CYGNO project. It has a volume of 50 L and features a structure composed of 34 copper rings with an electron collection area of 33×33 cm^2. Additionally, it includes a thin Triple-GEM, which amplifies the charge produced in the drift region, extending 50 cm in length. Each GEM foil is operated at 440 V.

The detector is filled with the He:CF_4 (60:40) gas mixture at atmospheric pressure, enclosed in a 10 mm thick plexiglass box. A drift field of 0.8 kV/cm is established by applying a 50 kV potential across the electrodes.

A Hamamatsu ORCA Fusion sCMOS camera (2304 × 2304 pixels) is integrated with four Hamamatsu R7378A PMTs, positioned at the corners of the readout area to detect secondary scintillation light, as illustrated in the schematic of Figure 9. This setup enables 3D trajectory reconstruction of the particle while ensuring precise localization of the region of interest. Real pictures of the LIME are displayed in Figure 10, showing both front and side views.

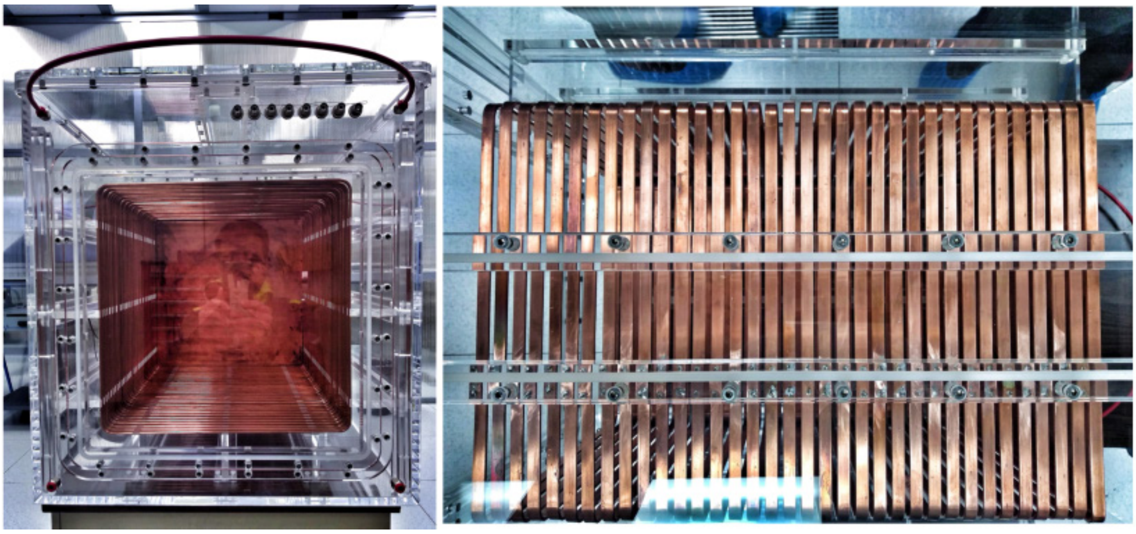
Figure 9 – Schematic of LIME prototype



Source: Extracted from CYGNO Collaboration (2026).

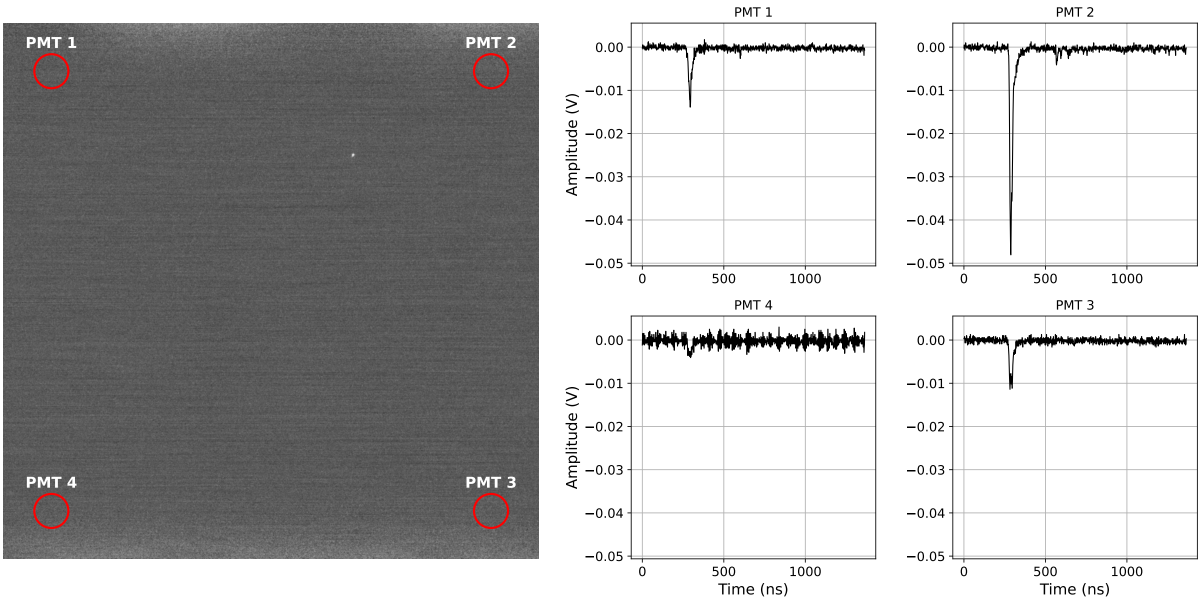
To illustrate the LIME optical readout system, Figure 11 shows a real acquired event featuring a localized energy deposition (visible as a small, bright circular spot) alongside its corresponding PMT waveforms. The geometric arrangement of the PMTs modulates the detected light amplitude based on the event's (x, y) position. In this specific event, the particle interacted closer to the top-right corner of the readout plane. Consequently, PMT 2 registers the highest signal amplitude due to its proximity. PMT 1 and PMT 3, being equidistant from the event, display similar, intermediate amplitudes, while PMT 4, located in the opposite corner, records the weakest signal.

Figure 10 – Pictures of the LIME detector, showing the front view (left) and side view (right).



Source: Extracted from (2).

Figure 11 – Camera field of view with the arrangement of the four PMTs, alongside the associated waveforms.



Source: Prepared by the author (2026).

3.4 BACKGROUND REJECTION STUDIES IN CYGNO

Background rejection strategies within the CYGNO collaboration evolved alongside the detector prototypes. Early studies using the LEMON prototype, a predecessor to LIME, relied on clustering algorithms to extract topological features from sCMOS images.

Baracchini et al. (76) characterized the ionization tracks using geometric and energy-based variables, including track length, transverse width, total energy, and energy density. The study performed a classical univariate cut-based analysis utilizing a threshold on the energy density variable. This approach achieved an 18% detection efficiency for 6 keV nuclear recoils while rejecting 96% of ^{55}Fe X-ray electron recoils.

More recently, Prajapati (77) expanded the ionization track characterization to a set of 16 features. His work evaluated classical cut-based baselines against a multivariate analysis using feature-based Machine Learning models, such as Random Forests, Gradient Boosting algorithms, and a MLP. The multivariate approaches outperformed univariate methodology, demonstrating that mapping non-linear interactions across multiple observables improves discrimination in the low energy energy regime. Furthermore, the study outlook indicated preliminary investigations into image-based Deep Learning as a potential future direction.

Motivated by these past CYGNO studies and the recent advancements of Deep Learning in other particle detectors, this work proposes a comparison between feature-based Machine Learning and image-based Deep Learning methodologies. By developing a custom Convolutional Neural Network from scratch and adapting transfer learning techniques from consolidated computer vision models, this study aims to evaluate the differences between manual feature extraction and end-to-end deep learning, thereby investigating the practical advantages of each approach for low-energy ER/NR discrimination in the LIME prototype, particularly regarding model interpretability and computational feasibility.

4 EVENT RECONSTRUCTION AND CHARACTERIZATION

The physical objective of background rejection resides in the ability to distinguish between ionization tracks that traverse the active gas volume of the detector, based on their shapes and energy deposition characteristics. Within the CYGNO experiment, this process begins with the capture of secondary scintillation light by the optical sensor. Electron recoils produce curved trajectories that increase in density toward the end of the track. In contrast, nuclear recoils result in short, straight, and localized tracks due to a higher energy density.

The raw sCMOS images contain not only the physical events of interest but also electronic noise and unavoidable tracks due to environmental radioactivity. To isolate the true events, Section 4.1 details the CYGNO reconstruction algorithm that preprocess the images and find candidate events by applying a version of the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clusterization method. Once clusters are identified, in Section 4.2 the events are characterized by performing feature extraction to compute a set of morphological and topological observables, along with energy measurements.

4.1 RECONSTRUCTION ALGORITHM

When an ionizing particle interacts within the active He:CF_4 (60:40) gas volume of the detector, primary ionization electrons are produced. These electrons drift under a uniform electric field toward the Triple-GEM amplification stage, where an electron avalanche occurs, inducing secondary scintillation light. The ORCA Fusion sCMOS camera is responsible for capturing this light to produce 2D images of the event.

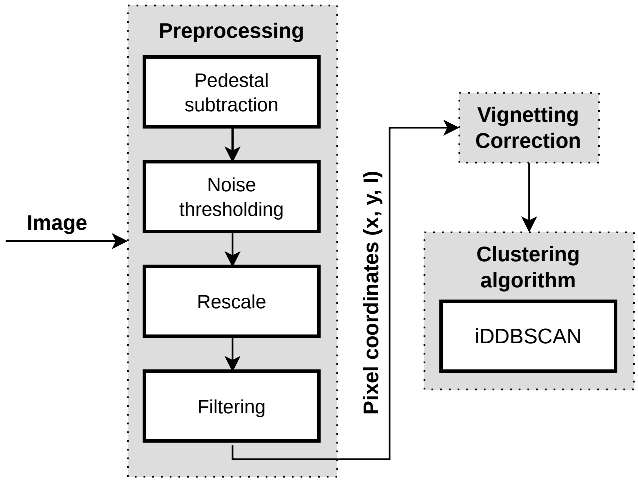
To ensure a reliable identification and characterization of these events, the raw images must be processed to remove noise artifacts before clusterization. Therefore, the CYGNO collaboration implemented a reconstruction algorithm (78) that consists of image preprocessing, vignetting effect correction, and an Intensity-based Directional DBSCAN (*iDDBSCAN*) (79). A complete overview of this sequence, from raw image processing and cluster identification, is illustrated in the flowchart of Figure 12.

4.1.1 Image preprocessing

The raw data from the sCMOS camera undergoes a preprocessing sequence divided into four main steps:

- **Pedestal subtraction:** The first step handles the electronic noise of the sCMOS sensor. The pedestal is the baseline electronic offset of the camera. To compute this, we use pedestal runs acquired by the LIME prototype with the GEMs powered off,

Figure 12 – Block diagram of the CYGNO reconstruction algorithm.



Source: Prepared by the author (2026).

ensuring no scintillation light is generated and that the image consists only of sensor noise. Each pedestal run tipically contains 100 images. We generate a pedestal map by averaging these images. For each pixel (i, j), we calculate the mean value μ_{ij} and its noise RMS (Root Mean Square), σ_{ij}. The pedestal baseline μ_{ij} is then subtracted from the corresponding pixels in the standard physics runs to produce the adjusted intensity I_{ij}:

$$I_{ij} = I_{\text{raw}}(i, j) - \mu_{ij} \tag{4.1}$$

Statistical noise fluctuations exist around this baseline, so the adjusted values I_{ij} fluctuate symmetrically around zero, producing both positive and negative ADC counts.

- **Zero suppression:** To reduce data volume and clear out sensor fluctuations, a threshold is applied to the adjusted intensities I_{ij}. Pixels whose intensity values do not exceed a statistical limit based on the local noise σ_{ij} are ignored by the clustering algorithm. We apply a threshold cut defined by the parameter n_σ, which is optimized based on the noise conditions of the specific dataset:

$$I_{ij} > n_\sigma \cdot \sigma_{ij} \tag{4.2}$$

This threshold ensures that only pixels containing meaningful charge depositions are kept, while sensor noise is rejected. Moreover, pixels in the region defined by $y < 304$ and $y > 2054$ are also masked and set to zero during this step, as this specific area of the Orca Fusion sensor is highly noisy, as detailed in (70), and would affect the trade-off between clustering performance and noise suppression efficiency.

- **Image rescale (Rebinning):** To optimize computational efficiency, a spatial rebinning operation is performed. The image is downsampled using a 4×4 pixel pooling window (*rebin 4*), which reduces the original resolution of 2304×2304 to 576×576 pixels. This operation aggregates pixel intensities into single macro-pixels, reducing the data grid dimensions for faster processing.

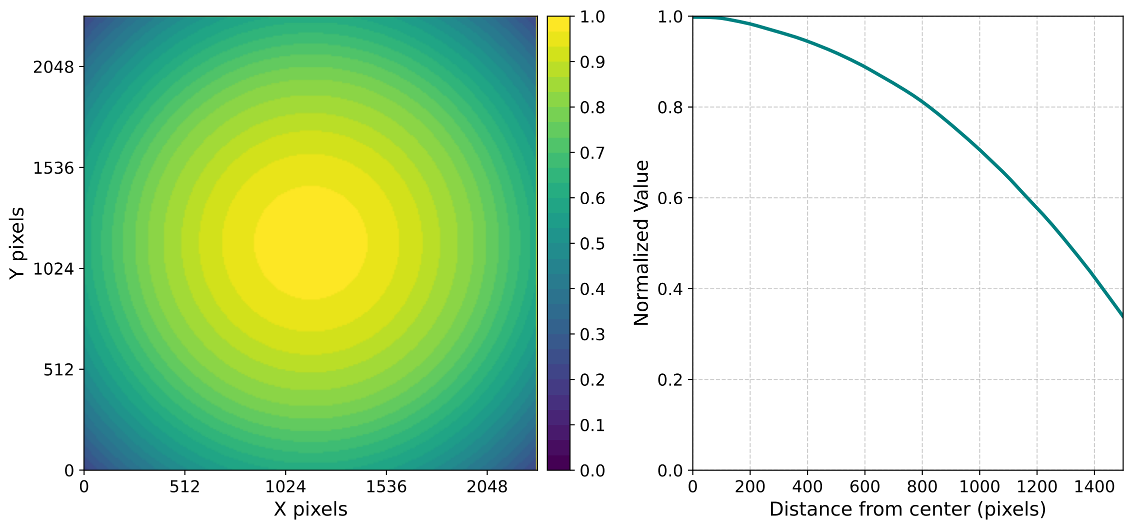
- **Median filtering:** A 2D spatial median filter is executed over the pixel matrix. This filter replaces each pixel intensity value with the median value calculated from its spatial neighborhood. This step removes noise spikes while maintaining the clear structural edges of the ionization tracks.

4.1.2 Vignetting effect correction

Optical vignetting is a geometric phenomenon where light intensity attenuates radially from the center of the camera lens toward its edges (80). Without correction, this attenuation distorts energy measurements, as a particle interaction would yield a smaller light integral near the detector edges than at the center of the readout area.

To adjust this distortion in light intensity, a vignetting calibration map is used. Images of a white paper under uniform illumination are captured, and pixel intensities are normalized to the central pixel. This correction map and its corresponding radial intensity profile are shown in Figure 13.

Figure 13 – Vignetting map (left) and the normalized intensity profile as a function of the distance from the sensor center (right).



Source: Prepared by the author (2026).

The pedestal-subtracted image is then corrected by dividing its pixel intensities by this calibration map:

$$I_{\text{corrected}}(x, y) = \frac{I_{\text{subtracted}}(x, y)}{M_{\text{vignetting}}(x, y)} \tag{4.3}$$

This ensures that a specific energy deposition produces an identical light yield independent of its 2D location.

4.1.3 iDDBSCAN

Following noise reduction and optical lens corrections, the image still contains particle tracks mixed with residual noise. To group the pixels corresponding to the truth event, the reconstruction algorithm developed by the CYGNO collaboration evolved through two implementations:

- **Intensity-based DBSCAN (iDBSCAN):** Built as an adaptation of the traditional DBSCAN (81), Intensity-based DBSCAN (iDBSCAN) (82) groups active pixels into clusters based on spatial proximity. While effective for short and dense tracks, iDBSCAN faces limitations with long, low-density trajectories left by cosmic tracks or high-energy electrons, frequently fragmenting them into separate clusters. As illustrated in Figure 14, this fragmentation often leads to incomplete track reconstruction.

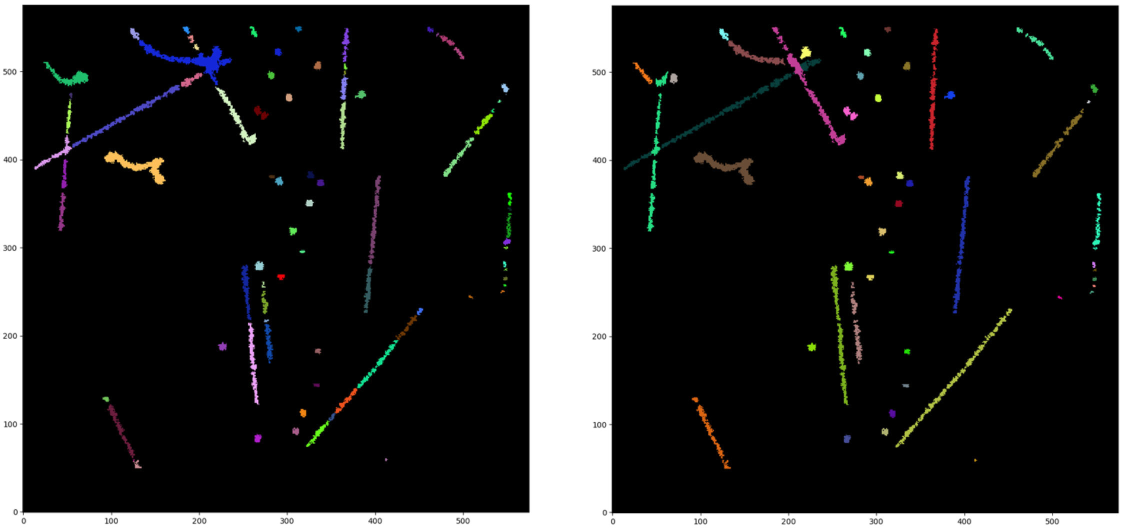
- **Directional iDBSCAN (iDDBSCAN) (79):** The current clusterization method resolves the fragmentation issue by including a directional analysis. The algorithm applies a Random Sample Consensus (RANSAC) routine to the initial clusters to search for extended linear arrangements. When a directional path is identified, a third-order polynomial curve is fitted to the track spine. As shown in Figure 15, any pixel located within a tracking width (w) around this fit line and within a directional search radius (ϵ_{dir}) of an existing point is merged into the main cluster. Residual noise pixels near the track are excluded, preventing the formation of spurious clusters.

4.2 FEATURE EXTRACTION

Following the clusterization step, the coordinate pixel arrays of the identified clusters and their associated light intensities are converted into physical variables. This feature extraction phase establishes the framework required to classify electron recoils and nuclear recoils across different energy ranges, representing a core component of this work. The separation of ER and NR tracks depends on the quality of this characterization, serving as the basis for subsequent machine learning classification algorithms.

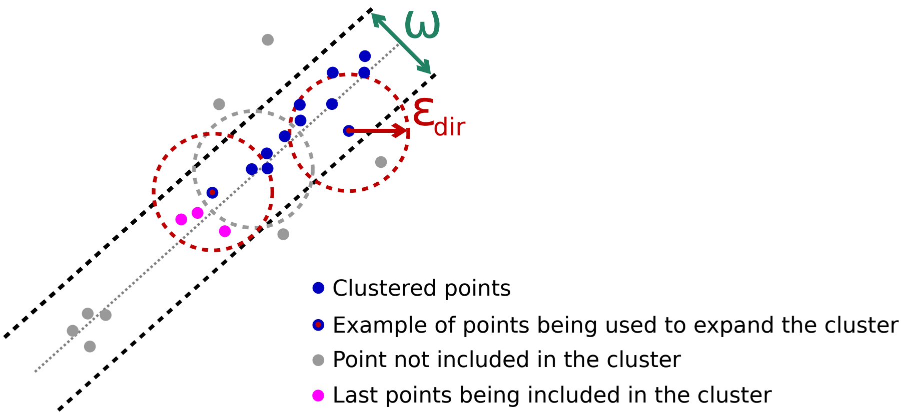
Most of the observables presented in this section were inspired from previous CYGNO studies (76, 77) and related literature (44, 45). However, new modified features

Figure 14 – Comparison between iDBSCAN (left) and iDDBSCAN (right) in an image acquired with a CYGNO detector, demonstrating the reduction of track fragmentation. Each cluster takes on a different color.



Source: Extracted from (79).

Figure 15 – Illustration of the iDDBSCAN directional search step, showing the track width (w) and directional radius (ϵ_{dir}) parameters.



Source: Adapted from (79).

are introduced, adapting standard spatial formulas to incorporate pixel charge weighting. Together, these form a set of features grouped to capture the scale, morphology, topology, and energy/spatial profiles of the events.

4.2.1 Ionization track shape preprocessing

Before extracting the features, the clusters identified by the reconstruction algorithm are preprocessed. The algorithm stores the cluster data in coordinate arrays known as the `redpix` branch. These coordinates are retrieved by mapping the *rebin 4* images (described in Section **5.5.5**) back to the original 2304×2304 sensor resolution.

This remapping procedure introduces additional low-intensity pixels around the cluster boundaries, which can increase the amount of residual background noise. To reduce this effect, a local threshold filter is applied. For each cluster, the pixel coordinates (x_i, y_i) and their respective charge intensities z_i are extracted under the following condition:

$$cx, cy, cz = \{x_i, y_i, z_i \mid z_i \geq z_{\text{threshold}}\} \tag{4.4}$$

where $z_{\text{threshold}} = 0.5$ ADC counts. To isolate the calculation from the full 2304×2304 sCMOS sensor area, a rectangular region of interest (ROI) is cropped around the coordinates with a safety pad ($p = 10$ pixels), defining a sub-grid of dimensions:

$$w_{\text{prelim}} = \max(cx) - \min(cx) + 2p \tag{4.5}$$

$$h_{\text{prelim}} = \max(cy) - \min(cy) + 2p \tag{4.6}$$

To fill small gaps inside the track caused by fluctuations in light collection, we apply a 3-pixel morphological closing operator. Even after applying iDDBSCAN, small noise clusters can still be aggregated around the main track. To eliminate these spurious clusters, a connected component is preserved only if its area satisfies $A \geq 15$ pixels and its peak charge reaches:

$$\max(z_{\text{island}}) \geq 30 \text{ ADC counts} \tag{4.7}$$

A visual comparison demonstrating this transition from the raw coordinate pixels to the processed track is shown in Figure 16.

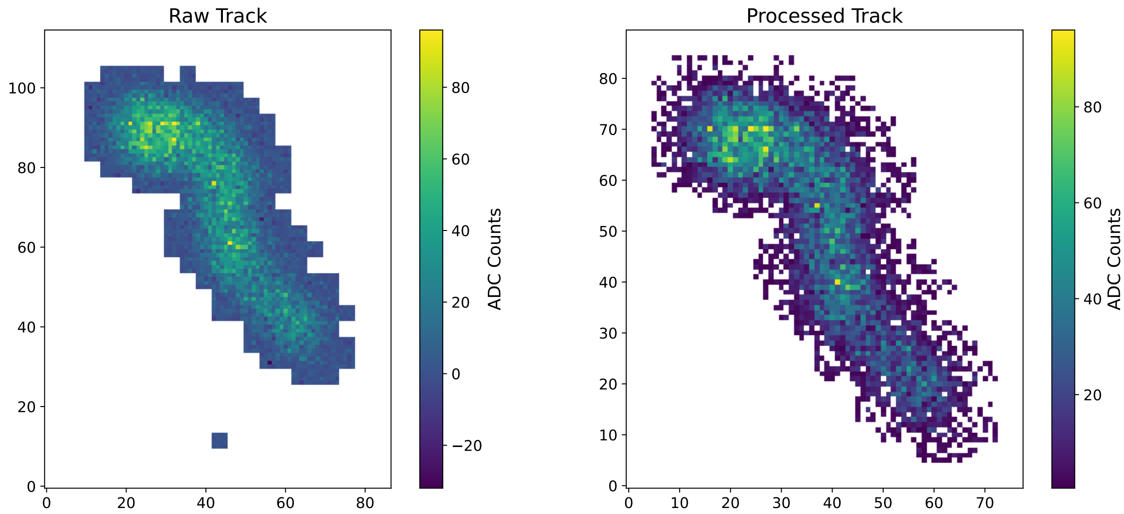
Once the core track is isolated, it serves as the definitive mask for computing topological and morphological observables. However, to calculate energy-dependent features, the cluster is expanded by 2 pixels to recover low-energy diffusion pixels that were excluded by the preliminary 0.5 ADC threshold. This is done to capture pixels with negative intensities resulting from the zero-suppression routine (Section **5.5.5**). Because the noise subtraction shifts the background variance symmetrically around zero, retaining these negative fluctuations is required to prevent a positive statistical bias in the energy variables. This dedicated array of pixel charges is denoted as $\mathbf{z}_{\text{energy}}$.

4.2.2 Energy and Global dimensions

This feature group describes the overall spatial extent and the total energy deposition of the cluster. These variables quantify the physical size, the boundary limits, and the charge-weighted center of mass of the electron cloud relative to the sensor plane.

- *Energy* (`Energy_calibrated`) (E_{cal}): This observable represents the primary ionization energy deposited by the particle, scaled into electron-equivalent energy units

Figure 16 – Visualization of the transition from raw coordinate matrix to the final integration zone, highlighting background noise suppression and track isolation.



Source: Prepared by the author (2026).

(keV_{ee}). It is calculated by dividing the integrated track charge (I) by the Light Yield calibration factor (LY). The integral I is defined as the sum of all pixel intensities (76), measured in ADC counts, within the expanded energy array:

$$I = \sum z_i \quad \text{for } z_i \in \mathbf{z}_{\text{energy}} \tag{4.8}$$

The formal experimental procedure used to extract this calibration factor from reference data is detailed in Section **5.3.2**.

- `size` (N_{pixels}): The total number of pixels contained within the expanded energy array ($\mathbf{z}_{\text{energy}}$). Because NRs deposit their energy over shorter and more concentrated trajectories due to their higher stopping power, they activate fewer pixels than ERs of equivalent energy.

- *Number of Hits* (`nhits`) (N_{hits}): The total number of active pixels composing the preprocessed track. Unlike N_{pixels}, this count includes only the pixels that survived the preliminary 0.5 ADC counts threshold. It provides a direct measure of the effective surface area illuminated by the particle's interaction.

- *Core Hits* (`nhits_highThr`): A subset of N_{hits} counting only the pixels whose individual charge deposition exceeds 10 ADC counts. By applying this cut, the variable separates the core of the track from its diffuse boundaries.

- *Root Mean Square* (`rms`): The standard deviation of the pixel charge intensities within the expanded energy array ($\mathbf{z}_{\text{energy}}$):

$$\text{rms} = \sqrt{\frac{1}{N_{\text{pixels}}} \sum_{i=1}^{N_{\text{pixels}}} (z_i - \bar{z})^2} \quad \text{for } z_i \in \mathbf{z}_{\text{energy}} \tag{4.9}$$

where $\bar{z}$ is the mean pixel intensity of the array. This variable quantifies the variance of energy deposition within the event.

- `xmin`, `xmax`, `ymin`, `ymax`: The minimum and maximum coordinate limits of the processed track bounding box along the axes (x, y) of the image. These provide the spatial boundaries of the event.
- `xmean`, `ymean`: The coordinates of the charge centroid of the ionization cloud. Rather than a purely geometric center based on the bounding box, these coordinates are calculated by weighting the spatial positions (x_i, y_i) by their respective intensities within the expanded energy array ($\mathbf{z}_{\text{energy}}$):

$$x_{\text{mean}} = \frac{\sum x_i \cdot z_i}{\sum z_i}, \quad y_{\text{mean}} = \frac{\sum y_i \cdot z_i}{\sum z_i} \quad \text{for } z_i \in \mathbf{z}_{\text{energy}} \tag{4.10}$$

The denominator in these equations corresponds to the integrated track charge (I).

4.2.3 Morphology and Track shape

The morphological characterization of an ionization track requires a direction-independent coordinate system. Because particles interact at arbitrary angles relative to the sensor plane, we extract the longitudinal and transverse axes via a charge-weighted Principal Component Analysis (PCA).

Let $\mathbf{P}$ be an $N_{\text{hits}} \times 2$ matrix containing the local pixel coordinates (x_i, y_i) of the processed track. A charge-weighted spatial covariance matrix $\mathbf{\Sigma}$ is constructed:

$$\mathbf{\Sigma} = \frac{1}{\sum z_i} \mathbf{P}^T \mathbf{W} \mathbf{P} - \mu^T \mu \tag{4.11}$$

where $\mathbf{W} = \text{diag}(z_1, z_2, \ldots, z_{N_{\text{hits}}})$ is the diagonal matrix of pixel intensities, and $\mu = [x_{\text{mean}}, y_{\text{mean}}]$ is the charge centroid vector.

The matrix $\mathbf{\Sigma}$ diagonalized through an eigenvalue decomposition:

$$\mathbf{\Sigma} \mathbf{v}_j = \lambda_j \mathbf{v}_j, \quad j \in \{1, 2\} \tag{4.12}$$

yielding the eigenvalues λ_1, λ_2 (with $\lambda_1 \geq \lambda_2$) and their corresponding orthogonal eigenvectors $\mathbf{v}_1, \mathbf{v}_2$.

The major eigenvector $\mathbf{v}_1 = [\cos\theta, \sin\theta]^T$ defines the longitudinal axis of the track. To guarantee orthogonality, the minor eigenvector (transverse axis) is set as $\mathbf{v}_2 = [-\sin\theta, \cos\theta]^T$.

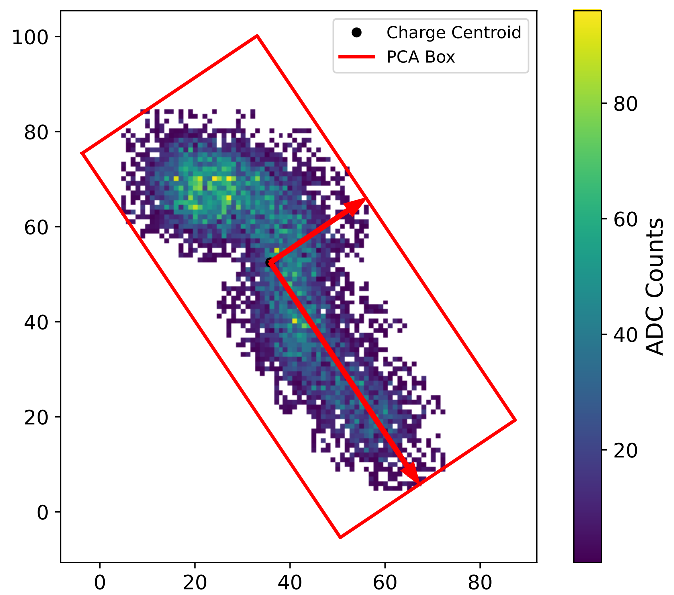
The coordinate points are then centered around the charge centroid μ and projected onto these axes, resulting in the longitudinal (u) and transverse (v) projections:

$$u_i = (x_i - x_{\text{mean}})\cos\theta + (y_i - y_{\text{mean}})\sin\theta \tag{4.13}$$

$$v_i = -(x_i - x_{\text{mean}})\sin\theta + (y_i - y_{\text{mean}})\cos\theta \tag{4.14}$$

These rotated variables define the independent principal axes necessary to compute all subsequent directional features. The result of this projection, along with its corresponding bounding box, is illustrated in Figure 17.

Figure 17 – PCA bounding box and principal axes superimposed on a preprocessed track.



Source: Prepared by the author (2026).

- *Length Along Principal Axis* (`LAPA`): Defined by the difference between the boundaries of the projected charge distribution onto the major axis (44):

$$\text{LAPA} = \max(\mathbf{u}) - \min(\mathbf{u}) \tag{4.15}$$

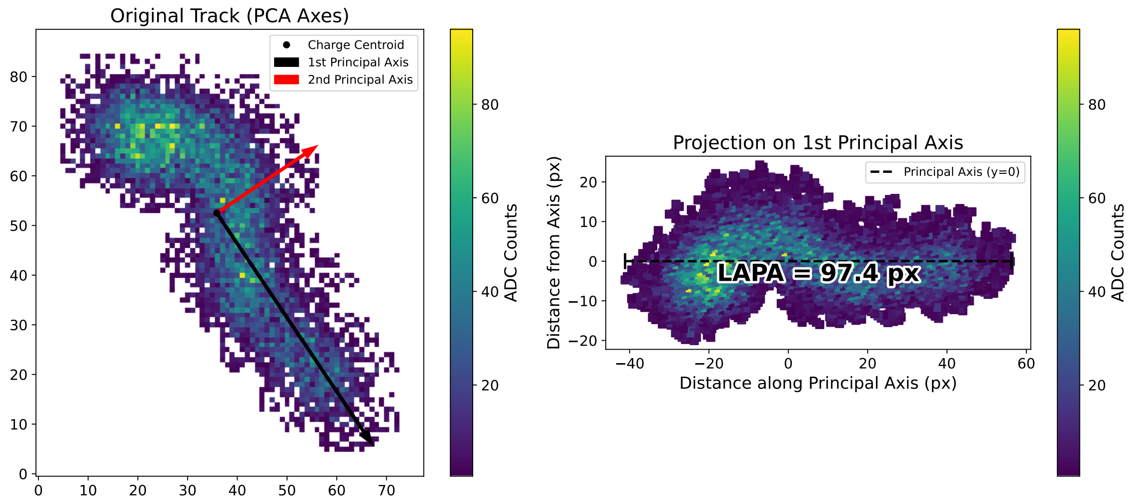
As shown in the right panel of Figure 18, it measures the maximum spatial extension of the track along its major axis.

- *Width*: Defined by the difference between the boundaries of the projected charge distribution onto the minor axis:

$$\text{Width} = \max(\mathbf{v}) - \min(\mathbf{v}) \tag{4.16}$$

This parameter captures the maximum transverse extension of the ionization cloud, measured along its minor axis.

Figure 18 – Track projection onto the first principal component axis.



Source: Prepared by the author (2026).

- *Skeleton Length* (`Skeleton_Length`): Defined as the total integrated path length of the particle's trajectory. To extract this measurement, we implement a technique called Skeletonization utilizing the `scipy` (83) and `scikit-image` (84) libraries. The calculation begins by applying a 9×9 median filter to the processed track mask to remove edge fluctuations. Internal voids caused by local charge variations are then filled using an area threshold of 7 pixels. The mask is smoothed with a discrete Gaussian filter ($\sigma = 1$) and thresholded at an intensity > 0.1.

  This layout undergoes thinning via `morphology.skeletonize` to compress the track into a one-pixel-wide backbone. If the initial length exceeds 50 pixels, an automated pruning loop executes for 5 iterations. This routine performs spatial convolutions using a discrete 3×3 connectivity kernel:

$$\mathbf{K}_{\text{neighbor}} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{4.17}$$

  Pixels detecting exactly one active neighbor are flagged as endpoints and iteratively removed to eliminate spurious branches. The step-by-step transformation from the raw mask to the finalized backbone is detailed in Figure 19.

  The sum of the remaining pixels yields the `Skeleton_Length`. Unlike `LAPA`, which measures the projected length, the skeleton captures the true trajectory of the track when it is not straight.

- *Slimness*: Represents the spatial aspect ratio of the track (76), computed from its

Figure 19 – Skeletonization process: from the processed track mask to the one-pixel-wide backbone.

Source: Prepared by the author (2026).

projected boundaries:

$$\text{Slimness} = \frac{\text{Width}}{\text{LAPA}} \tag{4.18}$$

Symmetric NRs yield a ratio close to 1, whereas extended ER tracks exhibit lower values.

- *PCA Slimness* (`Slimness_pca`): A charge-weighted alternative to the spatial aspect ratio, derived from the eigenvalues of the PCA covariance matrix:

$$\text{Slimness}_{\text{PCA}} = \sqrt{\frac{\lambda_2}{\lambda_1}} \tag{4.19}$$

The eigenvalues λ_1 and λ_2 represent the spatial variance of the charge distribution along the principal axes, therefore their square roots corresponds to the charge standard deviations. Unlike *Slimness*, which is computed from spatial track dimensions, *PCA Slimness* takes into account the pixel charge when defining the principal axes and quantifying the track extension. As a consequence, low-energy tracks with circular shapes may yield *Slimness* values greater than 1 when the LAPA is smaller than the Width. In contrast, *PCA Slimness* is derived from the ordered PCA eigenvalues ($\lambda_1 \geq \lambda_2$), ensuring that $\sqrt{\lambda_2/\lambda_1}$ is always bounded between 0 and 1. This makes *PCA Slimness* a more robust estimator of the track's true aspect ratio.

4.2.4 Longitudinal and Transverse charge profiles

Features in this category analyze the spatial distribution of charge projected along the principal axes of the track ($\mathbf{u}$ and $\mathbf{v}$) defined with the PCA. By projecting the 2D charge distribution onto the major and minor axes, we extract 1D profiles that map the directional energy deposition and the charge diffusion induced during electron drift. A

Gaussian function is then fitted to each directional profile, as illustrated in Figure 20. The physical meaning of the extracted fit parameters is described as follows:

- *Longitudinal RMS* (`longrms`) (σ_u): The charge-weighted standard deviation of the coordinates projected along the major axis (u_i), calculated as:

$$\sigma_u = \sqrt{\frac{\sum_{i=1}^{N_{\text{hits}}} z_i (u_i - \bar{u})^2}{\sum_{i=1}^{N_{\text{hits}}} z_i}} \tag{4.20}$$

  where $\bar{u}$ is the charge-weighted mean (centroid) of the longitudinal coordinates. This variable quantifies the longitudinal dispersion of the energy deposition relative to the charge centroid.

- *Lateral RMS* (`latrms`) (σ_v): The charge-weighted standard deviation of the coordinates projected along the minor axis (v_i), calculated as:

$$\sigma_v = \sqrt{\frac{\sum_{i=1}^{N_{\text{hits}}} z_i (v_i - \bar{v})^2}{\sum_{i=1}^{N_{\text{hits}}} z_i}} \tag{4.21}$$

  where $\bar{v}$ is the charge-weighted mean (centroid) of the transverse coordinates. This observable quantifies the transverse dispersion of the ionization cloud, reflecting the combined effects of the initial interaction and gas diffusion during drift.

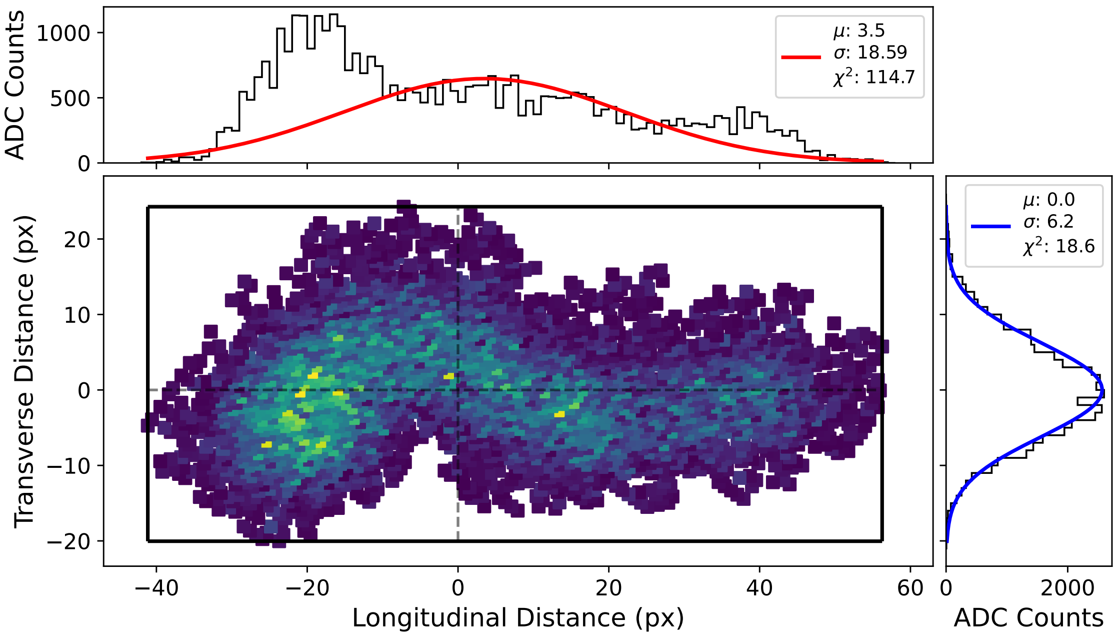
- A_u (`lgaussamp`) and A_v (`tgaussamp`): The amplitude parameters extracted from the Gaussian fits applied to the 1D longitudinal and transverse profile histograms, respectively. They represent the peak energy densities deposited along each principal axes.

- μ_u (`lgaussmean`) and μ_v (`tgaussmean`): The mean position parameters extracted from the longitudinal and transverse Gaussian fits, tracking the symmetry centers of the energy projections.

- $\sigma_{u,\text{fit}}$ (`lgausssigma`) and $\sigma_{v,\text{fit}}$ (`tgausssigma`): The standard deviation parameters extracted from the Gaussian fits (σ^T_{Gauss}). Because NR tracks exhibit less scattering than ER tracks, their transverse spread remains tighter, leading to lower values of Gaussian width.

- χ^2_u (`lchi2`) and χ^2_v (`tchi2`): The chi-squared goodness-of-fit statistics for the Gaussian fits. They quantify how closely the longitudinal and transverse charge profiles follow a Gaussian distribution. As shown in the marginal histograms of Figure 20, high-energy electron recoil tracks often deviate from a Gaussian shape along the longitudinal axis, resulting in an elevated χ^2_u. Conversely, the transverse charge density remains aligns with a Gaussian distribution, yielding a low χ^2_v.

Figure 20 – The 2D PCA aligned with its respective marginal 1D longitudinal and transverse charge profiles with a Gaussian fit.



Source: Prepared by the author (2026).

4.2.5 Charge density and Stopping power

The next group of features characterize the relationship between energy deposition and the spatial extent of the ionization track. By quantifying the concentration of charge relative to the track's trajectory, these metrics differentiate the high-density energy release of NRs from the diffuse, extended ionization patterns characteristic of ERs.

- *Maximum Density* (`MaxDen`): Defined as the most intense pixel after re-binning the track (44). We define a re-bin factor of 2. As displayed in Figure 21, this variable identifies high spatial energy density concentrations. NRs deposit energy at a higher rate, resulting in larger peak charge densities than the diffuse tracks of ERs.

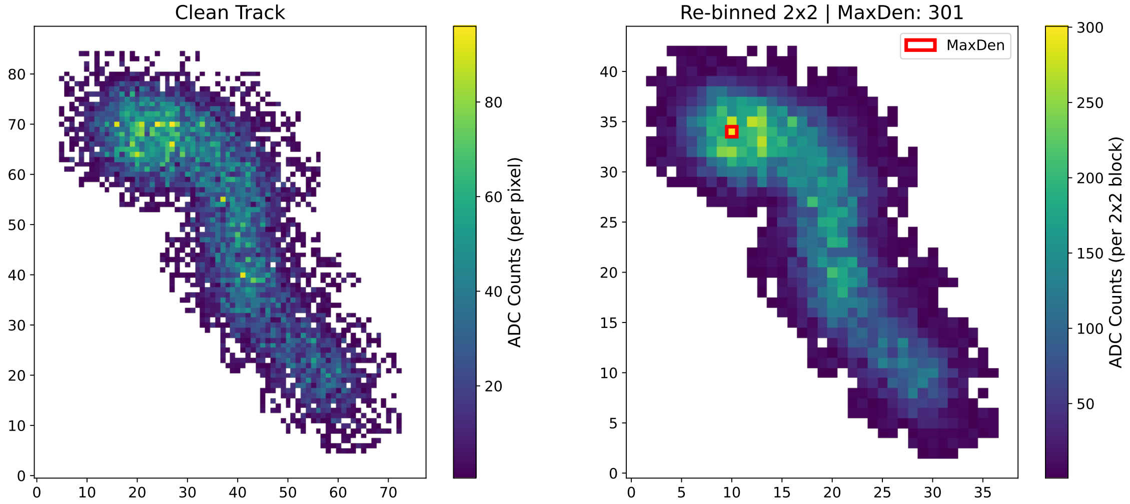
- η (`Eta`): A density metric defined as the ratio of `MaxDen` to the `Skeleton_Length` (77):

$$\eta = \frac{\text{MaxDen}}{\text{Skeleton\_Length}} \tag{4.22}$$

  NRs deposit their energy within a single, localized cluster, whereas ERs typically produce elongated tracks. Consequently, NRs exhibit higher η values, while ERs maintain lower values.

- ρ (`rho`): A charge-density metric that normalizes the combined longitudinal and

Figure 21 – Spatial 2 × 2 re-binning of the processed track to find peak charge deposition.



Source: Prepared by the author (2026).

transverse charge dispersion by the number of active pixels:

$$\rho = \frac{\sqrt{\sigma_u^2 + \sigma_v^2}}{N_{\text{hits}}} \tag{4.23}$$

This variable quantifies the charge density concentration relative to the track's surface area. Compact NRs produce low ρ values due to their high charge concentration, while the more diffuse energy deposition of ERs results in higher values.

- *Energy Loss Pattern* (`dE_dX`): The energy loss pattern, or calibrated stopping power, represents the energy deposition over the track extension along the major axis (LAPA):

$$\frac{dE}{dx} = \frac{E_{\text{cal}}}{\text{LAPA}} \quad [\text{keV}_{\text{ee}}/\text{pixel}] \tag{4.24}$$

  Due to a high stopping power, NRs result in an elevated stopping ratio compared to equivalent-energy ERs.

- *Energy Density* (`dE_dA`): The ionization energy density per unit area, represents the energy deposition across the active pixel surface area (N_{hits}):

$$\frac{dE}{dA} = \frac{E_{\text{cal}}}{N_{\text{hits}}} \quad \left[\text{keV}_{\text{ee}}/\text{pixel}^2\right] \tag{4.25}$$

  This variable helps distinguishing high-density NRs from lower-density and diffuse ERs.

4.2.6 Spatial distribution and Volume

This category quantifies the spatial distribution of the ionization track. These features differentiate the compact energy deposition of NRs from the more diffuse deposition of ERs.

- *Standard Deviation of Spatial Distribution* (`SDSD`): Defined as the radial dispersion of pixels relative to the geometric center (μ_{geom}):

$$\text{SDSD} = \sqrt{\frac{\sum_{i=1}^{N_{\text{hits}}} (\mathbf{p}_i - \mu_{\text{geom}})^2}{N_{\text{hits}}}} \tag{4.26}$$

  where $\mathbf{p}_i = (x_i, y_i)$ represents the position vector of the i-th pixel. Because NRs have short spatial profiles, their radial dispersion is lower than that of diffuse ERs. This variable was adapted from (44), where it was originally introduced as the *Standard Deviation of Charge Distribution*. However, because its calculation is based on spatial coordinates and not on pixel intensities, it has been renamed in this work.

- *Cylindrical Thickness* (`CylThick`): The lateral variance of pixels relative to the major principal axis:

$$\text{CylThick} = \frac{1}{N_{\text{hits}}} \sum_{i=1}^{N_{\text{hits}}} v_i^2 \tag{4.27}$$

  where v_i is the orthogonal distance from each pixel to the major axis. This variable measures how much a recoil track deviates from the trajectory approximated by the principal axis (44).

- *Weighted Cylindrical Thickness* (`CylThick_weighted`): The charge-weighted lateral variance, calculated as:

$$\text{CylThick}_{\text{weighted}} = \frac{\sum_{i=1}^{N_{\text{hits}}} z_i \cdot v_i^2}{\sum_{i=1}^{N_{\text{hits}}} z_i} \tag{4.28}$$

- *Spatial Uniformity* (`SpatialUnif`): Measures how uniform a recoil track is by using the Euclidean distance $d_{ik} = \|\mathbf{p}_i - \mathbf{p}_k\|$. It is defined by the standard deviation of the mean distances m_i^u:

$$\text{SpatialUnif} = \sqrt{\frac{1}{N_{\text{hits}}} \sum_{i=1}^{N_{\text{hits}}} (m_i^u - \bar{m}^u)^2}, \quad m_i^u = \frac{1}{N_{\text{hits}}} \sum_{k=1}^{N_{\text{hits}}} d_{ik} \tag{4.29}$$

  ERs produce significant shape variations in their ionization clouds, while NRs are generally uniform. Similarly to `SDSD`, this variable was adapted from (44), where it was named as *Charge Uniformity*, and has been renamed here to reflect its spatial formulation.

- *Weighted Spatial Uniformity* (`SpatialUnif_weighted`): The charge-weighted version of `SpatialUnif`, taking the individual pixel charges (z_i) into account:

$$\text{SpatialUnif}_{\text{weighted}} = \sqrt{\frac{\sum z_i (m_i^w - \bar{m}^w)^2}{\sum z_i}}, \quad m_i^w = \frac{\sum_k d_{ik} \cdot z_k}{\sum z_k} \tag{4.30}$$

Throughout this section, a total of 36 features have been defined to characterize the reconstructed events. However, for the ER/NR discrimination task, only 30 of these are retained as the input feature space for the feature-based machine learning models. The spatial variables (`xmin`, `xmax`, `ymin`, `ymax`, `xmean`, and `ymean`) are excluded from the training set. This exclusion ensures that the classification algorithms remain invariant to the specific location of the event on the image, so the models learn the morphology and energy profile of the tracks. Nevertheless, as detailed in the following chapter (Section **5.5.5**), the `xmean` and `ymean` will be reintroduced. They will be used as auxiliary variables to help in preprocessing the raw images used for training the image-based deep learning models.

5 SIMULATION AND DATASET GENERATION

The previous chapters established the operational principles of the LIME prototype and detailed the optical reconstruction and characterization of the ionization tracks. To discriminate nuclear recoils from the dominant electronic recoil background, supervised machine learning models require labeled events for training and evaluation. However, obtaining a sufficiently large experimental dataset with reliable event-by-event labels is not feasible, since calibration and background data contain contributions from multiple interaction sources and the experimental conditions do not provide direct access to the true recoil type of each event.

Monte Carlo (MC) simulations therefore provide a practical framework for generating labeled datasets. By simulating the particle interactions, the recoil type and deposited energy are known for each event, allowing controlled datasets to be constructed for supervised learning. Events can also be generated over a defined energy range, enabling the evaluation of detector response and classification performance as a function of energy.

However, the simulated events must reproduce the response of the experimental detector before they can be used reliably for this purpose. In this work, a dedicated digitization tuning procedure was developed to adjust the simulated detector response to the characteristics observed in experimental data. This procedure constitutes an important part of the dataset-generation pipeline, as it allows the simulated sCMOS images to reproduce the relevant features of the LIME response.

To establish this simulation-to-data correspondence, the remainder of this chapter is structured as follows: Section 5.1 defines the distinction between kinetic and measurable energy through the Ionization Quenching Factor. Section 5.2 details the software used to convert Geant4 event simulations into sCMOS sensor images. Sections 5.3 and 5.4 describe the energy calibration procedure and the digitization tuning developed in this work to reproduce the experimental response. Finally, Section 5.5 details the dataset composition, the energy response characterization, and the pre-filtering step to provide a clean dataset for the classification algorithms.

5.1 IONIZATION ENERGY AND QUENCHING FACTOR

In the context of DM and rare event searches, accurately modeling the detector's energy response requires a distinction between two physical energy scales: the kinetic energy deposited by a particle, denoted as nuclear recoil energy ($\mathrm{keV_{nr}}$ or simply keV), and the detectable energy that effectively produces ionization, denoted as electron-equivalent energy ($\mathrm{keV_{ee}}$).

When a charged particle traverses the TPC, it interacts with the active gas mixture, creating a trail of electron-ion pairs. The average energy required to create a single pair is known as the W-value. For induced electron recoils, the kinetic energy is dissipated through electronic excitation and ionization. Consequently, for electrons, the relationship between kinetic energy and measured ionization energy is 1:1, meaning keV $\approx$ keV$_{ee}$.

Conversely, when a heavy nucleus is scattered, a significant fraction of its kinetic energy is lost to non-ionizing elastic collisions with other atoms in the medium (85). This energy is dissipated as atomic motion and heat rather than electronic excitation. Because of this ionization deficit, a nuclear recoil will produce a smaller electron cloud than an electron recoil of the exact same kinetic energy.

To account for this effect in the simulations, the Ionization Quenching Factor (QF) is utilized. The QF provides a quantitative measure of the quenching effect, defined as the ratio of the detectable electronic energy loss to the total kinetic energy of the recoiling particle. For the specific constituent nuclei of CYGNO's He:CF_4 (60:40) gas mixture (Helium, Carbon, and Fluorine) the QF was evaluated using SRIM (Stopping and Range of Ions in Matter) software simulations (86).

The quenching factor as a function of the initial kinetic energy is illustrated in Figure 22. As demonstrated by the profiles, heavier nuclei (such as Fluorine and Carbon) suffer from a much more severe quenching effect compared to the lighter Helium nuclei. The resulting QF curves are implemented in the Geant4 toolkit (87), the software used to simulate the interactions of ERs and NRs, to adjust the energy depositions within the simulation.

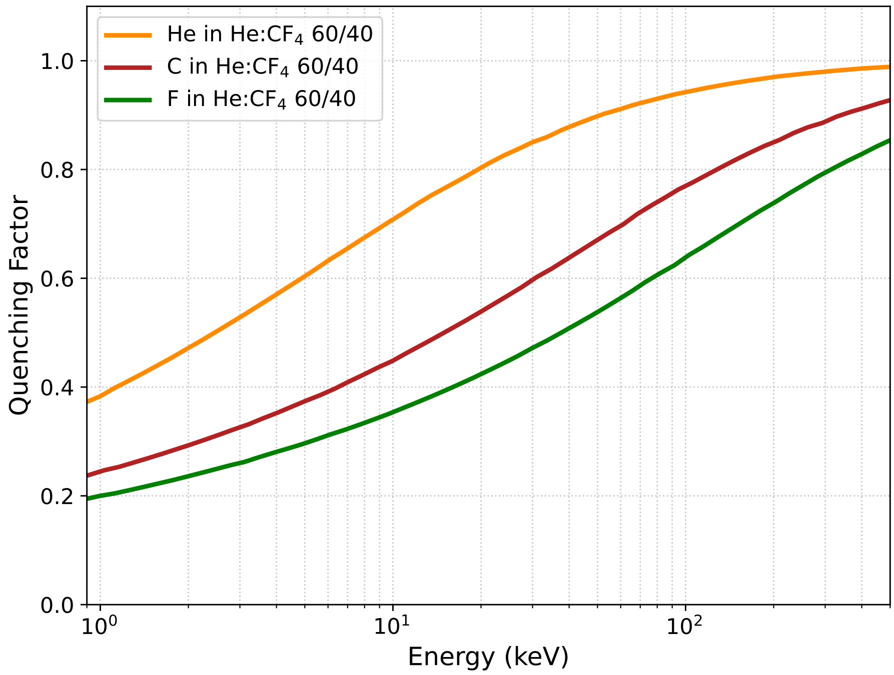
Because the optical sensor of the LIME prototype records the secondary scintillation light produced during the charge amplification process in the GEM stages, the raw kinetic energy (keV) of a particle is not directly accessible. Instead, the detector response is proportional to the ionization energy released in the gas and is therefore expressed in electron-equivalent energy. For this reason, all energy variables discussed throughout this work are reported in calibrated electron-equivalent energy (keV$_{ee}$).

5.2 DIGITIZATION SOFTWARE

The CYGNO digitization software (88) reproduces the detector response starting from the energy deposits generated by Geant4 simulations. This framework models the physical processes occurring from the primary ionization in the He:CF_4 (60:40) gas mixture to the final image production by the sCMOS sensor. A schematic representation of this simulation pipeline is illustrated in Figure 23.

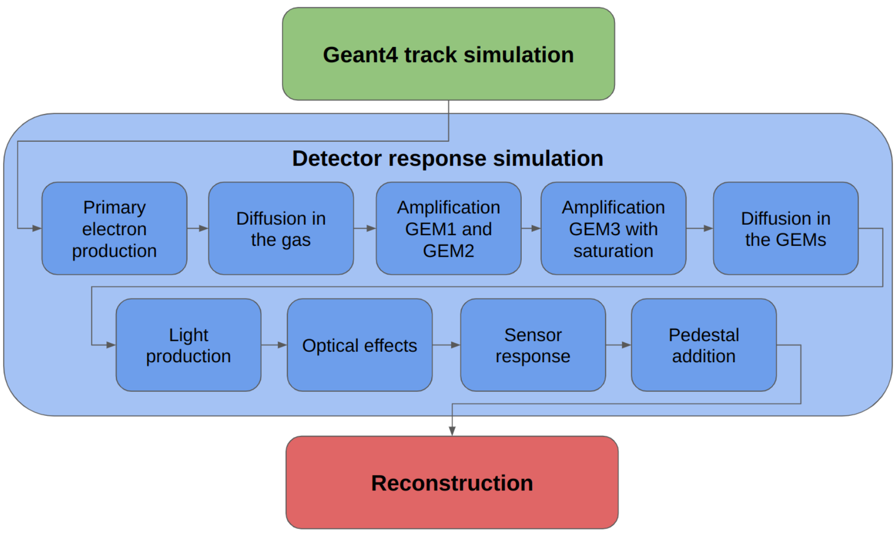
The main stages of the digitization process follow the TPC operating principle, previously described in Section **3.1.1** and illustrated in Figure 4. After the final stage, the digitized images are reconstructed using the exact same analysis code described in Section 4.1. These steps are described below (89):

Figure 22 – Ionization Quenching Factor curves for Helium (He), Carbon (C), and Fluorine (F) nuclei as a function of the initial recoil kinetic energy, parameterized for the CYGNO gas mixture.



Source: Adapted from (86).

Figure 23 – Schematic representation of the digitization framework.



Source: Extracted from (89).

- **Primary electrons production and absorption:** For each hit in the Geant4 simulation defined by an energy deposit ΔE at a specific position, the average number of primary ionization electrons produced, $\overline{N}_e$, is calculated as:

$$\overline{N}_e = \frac{\Delta E}{W_i} \tag{5.1}$$

where $W_i = 35$ eV/pair is the mean energy required to create an electron-ion pair in the gas mixture (W-value). The actual number of primary electrons, N_e^{pr}, is obtained by random sampling from a Poisson distribution. During the drift towards the GEMs along the z-direction, electrons might be captured by interactions with molecules in the gas volume, this is implemented as:

$$N_e = N_e^{pr} \cdot \exp\left(-\frac{z}{\lambda}\right) \tag{5.2}$$

where z is the distance to the amplification plane and λ is the attenuation length.

- **Electrons amplification and GEM gain:** The electrons reaching the GEMs are subsequently amplified. The average gain G of a single GEM is parameterized as a function of the GEM voltage drop ΔV:

  $$G = 0.034 \cdot \exp(0.021 \cdot \Delta V) \tag{5.3}$$

  Gain fluctuations are simulated only in the first amplification stage, where the gain of each electron is sampled from an exponential distribution. The number of electrons produced by the amplification of the first GEM, $N_e^{G_1}$, is calculated by:

  $$N_e^{G_1} = \sum_{k=1}^{N_e} \left(G_k^{G_1} \cdot \epsilon_{extr}^{G_1}\right) \tag{5.4}$$

  where $G_k^{G_1}$ is the actual gain of the GEM for each electron, sampled from an exponential distribution with a mean value of G^{G_1} (as given in the gain parametrization), and $\epsilon_{extr}^{G_1}$ is the extraction efficiency parameterized as:

  $$\epsilon_{extr} = 0.87 \cdot \exp(-0.002 \cdot \Delta V) \tag{5.5}$$

  For the second GEM stage, the number of produced electrons $N_e^{G_2}$ is calculated using the mean gain G^{G_2}:

  $$N_e^{G_2} = N_e^{G_1} \cdot G^{G_2} \cdot \epsilon_{extr}^{G_2} \tag{5.6}$$

  The third and final GEM stage, where gain saturation effects become dominant due to high charge density, is described in the subsequent item.

- **Electrons diffusion:** During the drift in the gas, electrons experience diffusion. The transverse ($\sigma_{x,y}$) and longitudinal (σ_z) diffusion effects are implemented by smearing the coordinates of each electron using Gaussian functions with standard deviations defined by:

  $$\sigma_{x,y} = \sqrt{\sigma_T^2 \cdot z + \sigma_{T0}^2} \tag{5.7}$$

$$\sigma_z = \sqrt{\sigma_L^2 \cdot z + \sigma_{L0}^2} \tag{5.8}$$

where σ_T and σ_L are the diffusion coefficients, while σ_{T0} and σ_{L0} represent constant contributions from the GEM amplification.

- **GEM gain saturation:** At the third GEM stage, high electron densities induce space-charge effects that partially shield the electric field inside the holes. To account for this effect, the number of electrons exiting a GEM channel, n_{out}, is adjusted following:

$$n_{out} = \frac{n_{in} \exp\left[\tilde{\alpha}(\Delta V - V_0)\right]}{1 + \beta\, n_{in} \left(\exp\left[\tilde{\alpha}(\Delta V - V_0)\right] - 1\right)} \tag{5.9}$$

where n_{in} is the number of electrons entering the hole, $\tilde{\alpha}$ and V_0 are parameterization constants, β determines the electron population at which amplification is lost and ΔV is the voltage drop applied at the GEM plane.

- **Light production and sCMOS sensor response:** Secondary scintillation photons are produced proportionally to the electrons emerging from the third GEM, N_e^{out}. The mean photon yield, $\overline{N_\gamma}$, is given by:

$$\overline{N_\gamma} = LY \cdot N_e^{out} \tag{5.10}$$

where LY is the light yield of the gas mixture. The actual photon count is sampled from a Poisson distribution. The fraction of photons collected by the optical system is determined by the solid angle acceptance Ω, yielding:

$$N_\gamma^{Pix} = N_\gamma^{tot} \cdot \Omega \tag{5.11}$$

To simulate the lens vignetting effect, previously described in Section **4.1.2**, the photon grid is multiplied by the detector response map $M_{\text{vignetting}}(x, y)$ shown in Figure 13:

$$M_{\text{vignetted}}(x, y) = M_{\text{photon}}(x, y) \cdot M_{\text{vignetting}}(x, y) \tag{5.12}$$

This is the same calibration map used by the reconstruction algorithm to compensate for the optical attenuation by applying the inverse operation.

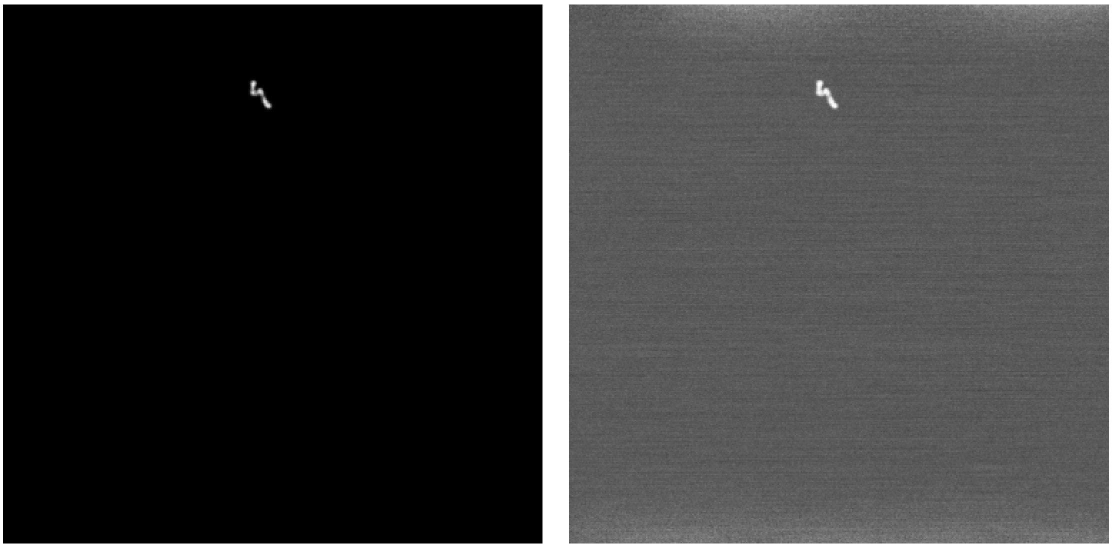
Finally, the camera converts the photon signal into digital counts according to

$$\text{ADC} = C_{\text{conv}} \cdot M_{\text{vignetted}}(x, y) \tag{5.13}$$

where $C_{\text{conv}} = 4$ is the conversion factor.

- **Pedestal addition:** The final simulated image is completed by superimposing the digitized track onto a real sCMOS pedestal image, acquired from pedestal runs (mentioned in Section **4.1.1**). This ensures that the simulated events contain the same sensor noise characteristics present in the experimental data before being processed by the reconstruction algorithm. A visual demonstration of this step is shown in Figure 24.

Figure 24 – Comparison of a simulated track before (left) and after (right) the superimposition of real pedestal noise.



Source: Prepared by the author (2026).

5.3 ENERGY CALIBRATION

The detector response is directly impacted by environmental and operational conditions, such as pressure, temperature, oxygen, humidity, and gas mixture purity. Figure 25 illustrates this temporal variation, demonstrating how the measured Light Yield fluctuates across different acquisition periods.

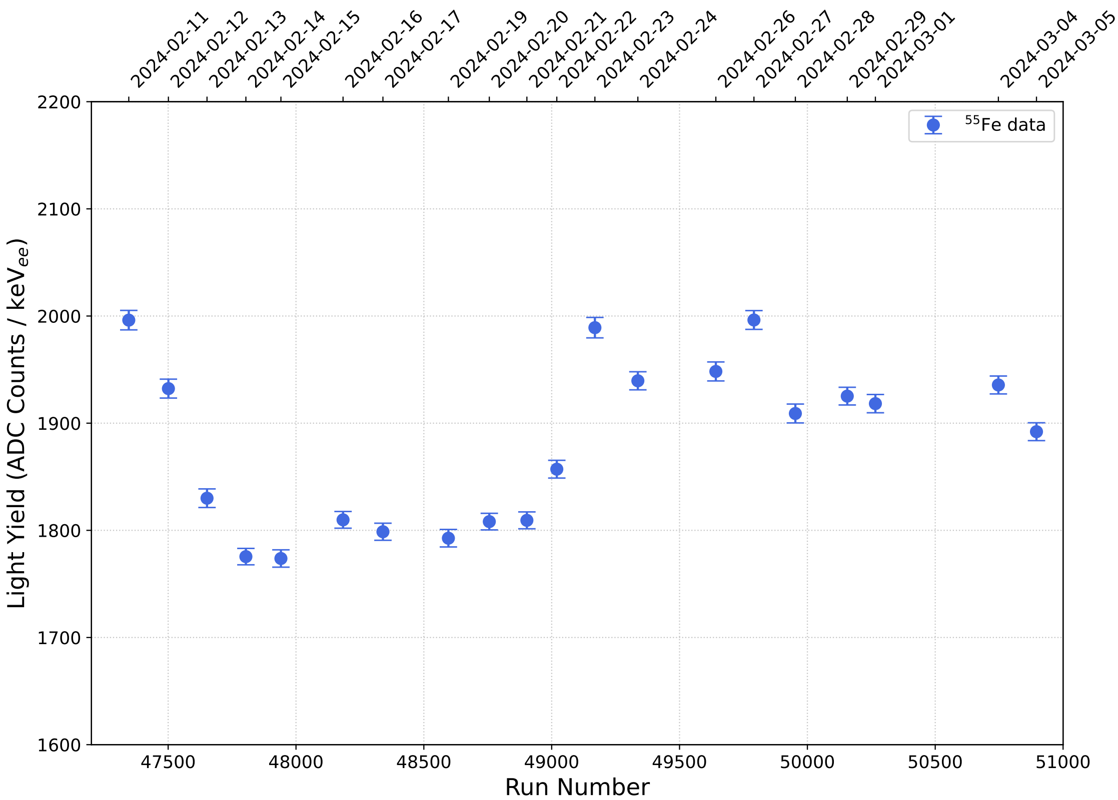
5.3.1 Daily Calibration experimental setup

To monitor the detector stability and establish an energy scale, the CYGNO collaboration conducted daily data acquisitions, referred to as Daily Calibration runs. These runs utilize an external radioactive ^{55}Fe source, which predominantly emits 5.9 keV X-rays. When these X-rays interact in the gas, they induce characteristic electron recoils. These events produce compact, symmetric, and circular charge distributions in the images acquired by the camera, commonly referred to as “spots”. An example of an image containing multiple ^{55}Fe spots is shown in Figure 26.

Figure 27 compares the cluster integral distribution of a run without the source to a calibration run using the ^{55}Fe source, illustrating the clear peak associated with the 5.9 keV energy.

The calibration routine maps the detector response along the drift direction, performing a z-scan across the field cage. Standard acquisitions tipically record 100 images at five position steps:

Figure 25 – Time evolution of the LIME prototype Light Yield, illustrating fluctuations due to environmental conditions.



Source: Prepared by the author (2026).

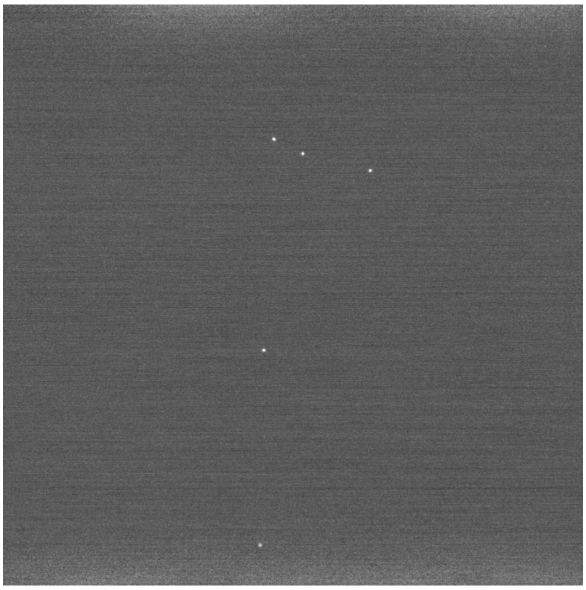
- **Step 1:** $z = 50$ mm
- **Step 2:** $z = 150$ mm
- **Step 3:** $z = 250$ mm
- **Step 4:** $z = 350$ mm
- **Step 5:** $z = 465$ mm

This z-scan process allows for the evaluation of longitudinal and transverse electron diffusion effects during drift. However, for the specific purpose of energy calibration, the procedure detailed in Section **5.3.2** uses the data acquired at **Step 3**. This central position ($z = 250$ mm) is chosen to avoid edge-field distortions, being a stable reference to convert the integrated ADC counts to energy.

5.3.2 Light Yield extraction procedure

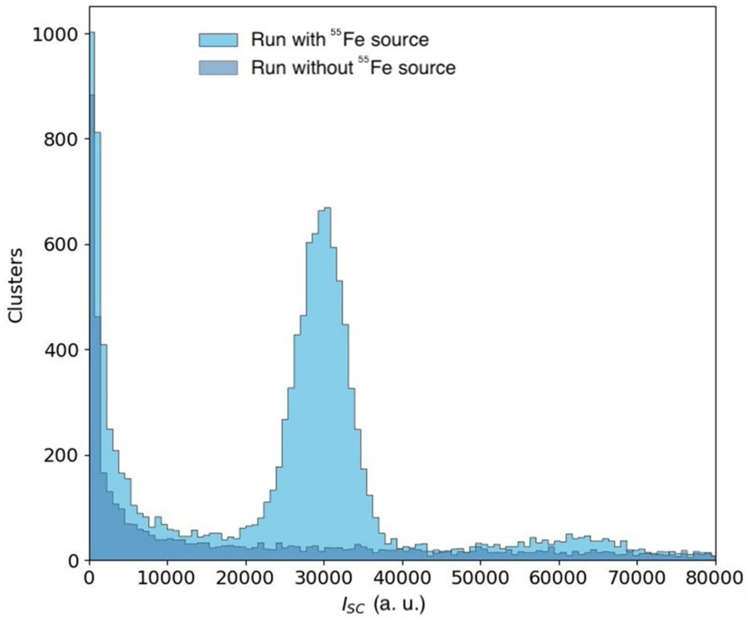
To convert the track integral (I), measured in ADC counts, into an energy estimate, a calibration procedure is performed. The cluster integral (I) is obtained by summing

Figure 26 – Example of an image acquired during a ^{55}Fe Daily Calibration run.



Source: Prepared by the author (2026).

Figure 27 – Comparison of the cluster integral distribution between a run with and without the ^{55}Fe source.



Source: Extracted from (90).

the pixel intensities within the track boundary, as defined in Section 4.2. To select clear calibration events and avoid distortions or background tracks, the following pre-selection cuts are applied:

1. **Radial cut:** Tracks must be located within the central region of the sensor. This cut minimizes the impact of lens vignetting (Figure 13), preventing distortions from affecting the calibration. The cut is defined by:

$$R = \sqrt{(x_{\text{mean}} - 1152)^2 + (y_{\text{mean}} - 1152)^2} < 700 \text{ pixels} \tag{5.14}$$

2. **Slimness cut:** As demonstrated in Section **5.3.1**, the events from the ^{55}Fe acquisitions produce circular and symmetric spots. To preferentially select these events and reject background tracks, a geometric condition is applied based on the *Slimness* defined in Section 4.2:

$$\text{Slimness} = \frac{\text{Width}}{\text{LAPA}} > 0.9 \tag{5.15}$$

The cluster integral distribution of the selected events is then histogrammed and fitted with a Gaussian function:

$$f(x \mid A, \mu, \sigma) = A \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{5.16}$$

where x represents the cluster integral in ADC counts, A is the amplitude of the peak, μ is the mean of the distribution, and σ is the standard deviation.

A second selection is then applied around the fitted peak. Events falling outside the interval $[\mu - 2\sigma, \mu + 2\sigma]$ are rejected. The mean of the remaining events corresponds to the mean cluster integral ($\bar{I}$).

To calculate the Light Yield (LY), the mean integral obtained specifically at Step 3 is divided by the known energy of the reference source:

$$LY = \frac{\bar{I}}{E_{\text{source}}} \quad [\text{ADC/keV}_{\text{ee}}] \tag{5.17}$$

For real experimental data, the reference energy is $E_{\text{source}} = 5.9$ keV, corresponding to the ^{55}Fe source. For the MC simulation, $E_{\text{source}} = 6.0$ keV simulated ERs are used.

The resulting LY defines the detector energy scale and is used to convert the raw integral (I) of any reconstructed track into the calibrated energy feature (E_{cal}):

$$E_{\text{cal}} = \frac{I}{LY} \quad [\text{keV}_{\text{ee}}] \tag{5.18}$$

5.4 DIGITIZATION PARAMETER TUNING

The digitization framework described in Section 5.2 depends on a set of physical models and detector-specific parameters. To ensure that the simulated dataset accurately mirrors experimental data and accounts for the environmental variations described in Section 5.3, the digitization software parameters had to be empirically tuned.

5.4.1 Optimization via Optuna

The tuning process aims to minimize the difference between the mean integral curve of the simulated events and the real Daily Calibration data across the five z-scan steps. The minimization is performed using Optuna (91), an open-source hyperparameter optimization framework that employs Bayesian optimization algorithms. This approach efficiently explores the parameter space by building a probability model of the objective function.

To align the simulation with realistic experimental conditions, the real calibration data was selected from the Americium-Beryllium (AmBe) Campaign. The AmBe Campaign consists of a series of data acquisitions using an AmBe radioactive source, which induces nuclear recoils. The selected calibration dataset includes runs 100026 to 100030, corresponding to Steps 1 to 5, respectively, acquired on December 19, 2024. Using calibration data acquired during the same campaign ensures a consistent MC/data match for further real data analysis.

For the MC simulation, 100 ER events with an energy of 6.0 keV were generated at the center of the image for each of the five z-positions. This specific energy was utilized as the closest available simulated energy to the 5.9 keV ^{55}Fe emission. The mean integral of the spots was extracted following the procedure described in Section **5.3.2**, applying a Gaussian fit to the distribution and calculating the mean within a $\pm 2\sigma$ window.

The optimization target is the Mean Squared Error (MSE) between the simulated and real mean integral values, defined as:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \left(\overline{I}_{MC,i} - \overline{I}_{Data,i} \right)^2 \tag{5.19}$$

where $N = 5$ is the number of z-steps, $\overline{I}_{MC,i}$ is the simulated mean integral, and $\overline{I}_{Data,i}$ is the experimental mean integral at step i.

Minimizing this objective function forces the integral of a 6.0 keV simulated event to match a 5.9 keV experimental event. This induces a gain suppression of about 1.7% in the simulation. However, this 0.1 keV energy difference does not significantly impact the tuning, since it is corrected during the final energy calibration step, when the procedure normalizes both the real data and the simulation by their true initial energies (E_{source}).

Although the optimization objective is defined exclusively by the integral MSE, eight additional observables, extracted as detailed in Section 4.2, are simultaneously evaluated: `size`, `nhits`, `lgaussamp`, `lgausssigma`, `tgaussamp`, `tgausssigma`, `LAPA`, and `width`. The objective of monitoring these features is to ensure that the circular shape and symmetrical energy distribution characteristic of ^{55}Fe spots are correctly modeled by the digitization.

The optimization focuses on three digitization parameters introduced in Section 5.2: the GEM gain scaling factor (`GEM_gain_factor`), the attenuation length λ (`absorption_l`), and the GEM gain saturation parameter β (`beta`). The search boundaries for these three variables were established based on previous digitization studies (89). Additionally, the diffusion coefficients σ_T (`diff_coeff_T`) and σ_L (`diff_coeff_L`), along with their constant contributions σ_{T0} (`diff_const_sigma0T`) and σ_{L0} (`diff_const_sigma0L`) are utilized as free variables for fine-tuning to match the additional observables. The default values, the defined search space for the parameters, and the final optimized values are summarized in Table 2.

Table 2 – Digitization parameters configured for the Optuna optimization, including default values, the search space, and the final tuned values.

Parameter	**Default Value**	**Search Space**	**Optimized Value**
`GEM_gain_factor`	0.03	[0.026, 0.038]	0.033
`absorption_l`	1350 mm	[600, 2500]	1300 mm
`beta`	0.8×10^{-5}	$[0.5, 5.0] \times 10^{-5}$	0.671×10^{-5}
`diff_const_sigma0T`	0.13475	-	0.105
`diff_coeff_T`	0.0143819	-	0.019
`diff_const_sigma0L`	0.0676	-	0.03
`diff_coeff_L`	0.0103483	-	0.0052

Source: Prepared by the author (2026).

The results of the optimization across the 9 evaluated observables are illustrated in Figure 28. The ratio panels demonstrate that the discrepancy between the MC simulation and the real experimental data is maintained within a 10% margin across all steps and features.

By applying the optimized digitization parameters obtained via the Optuna framework, the Light Yield extraction procedure described in Section **5.3.2** was executed to determine the final conversion factors. Figure 29 displays the resulting integral distributions and the respective Gaussian fits for both the experimental calibration data and the optimized MC simulation at Step 3. The final calculated Light Yield values are:

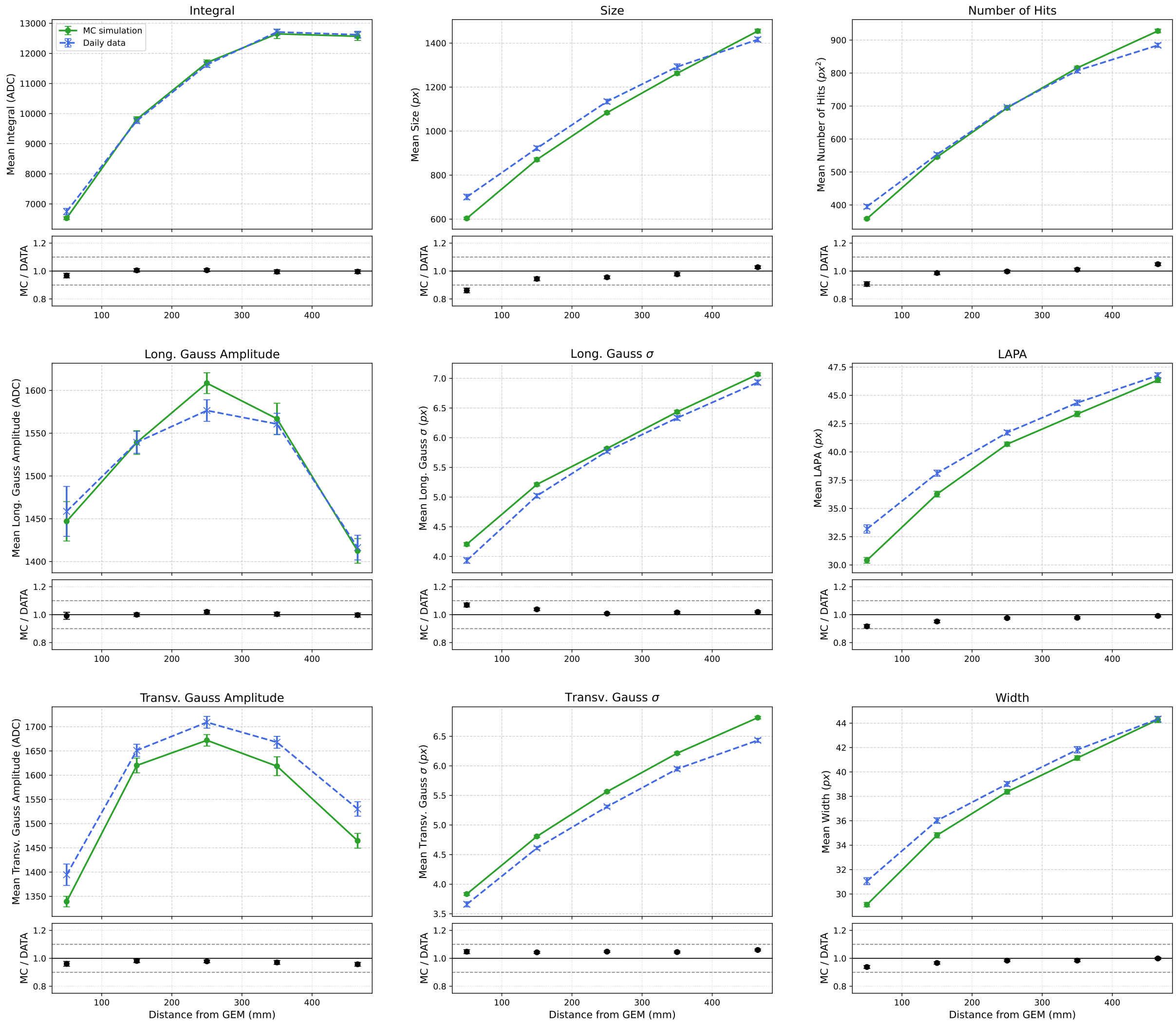
- **Daily Calibration data:** $LY_{\mathrm{Data}} = 1970.45 \pm 14.92 \quad [\mathrm{ADC/keV_{ee}}]$
- **MC Simulation:** $LY_{\mathrm{MC}} = 1945.36 \pm 14.89 \quad [\mathrm{ADC/keV_{ee}}]$

The close agreement between these values indicates that the tuned digitization parameters successfully reproduce the charge response of the detector.

5.5 DATASET PREPARATION

With the digitization parameters tuned to reproduce the experimental detector response, the final dataset can be generated and processed. This section details the sample

Figure 28 – Comparison of the 9 features across the 5 steps between real ^{55}Fe data and the tuned MC simulation.



Source: Prepared by the author (2026).

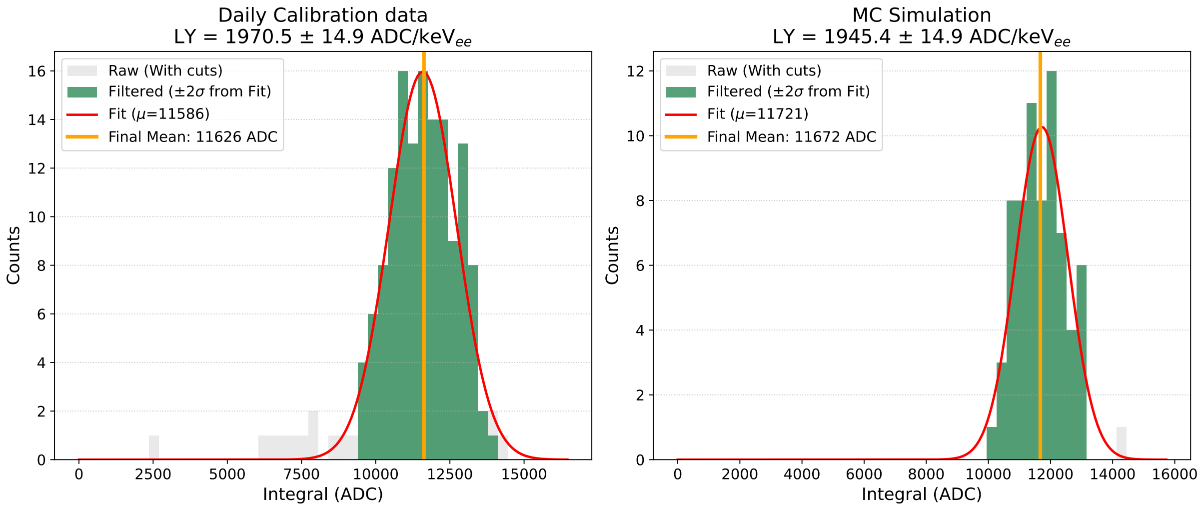
distribution across the discrete energy levels, the characterization of the energy response, and the feature-based pre-filtering applied to isolate physical tracks from sensor noise.

5.5.1 Sample distribution

To assess the machine learning models' performance across the low-energy thresholds critical for DM and rare event searches, the dataset was designed with discrete, controlled energy levels. Initially, utilizing the Geant4 toolkit, particles were injected into the He:CF_4 (60:40) active gas volume.

For the background class, Electron Recoils were simulated with kinetic energies matching the target values, as their detectable energy is identical to their initial kinetic energy. For the signal class, representing Nuclear Recoils of Helium (He), Carbon (C), and

Figure 29 – Light Yield extraction procedure applied to Daily Calibration data (left) and MC Simulation (right) at Step 3.



Source: Prepared by the author (2026).

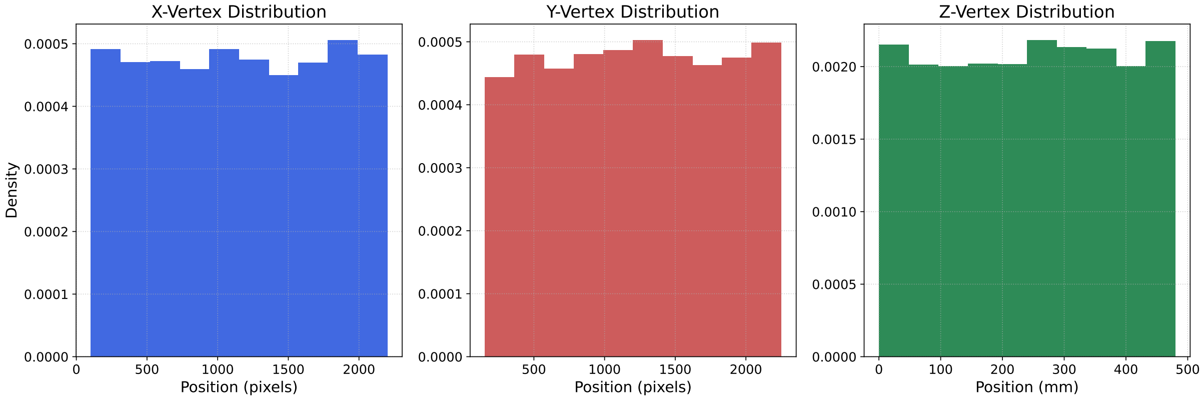
Fluorine (F), generating events at identical kinetic energies would result in significantly lower detectable energies due to the quenching effect, as previously detailed in Section 5.1. This would introduce a trivial discrimination mechanism, where models could separate classes based on light yield differences rather than topological features. To prevent this bias, the simulated energies in the Geant4 for the NR species were scaled up using the Quenching Factor curves (Figure 22). For example, a 10 keV$_{ee}$ ER is effectively compared against a 10 keV$_{ee}$ NR in terms of measured energy.

Therefore, the dataset was generated targeting fixed electron-equivalent energies: 3, 5, 10, 15, 20, and 25 keV$_{ee}$. Initially, 1000 events were generated for each particle species at each discrete keV$_{ee}$ step. To avoid introducing position-dependent biases, the interaction positions were uniformly distributed throughout the active volume of the TPC. This corresponds to a uniform probability density across the x and y dimensions of the 2304×2304 camera plane, as well as along the z-axis representing the drift direction. This spatial distribution is illustrated in Figure 30.

Following the primary generation, the simulated tracks are processed by the digitization software, as described in Section 5.2. The parameters of the digitization software were set according to the optimized configuration detailed in Section **5.4.1** and summarized in Table 2.

After digitization, a spatial filter is applied to the dataset. As previously detailed in Section **4.1.1**, the upper and bottom regions of the Orca Fusion sensor exhibit excessive noise, therefore pixels satisfying $y < 304$ and $y > 2054$ are masked and set to zero during the Zero Suppression stage. Consequently, simulated events whose coordinates fall within

Figure 30 – Histograms of the simulated tracks along the x, y, and z (drift) axes, demonstrating the uniform distribution across the detector volume.



Source: Prepared by the author (2026).

these regions would be ignored or fragmented by the clustering algorithm. To ensure a clean sample, these events are excluded from the dataset before the reconstruction phase.

The final dataset composition, after applying the geometric cuts, is summarized in Table 3.

Table 3 – Dataset composition for ERs and NRs.

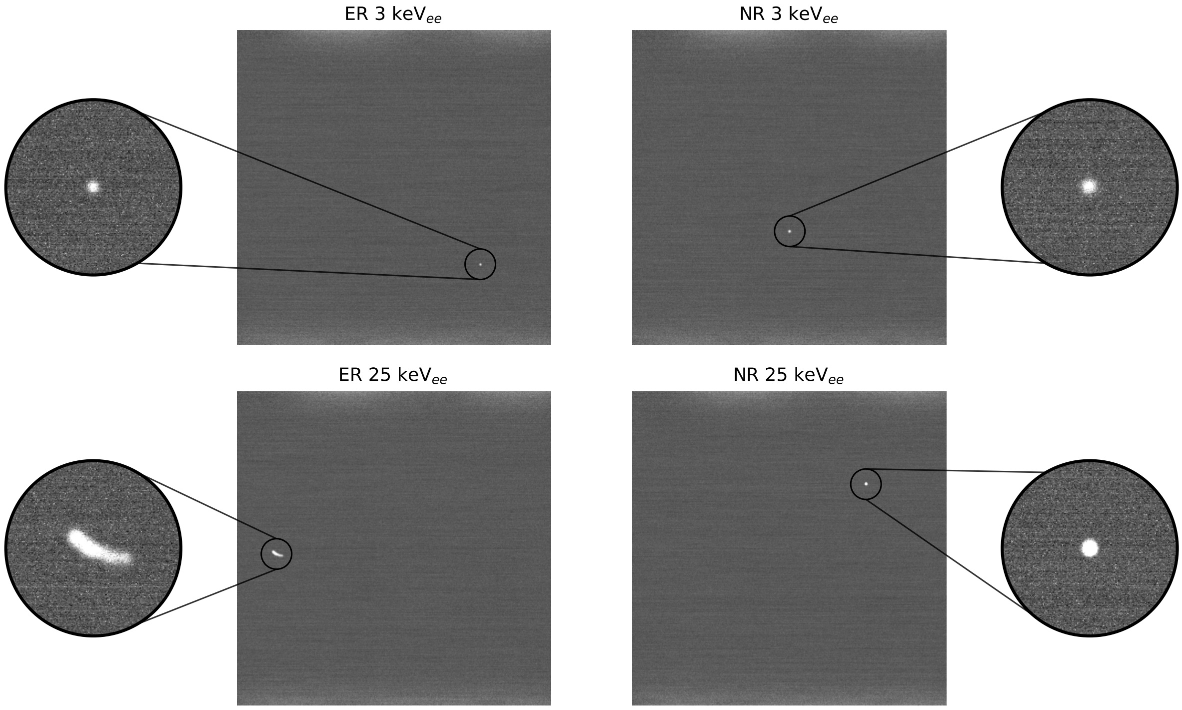
Energy (keV_{ee})	ER	NR (He)	NR (C)	NR (F)
3	826	837	839	842
5	833	838	842	842
10	796	829	833	828
15	817	837	847	828
20	816	838	841	834
25	813	840	841	855

Source: Prepared by the author (2026).

To visually illustrate the differences between the background and signal classes, Figure 31 displays a side-by-side comparison of typical ER and NR simulated events at both the lowest (3 keV_{ee}) and highest (25 keV_{ee}) evaluated energies. While at lower energies the events appear compact and circular, at higher energies, the topological distinction becomes highly pronounced. Electron recoils exhibit extended, tortuous, and diffuse ionization tracks, while nuclear recoils maintain a highly dense and circular charge profile. These differences form the basis for the pattern recognition performed by the ML algorithms.

Finally, the events are processed by the reconstruction algorithm, and the found clusters undergo the feature extraction process detailed in Section 4.2. The complete set of features extracted for each simulated event is saved into a single `.csv` file. This format

Figure 31 – Visual comparison of simulated ERs and NRs (represented by Helium) at 3 $\mathrm{keV_{ee}}$ and 25 $\mathrm{keV_{ee}}$.



Source: Prepared by the author (2026).

structures the dataset utilized for all subsequent ML analyses.

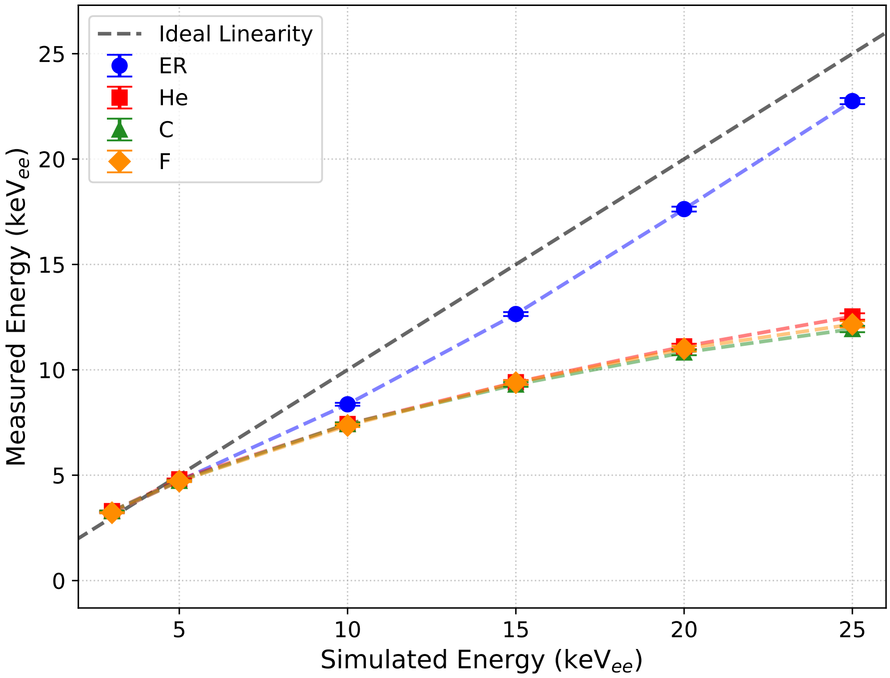
5.5.2 Energy response

One of the advantages of simulated data is the direct access to the true deposited energy of each event. By evaluating the correlation between this original energy simulated in Geant4 and the final calibrated energy obtained after reconstruction, we can characterize the detector response. Figure 32 shows the measured energy as a function of the simulated energy for both ER and NR species (He, C, and F).

For the ER class, the measured energy perfectly follows the ideal linearity reference line ($y = x$) in the low-energy region up to approximately 5 $\mathrm{keV_{ee}}$. As the simulated energy increases, a slight deviation from the ideal linearity is observed. This behavior is primarily attributed to charge saturation effects in the Triple-GEM stack, as detailed in Section 5.2. However, because higher-energy electrons undergo significant scattering and diffusion in the gas, their tracks become extended and tortuous. This spatial spread distributes the primary electrons across a wider area of the amplification plane, mitigating local saturation effects.

For the NR classes (He, C, and F), the energy response is also linear at lower energies. However, for simulated energies above 10 $\mathrm{keV_{ee}}$, a pronounced non-linearity is

Figure 32 – Measured energy as a function of the simulated energy for ER and NR, demonstrating linearity trends and charge saturation effects.



Source: Prepared by the author (2026).

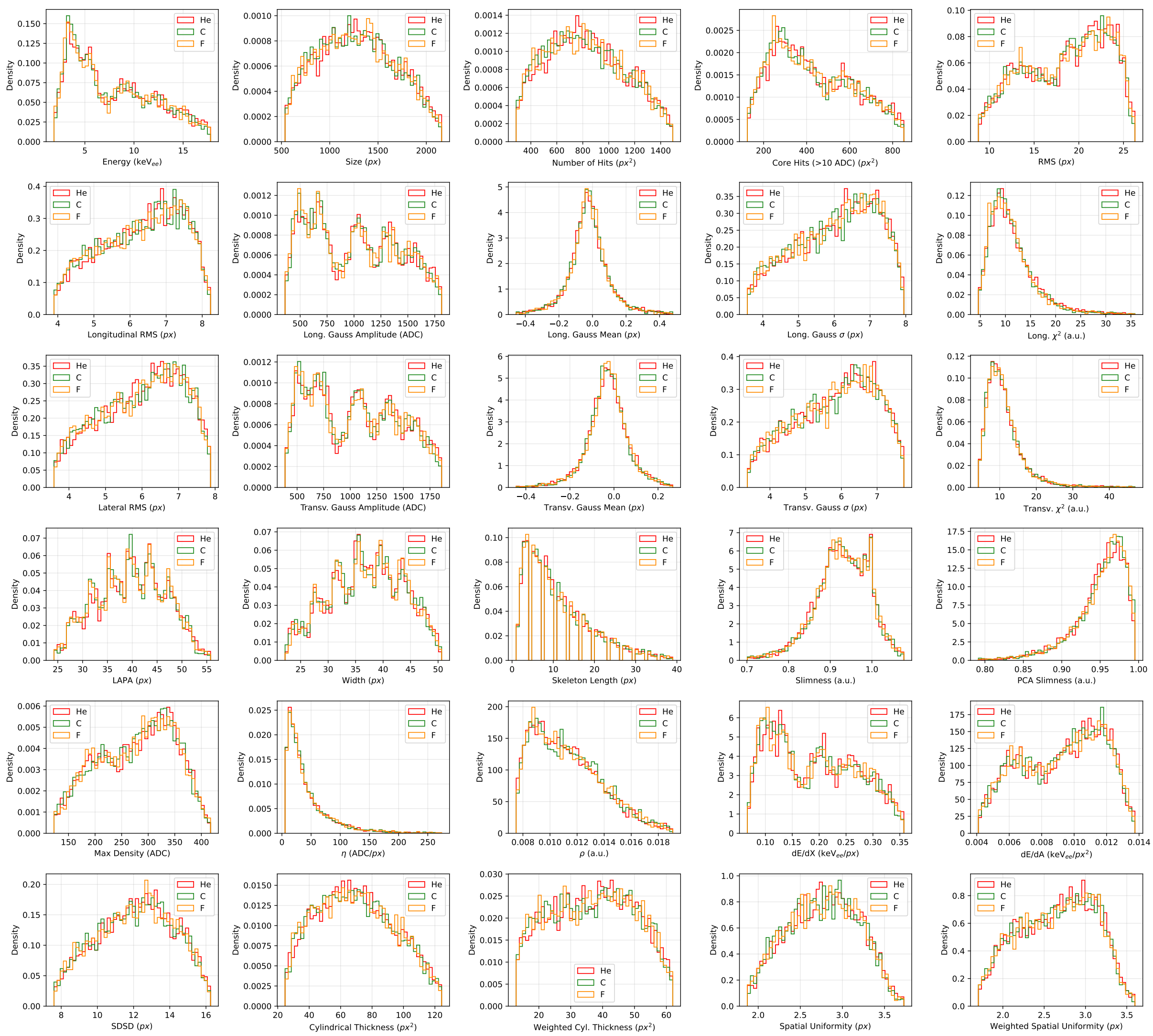
observed for all three NR species. Unlike ERs, NRs remain spatially compact even as the deposited energy increases, as illustrated in Figure 31. Consequently, this localized charge deposition saturates the GEMs, explaining the significant drop in the measured energy for NRs when compared to ERs.

5.5.3 Feature distributions

Once the dataset was generated, the distributions of the 30 extracted features were inspected to evaluate their discriminating power and identify potential redundancies. As a first step, the feature distributions of the different NR species were compared. In principle, each recoil species could be treated as a separate class, leading to a multiclass classification problem. However, as shown in Figure 33, the strong overlap between the distributions indicates that the features are unable to distinguish between these nuclei within the energy range of interest. This can also be noticed in the plot of the detector's energy response (Figure 32), where the curves for all NR species overlap. For this reason, Helium was selected as the representative species for the NR class. This simplification reduces redundancy in the classification problem and lowers the computational cost.

Having defined the ER and NR classes, we evaluate the 1D distributions of the 30 extracted features to assess their individual discrimination capability. Figure 34 presents

Figure 33 – Feature distribution comparison across Helium, Carbon, and Fluorine simulated nuclear recoils, demonstrating morphological equivalence.



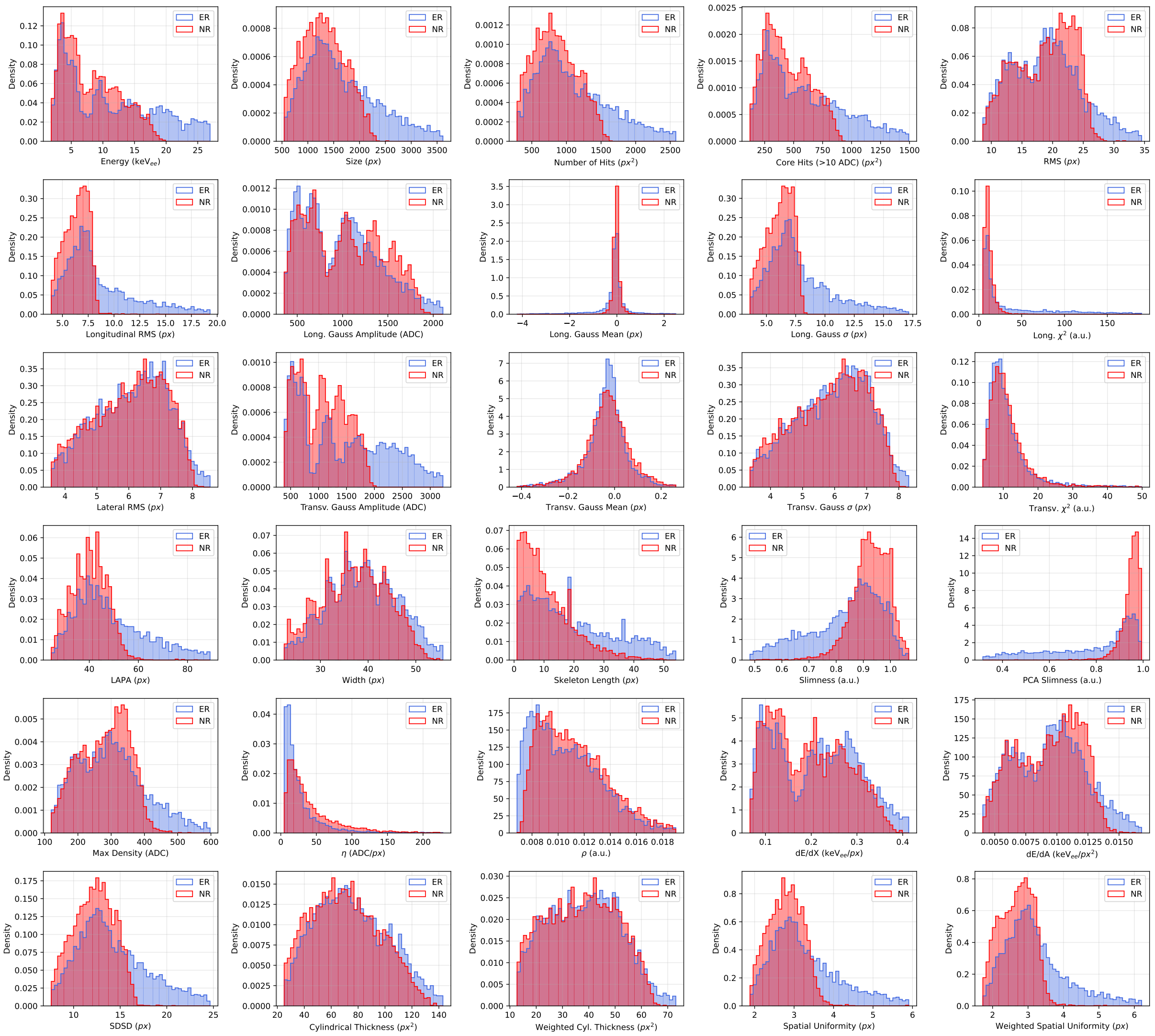
Source: Prepared by the author (2026).

the histograms for both classes.

Visual inspection of these distributions allows for a preliminary analysis of the best features for ER/NR separation. The most significant differences between the classes are observed in features related to track extension and charge diffusion: *Size*, *Number of Hits*, *Core Hits*, *Long. RMS*, *Long.* $\sigma_{u,fit}$, *LAPA*, *Slimness*, and *PCA Slimness*. This is physically expected, as ER tracks have spatial extension and diffusion that scale with energy, while NRs produce localized charge clusters.

Building on this expected result, features related to local energy concentration, such as dE/dX and dE/dA, should be promising discriminators. However, their distributions exhibit significant overlap between the ER and NR classes, which suggest a low separation

Figure 34 – Distribution of the extracted features for the simulated ER and NR classes.



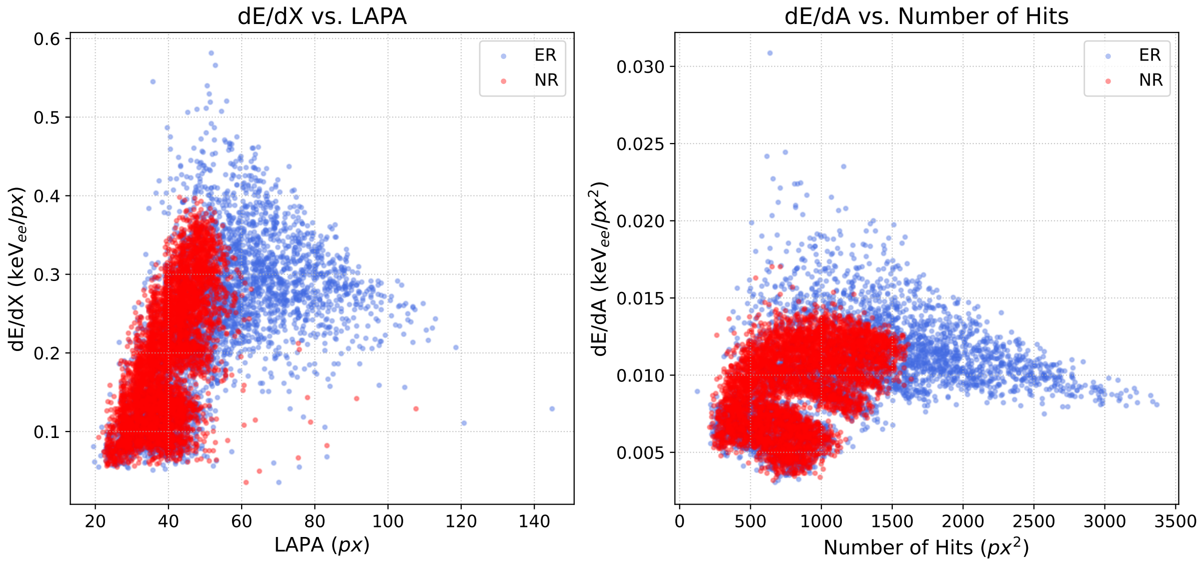
Source: Prepared by the author (2026).

power when evaluated individually. To investigate this further, Figure 35 illustrates a 2D feature space analysis, plotting dE/dX against $LAPA$ (left) and dE/dA against *Number of Hits* (right).

These plots reveal a clear region of separation between the classes that remains hidden in the 1D distributions. This demonstrates that individual feature analysis is insufficient for robust ER/NR discrimination and that the correlation between variables must be taken into account, motivating the use of multivariate ML algorithms.

Finally, it is important to point out that the maximum calibrated energy (shown in the first histogram of Figure 34) for the NR class is restricted to approximately 20 keV$_{ee}$. As reported in the previous section, this is a direct consequence of the detector's energy response saturation induced by the nuclear recoils.

Figure 35 – *dE/dX* vs. *LAPA* (left) and *dE/dA* vs. *Number of Hits* (right), illustrating separability between ER and NR events.



Source: Prepared by the author (2026).

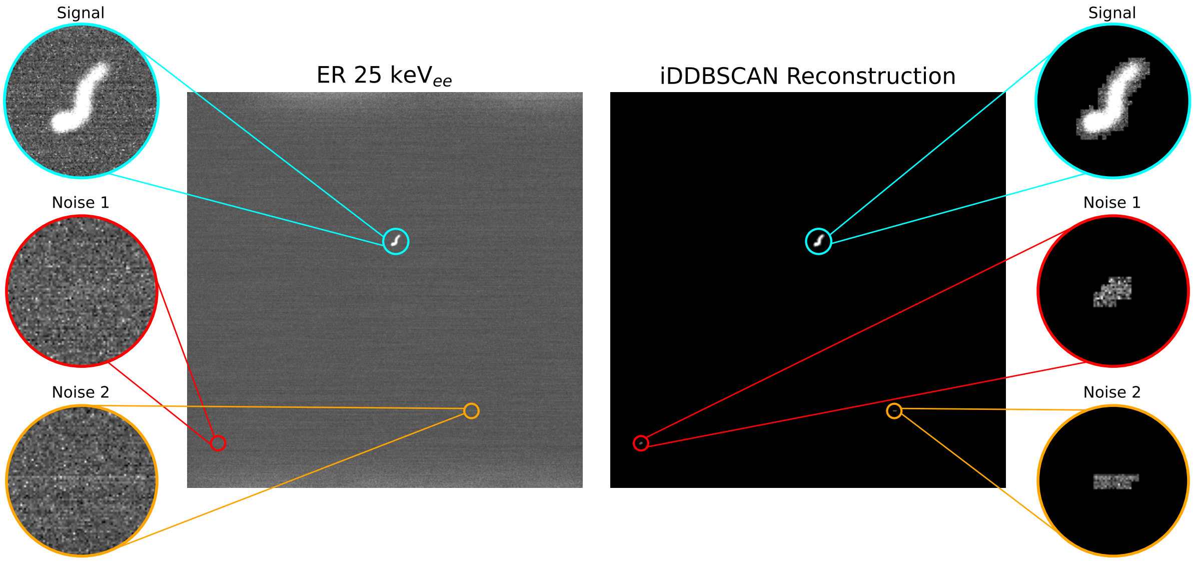
5.5.4 Clusterization filtering

The characterization analysis performed until now has focused on the reconstructed simulated events. However, even with the image preprocessing steps described in Section **4.1.1**, the iDDBSCAN reconstruction algorithm is not perfectly efficient and is susceptible to residual camera noise. As illustrated in the example in Figure 36, the algorithm successfully isolates the signal cluster (cyan) but also reconstructs spurious "fake" clusters originating from sensor noise (red and orange). Since the proposed pipeline processes all clusters identified by the reconstruction, these noise clusters overlap with the ER/NR distributions.

To evaluate strategies for eliminating these noise clusters, their feature distributions were analyzed. Noise data was obtained by running the reconstruction algorithm over pedestal runs, which contain only sensor noise without physical interactions, as introduced in Section **4.1.1**. An inspection of the feature distributions revealed that the ρ feature provides clear separability between physical tracks and sensor noise. Figure 37 displays the ρ distributions for the ER, NR, and Noise classes.

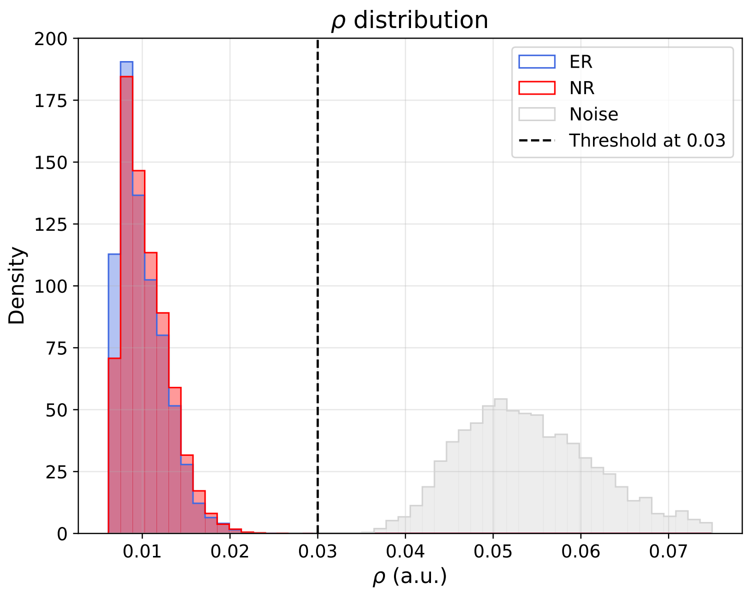
In this case, a simple 1D analysis was sufficient to resolve this problem, allowing for the complete filtering of noise clusters through a single threshold of $\rho < 0.3$ applied to the dataset. This step establishes a binary classification framework (ER vs. NR) for all subsequent ML algorithms.

Figure 36 – Visualization of the iDDBSCAN clustering process. In this example, the reconstruction identifies the true particle track (signal cluster) alongside two small noise clusters located in the lower part of the image.



Source: Prepared by the author (2026).

Figure 37 – ρ distribution, highlighting the separability between ER/NR tracks and sensor noise.



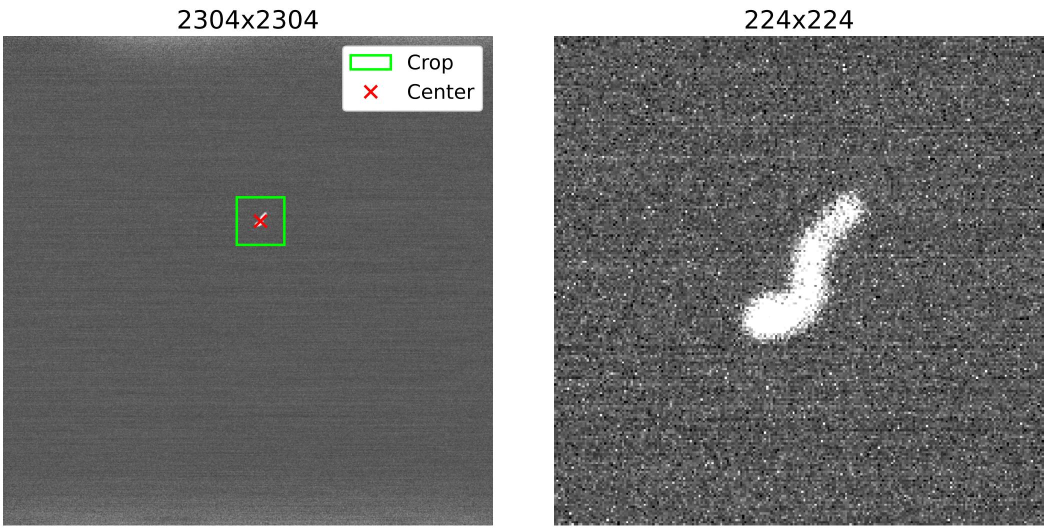
Source: Prepared by the author (2026).

5.5.5 Image preprocessing for Deep Learning

Unlike the feature engineering approach, where events were characterized through specific physical observables, image-based deep learning models learn directly from the raw pixel data. Therefore, to ensure consistency and prevent training bias, the image dataset was constructed using the exact same sample distribution as established in Section **5.5.1**. However, processing the full 2304×2304 image is computationally inefficient and expensive due to both the high sparsity of the signal and the large file size associated with such resolution.

To address this, we implement a cropping pipeline: using the charge centroid (`xmean`, `ymean`) of the reconstructed track, we isolate a 224×224 pixel ROI. This window size was selected to ensure full containment of the particle track topologies across the entire energy spectrum, while maintaining compatibility with standard input layer dimensions required by computer vision architectures (e.g., VGG, ResNet, and ViT). Figure 38 illustrates the original image with the cropping bounding box and the resulting crop.

Figure 38 – Deep Learning preprocessing pipeline. The cropping strategy focuses on the ROI centered on the charge centroid.

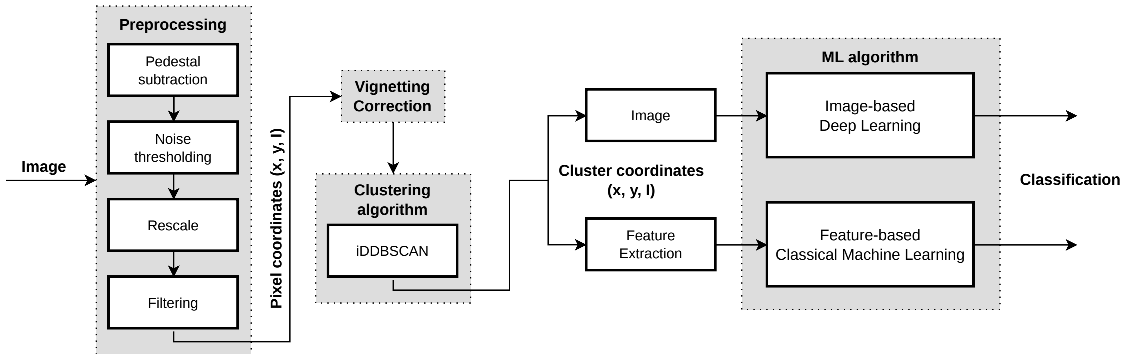


Source: Prepared by the author (2026).

6 MACHINE LEARNING STRATEGIES

Direct searches for DM depend on the capability to isolate rare nuclear recoils, the expected WIMP signal, from a dominant background of electron recoils induced by environmental and intrinsic detector radioactivity. As concluded in Chapter 5, the transition from raw event reconstruction and characterization to physical particle identification requires a multivariate classification framework. This chapter details the development of this discrimination strategy, integrating machine learning into the post-reconstruction stage as the final analysis block, as illustrated in Figure 39.

Figure 39 – Block diagram of the reconstruction algorithm, integrated with the Machine Learning decision block as the final stage.



Source: Prepared by the author (2026).

The proposed methodology explores two different approaches: feature-based models, introduced in Section 6.1, which use the variables extracted from the clusters, and image-based deep learning architectures, detailed in Section 6.2, which learn patterns directly from the pixels of the preprocessed images. A central objective is to compare these two strategies, evaluating whether feature engineering can match or outperform deep learning methods, while identifying their respective physical and computational trade-offs.

The remainder of this chapter is organized as follows: Section 6.3 introduces the evaluation metrics and their correspondence with the terminology commonly used in particle physics. Finally, Section 6.4 details the training and optimization procedures, including cross-validation, feature selection, and hyperparameter optimization.

6.1 FEATURE-BASED CLASSICAL MACHINE LEARNING

This approach uses the vector of the 30 extracted features as inputs. The models were implemented using the open-source `scikit-learn` (92) and `XGBoost` (93) libraries. To compare different classification strategies, a diverse set of algorithms was selected, including linear models, kernel-based methods, tree-based ensembles and neural networks.

6.1.1 Logistic Regression

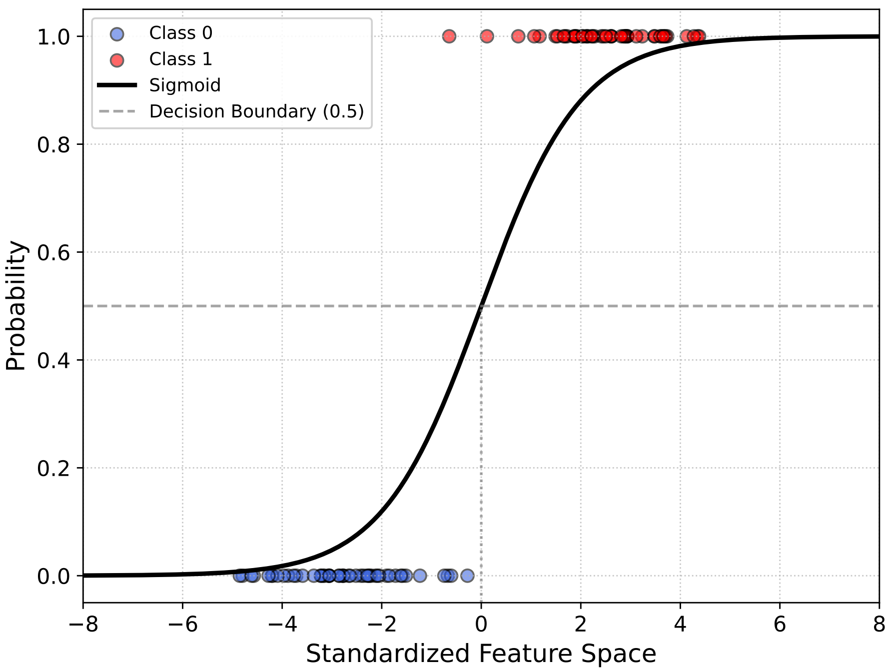
Logistic Regression (94) was selected as the linear baseline classifier. The model estimates the probability that an event belongs to the positive class by applying the logistic sigmoid function to a linear combination of the input features:

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^T\mathbf{x}+b)}} \tag{6.1}$$

where $\mathbf{w}$ represents the weight vector, b is the bias term, and $\mathbf{x}$ is the input feature vector.

As illustrated in Figure 40, this function bounds the continuous input space between 0 and 1. Binary predictions are obtained by applying a decision threshold to the estimated probability.

Figure 40 – Logistic Regression sigmoid curve mapping the model output into class probabilities.



Source: Prepared by the author (2026).

6.1.2 Support Vector Machine

The Support Vector Machine (SVM) is a classifier introduced by Cortes and Vapnik (95). Its objective is to find an optimal hyperplane that separates the data into two classes within the feature space. For a training dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^p$ is the feature vector and $y_i \in \{-1, 1\}$ is the class label, the hyperplane is defined by $\mathbf{w}^T\mathbf{x} + b = 0$.

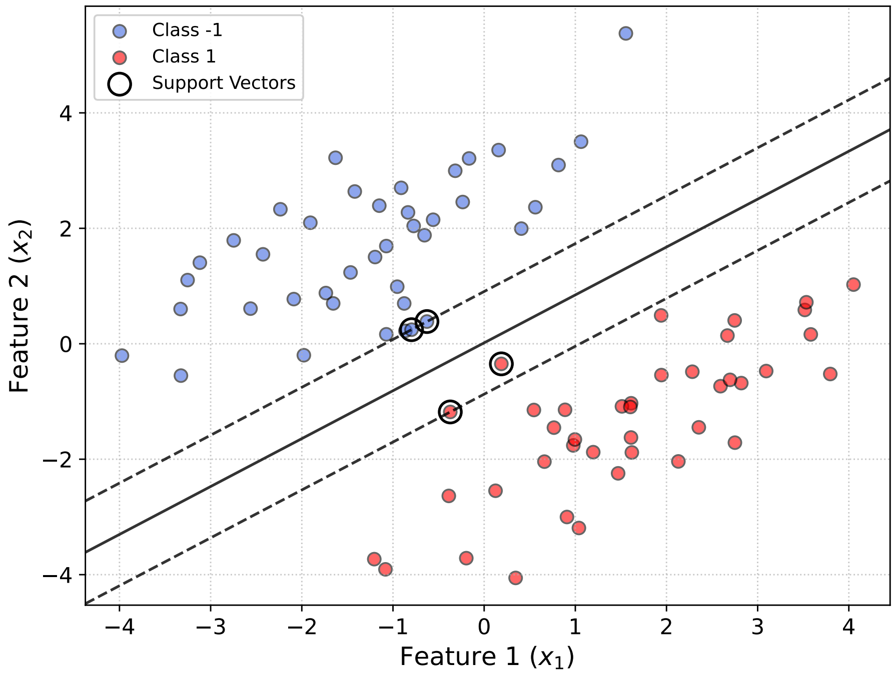
Figure 41 illustrates this concept in a two-dimensional feature space using a linear kernel. The optimal hyperplane maximizes the margin, defined as the distance between the boundary and the closest data points of each class. These boundary points are called support vectors. Maximizing this margin is equivalent to solving the following optimization problem:

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i \tag{6.2}$$

subject to the constraints $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i$ for all $i = 1, \ldots, N$.

The hyperparameter C controls the trade-off between maximizing the margin and penalizing classification errors. In this work, both linear and Radial Basis Function (RBF) kernels were evaluated.

Figure 41 – Linear kernel SVM decision boundary in a 2D feature space, highlighting the optimal separating hyperplane, the margins, and the support vectors.



Source: Prepared by the author (2026).

6.1.3 Random Forest

The foundation of tree-based ensemble methods is the Decision Tree (96). A decision tree is a non-parametric algorithm that partitions the feature space into distinct regions. At each internal node, the algorithm evaluates a binary conditional split based on a specific feature threshold, maximizing information gain. This sequence of splits terminates at the leaf nodes, which assign the final class prediction. Although decision trees are easy to interpret, they are prone to overfitting and can be sensitive to fluctuations in the training data.

To resolve this limitation, Random Forest, formalized by Breiman (97), employs the bootstrap aggregating (*bagging*) technique. As shown in the left panel of Figure 42, this technique constructs a ensemble by training multiple independent decision trees on modified versions of the dataset.

The ensemble is built by training B decision trees on different bootstrap samples of the original dataset. Each bootstrap sample is generated by random sampling with replacement, meaning that some events may appear multiple times while others may not be selected. A decision tree is then trained on each bootstrap sample. In addition, Random Forest considers only a random subset of the available features when evaluating each node split, reducing the correlation between individual trees.

The final prediction is obtained by combining the outputs of all trees through majority voting:

$$\hat{y}_{rf}(\mathbf{x}) = \text{majority vote}\{\hat{f}_b(\mathbf{x})\}_{b=1}^{B} \tag{6.3}$$

where $\hat{f}_b(\mathbf{x})$ is the binary prediction of the b-th tree.

6.1.4 Extreme Gradient Boosting

Gradient Boosting, introduced by Friedman (98), is a tree-based ensemble method that employs the boosting technique. Unlike Random Forest, which trains independent trees in parallel, Gradient Boosting builds trees in a sequential and additive manner, as illustrated in the right panel of Figure 42. The process starts with a base model, and at each iteration, a new decision tree is trained to correct the residual errors of the current ensemble.

This process can be interpreted as a gradient-descent optimization. At iteration m, a new tree $h_m(\mathbf{x})$ is fitted to the negative gradient of the loss function. The final prediction is obtained by combining the contributions of all trees:

$$F_M(\mathbf{x}) = F_0(\mathbf{x}) + \eta \sum_{m=1}^{M} h_m(\mathbf{x}) \tag{6.4}$$

where η is the learning rate and M is the number of boosting iterations.

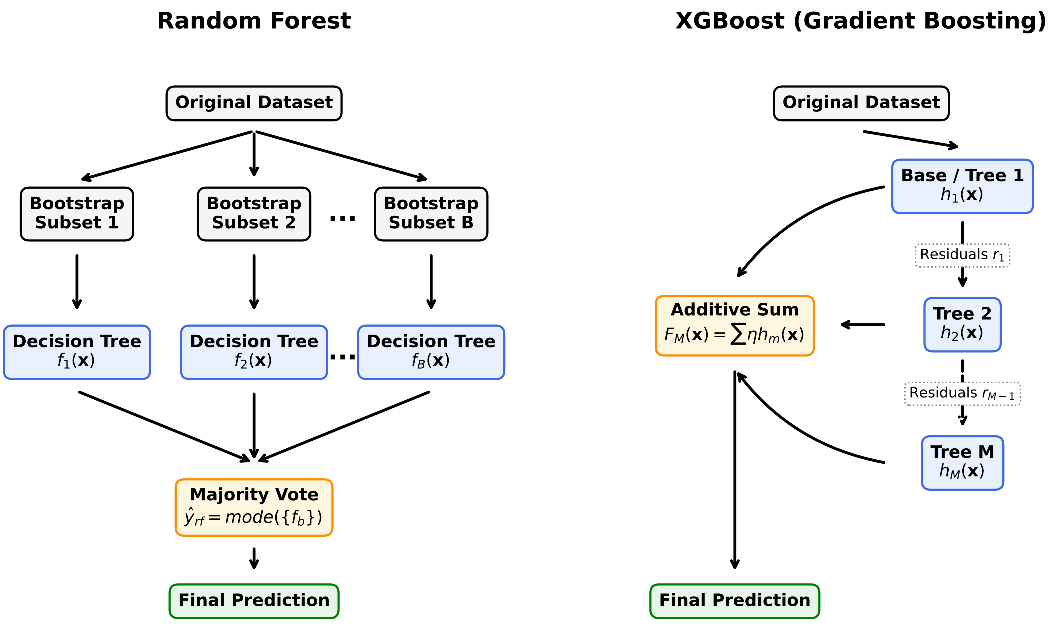
In this work, the Extreme Gradient Boosting (XGBoost) implementation (99) was employed, which provides an optimized and efficient architecture for this algorithm.

6.1.5 Multi-Layer Perceptron

The Multi-Layer Perceptron (100) is a fully connected feedforward neural network, illustrated in Figure 43, representing a multi-layer extension of the classic Perceptron (101). The model architecture is described as follows:

- **Input layer:** Receives the vector of input features. Each input is passed forward to the subsequent hidden layers.

Figure 42 – Comparison of tree-based models: Random Forest operating via parallel bagging (left) and Gradient Boosting operating via sequential error correction (right).



Source: Prepared by the author (2026).

- **Weight initialization:** Before training, weights ($\mathbf{w}$) and biases ($\mathbf{b}$) are initialized. Proper initialization prevents gradient vanishing or explosion (102), ensuring the model starts from a stable state.

- **Hidden layers:** Each layer performs a linear transformation followed by a non-linear activation ϕ:

$$\mathbf{z}^{(l)} = \mathbf{w}^{(l)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)}, \quad \mathbf{a}^{(l)} = \phi(\mathbf{z}^{(l)}) \tag{6.5}$$

  where $\mathbf{w}^{(l)}$ and $\mathbf{b}^{(l)}$ are the weight matrix and bias vector for layer l, and $\mathbf{a}^{(l-1)}$ is the activation from the preceding layer. Common activation functions such as ReLU (103) are employed to map complex non-linear interactions.

- **Output layer:** An output neuron applies an activation function for the desired task.

- **Backpropagation algorithm:** The network minimizes a loss function $J(\mathbf{w}, \mathbf{b})$ by employing the backpropagation algorithm (104). The choice of J depends on the task: Cross-Entropy is standard for classification, while Mean Squared Error (MSE) is typical for regression (105). This process propagates the error from the output layer back to the input via the chain rule. For a weight $w_{jk}^{(l)}$, the gradient is:

$$\frac{\partial J}{\partial w_{jk}^{(l)}} = \delta_j^{(l)} a_k^{(l-1)} \tag{6.6}$$

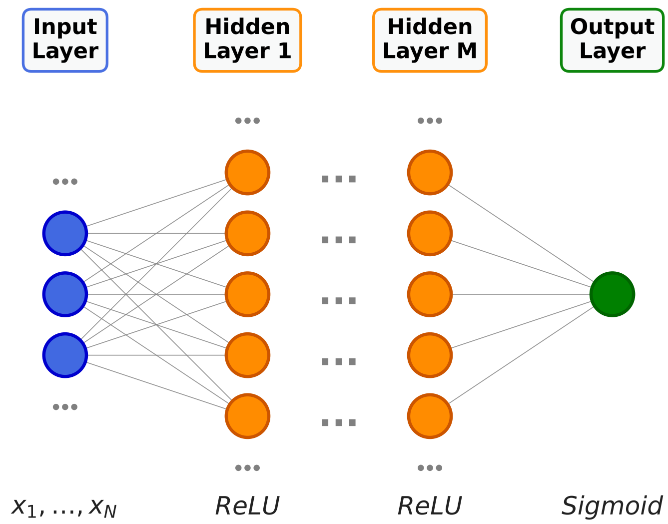
where $\delta_j^{(l)} = \frac{\partial J}{\partial z_j^{(l)}}$ is the propagated error term.

- **Weight update:** Parameters are adjusted iteratively via gradient descent (106):

$$w \leftarrow w - \alpha \frac{\partial J}{\partial w}, \quad b \leftarrow b - \alpha \frac{\partial J}{\partial b} \tag{6.7}$$

where α is the learning rate.

Figure 43 – Schematic representation of a feedforward Multi-Layer Perceptron.



Source: Prepared by the author (2026).

6.2 IMAGE-BASED DEEP LEARNING

Unlike the feature-based approach, image-based deep learning methods will process the cropped 224×224 track images. This approach bypasses the extensive manual feature engineering detailed in Chapter 4, allowing the models to autonomously learn discriminative features directly from the pixel intensities.

Historically, Deep Learning (DL) methodologies have demonstrated exceptional performance in visual pattern recognition, establishing benchmarks in large-scale classification tasks such as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (107). Notable architectures include VGGNet (108) and ResNet (109), which defined the standards for deep CNNs by scaling depth and introducing residual learning, respectively. More recently, Vision Transformers (ViT) (110) have emerged as an alternative architecture based on self-attention mechanisms.

Based on these developments, this section details the implementation of a custom CNN designed for the ionization track images produced by the LIME detector, alongside the adaptation of these recognized computer vision models through transfer learning techniques. All architectures were implemented and trained using the `PyTorch` framework (111).

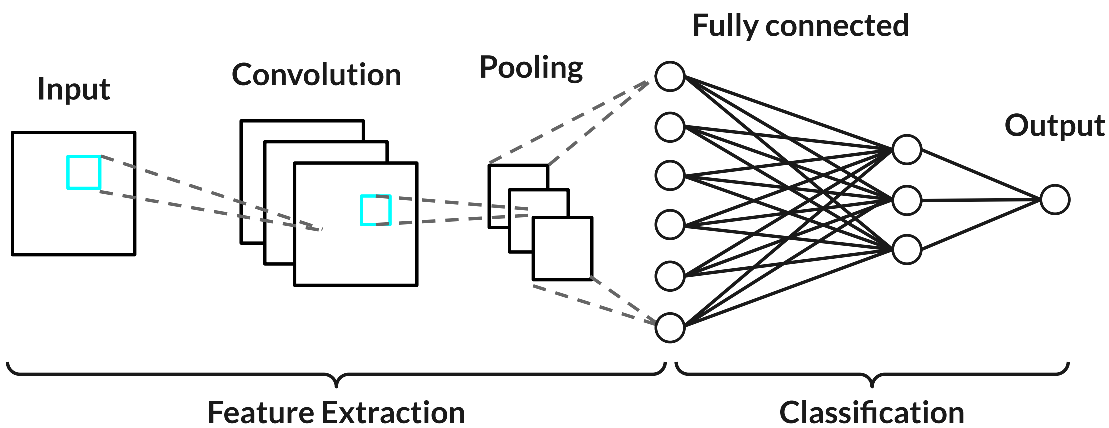
6.2.1 Convolutional Neural Networks

Formalized by LeCun et al. (112), CNNs have established themselves as highly efficient architectures for visual pattern recognition, surpassing traditional methods in large-scale image benchmarks (52, 113, 114, 115, 116). The architecture, shown in Figure 44 can be divided into two main stages:

1. **Feature extraction layers:** A sequence of convolutional and pooling operations. These layers apply filters across the input image to extract spatial features of the tracks.

2. **Classification layers:** A fully connected network that receives the flattened output from the previous stage. This structure operates as an MLP (Section **6.1.5**), using the learned feature maps as inputs.

Figure 44 – Standard architecture of a Convolutional Neural Network.



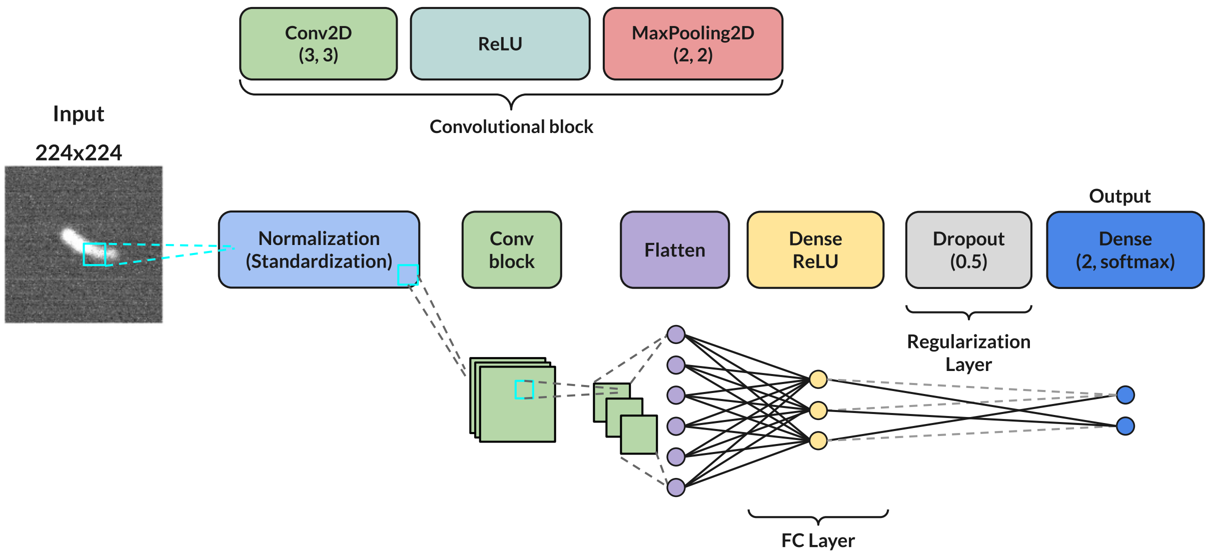
Source: Prepared by the author (2026).

Inspired by the VGGNet architecture (108), which demonstrated the effectiveness of stacking small (3×3) convolutional filters to capture spatial representations, a custom CNN was developed for the classification of LIME ionization tracks. Figure 45 illustrates the implementation of this architecture, which is organized as follows:

- **Normalization layer:** Standardizes the input pixel intensities before processing.

- **Convolutional blocks:** Each block applies a 3×3 convolution, a ReLU activation, and a 2×2 Max-Pooling operation. The number of filters doubles after each block.
- **Flatten layer:** Converts the final feature maps into a one-dimensional vector.
- **Classification head (FC layers):** A fully connected network that maps the extracted image features to the output classes.
- **Dropout:** Regularization layer that randomly deactivates neurons during training to reduce overfitting and improve model generalization.
- **Output layer:** Applies a softmax activation function to compute the probability distribution across the classes.

Figure 45 – Schematic overview of the proposed Custom CNN framework. This baseline architecture serves as the template for the variations evaluated in the architecture search.



Source: Prepared by the author (2026).

6.2.2 Transfer Learning and Fine-Tuning strategies

To benchmark the custom CNN approach, recognized computer vision architectures pre-trained on the ImageNet dataset were evaluated. These models were implemented using the `timm` (PyTorch Image Models) package (117), which provides a interface for different types of computer vision architectures. To maintain computational feasibility, compact versions of these networks were selected:

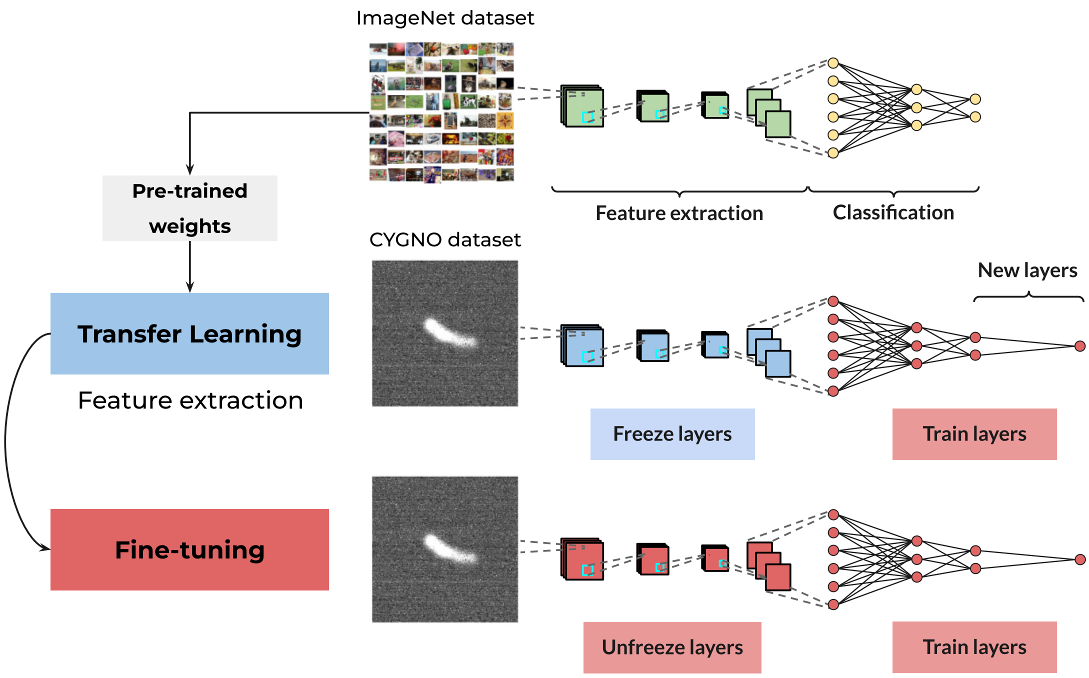
- **VGG11 (108):** A deep sequential CNN containing approximately 128.77 million parameters, used as a baseline pre-trained CNN.

- **ResNet18 (109):** A CNN utilizing residual skip connections, implemented to evaluate the effect of residual learning on track feature extraction. This architecture contains approximately 11.17 million parameters.
- **ViT-Tiny (110, 118, 119):** A Vision Transformer with approximately 5.43 million parameters, replacing local convolutions with self-attention mechanisms (120).

Each architecture was evaluated using two transfer learning strategies, illustrated in Figure 46:

1. **Feature Extraction (TL):** The pre-trained backbone layers are frozen, preserving the ImageNet weights. Only a new classification head is trained.
2. **Fine-Tuning (FT):** The network is initialized with pre-trained weights, but all layers are unfrozen. The entire architecture is then optimized on the target dataset using a reduced learning rate.

Figure 46 – Conceptual diagram comparing Feature Extraction and Fine-Tuning strategies.



Source: Prepared by the author (2026).

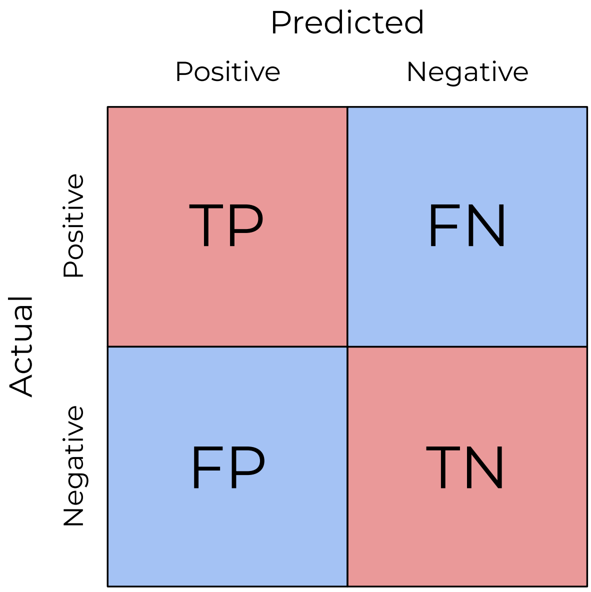
6.3 EVALUATION METRICS AND NOMENCLATURE

Before evaluating the performance of classification algorithms, we need to make a link between the terminology commonly used in machine learning and the metrics adopted in particle physics. For this, we can use a 2×2 confusion matrix, displayed in Figure 47. We define Nuclear Recoils as the positive class (signal) and Electron Recoils as the negative class (background).

The confusion matrix quadrants represent the following states:

- **True Positive (TP):** NR events correctly classified as signal.
- **False Positive (FP):** ER events incorrectly classified as signal.
- **True Negative (TN):** ER events correctly classified as background.
- **False Negative (FN):** NR events incorrectly classified as background.

Figure 47 – Confusion matrix for a binary classification problem.



Source: Prepared by the author (2026).

The objective of this work is to maximize the rejection of background ER events while keeping high efficiency for signal NR detection. For this reason, we focus on four metrics derived from the confusion matrix:

- **Signal Efficiency (ϵ^S):** Corresponds to the True Positive Rate (TPR). It measures the fraction of true NR events retained by the classifier:

$$\epsilon^S = \frac{TP}{TP + FN} \tag{6.8}$$

- **Background Efficiency (ϵ^B):** Corresponds to the False Positive Rate (FPR). It quantifies the fraction of ER events incorrectly classified as NR:

$$\epsilon^B = \frac{FP}{FP + TN} \tag{6.9}$$

- **Background Rejection (R_B):** Corresponds to the True Negative Rate (TNR). It measures the fraction of ER events correctly rejected by the classifier:

$$R_B = 1 - \epsilon^B = \frac{TN}{FP + TN} \tag{6.10}$$

- **Rejection Factor (R):** Defined as the inverse of the background efficiency, this metric measures how many background events are rejected for each background event that survives the selection (44):

$$R = \frac{1}{\epsilon^B} = \frac{FP + TN}{FP} \tag{6.11}$$

The trade-off between signal efficiency (ϵ^S) and background rejection (R_B) is evaluated using a modified Receiver Operating Characteristic (ROC) curve. While the standard ROC curve plots the TPR against the FPR, we adopt a representation plotting background rejection ($R_B = 1 - \epsilon^B$) against signal efficiency ($\epsilon^S = TPR$). This curve is constructed by varying the classification decision threshold across the model's output probability distribution, allowing for the direct assessment of background rejection levels for any fixed signal efficiency.

The **Area Under the Curve (AUC)** serves as a global metric of discrimination performance. An AUC of 1.0 represents perfect discrimination, and 0.5 corresponds to a random classifier performance.

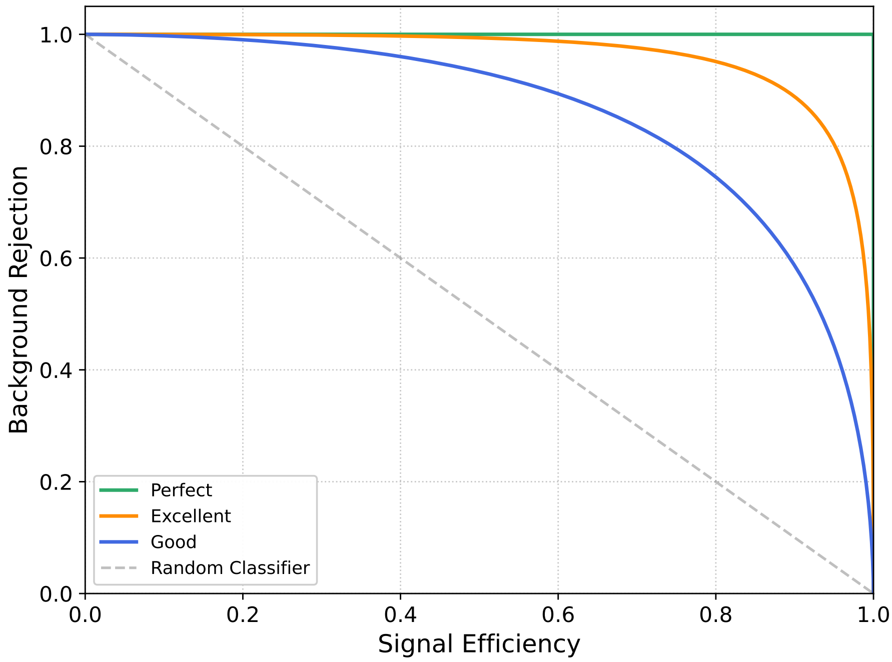
To complement the predictive metrics, the computational performance of each model is quantified through the **Inference Throughput** metric. Measured in events per second (events/s), throughput quantifies the execution speed of a model during the prediction phase. It is defined as:

$$\text{Throughput} = \frac{N_{\text{events}}}{\Delta t} \quad \text{(events/s)} \tag{6.12}$$

where N_{events} represents the total number of processed tracks and Δt is the total inference time in seconds.

Evaluating throughput is fundamental for the CYGNO experiment, as the classification pipeline must balance high background rejection and processing speed for offline data analysis.

Figure 48 – Background rejection vs. signal efficiency. The curves exemplify different discrimination levels: from random performance (AUC=0.5) up to perfect classification (AUC=1.0).



Source: Prepared by the author (2026).

6.4 TRAINING AND OPTIMIZATION PIPELINE

Following the application of the $\rho < 0.3$ filter to eliminate noise clusters (Section **5.5.4**), the dataset contains only physical tracks. Since these events were generated using simulations, their labels are known, allowing the discrimination task to be defined as a supervised binary classification problem. The target classes are mapped as follows:

- **Class 0**: Electronic Recoils, representing the background.
- **Class 1**: Nuclear Recoils, representing the signal.

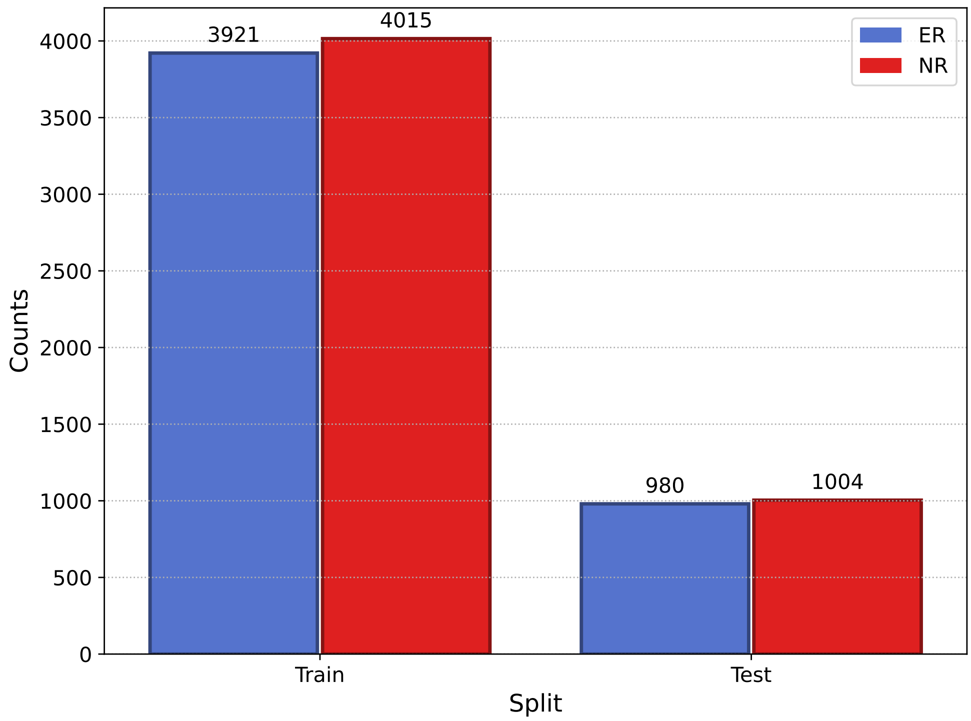
Due to computational feasibility, the training of the feature-based and image-based models was done in different hardware architectures. Classical machine learning models were executed on a environment equipped with an Intel Xeon Processor (Cascade Lake, 16 vCPUs), with no available GPU. On the other hand, deep learning and computer vision architectures rely on massive parallelization for tensor operations, so they highly benefit from GPU acceleration. Therefore, these models were trained on a local setup featuring an NVIDIA GTX 1070 8GB GPU and an Intel i7-7700 CPU.

To address structural differences between feature-based and image-based models, two independent optimization strategies were employed: Classical classifiers were optimized through feature selection followed by hyperparameter tuning, while deep learning models were optimized through an architecture search.

6.4.1 Data splitting and Stratified K-Fold Cross-Validation

To evaluate model generalization without data leakage, the dataset was initially partitioned into subsets: 80% as a training set used exclusively for model optimization and hyperparameter tuning, and 20% as held-out test set for the final benchmark evaluation. Figure 49 presents the sample distribution of both partitions, confirming the balance between ER and NR events.

Figure 49 – Sample distribution of ER and NR events across the training and test partitions.



Source: Prepared by the author (2026).

Model assessment on a single data split is susceptible to variance and evaluation bias (121, 122). To mitigate this, a 5-Fold Stratified Cross-Validation was implemented. The training data is divided into five equal-sized folds. In each iteration, the model is trained on four folds and subsequently evaluated on both the remaining validation fold and the independent test set. This process is repeated until all five folds have served as the validation set.

This strategy makes it possible to report the evaluation metrics as the mean $\pm$ standard deviation across the five iterations for the validation and test sets. Evaluating the test set iteratively along with the validation folds allows for a verification of training stability and generalization, an analysis that is not possible when training on a single data split.

6.4.2 Data normalization

Several machine learning algorithms are sensitive to the scale of the input variables (123). Models trained through gradient-based methods, such as MLPs and CNNs, as well as algorithms that rely on distance calculations, such as SVMs, generally benefit from features with comparable numerical scale. For this reason, the extracted 30 features were standardized using Z-score normalization (124):

$$x_{norm} = \frac{x - \mu}{\sigma} \tag{6.13}$$

where μ and σ denote the mean and standard deviation of each feature, respectively.

This process involves rescaling each feature such that it has a standard deviation of 1 and a mean of 0.

Tree-based methods, including Random Forest and XGBoost, are less affected by feature scaling because their decision rules are based on threshold splits on the feature space. Therefore, normalization was not applied to these models.

For the image-based deep learning models, image pixel intensities (I) were standardized as:

$$I_{norm} = \frac{I - \mu_{\text{img}}}{\sigma_{\text{img}}} \tag{6.14}$$

where μ_{img} and σ_{img} are the global mean and standard deviation of the image pixel intensities.

To avoid data leakage, all normalization parameters were calculated using only the training data within each cross-validation fold. The same parameters were then applied to the corresponding validation fold and to the independent test set.

6.4.3 Feature selection via the 1-Standard-Error rule

Dimensionality reduction is essential to remove redundant or noisy variables that do not contribute to physical discrimination. To ensure the selection process is independent of the model's final hyperparameter configuration, feature selection is performed initially using the default settings for each estimator. This process is achieved through two distinct methodologies, selected based on the internal properties of each estimator:

- **Recursive Feature Elimination (RFE):** Applied to Logistic Regression, Linear SVM, Random Forest, and XGBoost. The algorithm initializes with the complete feature set $\mathcal{F}_p = \{f_1, f_2, \ldots, f_p\}$, where $p = 30$. At each iteration k, the model is trained on the current subset, and an intrinsic importance score $I(f_i)$ is computed for each feature. For tree-based estimators, $I(f_i)$ is derived from the decrease in Gini impurity. For Logistic Regression and Linear SVM, it corresponds to the absolute magnitude of the standardized coefficients ($I(f_i) = |w_i|$). The feature with the lowest importance,

$$f^* = \arg\min_{f_i \in \mathcal{F}_k} I(f_i) \tag{6.15}$$

is removed from the current feature subset. The model is then retrained using the reduced set of features, and the procedure is repeated recursively until all subset sizes have been evaluated.

- **Sequential Feature Selection (SFS):** Applied to Non-linear SVM and MLP, which do not intrinsic feature importance metrics during training. SFS iteratively builds a feature subset by sequentially adding (Forward Selection) or removing (Backward Selection) features to optimize an objective function. In this work, the Forward approach was implemented. The process initializes with an empty set $\mathcal{S}_0 = \emptyset$. At step k, it evaluates all candidate features $f_j \notin \mathcal{S}_k$ and adds the feature f^+ that maximizes the cross-validation objective function J (ROC AUC):

$$f^+ = \arg\max_{f_j \notin \mathcal{S}_k} J(\mathcal{S}_k \cup \{f_j\}) \tag{6.16}$$

A limitation of standard forward selection is the *nesting effect*: once a feature is included in the subset, it cannot be removed later, even if subsequent feature combinations would lead to a better solution. To reduce this limitation, the Floating Search variant (125) was implemented using the `Mlxtend` library (126). After each inclusion step, the algorithm evaluates whether removing any previously selected feature (f^-) improves the objective function:

$$J(\mathcal{S}_{k+1} \setminus \{f^-\}) > J(\mathcal{S}_{k+1}) \tag{6.17}$$

where $f^- \in \mathcal{S}_{k+1}$. If the condition is satisfied, the feature is removed from the subset. This alternating inclusion-removal procedure allows the algorithm to revisit previous decisions and reduces the nesting effect.

To determine the final number of selected features, the *One-Standard-Error Rule* (1-SE rule) (127, 128) was adopted. Instead of selecting the subset with the best cross-validation performance, this criterion chooses the smallest subset whose performance remains within one standard error of the optimum. For an error metric, the selected subset satisfies

$$\text{Error}(k) \leq \text{Error}_{\text{min}} + \text{SE}_{\text{min}} \tag{6.18}$$

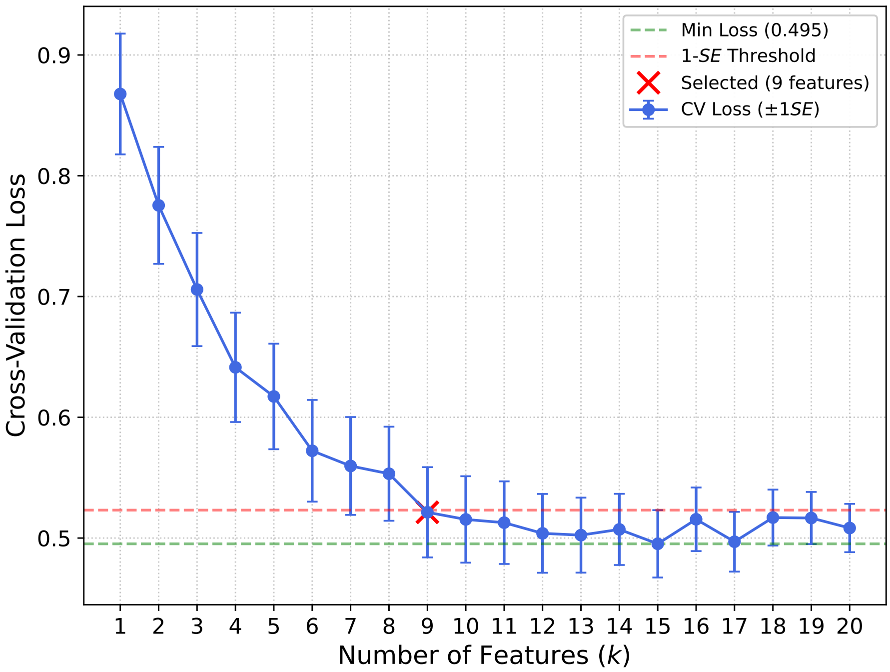
where $\text{Error}_{\text{min}}$ is the minimum observed error and SE_{min} is its associated standard error.

Since ROC AUC was used as the optimization metric in this work, the criterion becomes:

$$\mathrm{AUC}(k) \geq \mathrm{AUC}_{\max} - \mathrm{SE}_{\max} \tag{6.19}$$

Figure 50 illustrates the principle of the 1-SE rule. By favoring the smallest subset with performance comparable to the best-performing model, this approach reduces model complexity while preserving predictive performance. The specific feature subsets selected for each model are presented and analyzed in Chapter 7.

Figure 50 – Conceptual representation of the 1-Standard-Error rule, illustrating the selection of the most parsimonious feature subset within the error bound.



Source: Prepared by the author (2026).

6.4.4 Hyperparameter tuning

After selecting the optimal feature subsets, hyperparameter tuning was performed using the Optuna framework. For each classical model, 200 optimization trials were conducted using the Tree-structured Parzen Estimator (TPE) sampler, a Bayesian optimization algorithm that guides the search toward regions of the parameter space associated with better performance.

The objective function for this tuning stage was set to maximize the Cross-Validation ROC AUC, ensuring the models prioritized the probabilistic separability between ER and NR events. Table 4 details the hyperparameter search space explored for each classifier, including their default configurations used during the prior feature selection phase.

Table 4 – Hyperparameter search space and default values for Classical Machine Learning models using Optuna.

Model (FS Method)	**Parameter**	**Search Space**	**Default**
Logistic Regression (RFE)	C	$[10^{-2}, 10^{2}]$ (log-float)	1.0
	penalty	{'l2'}	'l2'
Linear SVM (RFE)	C	$[10^{-2}, 10^{2}]$ (log-float)	1.0
	kernel	{'linear'}	'linear'
Non-linear SVM (SFS)	C	$[10^{-2}, 10^{2}]$ (log-float)	1.0
	kernel	{'rbf'}	'rbf'
Random Forest (RFE)	*n_estimators*	$[50, 300]$ (int)	100
	max_depth	$[4, 10]$ (int)	None
	max_leaf_nodes	$[10, 30]$ (int)	None
XGBoost (RFE)	*n_estimators*	$[50, 300]$ (int)	100
	learning_rate	$[0.01, 0.4]$ (log-float)	0.3
	max_depth	$[4, 10]$ (int)	6
	max_leaves	$[10, 30]$ (int)	0
MLP (SFS)	*hidden_layer_sizes*	10 architectures*	(100,)
	α	$[10^{-4}, 10^{-1}]$ (log-float)	0.0001
	learning_rate_init	$[10^{-4}, 10^{-2}]$ (log-float)	0.001
	activation	{'relu', 'tanh'}	'relu'

*Architectures: (16,), (32,), (64,), (128,), (32,16), (64,32), (128,64), (64,32,16), (128,64,32), (128,64,32,16).

Source: Prepared by the author (2026).

6.4.5 Deep Learning architecture selection

In contrast to the classical models, whose configuration was optimized through hyperparameter tuning, the custom CNN architecture was selected through a grid search. Using the architecture illustrated in Figure 45 as a baseline, the following variations were tested:

- **Feature extraction layers:** Configurations using 4, 5, or 6 sequential convolutional blocks were evaluated. Each block consists of a 3×3 convolutional layer, a ReLU activation function, and a 2×2 Max-Pooling layer. The number of filters doubles after each block, starting from 16.

- **Fully Connected layers:** The classification head was optimized by varying both the depth (number of dense layers) and the width (neurons per layer). Four configurations were tested: $[64, 32]$, $[64, 32, 16]$, $[128, 64, 32]$, and $[128, 64, 32, 16]$.

For transfer learning, the ImageNet pre-trained architectures VGG11, ResNet18, and ViT-Tiny were adapted to the dataset. The original classification head was replaced with a two-neuron output layer followed by a Softmax activation, allowing the network to estimate the class probabilities for the ER and NR classes. Additionally, the input layer was modified to accept single-channel grayscale images instead of the standard three-channel RGB (Red, Green, Blue) format.

The training was monitored using the loss validation as the primary metric and, like the classical models, using the ROC AUC. To prevent overfitting, early stopping callbacks were implemented, interrupting the training when there is no improvement in loss for a predefined number of epochs, and saving the weights from the best epoch. Table 5 displays the training configuration. Results of the training are provided in Chapter 7.

Table 5 – Training configurations for image-based deep learning models.

General Configurations		
Parameter	**Value**	
Batch Size	64	
Criterion	CrossEntropyLoss	
Optimizer	Adam	
Scheduler	ReduceLROnPlateau	
Specific Configurations		
Hyperparameter	**CNN / Transfer Learning**	**Fine-tuning**
Epochs	100	20
Learning Rate	0.0001	0.00001
Early Stopping	15 epochs	10 epochs

Source: Prepared by the author (2026).

7 RESULTS AND DISCUSSION

The previous chapter presented the ML strategies investigated in this work, including feature-based and image-based approaches, together with the procedures adopted for feature selection, hyperparameter optimization, and model training. With the datasets prepared and validated, the next step is to evaluate the performance of the classification algorithms in the ER/NR discrimination task, comparing their predictive capability, background rejection performance, computational requirements, and interpretability.

To achieve this objective, this chapter is divided into two main sections. Section 7.1 presents the optimization of the feature-based and image-based models, including feature selection, hyperparameter tuning, and deep learning architecture search. Section 7.2 evaluates the optimized classifiers on the independent test set. This evaluation includes both the energy-dependent ROC analysis (Section **7.2.1**) and the rejection power study (Section **7.2.2**), providing a detailed assessment of ER/NR discrimination performance across the calibrated energy spectrum.

7.1 MODEL OPTIMIZATION AND TRAINING PERFORMANCE

This section details the results of the two-stage optimization pipeline (feature selection followed by hyperparameter tuning), and the architectural search for deep learning models.

7.1.1 Classical Machine Learning optimization

The classical machine learning models were optimized following the two-stage procedure described in Chapter 6.4. First, feature selection was performed through cross-validation. The selected feature subsets were then used in a Bayesian hyperparameter optimization.

7.1.1.1 Feature selection and Hyperparameter tuning

The feature selection process evaluated the predictive stability of each classical model as a function of the input dimensionality. Figure 51 illustrates the cross-validation ROC AUC profiles. The curves show a common initial trend across the estimators, with performance increasing as informative features are added and eventually reaching a plateau. For non-linear models, the inclusion of additional variables is followed by increased fluctuations in the cross-validation performance, indicating reduced stability as the feature dimensionality increases. The linear models, in contrast, do not exhibit the same behavior. This trend is consistent with the peaking phenomenon described in pattern recognition (129).

By applying the 1-SE rule, dimensionality was reduced from 30 features to subsets ranging from 7 to 13 observables. Both Linear and Non-Linear SVMs required only 7 features, while the MLP selected the largest subset, with a total of 13 features.

Following feature selection, the models were tuned following the Optuna optimization described in Section **5.4.1**. Table 6 details the optimized hyperparameters, the final feature subset sizes, and the computational cost required for the complete optimization procedure.

Table 6 – Optimization results: selected hyperparameters, final feature dimensions, and computational cost in CPU Core-hours.

Model	**Selected Hyperparameters**	**# Features**	**Cost (Core-h)**
Logistic Regression	C=4.44, *penalty*='l2'	8	0.16
Linear SVM	C=99.65, *kernel*='linear'	7	9.93
Non-linear SVM	C=82.46, *kernel*='rbf'	7	20.17
Random Forest	*n_estimators*=250, *max_depth*=10, *max_leaf_nodes*=30	11	2.25
XGBoost	*n_estimators*=158, *learning_rate*=0.17, *max_depth*=9, *max_leaves*=14	9	0.24
MLP	*hidden_layer_sizes*=(64, 32, 16), *activation*='relu', *learning_rate_init*=0.008, α=0.068	13	2.80

Source: Prepared by the author (2026).

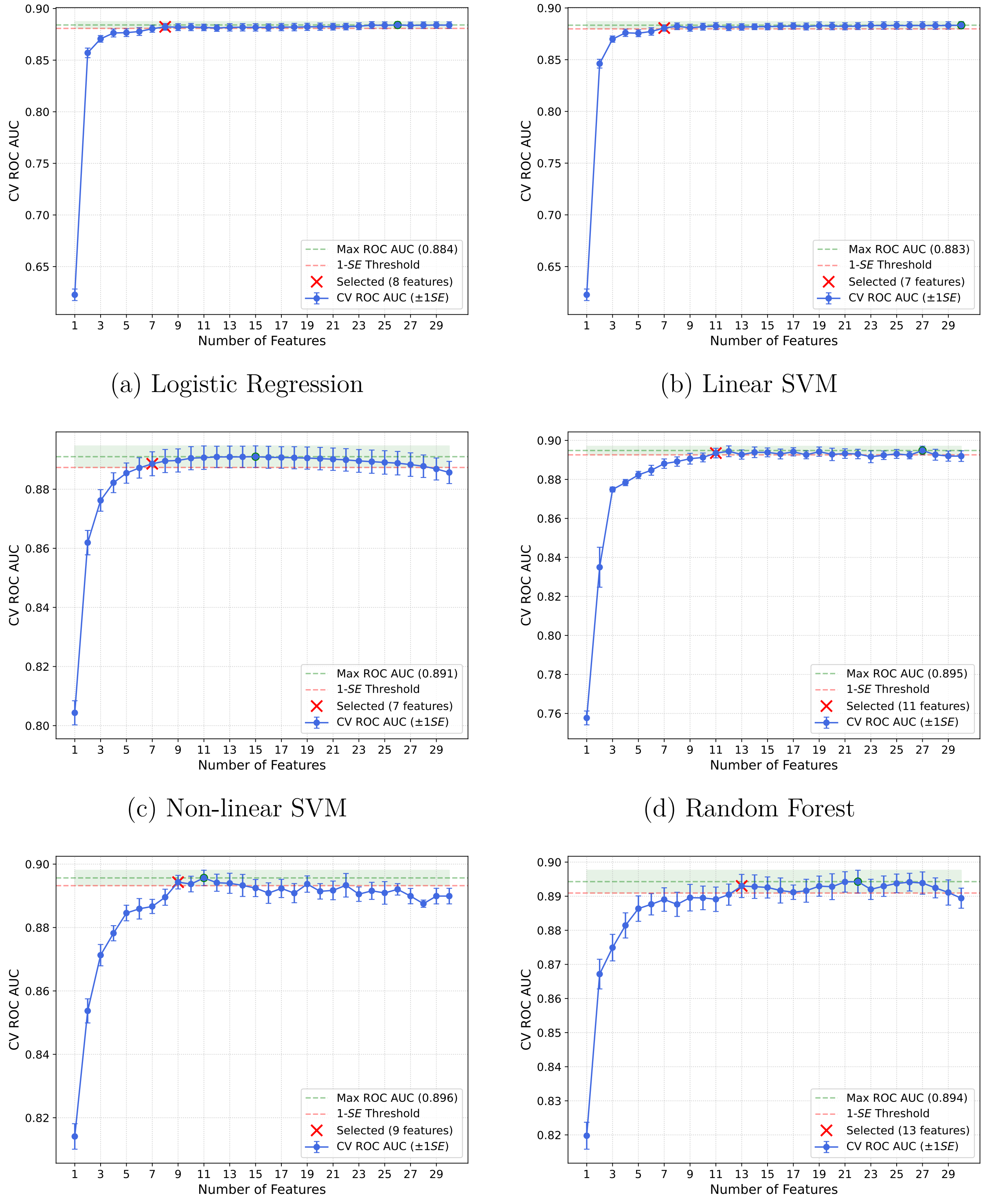
The computational cost metric highlights operational differences between the algorithms. Logistic Regression established the lower bound (0.16 Core-h) due to its straightforward linear formulation. Conversely, both SVM classifiers presented the highest computational cost, peaking at 20.17 Core-h. This occurs by two factors: the SFS algorithm, which requires exhaustive retraining across multiple candidate subsets, and the heavy computational cost introduced by `Scikit-Learn` to map the SVM geometric margins into class probabilities.

While Random Forest and the MLP required moderate resources (2.5 Core-h), XGBoost completed the full optimization pipeline in just 0.24 Core-h. This runtime rivals even the simplest linear baseline (Logistic Regression), demonstrating the efficiency of its gradient boosting implementation.

Table 7 compares the optimized models with baseline configurations using all 30 features and default hyperparameters. Although the mean ROC AUC of the optimized models does not surpass the baseline in all cases (such as in Logistic Regression, Linear SVM and Random Forest), the validation scores remain equivalent, while constraining model complexity. The largest effect is observed for the tree-based models, where the optimized configurations reduce the gap between training and validation performance, indicating improved generalization and less overfitting.

Furthermore, a slight performance gain is observed between linear and non-linear

Figure 51 – Feature selection ROC AUC profiles for classical models.



(a) Logistic Regression (b) Linear SVM

(c) Non-linear SVM (d) Random Forest

(e) XGBoost (f) MLP

Source: Prepared by the author (2026).

models. Optimized Logistic Regression and Linear SVM achieved validation ROC AUC values of 0.884 ± 0.007 and 0.881 ± 0.010, respectively. In contrast, non-linear models obtained higher mean validation scores, ranging from 0.892 ± 0.006 (Random Forest) to 0.900 ± 0.003 (MLP).

Table 7 – Comparison of Baseline vs. Optimized classical models evaluating the ROC AUC.

Model	**Train ROC AUC**	**Val ROC AUC**
Logistic Regression		
Baseline	0.887 ± 0.002	$\mathbf{0.884 \pm 0.007}$
Optimized	0.884 ± 0.002	0.883 ± 0.007
Linear SVM		
Baseline	0.887 ± 0.002	$\mathbf{0.883 \pm 0.008}$
Optimized	0.882 ± 0.002	0.881 ± 0.010
Non-linear SVM		
Baseline	0.898 ± 0.002	0.886 ± 0.008
Optimized	0.911 ± 0.002	$\mathbf{0.895 \pm 0.008}$
Random Forest		
Baseline	1.000 ± 0.000	$\mathbf{0.892 \pm 0.006}$
Optimized	0.919 ± 0.002	0.889 ± 0.008
XGBoost		
Baseline	1.000 ± 0.000	0.890 ± 0.005
Optimized	0.975 ± 0.001	$\mathbf{0.896 \pm 0.012}$
MLP		
Baseline	0.900 ± 0.003	0.889 ± 0.007
Optimized	0.911 ± 0.004	$\mathbf{0.900 \pm 0.003}$

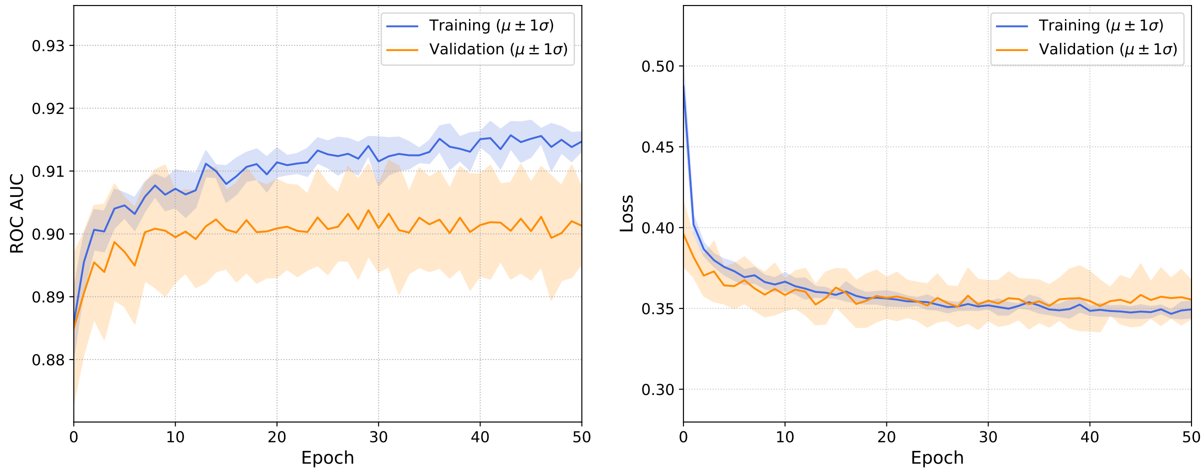
Source: Prepared by the author (2026).

To verify the training stability of the optimal MLP configuration, Figure 52 presents its cross-validation learning curves. The validation performance shows no significant improvement after approximately 20 training epochs and remains stable throughout the remaining epochs. This indicates that the model reaches a stable performance early in the training process.

7.1.1.2 Feature interpretability

Beyond reducing dimensionality, the feature selection stage provides information about which observables contribute most to ER/NR discrimination. Table 8 provides the Boolean selection matrix, summarizing the features selected by each optimized model. The features are ordered according to their overall selection frequency across all models, from the most frequently selected to the least selected. To further quantify their impact, Figure 53 extracts the intrinsic feature importance from the tree-based models. Minimal error bars confirm consistency across all five CV folds.

Figure 52 – Training and validation curves for the optimized MLP, with ROC AUC on the left and loss on the right.



Source: Prepared by the author (2026).

The ranking shows a preference for features related to energy density and charge profile. Observables such as `rms` and `dE_dA` were selected by all six estimators. Moreover, Figure 53 shows that `Slimness_pca` leads the classification logic in both Random Forest and XGBoost, dominating the importance metric with a wide margin over the second most important features (`longrms` and `dE_dA`, respectively). This result suggests that defining the track aspect ratio through charge-weighted variance (PCA) captures more discriminating information than using geometric boundaries (`Slimness`), as discussed in Section **4.2.3**.

The relevance of energy density and spatial combinations is consistent with both the visual inspection of simulated events presented in Section **5.5.1** and the 2D feature-space analysis discussed in Section **5.5.3**. As illustrated in Figure 31, nuclear recoils tend to produce compact and dense ionization patterns, whereas electron recoils generate longer and more diffuse tracks, particularly at higher energies. These differences are reflected in observables such as `dE_dA`, `Slimness_pca`, and the charge profile variables. While the 1D distributions of individual features often exhibit overlap between classes, their combination creates regions where ER and NR events become more separable. The feature selection algorithms identified and prioritized these multivariate relationships.

Conversely, features related to track dimensions, such as `LAPA`, `Width`, `size`, and `nhits`, were not selected by any evaluated algorithm. However, their information is preserved in other variables: `LAPA` and `Width` encode the `Slimness` ratio (selected by 3 models), while `nhits` appears in the denominator of `dE_dA` (selected by all 6 models). This suggests that the rejected observables provide redundant information for discrimination.

Table 8 – Feature selection boolean matrix. The ✓ symbol indicates selection by the respective optimal model. Features are ordered by their selection frequency.

Feature	**Logistic Regression**	**Linear SVM**	**Non-linear SVM**	**Random Forest**	**XGBoost**	**MLP**	**Total**
rms	✓	✓	✓	✓	✓	✓	6
dE_dA	✓	✓	✓	✓	✓	✓	6
lgaussamp	✓	✓		✓	✓	✓	5
lgausssigma			✓	✓	✓	✓	4
Slimness_pca			✓	✓	✓	✓	4
tgaussamp	✓	✓		✓	✓		4
CylThick	✓	✓			✓	✓	4
Slimness			✓	✓		✓	3
dE_dX			✓	✓	✓		3
MaxDen				✓	✓	✓	3
longrms	✓	✓		✓			3
SDSD	✓	✓					2
tgausssigma						✓	1
Energy_calibrated						✓	1
rho						✓	1
CylThick_weighted	✓						1
SpatialUnif						✓	1
tchi2						✓	1
latrms			✓				1
lchi2				✓			1
size							0
nhits							0
tgaussmean							0
lgaussmean							0
nhits_highThr							0
Skeleton_Length							0
Eta							0
LAPA							0
Width							0
SpatialUnif_weighted							0

Source: Prepared by the author (2026).

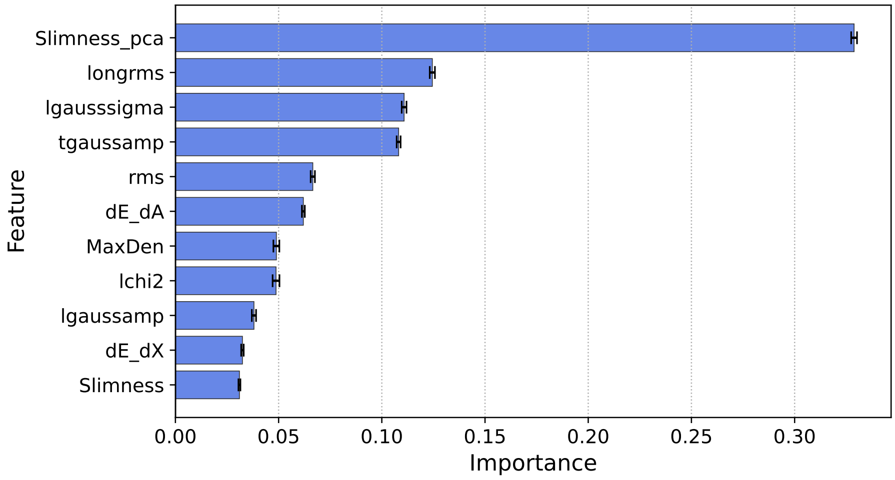
7.1.2 Deep Learning optimization

The architecture search for the Custom CNN and the transfer learning evaluation results are presented below. Table 9 summarizes the architecture search performed on the Custom CNN, while Table 11 benchmarks the optimal CNN architecture against the pre-trained transfer learning models.

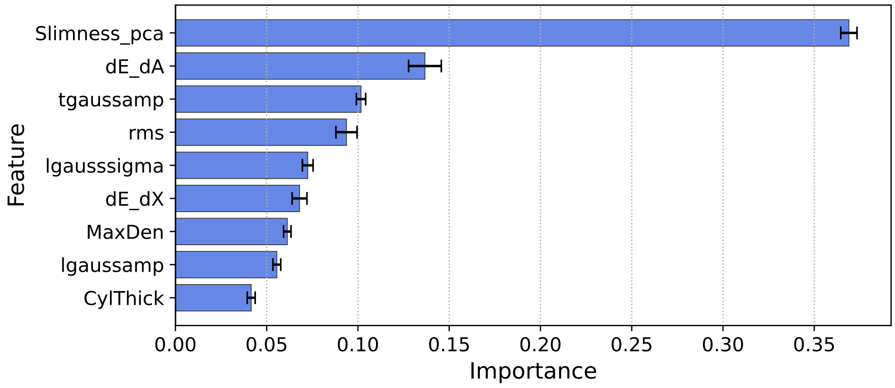
The results shows that deeper CNNs perform better for this task. Expanding the depth from 4 to 6 convolutional blocks improved the validation ROC AUC from 0.86 to 0.90 and reduced the mean validation loss from approximately 0.44 to 0.37.

A comparison can be made between the tested Custom CNN architectures and the optimized MLP. The MLP, using the [64, 32, 16] dense layer configuration and the set of 13 features, achieved a validation ROC AUC of 0.90, as shown in Table 7. In contrast, CNN architectures using the same dense layer configuration obtained validation ROC

Figure 53 – Intrinsic feature importance extracted from the optimized tree-based ensembles.



(a) Random Forest



(b) XGBoost

Source: Prepared by the author (2026).

AUC values of 0.86, 0.89, and 0.90 for four, five, and six convolutional blocks, respectively. The comparable performance obtained by the best CNN and MLP suggests that the handcrafted features capture a large fraction of the discriminative information present in the images.

Based on the lowest validation loss (0.37) and highest ROC AUC (0.90), the architecture with 6 convolutional blocks and four dense layers [128, 64, 32, 16] demonstrated the best performance and was adopted for the remainder of the analysis. The detailed layer-wise specifications and parameter counts of the architecture are provided in Table 10. The training curves for this optimal architecture are illustrated in Figure 54, which displays the average performance across the 5 cross-validation folds. The epoch axis corresponds to the minimum number of epochs reached across all folds before early stopping. The narrow standard deviation band indicates stable behavior across the cross-validation folds.

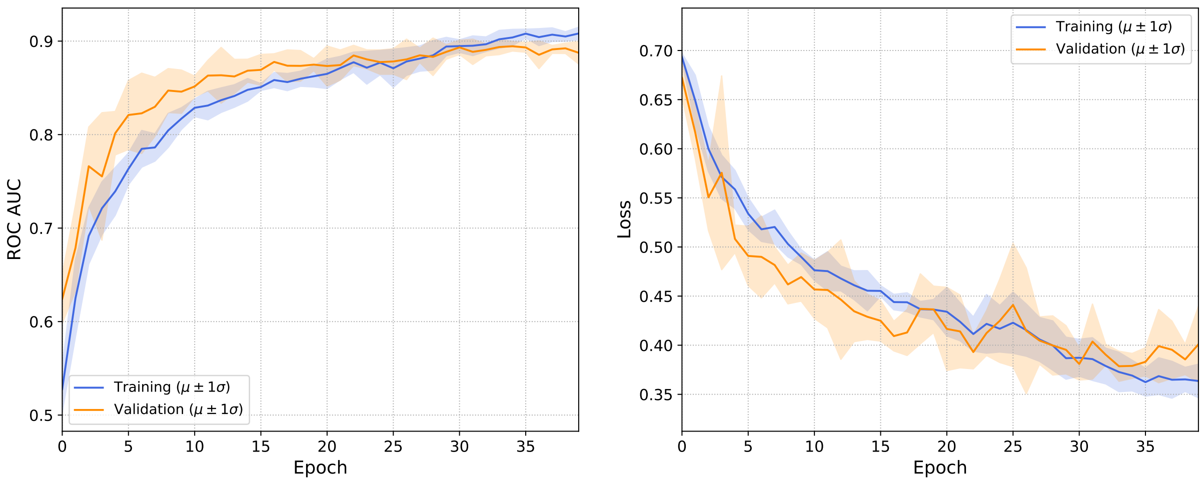
Table 9 – Performance and resource usage of Custom CNN architectures.

Architecture		**Resource Usage**				**Performance**			
Conv. Blocks	**Dense Layers**	**Params (M)**	**FLOPS (M)**	**Mem. (MB)**	**Cost (GPU-h)**	**Train Loss**	**Train ROC AUC**	**Val Loss**	**Val ROC AUC**
4	[64, 32]	1.70	182.24	777.25	0.28	0.37 ± 0.02	0.91 ± 0.02	0.44 ± 0.01	0.86 ± 0.01
	[64, 32, 16]	1.71	182.24	777.26	0.29	0.41 ± 0.03	0.88 ± 0.02	0.44 ± 0.01	0.86 ± 0.01
	[128, 64, 32]	3.32	183.86	802.88	0.27	0.40 ± 0.04	0.89 ± 0.02	0.44 ± 0.01	0.86 ± 0.01
	[128, 64, 32, 16]	3.32	183.86	802.89	0.31	0.41 ± 0.03	0.88 ± 0.03	0.44 ± 0.02	0.86 ± 0.01
5	[64, 32]	1.20	239.24	769.88	0.43	0.36 ± 0.02	0.91 ± 0.01	0.38 ± 0.01	0.89 ± 0.01
	[64, 32, 16]	1.20	239.24	769.89	0.51	0.38 ± 0.04	0.90 ± 0.02	0.39 ± 0.02	0.89 ± 0.01
	[128, 64, 32]	2.01	240.05	782.63	0.50	0.36 ± 0.02	0.91 ± 0.01	0.39 ± 0.02	0.89 ± 0.01
	[128, 64, 32, 16]	2.01	240.05	782.64	0.61	0.40 ± 0.03	0.88 ± 0.01	0.39 ± 0.01	0.89 ± 0.01
6	[64, 32]	1.87	296.54	780.76	0.58	0.34 ± 0.02	0.92 ± 0.01	0.38 ± 0.01	0.90 ± 0.01
	[64, 32, 16]	1.87	296.54	780.77	0.75	0.37 ± 0.02	0.90 ± 0.01	0.38 ± 0.01	0.89 ± 0.01
	[128, 64, 32]	2.17	296.84	786.14	0.57	0.34 ± 0.02	0.92 ± 0.01	0.38 ± 0.01	0.90 ± 0.01
	[128, 64, 32, 16]	2.17	296.84	786.15	0.80	0.35 ± 0.02	0.91 ± 0.01	$\mathbf{0.37 \pm 0.01}$	$\mathbf{0.90 \pm 0.01}$

Source: Prepared by the author (2026).

In addition, the training and validation loss curves remain closely aligned, showing no evidence of overfitting.

Figure 54 – Training and validation curves for the optimal Custom CNN architecture, with ROC AUC on the left and loss on the right.



Source: Prepared by the author (2026).

As shown in Table 11, we observe that the Custom CNN trained from scratch outperforms the transfer learning approaches. Evaluating the pre-trained models with frozen weights resulted in poor results, with the ResNet18 validation ROC AUC dropping to 0.65, for example. This indicates that pre-training on ImageNet does not encode features relevant to the topology of ionization tracks. Unfreezing the weights during fine-tuning improved performance, with VGG11, ViT-Tiny, and ResNet18 all reaching a validation ROC AUC of 0.80, but failing to match the 0.90 ROC AUC achieved by the Custom CNN.

Table 10 – Detailed architecture of the Custom CNN

Layer	Output Shape	Parameters
Input	(N, 1, 224, 224)	0
Conv2D (16, 3×3) + ReLU	(N, 16, 224, 224)	160
MaxPooling2D (2×2)	(N, 16, 112, 112)	0
Conv2D (32, 3×3) + ReLU	(N, 32, 112, 112)	4,640
MaxPooling2D (2×2)	(N, 32, 56, 56)	0
Conv2D (64, 3×3) + ReLU	(N, 64, 56, 56)	18,496
MaxPooling2D (2×2)	(N, 64, 28, 28)	0
Conv2D (128, 3×3) + ReLU	(N, 128, 28, 28)	73,856
MaxPooling2D (2×2)	(N, 128, 14, 14)	0
Conv2D (256, 3×3) + ReLU	(N, 256, 14, 14)	295,168
MaxPooling2D (2×2)	(N, 256, 7, 7)	0
Conv2D (512, 3×3) + ReLU	(N, 512, 7, 7)	1,180,160
MaxPooling2D (2×2)	(N, 512, 3, 3)	0
Flatten	(N, 4608)	0
Dense (128) + ReLU	(N, 128)	589,952
Dense (64) + ReLU	(N, 64)	8,256
Dense (32) + ReLU	(N, 32)	2,080
Dense (16) + ReLU	(N, 16)	528
Dropout (0.5)	(N, 16)	0
Dense (2) [Softmax]	(N, 2)	34
Total Parameters		**2,173,330**

*N denotes the batch size.

Source: Prepared by the author (2026).

Table 11 – Comparison with Transfer Learning and Fine-Tuning models.

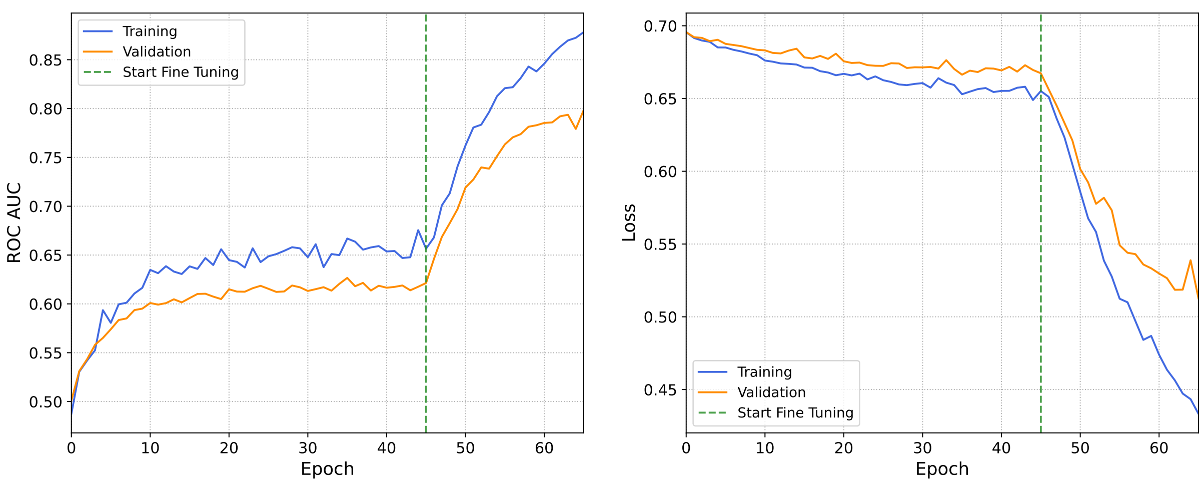
Model	Resource Usage				Performance			
	Params (M)	FLOPS (B)	Mem. (MB)	Cost (GPU-h)	Train Loss	Train ROC AUC	Val Loss	Val ROC AUC
Training from Scratch								
Custom CNN	2.17	0.30	786.15	0.80	0.35 ± 0.02	0.91 ± 0.01	0.37 ± 0.01	0.90 ± 0.01
Transfer Learning								
VGG11	0.01	7.55	1892.83	2.51	0.51 ± 0.01	0.81 ± 0.01	0.53 ± 0.01	0.78 ± 0.01
ViT-Tiny	<0.01	1.06	161.04	0.89	0.55 ± 0.01	0.77 ± 0.01	0.56 ± 0.00	0.76 ± 0.00
ResNet18	<0.01	1.74	464.43	0.72	0.65 ± 0.00	0.66 ± 0.01	0.66 ± 0.01	0.65 ± 0.01
Fine-Tuning								
VGG11	128.77	7.55	7642.11	0.93	0.47 ± 0.05	0.84 ± 0.04	0.52 ± 0.02	0.80 ± 0.01
ViT-Tiny	5.43	1.06	1930.51	0.28	0.47 ± 0.03	0.84 ± 0.03	0.52 ± 0.01	0.80 ± 0.01
ResNet18	11.17	1.74	1572.58	0.54	0.43 ± 0.01	0.88 ± 0.01	0.51 ± 0.01	0.80 ± 0.01

Source: Prepared by the author (2026).

A similar disparity is observed in computational resource requirements. Fine-tuning the VGG11 architecture requires more than 7.6 GB of memory, approaching the 8 GB VRAM limit of the GPU (NVIDIA GTX 1070) used in this study. Conversely, the Custom CNN requires only 786 MB of memory and 0.80 GPU-hours for training. In addition, it demands only 0.30 billion Floating-point Operations Per Second (FLOPS), while the VGG11 requires 7.55 billion FLOPS. This lower computational cost allows the use of larger batch sizes and ensures a lightweight training pipeline.

Finally, to illustrate the transition between training stages, Figures 55, 56, and 57 illustrate the training and validation trajectories for the ResNet18, VGG11, and ViT-Tiny architectures, respectively. The vertical green line separates transfer learning from fine-tuning. During the initial frozen phase, only the classification head is updated while the backbone retains the ImageNet weights. In this regime, the loss and ROC AUC curves exhibit limited improvement. After unfreezing the backbone, all architectures show a reduction in validation loss and an increase in ROC AUC, indicating adaptation to the LIME dataset. However, the massive parameter count of the pre-trained architectures (starting at 5.43M for ViT-Tiny compared to 2.17M for the Custom CNN) forces the network to overfit the training data. This is one of the reasons to employ the strategies detailed in Table 5: a reduced learning rate and a reduced early-stopping patience to prevent overfit.

Figure 55 – Training and validation trajectories for the ResNet18 architecture (Fold 2), with ROC AUC on the left and loss on the right.



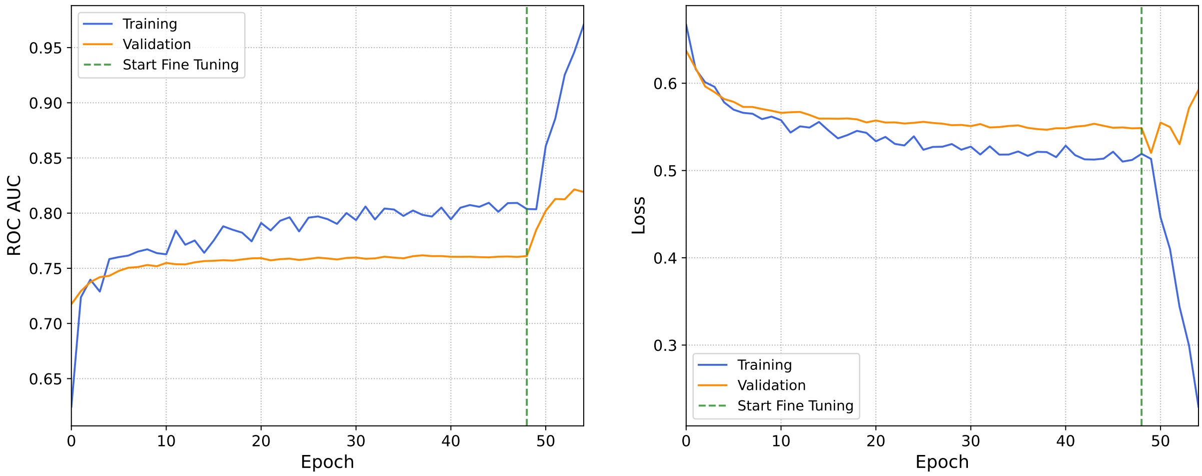
Source: Prepared by the author (2026).

7.2 BACKGROUND REJECTION EVALUATION

Following the optimization and architecture search phases, the best-performing model configurations from both the classical feature-based approach and the deep learning image-based approach are evaluated on the independent, held-out test set (20% of the data). The main goal of this section is to quantify the models' ability to identify NR candidate signals while rejecting the ER backgrounds across the energy spectrum.

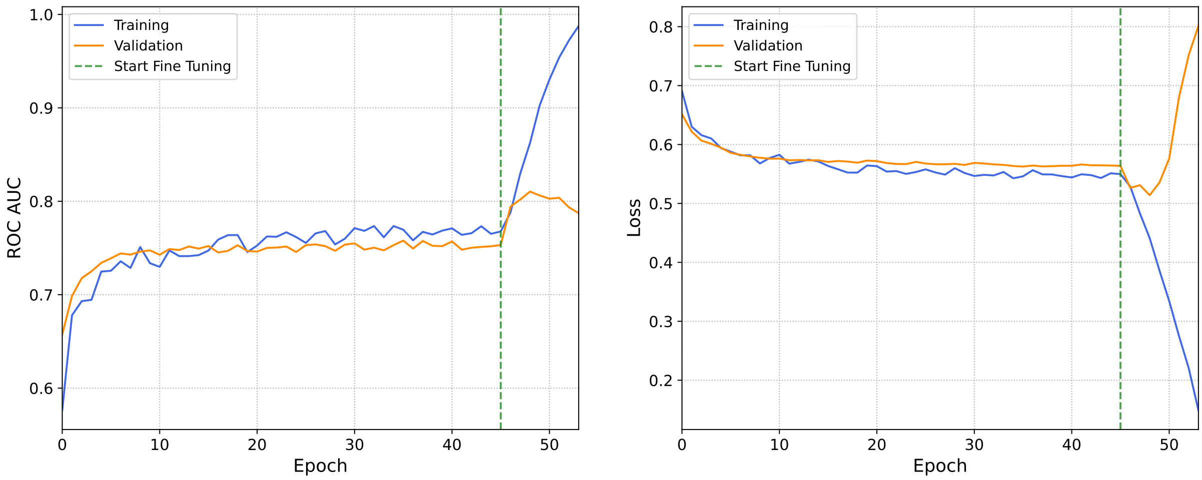
The test set performance, detailed in Table 12, demonstrates a clear convergence in discrimination capability. The Custom CNN and the optimized feature-based models (Non-linear SVM, Random Forest, XGBoost, and MLP) achieved closely aligned Test ROC AUC values ranging from 0.896 ± 0.004 to 0.902 ± 0.002. This near-equivalence

Figure 56 – Training and validation curves for the VGG11 architecture (Fold 5), with ROC AUC on the left and loss on the right.



Source: Prepared by the author (2026).

Figure 57 – Training and validation curves for the ViT-Tiny architecture (Fold 4), with ROC AUC on the left and loss on the right.



Source: Prepared by the author (2026).

confirms once again that the manually engineered features contain relevant discriminative information, matching the learning of the Custom CNN.

Although the predictive performance is similar, substantial differences emerge in computational efficiency. The Custom CNN, that demands spatial convolutions for every prediction, limits its processing speed to 2130 ± 350 events/s. Conversely, the feature-based models achieved throughputs orders of magnitude higher. XGBoost processes $(1.33 \pm 0.12) \times 10^5$ events/s, while the MLP reached the maximum rate of $(5.11 \pm 2.30) \times 10^5$ events/s. For this reason, these two models are the most promising candidates for offline data processing in the CYGNO experiment.

Furthermore, Table 12 also reports the test metrics for the transfer learning architectures. Even after fine-tuning, these models failed to exceed a ROC AUC of 0.82 (VGG11) and exhibited the lowest throughput rates of the study (< 1200 events/s). Given the inferior performance and higher computational cost, the transfer learning architectures are not considered in subsequent analyses. The remainder of this study uses the Custom CNN as the definitive image-based classifier, benchmarking its performance against the optimized feature-based algorithms established in the previous section.

Table 12 – Global comparison: performance and throughput on held-out dataset.

Model	Test ROC AUC	Throughput (event/s)
Feature-based Machine Learning		
Logistic Regression	0.887 ± 0.001	$(3.93 \pm 1.40) \times 10^5$
Linear SVM	0.882 ± 0.001	$(1.71 \pm 0.20) \times 10^4$
Non-linear SVM	0.898 ± 0.001	8481 ± 100
Random Forest	0.900 ± 0.001	$(6.46 \pm 0.21) \times 10^4$
XGBoost	0.901 ± 0.002	$(1.33 \pm 0.12) \times 10^5$
MLP	0.902 ± 0.002	$(5.11 \pm 2.30) \times 10^5$
Image-based Deep Learning		
Training from Scratch		
Custom CNN	0.896 ± 0.004	2130 ± 350
Transfer Learning		
VGG11	0.788 ± 0.003	391 ± 13
ResNet18	0.646 ± 0.003	$(1.12 \pm 0.04) \times 10^3$
ViT-Tiny	0.756 ± 0.001	906 ± 28
Fine-Tuning		
VGG11	0.820 ± 0.009	375 ± 6
ResNet18	0.803 ± 0.003	$(1.14 \pm 0.03) \times 10^3$
ViT-Tiny	0.794 ± 0.007	919 ± 16

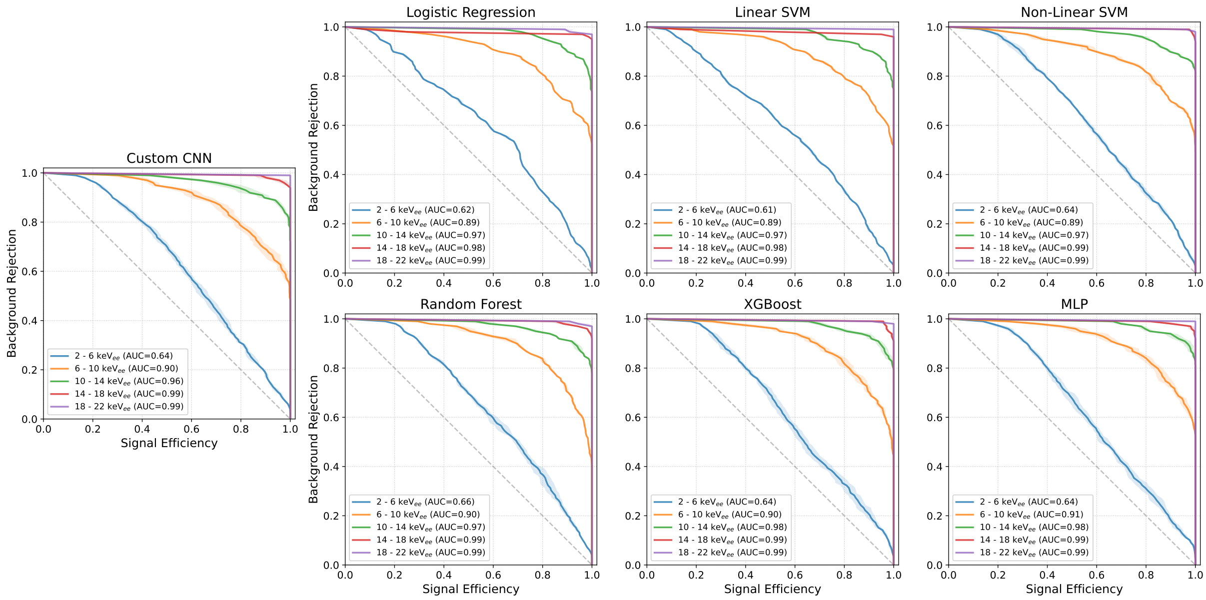
Source: Prepared by the author (2026).

7.2.1 Energy-dependent ROC curve analysis

The evaluation on the test set based on Table 12, while useful for a general comparison of model performance, does not provide a detailed analysis, as it aggregates results across all different energy levels in the test dataset. In this study, the calibrated (measured) energy was used instead of the discrete simulated energy values, in order to account for the non-linear detector response and saturation effects, which can introduce biases if the true simulated energy is directly used. To provide a more detailed evaluation, the test set was segmented into continuous calibrated energy bins of 4 $\mathrm{keV_{ee}}$ (e.g., 2–6 $\mathrm{keV_{ee}}$, 6–10 $\mathrm{keV_{ee}}$).

Figure 58 presents a grid of ROC curves for each optimized model across the calibrated energy spectrum. The ROC profiles show a clear dependence on energy. At low energies (≤ 6 keV$_{ee}$), the performance approaches random classification (AUC ≈ 0.5), since both electronic and nuclear recoils produce short tracks with limited ionization, making their topological differences less distinguishable, as previously illustrated in Fig. 31. In addition, the effects of diffusion further reduce the spatial information available for discrimination. In contrast, at higher energies (≥ 18 keV$_{ee}$), the improved spatial definition of the recoil tracks yields near-perfect separability (AUC ≈ 1.0) across all algorithms.

Figure 58 – ROC curves evaluated across calibrated energy bins for the Custom CNN (left) and the classical feature-based models (right).

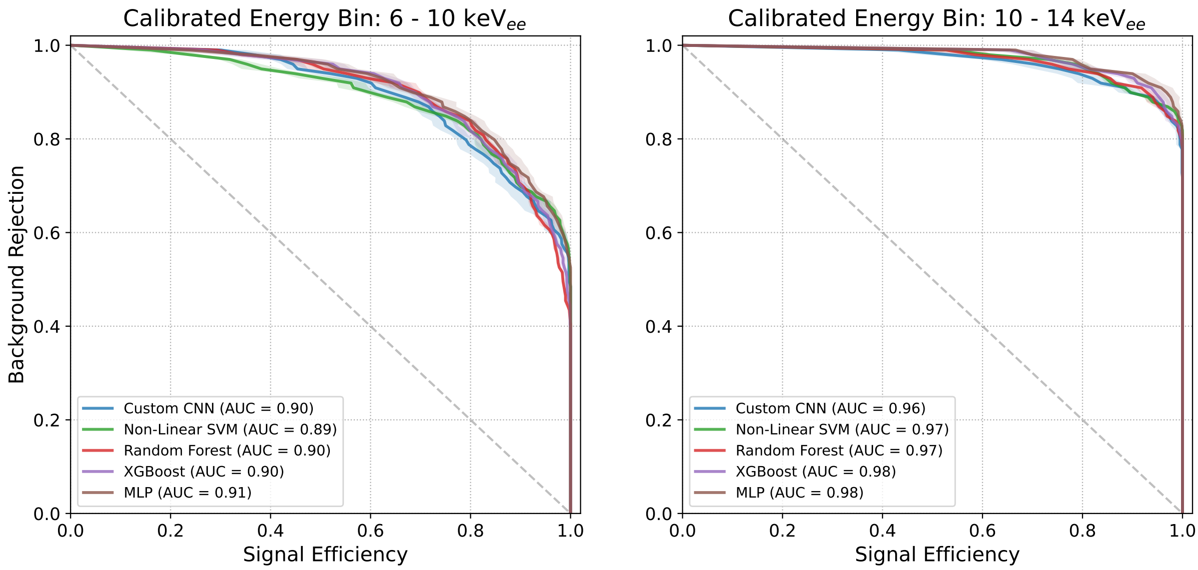


Source: Prepared by the author (2026).

Because extreme energies represent trivial regimes (either indistinguishable or perfectly separable), the most informative comparison between the models occurs in the intermediate energy regions. Figure 59 therefore compares the algorithms in the 6–10 keVee and 10–14 keVee energy bins. For readability, only the top-performing non-linear models are displayed. The linear classifiers (Logistic Regression and Linear SVM) are omitted since they consistently achieved lower ROC AUC values than the models shown.

In experimental DM searches, the achievable background rejection depends on the selected signal efficiency operating point. Nevertheless, the intermediate-energy ROC curves reveal a consistent ranking among the evaluated architectures. The feature-based models, particularly XGBoost and the MLP, produce nearly overlapping profiles that remain comparable to, and in some regions above, the performance of the Custom CNN across the signal efficiency range.

Figure 59 – Background Rejection vs. Signal Efficiency curves within the intermediate energy bins (6–10 and 10–14 keV_{ee}).



Source: Prepared by the author (2026).

This behavior shows that classical classifiers trained on the extracted observables achieve a level of separability comparable to that obtained directly from image-based analysis. The practical impact of this observation is further investigated in the next section through the rejection power metric, which evaluates the models at fixed signal-efficiency operating points.

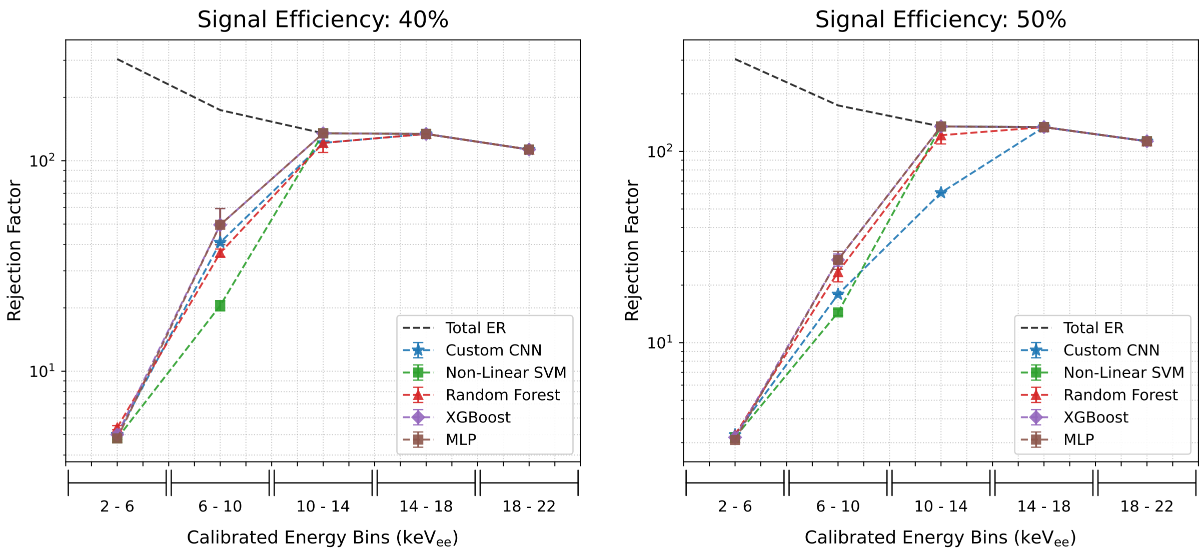
7.2.2 Rejection power

While the ROC AUC provides a generalized performance metric, experimental searches for rare events dictate specific operational thresholds. The Rejection Factor ($R = 1/\epsilon^B$) serves as a fundamental measure to assess particle discrimination power. To evaluate this metric, a probability threshold is determined on the test set to secure a fixed Signal Efficiency (e.g., 40% or 50%). The rejection power is then calculated based on the fraction of ERs that surpass this threshold.

Mathematically, a complete background rejection results in an infinite Rejection Factor. To prevent division by zero, the metric is bounded by the total number of ER samples available within each respective energy bin of the test set. This statistical limit is denoted as “Total ER” in the plots. Consequently, the measured rejection power in this study is capped near the order of 10^2, reflecting the size of the current simulation rather than the maximum discrimination capability of the algorithms. An advantage of the Rejection Factor is that validating higher performance thresholds (e.g., $R \approx 10^4$ or 10^5) does not require generating NR datasets. Future benchmarks will only demand massive, dedicated background simulations (ER), avoiding the computational cost of simulating energy-equivalent NR tracks.

Figure 60 illustrates the Rejection Factor of the optimized models across the calibrated energy spectrum. The exact numerical results are detailed in Tables 13 and 14 for the 40% and 50% signal efficiencies, respectively.

Figure 60 – Rejection Factor ($1/\epsilon^B$) as a function of calibrated energy bins at fixed Signal Efficiencies of 40% (Left) and 50% (Right). The black dashed line represents the maximum statistical limit defined by the test set size.



Source: Prepared by the author (2026).

Table 13 – Rejection Factor across different energy bins (keV$_{ee}$) at 40% signal efficiency.

Model	**Energy Bins (keV$_{ee}$)**				
	2–6	**6–10**	**10–14**	**14–18**	**18–22**
Logistic Regression	3.9 ± 0.1	24.9 ± 0.0	135.0 ± 0.0	44.7 ± 0.0	113.0 ± 0.0
Linear SVM	3.6 ± 0.1	29.0 ± 0.0	135.0 ± 0.0	67.0 ± 0.0	113.0 ± 0.0
Non-linear SVM	4.8 ± 0.1	20.5 ± 2.6	135.0 ± 0.0	134.0 ± 0.0	113.0 ± 0.0
Random Forest	5.4 ± 0.3	36.5 ± 3.5	121.5 ± 27.0	134.0 ± 0.0	113.0 ± 0.0
XGBoost	5.0 ± 0.3	49.6 ± 21.6	135.0 ± 0.0	134.0 ± 0.0	113.0 ± 0.0
MLP	4.8 ± 0.5	49.6 ± 21.6	135.0 ± 0.0	134.0 ± 0.0	113.0 ± 0.0
Custom CNN	5.0 ± 0.3	40.9 ± 11.0	121.5 ± 27.0	134.0 ± 0.0	113.0 ± 0.0

Source: Prepared by the author (2026).

The energy dependence of the rejection factor follows the same trend observed in the ROC analysis. In the lowest energy bin (2–6 keVee), all models exhibit limited discrimination power, with rejection factors below 5.5. In contrast, above 14 keVee, most

Table 14 – Rejection Factor across different energy bins (keV$_{ee}$) at 50% signal efficiency.

Model	**Energy Bins (keV$_{ee}$)**				
	2–6	**6–10**	**10–14**	**14–18**	**18–22**
Logistic Regression	3.0 ± 0.0	17.4 ± 0.0	135.0 ± 0.0	44.7 ± 0.0	113.0 ± 0.0
Linear SVM	2.9 ± 0.1	19.1 ± 2.1	135.0 ± 0.0	62.5 ± 8.9	113.0 ± 0.0
Non-linear SVM	3.2 ± 0.1	14.4 ± 1.2	135.0 ± 0.0	134.0 ± 0.0	113.0 ± 0.0
Random Forest	3.3 ± 0.2	23.4 ± 5.8	121.5 ± 27.0	134.0 ± 0.0	113.0 ± 0.0
XGBoost	3.2 ± 0.3	27.1 ± 4.5	135.0 ± 0.0	134.0 ± 0.0	113.0 ± 0.0
MLP	3.1 ± 0.2	27.1 ± 6.5	135.0 ± 0.0	134.0 ± 0.0	113.0 ± 0.0
Custom CNN	3.3 ± 0.2	17.9 ± 1.3	60.8 ± 13.5	134.0 ± 0.0	113.0 ± 0.0

Source: Prepared by the author (2026).

classifiers reach the statistical limit imposed by the available ER sample, indicating that the separation between electron and nuclear recoils becomes trivial in this energy bin.

The largest differences between architectures appear in the intermediate energy bins (6–10 and 10–14 keVee), where discrimination remains challenging. At 50% signal efficiency, XGBoost and the MLP achieve rejection factors of 27.1 ± 4.5 and 27.1 ± 6.5 in the 6–10 keVee interval, compared to 17.9 ± 1.3 for the Custom CNN. A similar trend is observed at 40% signal efficiency, where both models reach 49.6 ± 21.6, while the CNN achieves 40.9 ± 11.0. In the 10–14 keV$_{\mathrm{ee}}$ range, several feature-based models reach the statistical limit of the dataset, whereas the CNN remains below this bound at the 50% efficiency operating point. These results are consistent with the ROC analysis and indicate that the feature-based representation retains the information required for ER/NR discrimination.

Among the evaluated classifiers, XGBoost and the MLP provide the most favorable combination of background rejection and computational efficiency. Their performance in the intermediate energy region, where ER/NR discrimination is most challenging, remains comparable while exceeding that of the image-based approach.

Moreover, the XGBoost model achieves this performance using only 9 input features, whereas the MLP relies on 13 features, corresponding to the largest feature subset among the evaluated models. This lower input dimensionality, combined with comparable classification performance, provides an advantage in terms of computational efficiency and model simplicity when compared to the MLP.

Another advantage of XGBoost is its interpretability, as it provides a direct estimate of feature importance, allowing the classification performance to be linked to specific observables, as discussed in Section *7.1.1.2*. This capability facilitates the physical interpretation of the discrimination process and provides insight into the properties of ionization tracks that contribute to ER/NR separation.

7.3 FINAL REMARKS

The results presented throughout this chapter indicate that the discrimination performance is determined not only by the classification algorithm, but also by the information contained in the reconstructed track observables. Despite the diversity of the evaluated approaches, the best-performing feature-based and image-based models reached comparable performance levels. In particular, the optimized feature-based models consistently matched or exceeded the Custom CNN in the intermediate energy region, while requiring substantially lower computational resources. This indicates that the engineered observables retain the main information required for ER/NR discrimination in the current detector configuration.

The feature selection and interpretability analyses further showed that charge density and energy deposition observables, particularly `rms`, `dE_dA`, and `Slimness_pca`, play an important role in the classification. The consistency of these observables across different classifiers indicates that the discrimination is mainly related to differences in track compactness and charge distribution. These results provide a basis for the discussion of the limitations of the present approach and possible improvements presented in the conclusion.

8 CONCLUSION

This work presented the development and evaluation of ML frameworks for ER/NR discrimination in the CYGNO experiment, using the LIME prototype as a case study. The primary objective was to investigate whether the topology of optically reconstructed ionization tracks contains sufficient information to reject ER backgrounds while preserving NR candidates in the low-energy region relevant for rare-event searches.

To address this problem, a complete analysis pipeline was established, including detector simulation, digitization tuning, image processing, feature extraction, and supervised classification. A dedicated calibration procedure was implemented to align the MC simulations with experimental data, enabling the generation of realistic labeled datasets for training and validation. From the reconstructed tracks, a set of morphological and energy-related observables was extracted and used to train feature-based machine learning algorithms. In parallel, image-based deep learning architectures were developed to evaluate if discrimination could be improved through direct learning from raw images.

The results showed that both approaches can provide effective ER/NR discrimination. The energy-dependent analysis showed that the differences between the approaches are more evident in the intermediate energy region, where ER and NR events are more difficult to separate. In the 6–10 keV$_{ee}$ interval, XGBoost and the MLP achieved rejection factors of 27.1 ± 4.5 and 27.1 ± 6.5, respectively, at 50% signal efficiency, compared with 17.9 ± 1.3 for the Custom CNN. At 40% efficiency, both feature-based models reached 49.6 ± 21.6, while the CNN achieved 40.9 ± 11.0. In the 10–14 keV$_{ee}$ interval, XGBoost and the MLP reached the maximum rejection level allowed by the available test statistics, whereas the Custom CNN remained below this limit.

The feature selection results provide further insight into the observed performance. Observables related to charge density and energy deposition, particularly `rms`, `dE_dA`, and `Slimness_pca`, were consistently selected by different classifiers. Their consistent selection suggests that ER/NR discrimination is primarily driven by differences in track compactness and charge distribution. These findings provide a direct connection between the machine learning decision process and the underlying detector physics.

Computational efficiency provides an additional advantage to the feature-based approach. While the Custom CNN processed approximately 2×10^3 events per second, the optimized feature-based models achieved throughputs above 10^5 events per second. XGBoost provided a particularly favorable combination of discrimination performance, computational efficiency, and interpretability while using only 9 input features. These characteristics make it a practical candidate for applications in which event processing rate and model interpretability are relevant considerations.

The results also indicate that further improvements are unlikely to depend only on increasing the complexity of the classification models. The similar performance obtained by different architectures suggests that the current representations capture most of the information available from the reconstructed optical tracks. Improvements may therefore require changes in the detector response or the inclusion of complementary information. One relevant limitation identified in this work is the non-linear detector response associated with GEM saturation. The present analysis was performed using the nominal LIME configuration, with the GEMs operated at 440 V. During LIME operation, additional data-taking campaigns were performed at lower GEM voltage (420 V) and reduced drift field (0.5 kV/cm), providing alternative configurations in which the effect of reduced saturation on ER/NR discrimination could be investigated.

The optical readout can also benefit from subsequent detector developments. The ORCA-Quest camera, planned for CYGNO–04, provides lower readout noise and single-photon sensitivity compared with the ORCA-Fusion used in the present study. Its improved noise performance may enhance the reconstruction of low-energy events and preserve track information that is difficult to recover under the current conditions. A repetition of the present analysis using data or simulations representative of the new camera would therefore provide a direct assessment of its impact on classification performance.

Another important direction is the use of information from the PMT waveforms together with the camera images. The PMTs provide temporal information related to the development of the event along the drift direction, which is not contained in the two-dimensional camera image. Recent developments in PMT signal simulation for CYGNO (130) provide a basis for extending the present framework to this additional data source. In such a multimodal approach, features extracted from the PMT waveforms could be combined with the image-based observables investigated in this work. The spatial variables excluded from the present analysis, such as `xmin`, `xmax`, `ymin`, `ymax`, `xmean`, and `ymean`, could also be reconsidered, since the additional temporal information may allow spatially dependent effects to be modeled without relying exclusively on position-independent classification.

In summary, this work demonstrates that machine learning techniques provide an effective framework for ER/NR discrimination in CYGNO. The results show that a feature-based approach can achieve high performance while maintaining high computational efficiency and physical interpretability. These characteristics make it a promising strategy for future applications in directional DM searches and other rare-event experiments based on optical TPC technologies.

REFERENCES

1 Gerard Jungman, Marc Kamionkowski, and Kim Griest. Supersymmetric dark matter. *Physics Reports*, 267(5):195–373, 1996. ISSN 0370-1573. doi: https://doi.org/10.1016/0370-1573(95)00058-5. URL `https://www.sciencedirect.com/science/article/pii/0370157395000585`.

2 Fernando Domingues Amaro et al. The cygno experiment. *Instruments*, 6(1), 2022. ISSN 2410-390X. doi: 10.3390/instruments6010006. URL `https://www.mdpi.com/2410-390X/6/1/6`.

3 S. E. Vahsen et al. Cygnus: Feasibility of a nuclear recoil observatory with directional sensitivity to dark matter and neutrinos, 2020. URL `https://arxiv.org/abs/2008.12587`.

4 Tom W.B. Kibble. The standard model of particle physics. *European Review*, 23(1): 36–44, 2015. doi: 10.1017/S1062798714000520.

5 Jianglai Liu, Xun Chen, and Xiangdong Ji. Current status of direct dark matter detection experiments. *Nature Physics*, 13(3):212–216, 03 2017. ISSN 1745-2481. doi: 10.1038/nphys4039. URL `https://doi.org/10.1038/nphys4039`.

6 Gianfranco Bertone, Dan Hooper, and Joseph Silk. Particle dark matter: evidence, candidates and constraints. *Physics Reports*, 405(5):279–390, 2005. ISSN 0370-1573. doi: https://doi.org/10.1016/j.physrep.2004.08.031. URL `https://www.sciencedirect.com/science/article/pii/S0370157304003515`.

7 J. R. Bond, G. Efstathiou, and J. Silk. Massive neutrinos and the large-scale structure of the universe. *Phys. Rev. Lett.*, 45:1980–1984, Dec 1980. doi: 10.1103/PhysRevLett.45.1980. URL `https://link.aps.org/doi/10.1103/PhysRevLett.45.1980`.

8 John Ellis, J.S. Hagelin, D.V. Nanopoulos, K. Olive, and M. Srednicki. Supersymmetric relics from the big bang. *Nuclear Physics B*, 238(2):453–476, 1984. ISSN 0550-3213. doi: https://doi.org/10.1016/0550-3213(84)90461-9. URL `https://www.sciencedirect.com/science/article/pii/0550321384904619`.

9 P. SIKIVIE. Dark matter axions. *International Journal of Modern Physics A*, 25 (02n03):554–563, 2010. doi: 10.1142/S0217751X10048846. URL `https://doi.org/10.1142/S0217751X10048846`.

10 John Preskill, Mark B. Wise, and Frank Wilczek. Cosmology of the invisible axion. *Physics Letters B*, 120(1):127–132, 1983. ISSN 0370-2693. doi: https://doi.org/10.1016/0370-2693(83)90637-8. URL `https://www.sciencedirect.com/science/article/pii/0370269383906378`.

11 Scott Dodelson and Lawrence M. Widrow. Sterile neutrinos as dark matter. *Phys. Rev. Lett.*, 72:17–20, Jan 1994. doi: 10.1103/PhysRevLett.72.17. URL `https://link.aps.org/doi/10.1103/PhysRevLett.72.17`.

12 Felix Kahlhoefer. Review of lhc dark matter searches. *International Journal of Modern Physics A*, 32(13):1730006, 2017. doi: 10.1142/S0217751X1730006X. URL `https://doi.org/10.1142/S0217751X1730006X`.

13 Jennifer M. Gaskins. A review of indirect searches for particle dark matter. *Contemporary Physics*, 57(4):496–525, 2016. doi: 10.1080/00107514.2016.1175160. URL https://doi.org/10.1080/00107514.2016.1175160.

14 E. Baracchini et al. Cygno: a cygnus collaboration 1 m^3 module with optical readout for directional dark matter search, 2019. URL https://arxiv.org/abs/1901.04190.

15 Mark W. Goodman and Edward Witten. Detectability of certain dark-matter candidates. *Phys. Rev. D*, 31:3059–3063, Jun 1985. doi: 10.1103/PhysRevD.31.3059. URL https://link.aps.org/doi/10.1103/PhysRevD.31.3059.

16 Igor Abritta Costa. *Optimization of the clustering algorithm of the CYGNO experiment*. Doctorate thesis, Universidade Federal de Juiz de Fora, Faculty of Engineering, 2020. Post-Graduation Program in Electrical Engineering.

17 E. Aprile et al. Dark matter results from 225 live days of xenon100 data. *Phys. Rev. Lett.*, 109:181301, Nov 2012. doi: 10.1103/PhysRevLett.109.181301. URL https://link.aps.org/doi/10.1103/PhysRevLett.109.181301.

18 Rita Bernabei et al. First model independent results from dama/libra–phase2. *Universe*, 4(11), 2018. ISSN 2218-1997. doi: 10.3390/universe4110116. URL https://www.mdpi.com/2218-1997/4/11/116.

19 Lucy Reading-Ikkanda. Trouble detected in infamous dark matter signal. Quanta Magazine, 04 2018. URL https://www.quantamagazine.org/trouble-detected-in-infamous-dark-matter-signal-20180412/. Acessed in: 20 June 2026.

20 Julien Billard et al. Direct detection of dark matter—appec committee report*. *Reports on Progress in Physics*, 85(5):056201, apr 2022. doi: 10.1088/1361-6633/ac5754. URL https://doi.org/10.1088/1361-6633/ac5754.

21 S.P. Ahlen, F.T. Avignone, R.L. Brodzinski, A.K. Drukier, G. Gelmini, and D.N. Spergel. Limits on cold dark matter candidates from an ultralow background germanium spectrometer. *Physics Letters B*, 195(4):603–608, 1987. ISSN 0370-2693. doi: https://doi.org/10.1016/0370-2693(87)91581-4. URL https://www.sciencedirect.com/science/article/pii/0370269387915814.

22 A. Aguilar-Arevalo et al. Search for low-mass wimps in a 0.6 kg day exposure of the damic experiment at snolab. *Phys. Rev. D*, 94:082006, Oct 2016. doi: 10.1103/PhysRevD.94.082006. URL https://link.aps.org/doi/10.1103/PhysRevD.94.082006.

23 Michael Crisler et al. Sensei: First direct-detection constraints on sub-gev dark matter from a surface run. *Phys. Rev. Lett.*, 121:061803, Aug 2018. doi: 10.1103/PhysRevLett.121.061803. URL https://link.aps.org/doi/10.1103/PhysRevLett.121.061803.

24 Jungsic Park. Status of the COSINE Experiment. In *Park, Jungsic "Status of the COSINE Experiment" in 12th "Patras" Workshop on Axions, WIMPs and WISPs, PATRAS 2016 / Kim, Younglm, Lindner, Axel, Semertzidis, Yannis K. (eds.), Verlag Deutsches Elektronen-Synchrotron : 2017 ; PATRAS 2016 : 12th "Patras" Workshop on Axions, WIMPs and WISPs, 2016-06-20 - 2016-06-24, Jeju Island,*

pages 125–128. 12th Patras Workshop on Axions, WIMPs and WISPs, Jeju Island (South Korea), 20 Jun 2016 - 24 Jun 2016, Verlag Deutsches Elektronen-Synchrotron, 2017. doi: 10.3204/DESY-PROC-2009-03/Park_Jungsic. URL `https://bib-pubdb1.desy.de/record/395492`.

25 N.W. Ashcroft and N.D. Mermin. *Solid State Physics*. HRW international editions. Holt, Rinehart and Winston, 1976. ISBN 9780030839931. URL `https://books.google.com.br/books?id=1C9HAQAAIAAJ`.

26 A. Broniatowski et al. Cryogenic ge detectors with interleaved electrodes: Design and modeling. *Journal of Low Temperature Physics*, 151(3):830–834, 05 2008. ISSN 1573-7357. doi: 10.1007/s10909-008-9754-1. URL `https://doi.org/10.1007/s10909-008-9754-1`.

27 D.S. Akerib et al. Surface event rejection using phonon information in cdms. *Nuclear Physics B - Proceedings Supplements*, 173:137–140, 2007. ISSN 0920-5632. doi: https://doi.org/10.1016/j.nuclphysbps.2007.08.038. URL `https://www.sciencedirect.com/science/article/pii/S0920563207006639`. Proceedings of the 7th UCLA Symposium on Sources and Detection of Dark Matter and Dark Energy in the Universe.

28 P. Agnes et al. Results from the first use of low radioactivity argon in a dark matter search. *Phys. Rev. D*, 93:081101(R), Apr 2016. doi: 10.1103/PhysRevD.93.081101. URL `https://link.aps.org/doi/10.1103/PhysRevD.93.081101`.

29 C. E. Aalseth et al. Darkside-20k: A 20 tonne two-phase lar tpc for direct dark matter detection at lngs. *The European Physical Journal Plus*, 133(3):131, 03 2018. ISSN 2190-5444. doi: 10.1140/epjp/i2018-11973-4. URL `https://doi.org/10.1140/epjp/i2018-11973-4`.

30 P.-A. Amaudruz et al. Measurement of the scintillation time spectra and pulse-shape discrimination of low-energy β and nuclear recoils in liquid argon with deap-1. *Astroparticle Physics*, 85:1–23, 2016. ISSN 0927-6505. doi: https://doi.org/10.1016/j.astropartphys.2016.09.002. URL `https://www.sciencedirect.com/science/article/pii/S0927650516301232`.

31 M. Kuźniak et al. Deap-3600 dark matter search. *Nuclear and Particle Physics Proceedings*, 273-275:340–346, 2016. ISSN 2405-6014. doi: https://doi.org/10.1016/j.nuclphysbps.2015.09.048. URL `https://www.sciencedirect.com/science/article/pii/S2405601415005374`. 37th International Conference on High Energy Physics (ICHEP).

32 K. Abe et al. Light wimp search in xmass. *Physics Letters B*, 719(1):78–82, 2013. ISSN 0370-2693. doi: https://doi.org/10.1016/j.physletb.2013.01.001. URL `https://www.sciencedirect.com/science/article/pii/S0370269313000117`.

33 D. Akerib et al. First results from the lux dark matter experiment at the sanford underground research facility. *Phys. Rev. Lett.*, 112:091303, Mar 2014. doi: 10.1103/PhysRevLett.112.091303. URL `https://link.aps.org/doi/10.1103/PhysRevLett.112.091303`.

34 J. Angle et al. Search for light dark matter in xenon10 data. *Phys. Rev. Lett.*, 107: 051301, Jul 2011. doi: 10.1103/PhysRevLett.107.051301. URL `https://link.aps.org/doi/10.1103/PhysRevLett.107.051301`.

35 F Aubin et al. Discrimination of nuclear recoils from alpha particles with superheated liquids. *New Journal of Physics*, 10(10):103017, oct 2008. doi: 10.1088/1367-2630/10/10/103017. URL `https://doi.org/10.1088/1367-2630/10/10/103017`.

36 M. Barnabé-Heider et al. Response of superheated droplet detectors of the picasso dark matter search experiment. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 555(1):184–204, 2005. ISSN 0168-9002. doi: https://doi.org/10.1016/j.nima.2005.09.015. URL `https://www.sciencedirect.com/science/article/pii/S0168900205018255`.

37 C. Amole et al. Improved dark matter search results from pico-2l run 2. *Phys. Rev. D*, 93:061101(R), Mar 2016. doi: 10.1103/PhysRevD.93.061101. URL `https://link.aps.org/doi/10.1103/PhysRevD.93.061101`.

38 C. Amole et al. Dark matter search results from the pico-60 cf_3I bubble chamber. *Phys. Rev. D*, 93:052014, Mar 2016. doi: 10.1103/PhysRevD.93.052014. URL `https://link.aps.org/doi/10.1103/PhysRevD.93.052014`.

39 W J Bolte et al. A bubble chamber for dark matter detection (the coupp project status). *Journal of Physics: Conference Series*, 39(1):126, may 2006. doi: 10.1088/1742-6596/39/1/027. URL `https://doi.org/10.1088/1742-6596/39/1/027`.

40 M. Tanabashi et al. Review of particle physics. *Phys. Rev. D*, 98:030001, Aug 2018. doi: 10.1103/PhysRevD.98.030001. URL `https://link.aps.org/doi/10.1103/PhysRevD.98.030001`.

41 D Santos et al. Mimac: Micro-tpc matrix of chambers for dark matter directional detection. *Journal of Physics: Conference Series*, 469(1):012002, dec 2013. doi: 10.1088/1742-6596/469/1/012002. URL `https://doi.org/10.1088/1742-6596/469/1/012002`.

42 J.B.R. Battat et al. Low threshold results and limits from the drift directional dark matter detector. *Astroparticle Physics*, 91:65–74, 2017. ISSN 0927-6505. doi: https://doi.org/10.1016/j.astropartphys.2017.03.007. URL `https://www.sciencedirect.com/science/article/pii/S0927650517300075`.

43 J Castel et al. The trex-dm experiment at the canfranc underground laboratory. *Journal of Physics: Conference Series*, 1468(1):012063, feb 2020. doi: 10.1088/1742-6596/1468/1/012063. URL `https://doi.org/10.1088/1742-6596/1468/1/012063`.

44 M. Ghrear, S.E. Vahsen, and C. Deaconu. Observables for recoil identification in high-definition gas time projection chambers. *Journal of Cosmology and Astroparticle Physics*, 2021(10):005, oct 2021. doi: 10.1088/1475-7516/2021/10/005. URL `https://doi.org/10.1088/1475-7516/2021/10/005`.

45 N.S. Phan, R.J. Lauer, E.R. Lee, D. Loomba, J.A.J. Matthews, and E.H. Miller. Gem-based tpc with ccd imaging for directional dark matter detection. *Astroparticle Physics*, 84:82–96, 2016. ISSN 0927-6505. doi: https://doi.org/10.1016/j.astropartphys.2016.08.006. URL `https://www.sciencedirect.com/science/article/pii/S0927650516301190`.

46 J. Billard, F. Mayet, and D. Santos. Low energy electron/recoil discrimination for directional dark matter detection. *Journal of Cosmology and Astroparticle Physics*, 2012(07):020, jul 2012. doi: 10.1088/1475-7516/2012/07/020. URL `https://doi.org/10.1088/1475-7516/2012/07/020`.

47 P. Baldi, P. Sadowski, and D. Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5(1):4308, 07 2014. ISSN 2041-1723. doi: 10.1038/ncomms5308. URL `https://doi.org/10.1038/ncomms5308`.

48 Peter Sadowski, Balint Radics, Ananya, Yasunori Yamazaki, and Pierre Baldi. Efficient antihydrogen detection in antimatter physics by deep learning, 2017. URL `https://arxiv.org/abs/1706.01826`.

49 Pierre Baldi, Kevin Bauer, Clara Eng, Peter Sadowski, and Daniel Whiteson. Jet substructure classification in high-energy physics with deep neural networks. *Phys. Rev. D*, 93:094034, May 2016. doi: 10.1103/PhysRevD.93.094034. URL `https://link.aps.org/doi/10.1103/PhysRevD.93.094034`.

50 R. Acciarri et al. Convolutional neural networks applied to neutrino events in a liquid argon time projection chamber. *Journal of Instrumentation*, 12(03):P03011, mar 2017. doi: 10.1088/1748-0221/12/03/P03011. URL `https://doi.org/10.1088/1748-0221/12/03/P03011`.

51 J. Renner et al. Background rejection in next using deep neural networks. *Journal of Instrumentation*, 12(01):T01004, jan 2017. doi: 10.1088/1748-0221/12/01/T01004. URL `https://doi.org/10.1088/1748-0221/12/01/T01004`.

52 Christian Szegedy et al. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. doi: 10.1109/CVPR.2015.7298594.

53 J. Schueler, M. Ghrear, S. E. Vahsen, P. Sadowski, and C. Deaconu. Deep learning for improved kev-scale recoil identification in high resolution gas time projection chambers, 2022. URL `https://arxiv.org/abs/2206.10822`.

54 David R. Nygren. The time projection chamber. Report LBL-7577, Lawrence Berkeley National Laboratory, 1978.

55 Juan J. Gomez-Cadenas, Francesc Monrabal Capilla, and Paola Ferrario. High pressure gas xenon tpcs for double beta decay searches. *Frontiers in Physics*, Volume 7 - 2019, 2019. ISSN 2296-424X. doi: 10.3389/fphy.2019.00051. URL `https://www.frontiersin.org/journals/physics/articles/10.3389/fphy.2019.00051`.

56 E. Baracchini et al. Stability and detection performance of a gem-based optical readout tpc with he/cf4 gas mixtures. *Journal of Instrumentation*, 15(10):P10001,

oct 2020. doi: 10.1088/1748-0221/15/10/P10001. URL https://doi.org/10.1088/1748-0221/15/10/P10001.

57 A. Pansky, A. Breskin, A. Buzulutskov, R. Chechik, V. Elkind, and J. Va'vra. The scintillation of cf4 and its relevance to detection science. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 354(2):262–269, 1995. ISSN 0168-9002. doi: https://doi.org/10.1016/0168-9002(94)01064-1. URL https://www.sciencedirect.com/science/article/pii/0168900294010641.

58 Rita Roque, R. Daniel Mano, F. Amaro, Cristina Monteiro, and J.M.F. Santos. Electroluminescence studies in he-cf4 -isobutane mixtures, 06 2021.

59 F. D. Amaro et al. He-cf4-ch4 ternary mixtures as target gas for the cygno directional dark matter experiment, 2026. URL https://arxiv.org/abs/2606.01512.

60 F. Sauli. Gem: A new concept for electron amplification in gas detectors. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 386(2):531–534, 1997. ISSN 0168-9002. doi: https://doi.org/10.1016/S0168-9002(96)01172-2. URL https://www.sciencedirect.com/science/article/pii/S0168900296011722.

61 Fabio Sauli. Micro-pattern gas detectors. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 477(1):1–7, 2002. ISSN 0168-9002. doi: https://doi.org/10.1016/S0168-9002(01)01903-9. URL https://www.sciencedirect.com/science/article/pii/S0168900201019039. 5th Int. Conf. on Position-Sensitive Detectors.

62 S. Bachmann, A. Bressan, S. Kappler, B. Ketzer, M. Deutel, L. Ropelewski, F. Sauli, and E. Schulte. Development and applications of the gas electron multiplier. *Nuclear Instruments and Methods in Physics Research. Section A, Accelerators, Spectrometers, Detectors and Associated Equipment*, 471(1-2):p. 115–119, set 2001. ISSN 0168-9002.

63 D. Pinci. *A triple GEM detector for the muon system of the LHCb experiment.* Phd thesis, University of Cagliari, 2006.

64 V. Brio, V. Bellini, E. Cisbani, C. Petta, L. Re, and C. Sutera. Simulations of triple-gem tracker's response for experiments at jefferson lab. *Radiation Effects and Defects in Solids*, 174(11-12):1031–1046, 2019. doi: 10.1080/10420150.2019.1683841. URL https://doi.org/10.1080/10420150.2019.1683841.

65 P. R. B. Marinho. *Desenvolvimento de detectores Sensíveis à posição Multifilares e Multi-GEM para Obtenção de Imagens de Raios-X.* Doctorate thesis, CBPF, Rio de Janeiro, 2006.

66 V.C. Antochi et al. Performance of an optically read out time projection chamber with ultra-relativistic electrons. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 999:165209, 2021. ISSN 0168-9002. doi: https://doi.org/10.1016/j.nima.2021.165209. URL https://www.sciencedirect.com/science/article/pii/S0168900221001935.

67 Hamamatsu Photonics. *ORCA-Flash4.0 LT+ Digital CMOS camera C13440-20CU*. Hamamatsu Photonics, 2018. URL https://www.hamamatsu.com/eu/en/product/cameras/cmos-cameras/C13440-20CU.html. Acessado em: 20 de junho de 2026.

68 Hamamatsu Photonics. *ORCA-Fusion Digital CMOS camera C14440-20UP*. Hamamatsu Photonics, 2021. URL https://www.hamamatsu.com/eu/en/product/cameras/cmos-cameras/C14440-20UP.html.

69 Hamamatsu Photonics. *ORCA-Quest qCMOS camera C15550-22UP*. Hamamatsu Photonics, 2025. URL https://www.hamamatsu.com/eu/en/product/cameras/qcmos-cameras/C15550-22UP.html.

70 B D Almeida et al. Noise assessment of cmos active pixel sensors for the cygno experiment. *Measurement Science and Technology*, 34(12):125145, sep 2023. doi: 10.1088/1361-6501/acf7e1. URL https://doi.org/10.1088/1361-6501/acf7e1.

71 *Photomultiplier Tube R7378A Datasheet*. Hamamatsu Photonics, 2017.

72 M.M.F.R. Fraga, F.A.F. Fraga, S.T.G. Fetal, L.M.S. Margato, R.Ferreira Marques, and A.J.P.L. Policarpo. The gem scintillation in he–cf4, ar–cf4, ar–tea and xe–tea mixtures. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 504(1):88–92, 2003. ISSN 0168-9002. doi: https://doi.org/10.1016/S0168-9002(03)00758-7. URL https://www.sciencedirect.com/science/article/pii/S0168900203007587. Proceedings of the 3rd International Conference on New Developments in Photodetection.

73 CAEN S.p.A. *UM3152: V1742/VX1742/DT5742 32 Ch. 12-bit 5 GS/s Switched Capacitor Digitizer User's Manual*, 2024. URL https://www.caen.it.

74 CAEN S.p.A. *UM1934: V1720/VX1720/DT5720 8 Ch. 12-bit 250 MS/s Digitizer User's Manual*, 2024. URL https://www.caen.it.

75 Fernando Domingues Amaro et al. Bayesian network 3d event reconstruction in the cygno optical tpc for dark matter direct detection. *The European Physical Journal C*, 85(11):1261, 11 2025. ISSN 1434-6052. doi: 10.1140/epjc/s10052-025-14965-6. URL https://doi.org/10.1140/epjc/s10052-025-14965-6.

76 E Baracchini et al. Identification of low energy nuclear recoils in a gas time projection chamber with optical readout. *Measurement Science and Technology*, 32 (2):025902, dec 2020. doi: 10.1088/1361-6501/abbd12. URL https://doi.org/10.1088/1361-6501/abbd12.

77 A. Prajapati. *Multivariate Analysis for Background Rejection in CYGNO/INITIUM Experiment*. Phd thesis, Gran Sasso Science Institute, 2024. Astroparticle Physics, Cycle XXXV.

78 CYGNUS-RD Collaboration. reconstruction, 2026. URL https://github.com/CYGNUS-RD/reconstruction. GitHub repository.

79 F D Amaro et al. Directional idbscan to detect cosmic-ray tracks for the cygno experiment. *Measurement Science and Technology*, 34(12):125024, sep 2023. doi: 10.1088/1361-6501/acf402. URL https://doi.org/10.1088/1361-6501/acf402.

80 Andy Rowlands. *Physics of Digital Photography*. 2053-2563. IOP Publishing, 2017. ISBN 978-0-7503-1242-4. doi: 10.1088/978-0-7503-1242-4. URL `https://doi.org/10.1088/978-0-7503-1242-4`.

81 Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

82 E. Baracchini et al. A density-based clustering algorithm for the cygno data analysis. *Journal of Instrumentation*, 15(12):T12003, dec 2020. doi: 10.1088/1748-0221/15/12/T12003. URL `https://doi.org/10.1088/1748-0221/15/12/T12003`.

83 SciPy 1.0 Contributors. *SciPy: Open source scientific tools for Python*, 2026. URL `https://www.scipy.org/`.

84 van der Walt et al. *scikit-image: image processing in Python*, 2014. URL `https://scikit-image.org/`.

85 J Lindhard, V Nielsen, M Scharff, and P V Thomsen. Integral equations governing radiation effects. (notes on atomic collisions, iii). *Kgl. Danske Videnskab., Selskab. Mat. Fys. Medd.*, Vol: 33: No. 10, 01 1963. URL `https://www.osti.gov/biblio/4701226`.

86 Flaminia Di Giambattista. *Background study of the LIME prototype at underground LNGS for the CYGNO experiment*. Phd thesis, Gran Sasso Science Institute, Italy, 2024.

87 S. Agostinelli et al. Geant4—a simulation toolkit. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506(3):250–303, 2003. ISSN 0168-9002. doi: https://doi.org/10.1016/S0168-9002(03)01368-8. URL `https://www.sciencedirect.com/science/article/pii/S0168900203013688`.

88 CYGNUS-RD Collaboration. digitizationpp, 2026. URL `https://github.com/CYGNUS-RD/digitizationpp`. GitHub repository.

89 F.D. Amaro et al. Simulation of the cygno gaseous tpc optical readout. *Journal of Instrumentation*, 21(06):P06024, jun 2026. doi: 10.1088/1748-0221/21/06/P06024. URL `https://doi.org/10.1088/1748-0221/21/06/P06024`.

90 Fernando Domingues Amaro et al. Modeling the light response of an optically readout gem based tpc for the cygno experiment. *The European Physical Journal C*, 86(2):123, 02 2026. ISSN 1434-6052. doi: 10.1140/epjc/s10052-026-15318-7. URL `https://doi.org/10.1140/epjc/s10052-026-15318-7`.

91 Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

92 F. Pedregosa et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

93 Tianqi Chen et al. *xgboost: Extreme Gradient Boosting*, 2026. URL `https://github.com/dmlc/xgboost`. R package version 3.3.0.1.

94 David W. Hosmer Jr., Stanley Lemeshow, and Rodney X. Sturdivant. *Applied Logistic Regression*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 3rd edition, 2013. ISBN 9780470582473. doi: 10.1002/9781118548387.

95 C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, 1995. doi: 10.1007/BF00994018.

96 J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 03 1986. ISSN 1573-0565. doi: 10.1007/BF00116251. URL `https://doi.org/10.1007/BF00116251`.

97 L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

98 Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232, 2001. doi: 10.1214/aos/1013203451. URL `https://doi.org/10.1214/aos/1013203451`.

99 Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL `http://doi.acm.org/10.1145/2939672.2939785`.

100 Hannes Schulz and Sven Behnke. Deep learning. *KI - Künstliche Intelligenz*, 26(4): 357–363, 11 2012. ISSN 1610-1987. doi: 10.1007/s13218-012-0198-z. URL `https://doi.org/10.1007/s13218-012-0198-z`.

101 Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. doi: 10.1037/h0042519. URL `https://doi.org/10.1037/h0042519`.

102 Hong Hui Tan and King Hann Lim. Vanishing gradient mitigation with deep learning neural network optimization. In *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, pages 1–4, 2019. doi: 10.1109/ICSCC.2019.8843652.

103 Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 807–814, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

104 Raúl Rojas. *The Backpropagation Algorithm*, pages 149–182. Springer Berlin Heidelberg, Berlin, Heidelberg, 1996. ISBN 978-3-642-61068-4. doi: 10.1007/978-3-642-61068-4_7. URL `https://doi.org/10.1007/978-3-642-61068-4_7`.

105 Katarzyna Janocha and Wojciech Marian Czarnecki. On loss functions for deep neural networks in classification, 2017. URL https://arxiv.org/abs/1702.05659.

106 Sebastian Ruder. An overview of gradient descent optimization algorithms, 2017. URL https://arxiv.org/abs/1609.04747.

107 Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

108 Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL https://arxiv.org/abs/1409.1556.

109 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

110 Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

111 Adam Paszke et al. Pytorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703.

112 Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 12 1989. ISSN 0899-7667. doi: 10.1162/neco.1989.1.4.541. URL https://doi.org/10.1162/neco.1989.1.4.541.

113 Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017. ISSN 0001-0782. doi: 10.1145/3065386. URL https://doi.org/10.1145/3065386.

114 Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.

115 Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. URL https://arxiv.org/abs/1905.11946.

116 Zhuang Liu et al. A convnet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976, 2022. doi: 10.1109/CVPR52688.2022.01167.

117 Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

118 Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers, 2022. URL https://arxiv.org/abs/2106.10270.

119 Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/touvron21a.html`.

120 Ashish Vaswani et al. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

121 Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, page 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603638.

122 Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.*, 10(7):1895–1923, 1998. ISSN 0899-7667. doi: 10.1162/089976698300017197. URL `https://doi.org/10.1162/089976698300017197`.

123 João Manoel Herrera Pinheiro et al. The impact of feature scaling in machine learning: Effects on regression and classification tasks. *IEEE Access*, 13: 199903–199931, 2025. doi: 10.1109/ACCESS.2025.3635541.

124 Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 3rd edition, 2011. ISBN 978-0-12-381479-1. URL `https://shop.elsevier.com/books/data-mining-concepts-and-techniques/han/978-0-12-381479-1`.

125 P. Pudil, J. Novovičová, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119–1125, 1994. ISSN 0167-8655. doi: https://doi.org/10.1016/0167-8655(94)90127-9. URL `https://www.sciencedirect.com/science/article/pii/0167865594901279`.

126 Sebastian Raschka. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *Journal of Open Source Software*, 3(24):638, 2018. doi: 10.21105/joss.00638. URL `https://doi.org/10.21105/joss.00638`.

127 Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition, 2009. ISBN 9780387848570. doi: 10.1007/978-0-387-84858-7.

128 Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC, New York, 2015. ISBN 9781498712163. doi: 10.1201/b18401.

129 Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern Recognition*. Academic Press, 4th edition, 2008. ISBN 9781597492720.

130 Luan Gomes Matosinhos Carvalho. Characterization and simulation of photomultiplier tube signals for the cygno experiment, 2025. URL https://repositorio.ufjf.br/jspui/handle/ufjf/20258.

APPENDIX A – PUBLICATIONS

During the master's program, the author contributed to publications within the CYGNO Collaboration, including journal articles and conference proceedings.

Journal Articles

1. Amaro, F. D. *et al.*, including L. G. M. Carvalho. *Modeling the light response of an optically readout GEM based TPC for the CYGNO experiment.* European Physical Journal C, v. 86, p. 123, 2026.

2. Amaro, F. D. *et al.*, including L. G. M. Carvalho. *Fast reconstruction-based ROI triggering via anomaly detection in the CYGNO optical TPC.* Machine Learning: Science and Technology, v. 7, p. 025058, 2026.

3. Amaro, F. D. *et al.*, including L. G. M. Carvalho. *Upgrade of the trigger and data acquisition system for continuous imaging and multi-camera operation in CYGNO.* Journal of Instrumentation, v. 21, p. T04004, 2026.

4. Amaro, F. D. *et al.*, including L. G. M. Carvalho. *The CYGNO experiment: a gaseous TPC with optical readout for rare events searches.* Journal of Instrumentation, v. 21, p. C06006, 2026.

5. Amaro, F. D. *et al.*, including L. G. M. Carvalho. *Simulation of the CYGNO gaseous TPC optical readout.* Journal of Instrumentation, v. 21, p. P06024, 2026.

6. Amaro, F. D. *et al.*, including L. G. M. Carvalho. *Bayesian network 3D event reconstruction in the CYGNO optical TPC for dark matter direct detection.* European Physical Journal C, v. 85, p. 1261, 2025.

7. Antonietti, R. *et al.*, including L. G. M. Carvalho. *The CYGNO experiment.* Nuovo Cimento C, v. 48, p. 95, 2025.

Conference Proceedings

1. Oppedisano, G. M. *et al.*, including L. G. M. Carvalho. *Trigger Optimization and Event Classification for Dark Matter Searches in the CYGNO Experiment Using Machine Learning.* In: *14th Young Researchers Meeting*, Gran Sasso Science Institute, Italy, 2025. Proceedings of Science (PoS), v. 519, p. 003, 2025.

Accepted Conference Papers

1. Amaro, F. D. *et al.*, including L. G. M. Carvalho. *Stochastic Modeling of Noise in a PMT-Based Low-Light Detection System.* Accepted for presentation and publication in: *XXVI Congresso Brasileiro de Automática (CBA 2026).*

2. Amaro, F. D. *et al.*, including L. G. M. Carvalho. *U-Net for Pixel Selection in Low-Energy Events in the CYGNO Optical TPC: A Preliminary Study.* Accepted for presentation and publication in: *XXVI Congresso Brasileiro de Automática (CBA 2026).*