UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULTY OF ENGINEERING
POST-GRADUATION PROGRAM IN ELECTRICAL ENGINEERING

Igor Abritta Costa

Optimization of the clustering algorithm of the CYGNO experiment

Juiz de Fora
2020

Igor Abritta Costa

Optimization of the clustering algorithm of the CYGNO experiment

Thesis presented to the Post-Graduation Program in Electrical Engineering, concentration area: Electronic Systems, from the Faculty of Engineering of the Federal University of Juiz de Fora as a partial requirement for obtaining the Doctor degree.

Advisor: Prof. Dr. Rafael Antunes Nóbrega

Coadvisor: PhD. Davide Pinci

Juiz de Fora

2020

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

Costa, Igor Abritta.

Optimization of the clustering algorithm of the CYGNO experiment / Igor Abritta Costa. – 2020.

69 f. : il.

Advisor: Rafael Antunes Nóbrega

Coadvisor: Davide Pinci

Thesis (Doctorate) – Universidade Federal de Juiz de Fora, Faculty of Engineering. Post-Graduation Program in Electrical Engineering, 2020.

1. Preprocessing. 2. Image Analysis. 3. DBSCAN. 4. Particle physics experiment. I. Nóbrega, Rafael Antunes, adv. II. Pinci, Davide, coadv. III. Título.

Igor Abritta Costa

Optimization of the clustering algorithm of the CYGNO experiment

Thesis presented to the Post-Graduation Program in Electrical Engineering, concentration area: Electronic Systems, from the Faculty of Engineering of the Federal University of Juiz de Fora as a partial requirement for obtaining the Doctor degree.

Approved at 16 of November of 2020

EXAMINATION BOARD

Prof. Dr. Rafael Antunes Nóbrega - Advisor
Universidade Federal de Juiz de Fora

Prof. Dr. Davide Pinci - Coadvisor
Istituto Nazionale di Fisica Nucleare Sezione di Roma, INFN

Prof. Dr. Ernesto Kemp
Universidade Estadual de Campinas

Leandro Rodrigues Manso Silva

Prof. Dr. Leandro Rodrigues Manso Silva
Universidade Federal de Juiz de Fora

Prof. Dr. Giovanni Mazzitelli
Istituto Nazionale di Fisica Nucleare Laboratori Nazionali di Frascati, INFN

Prof. Dr. Augusto Santiago Cerqueira
Universidade Federal de Juiz de Fora

Aos meus pais, meus irmãos, à minha namorada, aos meus familiares, aos meus amigos.

ACKNOWLEDGMENT

Agradeço primeiramente a Deus por me iluminar e me guiar no caminho que escolhi traçar.

Aos meus pais, Marcelino e Isabel, por serem sempre meu porto-seguro e por estarem sempre ao meu lado, me guiando no caminho certo. Pelo respeito, afeto e amor incondicional dedicados a mim. Aos meus irmãos José e Lucas, às minhas cunhadas Camila e Luiza e às minhas sobrinhas Manuela e Helena. Muito obrigado por tudo, amo vocês.

À minha companheira de vida e jornada, Clara, por acreditar em mim e me dar forças para vencer cada obstáculo, te amo.

To my noble teachers, who with professionalism, dedication and commitment contributed to my professional and personal growth. Especially to my advisor, Rafael and my co-advisor Davide, who were essential in this project. My sincere thanks.

To the NIPS teammates and friends of UFJF lab. For very fruitful discussions, help and good humor. It is always good to have someone to share joy and despair.

To all CYGNO employees, the Baita teammates and my Italian friends, who certainly made my journey in Italy much more pleasant and profitable, *grazie mille per tutto.*

Finally, I would like to thank CAPES (Coordination for the Improvement of Higher Education Personnel), the Federal University of Juiz de Fora, the Faculty of Engineering, the Istituto Nazionale di Fisica Nucleare (INFN) and Sapienza Università di Roma for all the support and tools necessary to the development of this work.

RESUMO

Em geral, a tarefa de agrupar objetos em imagens pode ser simples e vários algoritmos foram desenvolvidos para esse fim. No entanto, o desempenho de tais algoritmos precisa ser entendido nos ambientes específicos da aplicação e, adicionamente, quando se trata de identificação de eventos raros, com baixa relação sinal-ruído, torna-se ainda mais necessário o estudo e, eventualmente, a otimização desses algoritmos considerando as particularidades do problema enfrentado, como no caso do experimento CYGNO que está desenvolvendo um novo sistema de detecção de partículas baseado em TPC com uma *Triple-GEM* acoplada a um sensor CMOS de baixo ruído e alta resolução espacial. Neste contexto, dois dos algoritmos de agrupamento mais citados na literatura científica conhecidos como *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) e *Nearest Neighbor Clustering* (NNC) foram avaliados no ambiente do CYGNO. Fazendo-se uso deste estudo, este trabalho de tese oferece uma proposta de adaptação do algoritmo do DBSCAN, denominada intensidade-DBSCAN (iDBSCAN), e faz um estudo comparativo entre os métodos estudados. Uma descrição do algoritmo iDBSCAN, incluindo teste e validação de seus parâmetros, e uma comparação com o próprio DBSCAN e o NNC utilizando-se de dados adquiridos com um dos protótipos do detector CYGNO serão apresentadas. Os resultados mostram que a versão adaptada do DBSCAN é capaz de fornecer eficiência de detecção similar aos algoritmos clássicos avaliados e, ao mesmo tempo, melhorar a resolução de energia e a rejeição de fundo do detector.

Palavras-chave: Processamento de Imagens. DBSCAN. Analise de Images. Experimento de Física de Partículas.

ABSTRACT

In general, the task of clustering objects in images might be simple and several algorithms have been developed for this purpose. However, the performance of such algorithms needs to be understood in the specific environments of the application and, additionally, when it comes to the identification of rare events, with low signal-to-noise ratio, it becomes even more necessary to study and, eventually, optimization of these algorithms considering the particularities of the problem faced, as in the case of the CYGNO experiment that is developing a new detection system based on a TPC Triple GEM detector coupled to a low noise and high spatial resolution CMOS sensor. In this context, two of the most commonly mentioned clustering algorithms in the scientific literature known as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Nearest Neighbor Clustering (NNC) were evaluated in the CYGNO environment. Using this study, this thesis work offers a proposal to adapt the DBSCAN algorithm, called intensity-DBSCAN (iDBSCAN), and makes a comparative study between the methods studied. A description of the iDBSCAN algorithm, including testing and validating its parameters, and a comparison with the DBSCAN itself and the NNC using data acquired with one of the CYGNO detector prototypes will be presented. The achieved results show that the adapted version of DBSCAN is capable of providing a detection efficiency as good as those obtained with the classical algorithms and, at the same time, improve the energy resolution and background rejection of the detector.

Keywords: Preprocessing. Image Analysis. DBSCAN.Particle physics experiment.

LIST OF ILLUSTRATIONS

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AP	Affinity Propagation
CFT	Clustering Feature Tree
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DM	Dark Matter
EN	Electronic noise
ER	Electron Recoils
GEM	Gas Electron Multipliers
iDBSCAN	intensity-based DBSCAN
LEMOn	Large Elliptical MOdule
LHC	Large Hadron Collider
MPGD	Micro Pattern Gas Detector
MSGC	Microstrip Gas Chamber
MWPC	Multi-Wire Proportional Chamber
NNC	Nearest Neighbor Clustering
NRAD	Natural radioactivity
PCA	Principal Component Analysis
PMT	Photomultiplier Tube
SMP	Standard Model Particle
sCMOS	scientific CMOS
TPC	Time Projection Chamber
WIMP	Weakly Interacting Massive Particles

CONTENTS

1 INTRODUCTION

One of the most important tasks in the big data world is the classification of a large amount of information generated by many sources. Particle physics experiments are contributing a lot with the production of more and more data that needs to be analyzed, understood and presented in a clear way. However, when dealing with raw data it is usual to have a long path to go through until reaching the classification task and these steps are as crucial as the classification itself. Consequently, they have received a lot of attention by researchers and many tools were developed and are still being developed to fit in and solve different scenarios.

In this context, algorithms that are able to search for similar groups of pixels in an image, called clustering algorithms, are being deeply studied in the last years. In general the signal clustering in images is simple, and there are several algorithms developed for this purpose (1). However, when it comes to the identification of rare events with low signal-to-noise ratio there is a requirement for high efficiency and high background rejection. Therefore, the search for a tool that best fits the reality of a given experiment requires further and specific studies which may eventually lead to proposals of new methods or of improvements on already known algorithms, like (2, 3, 4, 5).

Clustering algorithms are normally developed to look for underlying patterns in a data set with the intention of grouping correlated samples. Usually, this search proceeds without constraints and completely unsupervised (6, 7). However, the addition of some prior information on the cluster discovery process had been widely discussed (8), since its use concedes the possibility to insert expertise knowledge about the subject; for example by specifying some expected attributes or even not possible features (9) leading to a better performance of the algorithm in that particular environment.

This work offers a performance study of two widely used clustering algorithms, known as Nearest Neighbor Clustering (NNC) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (10), applied to an experiment called CYGNO (11), that aims to find evidences of dark matter, and proposes an adapted version of the DBSCAN algorithm with the intention of improving its performance in such environment. The impact of these algorithms on the detector was assessed in detail using three different databases: one for electronic noise, another for environment noise and another for a signal generated by a radioactive source that produces low energy photons (5.9 keV), in the detector's region of interest.

1.1 MOTIVATION

Experiments that seek to unveil dark matter, as is the case with CYGNO, usually operate at the frontier of knowledge, requiring the use of the best and latest technologies.

This is true for hardware that builds the detector and also for data analysis algorithms. In the specific case of the CYGNO experiment, which uses a scientific CMOS (sCMOS) sensor as a readout system and, consequently, outputs images, where most of the pixels are activated by electronic noise or natural radioactive events and only a small portion is expected to be due to the signal of interest, the use of pre-processing and clustering algorithms are of paramount importance. In particular, the ability of rejecting background events is one of the most important parameters of this type of experiment where it is essential to maximize the signal-to-noise ratio of its measurement apparatus.

This work proposes a first comparative performance study between two classic clustering algorithms in the CYGNO environment. Such study led to the development of an adapted version of the DBSCAN algorithm, called intensity-based DBSCAN (iDBSCAN), capable of improving the noise background rejection and the energy resolution of the experiment.

1.2 WHAT WAS DONE

In this work the most important clustering methods were studied, two of them (NNC and DBSCAN) were evaluated using acquired data from a CYGNO's detector and a modification of the DBSCAN algorithm was proposed with the aim of improving the background rejection of the experiment's clustering stage. Also, this thesis proposes a comparative study on the impact of NNC, DBSCAN and iDBSCAN on two crucial detector's parameters: background rejection and energy resolution, measured in the energy range of a few keV. For such, low energy particles (5.9 keV photons) produced by a ^{55}Fe radiation source, natural radioactive events and noise acquisition data were employed.

1.3 TEXT STRUCTURE

This document will be organized as follows: chapter 2 will present a literature review about clustering algorithms and dark matter detection and chapter 3 gives a introduction of the CYGNO experiment and the detector; chapter 4 will present the experimental setup and chapter 5 will describe the usual CYGNO's data analysis procedure, giving a main focus on the description of iDBSCAN and validation of its in-use parameters; chapter 6 will be used to compare the iDBSCAN algorithm to the NNC and DBSCAN ones by assessing their impact on the detector's performance; and, finally, chapter 7 will offer the work's final conclusions.

2 LITERATURE REVIEW

In this chapter, the recent history and review of two subjects that are central to the development of this thesis will be presented: (1) clustering methods focusing on density-based methods; and (2) direct dark matter detection with an emphasis on the Micro Pattern Gas Detector (MPGD) technology.

2.1 CLUSTERING ALGORITHMS

Data Clustering or Cluster Analysis is a multivariate data mining technique that, using numerical methods and only the information of the available variables, aims to automatically group the n cases of the database into k groups by unsupervised learning, generally called clusters or groupings. In Literature, cluster analysis can also be called Clustering, Q-analysis, Typology, Classification Analysis or Numerical Taxonomy.

Unlike the concept of classification, clustering technique is more "primitive", in which no assumptions are made regarding groups. Unlike classification, clustering does not have predefined classes and examples of labeled training classes, and thus performs a form of unsupervised learning.

The first published record on a Clustering method was made in 1948, with the work of (12) on the Hierarchical Method of Complete Liaison. Since then, more than a hundred different clustering algorithms have been defined. Any clustering method is defined by a specific algorithm that determines how the cases will be divided into different clusters. All the proposed methods are based on the idea of distance or similarity between the observations. They also define the relevance of the objects for each cluster according to the similarity between each element and the other ones that belong to the group.

The basic idea is that elements that make up the same cluster must be highly similar, but must be very dissimilar from objects in other clusters. In other words, all clustering is done with the objective of maximizing homogeneity within each cluster and maximizing heterogeneity between clusters. As examples of areas interested in the problem of clustering, we can mention: data mining, statistics, engineering, machine learning (13), medicine, marketing, administration and biology. Inside the commented areas we can mention that clustering methods can be helpful in problems of pattern recognition, data analysis, image processing, market research, purchase pattern, physical and chemical specifications of oils, analysis of disease symptoms, characteristics of living things, gene functionality, the composition of soils, aspects of the personality of individuals, customer profiles, marketing, image segmentation, document grouping, information technology, workforce management and planning, genome data studies in biology, particle physics, among many others. During the past decades, thousands of clustering algorithms have been proposed in the literature from many different fields (14, 13). These clustering algorithms can be roughly classified into different groups, including:

- hierarchical clustering algorithms, such as Single-Link, Average-Link and Complete-Link methods, etc. (14);
- partitional clustering algorithms such as k-Means, k-Medoids, EM clustering, k-Harmonic Means, etc. (15);
- density-based clustering algorithms such as DBSCAN, DENCLUE, OPTICS, etc. (14), which will be the focus of this thesis;
- grid-based clustering algorithms such as STING, WaveCluster, etc. (14);
- spectral clustering algorithms (16);
- and many other clustering algorithms such as Affinity Propagation (AP) (17).

The great advantage of using Clustering techniques is that, by grouping similar data, the peculiar characteristics of each of the identified groups can be described more efficiently and effectively. This provides a greater understanding of the original data set, in addition to enabling the development of classification schemes for new data and discovering interesting correlations between data attributes that would not be easily visualized without the use of such techniques. Alternatively, Clustering can be used as a pre-processing step for other algorithms, such as characterization and classification, that would work on the identified clusters.

However, it is important to note that the data clustering field has evolved very far beyond the capability of any text books or surveys proposed in the literature. Therefore, more and more research efforts are still constantly required in order to provide more systematic and comprehensive surveys about the field.

2.1.1 Clustering algorithm requirements

An ideal Clustering method should meet the following requirements (10, 18, 19):

1. discover clusters with arbitrary shape - the shape of the clusters, considering the Euclidean space, can be spherical, linear, elongated, elliptical, cylindrical, spiral, etc.;
2. identify clusters of varying sizes;
3. accept the various types of possible variables - the methods need to be able to handle the variables of the types: scaled at intervals, binary, nominal (categorical), ordinal, scaled in proportion, or at combinations of these types of variables;

4. the order in which objects are presented is insensitive - the same set of objects when presented with different orderings must provide the same results;

5. work with objects with any number of attributes (dimensions) - human eyes are good at judging the quality of Clustering with up to three dimensions, so methods must efficiently handle objects with high dimensions and provide intelligible results where human visualization is impossible.

6. be scalable to handle any number of objects - a large data base can contain millions of objects. The method must be fast and scalable with the number of dimensions and the number of objects to be clustered;

7. provide interpretable and usable results - the descriptions of the clusters must be easily assimilated. Users expect the results of the Clusters to be interpretable, understandable and usable, therefore it is important to have simple representations;

8. be robust in the presence of noise - most databases in the real world contain noises or missing, unknown or wrong data. Their existence should not affect the quality of the clusters obtained;

9. require minimal knowledge to determine the input parameters - the appropriate values are often unknown and difficult to determine, especially for sets of high dimensional objects and large numbers of objects. In some methods, the results of Clustering are quite sensitive to the input parameters;

10. accept restrictions - real-world applications may need to group objects according to various types of restrictions. The method must find groups of data with behavior that satisfies the specified restrictions;

11. number of clusters - finding the natural number of clusters in a set of objects is a difficult task. Many methods need a reference value.

However, no Clustering technique is able to adequately addresses all of these points, although considerable work has been done to address each point separately. Thus, there are methods suitable for large quantities of objects and others for small quantities; methods in which the number of clusters has to be provided by the user and others in which there is no such requirement; methods more suitable for clusters of spherical or convex shape and others that the shape of the cluster is not relevant; etc.

2.1.2 Density-based clustering

In density-based Clustering methods, clusters are defined as dense regions, separated by less dense regions that represent noise. Dense regions can be arbitrarily shaped and points within a region can also be arbitrarily distributed. So density-based methods are suitable for discovering arbitrarily shaped clusters, such as elliptical, cylindrical, spiral, etc. even those completely surrounded by another “cluster” and also they are experts in identifying and filtering noise (19). Density-based methods differ by the way the clusters grow: some determine the clusters according to the density of the objects’ surroundings, others work according to some density function.

Another advantage of density-based clustering compared with other traditional clustering techniques is that density-based clustering algorithms do not need the number of clusters k to be specified beforehand. It is a significant advantage when dealing with complex datasets where determining the number of clusters beforehand is a non-trivial task.

Since the first density-based clustering algorithm DBSCAN (20) was proposed, density-based clustering algorithms have attracted considerable research efforts due to their many attractive benefits, e.g., robustness again noise and the ability to detect arbitrarily-shaped clusters described above. There are in the literature many density-based clustering algorithms following different density notions, e.g., the cardinality of neighborhood of an object (20), the influence of an object in its neighborhood (21), and different research directions, e.g., subspace clustering (22), network clustering (23), data stream clustering (24). Among them, the density-based notion of DBSCAN is perhaps one of the most successful paradigms. In the literature, there are many algorithms that have been proposed based on the DBSCAN paradigm, e.g., GDBSCAN (25), SUBCLU (22).

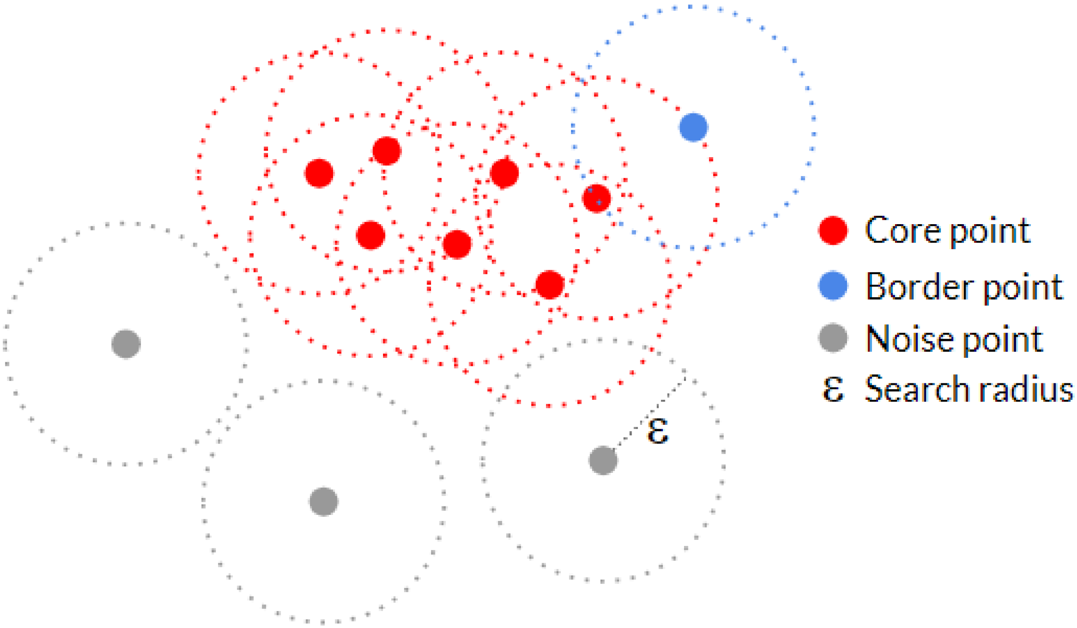
2.1.2.1 DBSCAN

In density-based clustering, clusters are regarded as areas of high object density in the data space separated by areas of lower object density. The algorithm DBSCAN, short for 'Density Based Spatial Clustering of Application with Noise’, is a non-parametric density-based clustering method proposed by (20) that formalizes a density notion for clustering using two parameters: ϵ denoting a volume and N_{min} denoting a minimal number of objects. An object belongs to a cluster if it has at least N_{min} objects inside its ϵ neighborhood.

(20) wrote that the notion of clusters in the DBSCAN algorithm is applied to Euclidean spaces of two and three dimensions, as for any characteristic space of high dimension. The DBSCAN method is applicable to any database containing data from a metric space (26) and this approach works with any distance function, so that an appropriate function can be chosen for any given application.

The key idea of the DBSCAN method is that, for each point in a cluster, the neighborhood for a given radius contains at least a certain number of points, that is, the density in the neighborhood must exceed a threshold. Figure 1 illustrate how DBSCAN works. The red points are core points, because the area surrounding these points in an ϵ radius contains at least N_{min} points (including the point itself). In this particular case they form a single cluster because they are all reachable from one to another. The blue point is a border point that can be reachable from one core point and thus belong to the cluster as well. At last, the gray points are noise points that are neither a core points nor directly-reachables.

Figura 1 – Illustrative image on how DBSCAN works



Source: Prepared by the author (2020).

According to explanations above, DBSCAN is designed to discover clusters and noise in a spatial database. This qualifies the method in directly classifying noise, which by other clustering methods, such as k-means, hierarchical, CLARANS, etc., would be mandatory placed in some cluster, not necessarily formed only by noise.

Ideally, it is necessary to know the appropriate parameters ϵ and N_{min} as it is possible to recover all the points that are reachable by density from a given point using the correct parameters; the DBSCAN algorithm is thus very sensitive to the parameters defined by the user (19).

The DBSCAN algorithm has attracted much research interest during the last decades with many extensions and applications in various fields, e.g., (27, 22, 28, 29, 30, 31, 32).

2.1.2.2 DBSCAN extensions

Among many different extensions of DBSCAN, density-based clustering algorithms for complex data have become an emerging research topic with many proposed techniques in the literature, e.g., (33, 28, 31, 23, 34, 35, 29, 36). However, the rapid growth of advanced data acquisition methods in many fields, e.g., medicine, biology and environment, has continuously produced a large amount of data with increasing volume and complexity, e.g., stream, time-series, graph or uncertain data. As a consequence, many challenges have been constantly arisen in order to provide efficient and effective data mining algorithms to extract knowledge from these data, in particular density-based clustering algorithms.

Recently, interactive exploring of data has become a significant feature in many data mining algorithms, especially for complex data, e.g., (37, 38), since it allows domain experts to be involved into the clustering process to improve the performance and outcome. However, throughout the literature review, all the existing extensions of DBSCAN only work in a batch scheme. They produce a single result at the end and do not allow user interaction during their runtime. Providing an interactive extension of DBSCAN, therefore, is another challenge and is extremely useful for many applications, e.g., the segmentation of white matter structure in human brain (39), characters recognition (40) or image clustering (38).

2.1.3 Comparing different clustering algorithms on toy dataset

Many clustering algorithms have been developed over the years, each one having its advantages and disadvantages. In (41) there is a qualitative comparison between DBSCAN and some of the most well-known clustering algorithms in the literature:

- MiniBatchKMeans (42) is a cluster method that tries to separate samples in n groups of equal variance and uses the number of clusters as an input parameter; its uses is indicated for general-purpose and its performance tends to decrease if the sample has too many clusters.

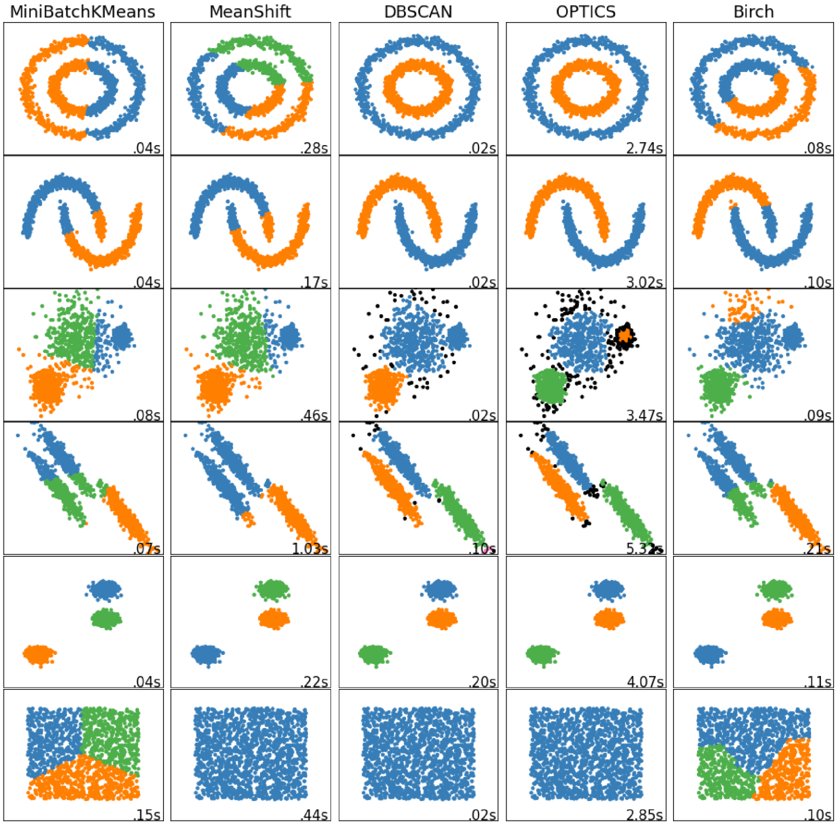
- MeanShift (43) is a centroid based algorithm that aims to discover blobs in a smooth density of samples; works well for a sample with many clusters and uses the distances between points as a metric.

- OPTICS (44) algorithm are very similar to the DBSCAN method and can be considered a generalization of DBSCAN; the main difference is the fact that OPTICS uses a reachability graph to choose different values of *eps* allowing the extraction of clusters with variable density within a single data set.

- Birch (45) is indicated to large dataset and dataset with outlier; this method uses the Euclidean distance between points and builds a tree data structure, called Clustering Feature Tree (CFT), with the cluster centroids being read off the leaf.

The example shown in Figure 2 is a comparison between different clustering algorithms in order to illustrate some of their different characteristics. The parameters of each of these dataset-algorithm pairs has been tuned to produce good clustering results.

Figura 2 – Comparison between some clustering techniques on toy datasets



Source: Extracted from (41).

As can be seen, KMeans, MeanShift and Birch do not perform well when the populations to be clustered are not well separated. Another important point to comment on is the last dataset, which is a case of null hypothesis, where the population is uniformly distributed and should not be clustered. However, using the same parameters for the clustering algorithms of the previous line, it is observed that all algorithms found one or more clusters.

In these particular cases, the biggest difference between the DBSCAN and OPTICS algorithms is in relation to the processing time. DBSCAN was on average 110 times faster than OPTICS, in addition DBSCAN has presented processing times as good as the other algorithms tested in this example.

2.2 DARK MATTER DIRECT DETECTION

Since the seventies, particle physicists have described the fundamental structure of matter using an elegant series of equations called Standard Model. The model tries to describe everything that can be observed in the universe from some basic blocks called fundamental particles (46).

However, there are some observational facts that the Standard Model in its original form cannot explain. One of this observations is that astrophysicists have noticed that galaxies appear to have much more mass than can be seen with the telescope (five times more, to be exact). This extra mass is invisible, but it is noticeable through the gravity it exerts and it is called by now Dark Matter (DM). Then, the need arises to understand the characteristics and how this new type of matter interacts with the usual matter. In this way, a new path is needed, since the known Standard Model does not include such material and also new technologies capable of taking the science further to the frontier of knowledge.

Therefore, for the experimental proof of the theories related to the Dark Matter particles, equipments that are capable of detecting particles at very low energies have been built in the last years, recreating an environment where it could be possible to observe the interaction processes of the DM.

Several astronomical observations points to the existence of dark matter, such as the high speed of rotation of galaxies, anisotropy of cosmic background radiation and the study of the effect of gravitational lenses (47, 48). Many models try to explain the nature of dark matter and there is a vast literature on candidates for it (49, 50, 51, 52, 53). One of the most accepted candidates by scientists for dark matter is known as Weakly Interacting Massive Particles (WIMP). In general, they are presented in extension theories of the Standard Model and provide the correct value for the abundance of DM density. Calculations show that WIMPs may have remained since the first moments of the universe in a sufficient number to present a significant fraction of the relic density of dark matter. Since then, there has been hope in detecting WIMPs directly by observing their elastic scattering on targets, such as atomic nuclei.

There is a convenient classification for candidates of dark matter. They can be classified into hot, warm or cold dark matter. These names reflect their typical speeds at the beginning of the universe, more precisely at the time of recombination, that is, the higher the speed, the more "hot"the matter would be. Light neutrinos are the best candidates for hot dark matter and are the only candidates that have proven existence so far. Supersymmetric particles as WIMPs (50, 54) or axions (52) appears as possible candidates for cold dark matter. Sterile neutrinos (53, 55) are candidates for warm dark matter.

This thesis is developed in the context of the direct detection of WIMPs, assuming the premise that they are one of the most promising candidates for dark matter, since this is the approach of the CYGNO Experiment, which will be discussed in more detail in section 3.1.

Physical existence beyond the standard model is based, among others, on numerous indirect observations of the existence of dark matter. Therefore, in order to observe and study this phenomenon, several experiments for direct detection of dark matter began in recent years (56). The indirect detection of dark matter is possible by astrophysical and cosmological observations, such as its self-interacting strength from colliding galaxy clusters (57, 58, 59) or the Universe's ionization history (60).

In the other hand, the direct detection of these kind of particle, in order to measure its properties (mass, coupling and interaction cross section), can be divided today in three different approaches:

- detecting the dark matter particles produced in hadron colliders, like Large Hadron Collider (LHC) (61);
- by detecting the decay of dark matter via *annihilation* processes in regions that contain a high density of dark matter (62).
- and by direct detection of WIMP through the scattering process in ultra sensitive experiments with very low levels of background, like CYGNO (63).

The possibility of direct detection of dark matter particles by observing the interaction of WIMPs was first discussed in 1985 (64). Since WIMPs carries no electric charge, a low probability of interaction of these particles with atomic electrons is expected. However, it is possible that elastic scattering occurs with the atomic nucleus. That is, a nuclear recoil can be generated by the momentum transfer, which is detectable (65). The total energy loss of recoil in a WIMP detector can be described by Equation 2.2.

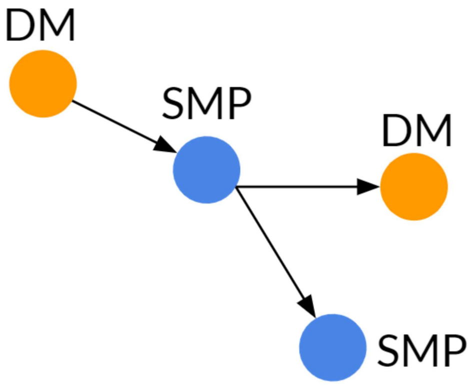
$$\left(\frac{dE}{dx}\right)_{tot} = \left(\frac{dE}{dx}\right)_{elec} + \left(\frac{dE}{dx}\right)_{nucl} \tag{2.1}$$

In the interaction between a WIMP particle and a nucleus, most of the energy is dissipated in the form of heat, which can lead to an atomic motion. And the rest are electronic energy losses, which can excite or ionize atoms. This atomic excitation can cause scintillation light, and its detection is possible using photosensors. However, since a very small portion of the signal is generated, the number of photons available for detection is low.

2.2.1 Detectors for WIMP Searches

The direct detection experiments seek to obtain information through the interaction between the known particles and the Dark Matter through collision, schematized in Figure 3, which is the spreading between the DM particle and the atomic nucleus of the target material, measuring the energy released in the process. After the ionization of the material, when the electron fills the gap in the atomic electrosphere, the emitted photon is detected by a photosensor or the ejected electron can be accelerated by an external field and at the end it interacts with other material emitting photons that are detected, such as the case of the Xenon experiment (66).

Figura 3 – Simplified diagram of direct detection. DM particles interact with one Standard Model Particle (SMP), resulting in another particle of DM and another of the standard model

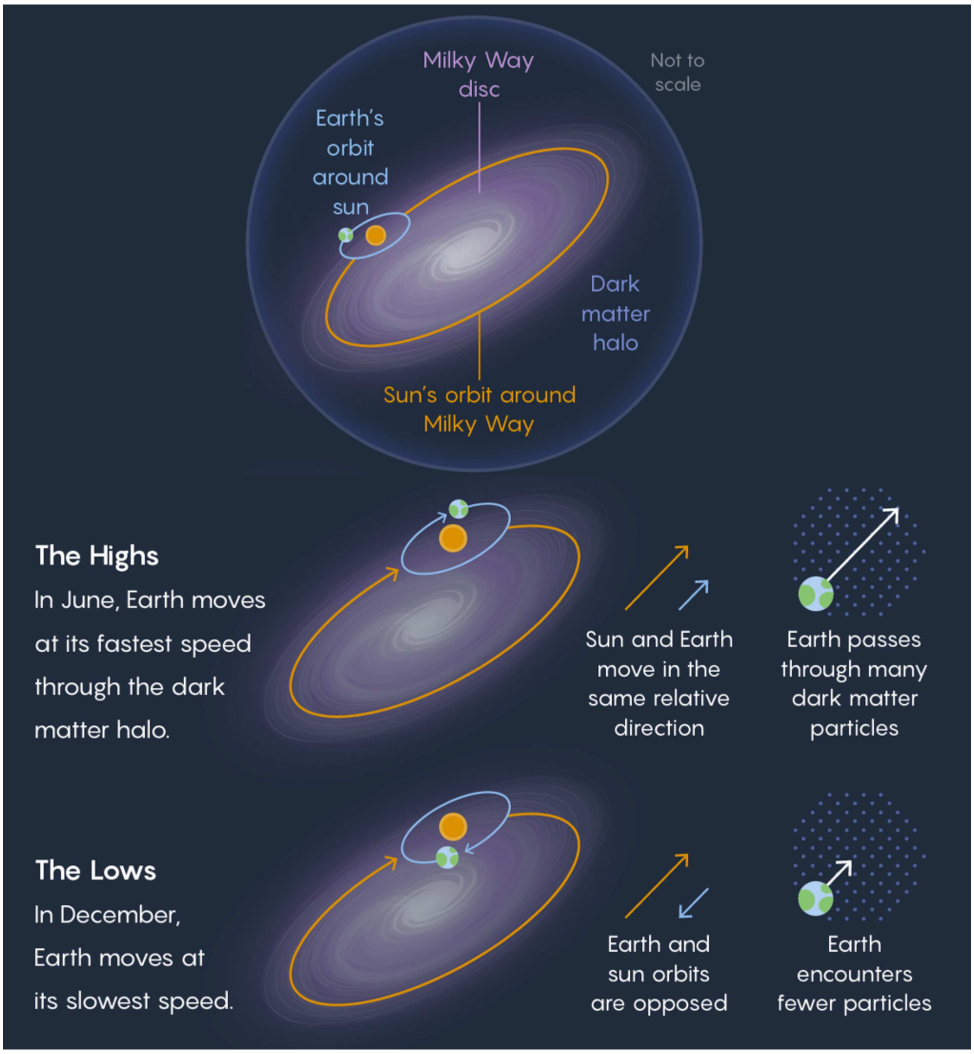


Source: Prepared by the author (2020).

Taking into account the selection criteria for the events attributed to the DM, there are some uncertainties that strongly influence the characteristics of the experimental apparatus. Some of these apparatus assume possible DM characteristics necessary for signal filtering. The main examples are the DAMA (67) and GoGeNT (68) experiments, which deal with the measurement of a signal that varies over the year, called annual modulation, due to the variation in the speed of the planet along its rotation around the Sun and the movement in the galaxy, as illustrated in Figure 4.

Knowing that, the collision between the DM particles and the target material, in that laboratory, would detect a more energetic event at one time of the year, since the detector on our planet would be moving along the solar system with a higher speed, so that the relative speed between the target material and the DM particles would be added, increasing the energy deposited in the collision, whereas, the opposite would happen in the other part of the year. This effect was detected by DAMA researchers, but as no other direct detector experiment found a similar signal, there are still some objections about

Figura 4 – Simplified diagram of the seasonal search of dark matter


Milky Way disc
Not to scale
Earth's orbit around sun
Dark matter halo
Sun's orbit around Milky Way
The Highs
In June, Earth moves at its fastest speed through the dark matter halo.
Sun and Earth move in the same relative direction
Earth passes through many dark matter particles
The Lows
In December, Earth moves at its slowest speed.
Earth and sun orbits are opposed
Earth encounters fewer particles

Source: Illustration made by Lucy Reading-Ikkanda for Quanta Magazine.

their discovery and many other experiments are being developed in order to search for signals of dark matter.

Numerous detectors around the world have been developed with different designs aiming to detect WIMP particles. We can categorize them as follows:

Inorganic Crystal Detectors This type of detector works with the idea of using a high purity crystal in order to detect dark matter-induced charge signals with a very good resolution. The first experiment attempt to directly detect WIMPs used a 0.72kg Germanium crystal (69). After that, other experiments designed their detectors using Silicon crystals, which is the case of DAMIC (70) and SENSEI (71), that was able to improve their sensitivity to WIMPs at lower masses. Following the same principle, DAMA/LIBRA has published its results showing an annual modulation signature (72) using an array of high-purity NaI(Tl) scintillator crystals. The same type of crystal was also used by COSINE-100 experiment (73).

Cryogenic Detectors This type of detector aims to detect dark matter measuring the temperature after a particle interaction, knowing that most of the energy is released in the form of phonons. A particularity of this kind of detector is the necessity to operate at cryogenic temperatures, typically ≤ 50 mK. Some experiments that uses this technology are EDELWEISS (74) and CDMS (75).

Noble Liquid Detectors Detectors using liquid noble gases (argon or xenon) as WIMP target either measures only the primary scintillation signal (single phase detectors) or detects the primary scintillation light as well as the ionization signal in a dual-phase Time Projection Chamber (TPC) using Photomultiplier Tube (PMT). Examples of experiments using this approach are DarkSide-50 (76), DEAP (77), LUX (78) and XENON10 (79).

Bubble Chambers This approach works using superheated liquids (e.g. CF_3I, C_3F_8, C_4F_{10}, etc) as WIMP targets. When the particle interacts inside the liquid, kept at a temperature just below its boiling point, energy deposition leads to a phase transition starting the bubble formation. This effect is typically read out by means of cameras, which allows 3d-reconstruction of the event tracks, and make use of acoustic sensor, which can help in the particle detection and discrimination. After each event, it is necessary to remove the bubble by a compressing and decompressed process, which implies a long detector deadtime. Bubble chamber experiments are PICASSO (80, 81), PICO-2L (82) that reached a threshold down to 3.3 keV, PICO-60 (83) and COUPP (84)

Directional Detectors This last method aims to construct a detector capable of identifying the directionality of the WIMP particle, which in principle would allow the discrimination of a WIMP signal from the background with just 30 WIMP collected events (85). The operation principle of this approach is a TPC gas detector readout by a high-granularity sensor. The WIMP particle interacts with the gas, producing electrons that are drifted by the electric field until reaching the anode where they are collected. Typically the threshold of this detector is around 20 keV$_{ee}$, but the MIMAC prototype (86) had already achieved 2keV$_{ee}$. DRIFT-II is one of the most sensitive direction detectors, with 1m^3-scale (87).

A summarized table with a list of the leading direct detection experiments on WIMP interactions extract from (56) and updated with CYGNO experiment is showed in Table 1, together with the specification of type, target, mass and the most relevant publication for each experiment.

Over the past years, direct DM experimental programs has been focused on WIMPs, mostly above 10 GeV mass. Figure 5 shows the limits for Spin-Independent (SI) WIMP

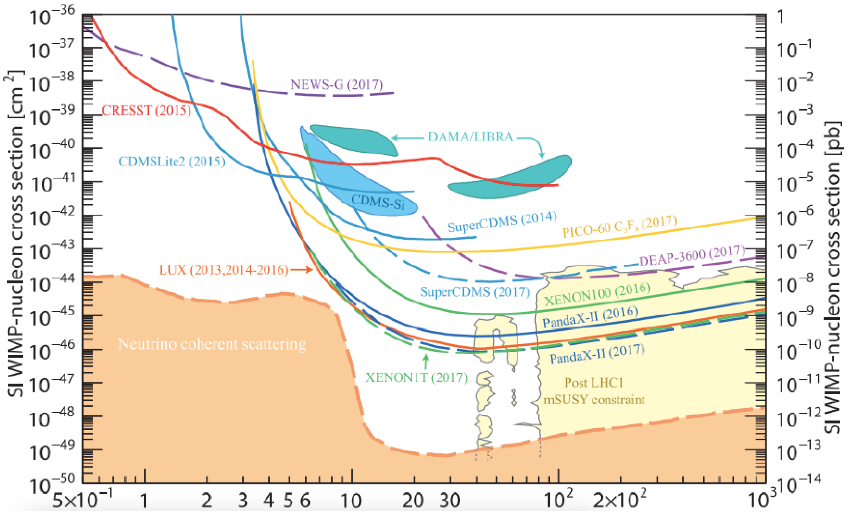
Tabela 1 – Alphabetical list of some of the leading direct detection experiments that published results on WIMP interactions

Experiment	Type	Target	Mass [kg]	Laboratory	Ref.
ANAIS-112	Crystal	NaI	112	Canfranc	(88)
CDEX-10	Crystal	Ge	10	CJPL	(89)
CDMSLite	Cryogenic	Ge	1.4	Soudan	(90)
COSINE-100	Crystal	NaI	106	YangYang	(91)
CRESST-II	Cryogenic	$CaWO_4$	5	LNGS	(92)
CRESST-III	Cryogenic	$CaWO_4$	0.024	LNGS	(93)
CYGNO Phase-I	Directional	$HeCF_4$	1.6	LNGS	(63)
DAMA/LIBRA-II	Crystal	NaI	250	LNGS	(72)
DarkSide-50	TPC	Ar	46	LNGS	(94)
DEAP-3600	SinglePhase	Ar	3300	SNOLAB	(95)
DRIFT-II	Directional	CF_4	0.14	Boulby	(87)
EDELWEISS	Cryogenic	Ge	20	LSM	(96)
LUX	TPC	Xe	250	SURF	(97)
NEWS-G	Gas Counter	Ne	0.283	SNOLAB	(98)
PandaX-II	TPC	Xe	580	CJPL	(99)
PICASSO	Superheated Droplet	C_4F_{10}	3.0	SNOLAB	(100)
PICO-60	Bubble Chamber	C_3F_8	52	SNOLAB	(101)
SENSEI	CCD	Si	9.5×10^{-5}	FNAL	(102)
SuperCDMS	Cryogenic	Si	9.3×10^{-4}	SNOLAB	(103)
XENON100	TPC	Xe	62	LNGS	(104)
XENON1T	TPC	Xe	1995	LNGS	(105)
XMASS	Single phase	Xe	832	Kamioka	(106)

Source: Extract from (56).

nucleon coupling depending on WIMP mass for many experiments. These experiments searches for nuclear recoils due to the elastic scattering of WIMPs inside the active volume of the detector that have low energy (10-100 keV). However, due to the rarity of the expected interactions, it is necessary to control, minimise or even reject any source of background that are indistinguishable from the DM signal in the data analysis part.

Figura 5 – WIMP cross section limits (normalized to a single nucleon) for Spin-Independent coupling versus mass

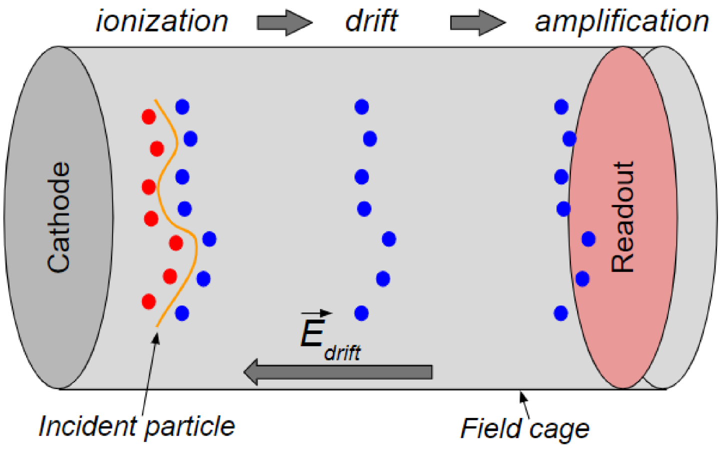


Source: Extracted from Particle Data Group 2017 (107).

2.2.2 Time projection chamber (TPC)

Introduced in 1976 by D.R. Nygren (108, 109), the TPC idea was possible thanks to the development of the Multi-Wire Proportional Chamber (MWPC). As shown in Figure 6, the TPC consists of a field cage filled with gas, which is the sensitive volume of the detector. The endcaps of the field cage are usually called anode (positive terminal) and cathode (negative terminal), used to create an electric potential difference strong enough to, in case of ionization of the gas atoms, drift the ions to the cathode and the electrons to the anode, where the readout system is placed. Other cathode/anode configurations are possible as the one which uses two anodes as endcaps with a central cathode that divides the volume into two identical halves. The working principle of the TPC is the following: if the particle that passes through the gas has enough energy to ionize it, a track of electrons and ions is produced. Due to the electric field, electrons migrate towards the anode and ions towards the cathode. The charge measured at the anode terminal is supposed to be proportional to the energy of the particle (110). For optimum operation. the amplification and readout stage of a TPC is, nowadays, based on Micro-Pattern Gas Detectors, replacing the classic MWPC.

Figura 6 – The TPC working principle



Source: Prepared by the author (2020).

The TPC is one of the main detectors used in problems where it is important to reconstruct the traces of the particles, for the direction search, as it allows a complete 3D picture of the ionization deposited in a gas volume. The x-y coordinates are reconstructed directly by the readout plane by making use of its segmentation, usually with a resolution of few μm, while the z coordinate might be estimated if the drift velocity (v_d) of the electrons in the gas volume is known.

$$z = v_d(t_1 - t_0)$$

where t_1 is the arrival time at the anode and t_0 is the interaction time. If t_0 is not known, the exact position of the track can not be known but it is still possible to reconstruct its z-profile if time resolution is good enough to discriminate the time of arrival of the electrons produced by a single track.

Additionally, the charge collected by the anode (or cathode) tends to be proportional to the ionization energy loss (dE/dx) produced by the incident particle and its measurement offers an important parameter to be used for particle identification.

Finally, a magnetic (B) field might be applied parallel to the electrical (E) field to bend the trajectory of charged particles. The resulting curvature allows for measurement of particle momenta. In addition, a magnetic field reduces the diffusion of the electrons on its way to the anode, which ensures a better x-y resolution. In order to ensure reliable operation conditions, like a constant drift velocity and constant gain, a good homogeneity of E and B fields is required and should, therefore, be monitored.

2.2.3 Micro-pattern gas detectors (MPGD)

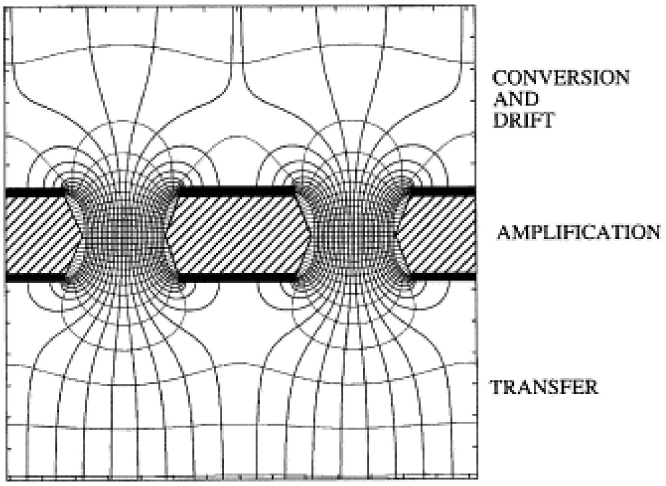
CYGNO makes use of the Micro-Pattern Gas Detectors (MPGD) technology to amplify the signal generated by a TPC. In particle physics, a MPGD is a high-granularity gaseous detectors with small (below 1 mm) distances between its anode and cathode electrodes that can reveal the presence of particles. Before the invention of the MPGD technology, TPC was readout by MWPC, developed in 1968 by the physicist Georges Charpak (111). Such technology represented a revolution in the field of particle detection, inserting it in the electronics era, replacing technologies such as cloud and bubble chambers. Its invention has earned Georges Charpak a Nobel Prize in Physics in 1992.

Along the past years, several types of electric field patterns have been developed, such as multiwire, single wire, strips, holes, parallel plate and grooves (112), in order to produce an enhanced field region, which is where multiplication takes place. The use of strip patterns was very popular and the detectors that use this structure were called Microstrip Gas Chamber (MSGC) (113). Due to the narrower spacing between the anode strips this technology was able to increase the capacity rate of the detectors by two orders of magnitude, which made it a very attractive technology for many applications. However, the stability of the detector remained a problem, since there were still discharges capable of modifying the field shape locally. Such discharges are mainly induced by strongly ionizing particles or high particle rates, which can damage anode strips. This problem was overcome with the introduction of the Gas Electron Multiplier (GEM) technology (114), which was used as a pre-amplification stage for such detectors. When the GEM was coupled to an MSGC or MWPC, the additional gain provided by the GEM allowed the operation of the combined detector at reduced voltages, reducing the probability of discharge and increasing reliability.

2.2.3.1 Gas Electron Multiplier (GEM):

In 1997 an amplification device was introduced and started to be used in gas detectors, this equipment was named GEM (114). It consists of a very thin insulating sheet, typically Kapton, covered on both sides with a thin metallic layer perforated by an array of small holes, in the case of CYGNO, spaced by 140 μm and with 70 μm diameter each. When a potential difference is applied between the two metallic layers of the GEM, an intense electric field is created inside each hole, illustrated in Figure 7. This electric field, once immersed in a suitable gas solution, is strong enough to promote an avalanche of electrons. The gain achieved with a single-GEM is proportional to the potential difference applied to it. Besides that, the GEM is able to avoid the distortion of the electric field caused by positive ions once they are collected on the GEM copper cladding. The output of the GEM is usually acquired in two ways: reading the signals induced on electrodes or, in the case of CYGNO, using an optical sensor to record the light emitted by the de-excitation of the gas molecules.

Figura 7 – Schematics and fields of the gas electron multiplier



Source: Extracted from (115).

In order to obtain greater gains even and, at the same time, work with lower voltage values, it is possible to cascade modules of GEMs. Two or more GEM layers can be used in sequence, forming what is called double-GEM, triple-GEM, quad-GEM, etc. In multi-GEM detectors, the gain is distributed among its layers, and each of them can provide gains in the order of 100. Therefore, a triple-GEM detector can achieve gains in the order of 10^6 (116). In fact, the use of three cascading GEMs (triple-GEM) has become a standard in many applications (117, 118, 119), as used by the CYGNO experiment.

3 CYGNO

This section is dedicated to set and provide an overview about what the object of study is and where this study is carried out, that is, a brief explanation will be made about the CYGNO experiment.

3.1 CYGNO EXPERIMENT

In experiments related to the direct detection of dark matter, it is essential to know the background radiation, both internal to the used detectors, and of the environment, since signals produced by them could be confused with signals generated by WIMPs in the energy scale of interest (below 30 keV) (120). Thus, it is important to develop techniques whose objective is to separate the signals generated in the detectors due to nuclear recoils (possibly generated by WIMPs) from the signals due to electronic recoils (in general, generated by the background radiation). Therefore, in order to suppress cosmogencic backgrounds, these kind of experiments are typically located underground and also are manufactured using excellent radio-purity materials, some of them with the capability to discriminate a nuclear recoil from other interactions.

3.2 DM SEARCHES WITH CYGNO

The Milky Way presents a rotation movement around the galactic center in a clockwise direction (from the galactic north pole). This movement presents, like other spiral galaxies, irregularities in relation to what is predicted based on the total visible mass (formed by stars, gases and other components) and what is actually measured. It is noted that the regions furthest from the galaxy rotate at higher speeds than would be predicted by Kepler's Laws. Therefore, it is concluded that the rotation speed does not necessarily decrease with distance, but remains practically constant from the disk (121).

The rotation curve describes the rotation speed of the stars in the galaxy as a function of their distance from the center. This speed is directly related to the amount of matter that is found inside this orbit, being possible, therefore, to infer the mass of the galaxy through the movement of its components. As the rotation curve of the Milky Way reveals, the speed in its outer parts is greater than expected, which implies that a large amount of matter exists beyond the disk, far beyond from what can be observed. Consequently, it is believed that the anomaly is caused by dark matter, directly undetectable and whose nature is unknown (122).

The Sun describes an orbit around the galactic center at a speed of about 220 kilometers per second and its velocity vector points to the Cygnus constellation. The signal due to the WIMP scattering expected in the CYGNO detector is due to Earth's relative motion with respect to the galactic halo and the DM wind is apparently coming from the Cygnus constellation, which is expected to be observable on our planet. Determining the dark matter particle direction coming from space (123) can provide a correlation with an astrophysical source that does not resemble any background noise and therefore provide the necessary information for an unambiguous identification of the dark matter signal. In addition, measuring the directionality of these particles can help discriminate between different models of dark matter (124, 125) and provide more information about the properties of WIMPs, which would not be possible with non-directional detectors.

With this in mind, the CYGNO collaboration proposes a differentiated approach to tackle this problem. The detector will use a high-resolution TPC with a low density light nuclei target, such as Helium and Fluorine gases, to increase the sensitivity to WIMP masses, while at the same time maintaining directionality information and an good ability to reject background noise, even at low energies. With the use of Helium, it is possible to work at atmospheric pressure, which in addition to reducing costs with the production of equipment that withstand different pressures, also guarantees a reasonable volume to target mass ratio. The characteristics described above enables CYGNO to explore new particle physics cases that need a high capacity to discriminate nuclear recoils from other particles, in addition to the need to know their direction of arrival, which is the case of elastic scattering of sub-GeV DM (123) and of solar neutrinos (126, 127).

3.3 CYGNO DETECTOR

CYGNO is a project that aims to develop a MPGD detector based on TPC with a triple Gas Electron Multipliers (GEM) and a sCMOS optical readout that delivers a high precision 3D tracking capabilities, sensitive to the direction of the recoiling nuclei and electrons for Dark Matter searches at low (1-10GeV) WIMP masses down to the Neutrino Floor. Figure 8 shows a 3D drawing of the CYGNO detector.

In order to reach the final objective, this project should go trough several development phases that can be divided as follows:

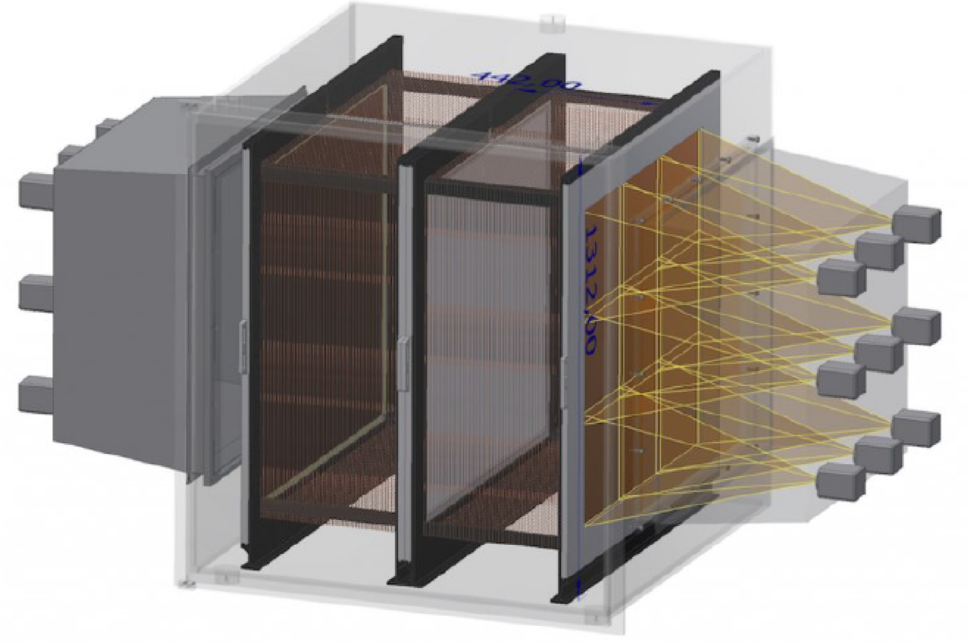
PHASE-0 is the current one and it is focus on the detector development, using prototypes as a test platform to understand the detector's characteristics; during this phase, many tools are under development in order to simulate and analyse the acquired data;

PHASE-1 aims to built the 1 m^3 demonstrator;

PHASE-2 is expected to develop a 30-100 m^3 detector.

In recent years, the CYGNO collaboration has been testing different prototypes (NITEC (128), ORANGE (129, 130), LEMOn (11, 131, 132, 133)), varying the radioactive

Figura 8 – 3D Drawing of the CYGNO detector



Source: Extracted from CYGNO Collaboration (2019).

sources (electron beam test facility, neutron beams) and some of the general operating conditions in order to understand the nuances of the project in order to develop the final detector.

The main idea of all prototypes is basically the same: an acrylic box filled with gas and with at least one transparent side to allow the camera to look inside the sensitive area. A drift field is implemented to lead the electrons to the camera side of the box where a Triple-GEM (134) is placed in order to amplify the signal produced in the ionization process. The camera is placed outside the box to take pictures of the light signal produced by the avalanche process of the GEM stage.

By now, there are four prototypes: ORANGE, MANGO, LEMOn and LIME. ORANGE and MANGO have a small drift volume, LEMOn is a 7 liter active drift volume and is better described in 4.1 as it is the detector used in this work, and the most recent one, LIME which represents exactly 1/9 of the CYGNO demonstrator of 1 m^3.

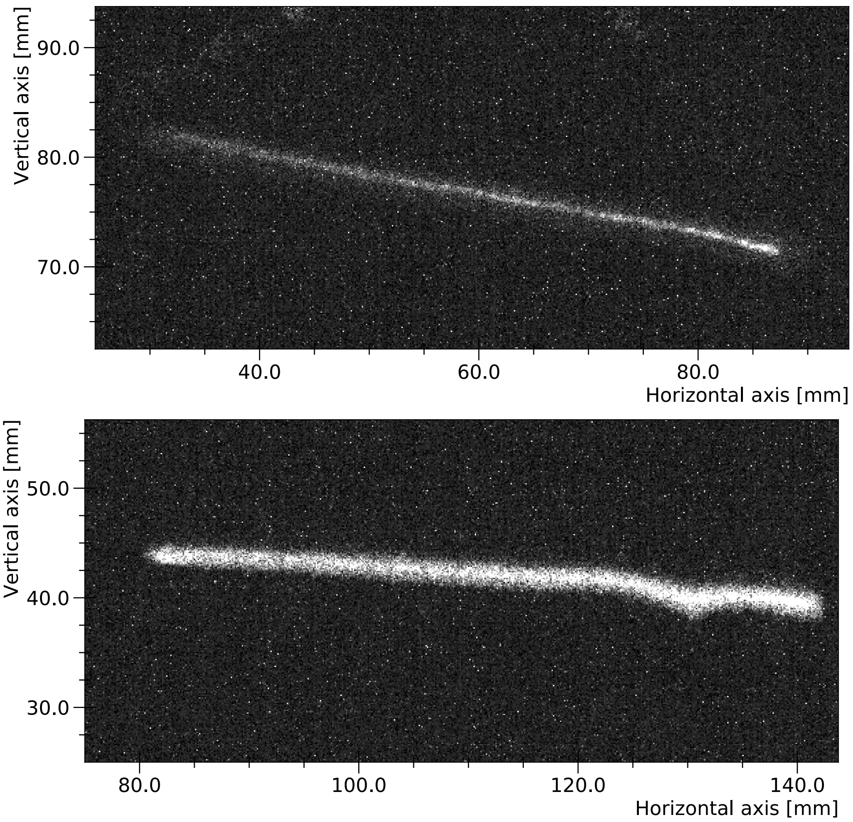
3.3.1 Expected recoil's signature in the detector

The detector was designed so that the main expected signal signatures come from nuclear and electron recoils. In the case of nuclear recoil, it is necessary that a neutral particle elastically scatters in a gas nuclei, which will cause its movement and consequent release of energy within the detector. The arrival direction of the particle that causes the scattering can be measured using the direction information of the particles that have left signals in the detector. This measurement provides relevant information to identify the following particles:

- WIMP-like Dark Matter;
- Environmental fast neutrons;
- Sub-GeV DM produced in Supernova;
- Solar neutrinos via Coeherent Scattering.

Figure 9 shows examples of nuclear recoils produced by a AmBe source placed near to the LEMOn detector.

Figura 9 – Examples of nuclear recoils inside LEMOn detector



Source: Prepared by the author (2020).

However, due to the fact that electrons participate in both weak and electromagnetic iterations, electron recoils can be induced by several types of particles and can be mistakenly identified as nuclear recoil. The particle released energy and direction information can be exploit in order to identify the signals and reject possible background noise. Some examples of relevant electron recoils signatures are:

- non-WIMP sub-GeV Dark Matter;

- Solar neutrinos via Elastic Scattering.

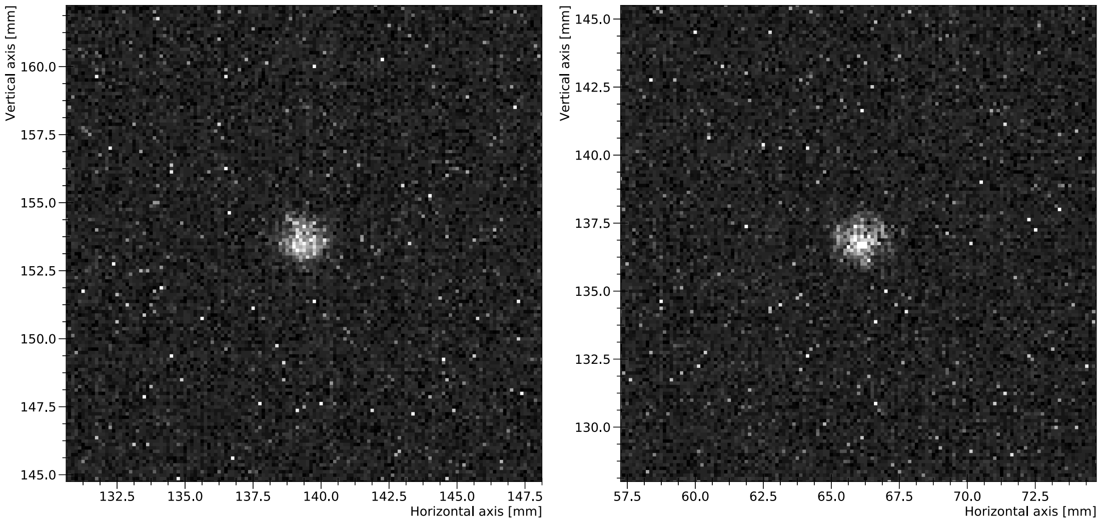
3.3.2 Energy Calibration of the detector

The CYGNO detector readout concept is a sCMOS camera that will take pictures of the interaction inside the detector. The camera will register the light (photons) produced by the electron scattering. To know the conversion factor between the amount of collected photons and the energy released by an interacting particle, a calibration is needed and one way to achieve it is by inducing well-known signals in the detector using radioactive sources.

As explained by (135), a calibration source need to fulfill few requirements: uniform coverage of the sensitive region, emitted particle in the region of interest of the detector, generated tracks need to be fully contained inside the detector sensitive area, and last, the radioactive isotope source needs to have a proper half-life. Examples of commonly used radioactive sources are: ^{60}Co, ^{55}Fe, ^{83m}Kr and ^{192}Ir.

In this thesis one of the data sets analyzed was taken using a ^{55}Fe source. This radioactive source is commonly used in low energy calibrations (136, 137) due to the fact that it emits particles in the order of few keV with low background events. The ^{55}Fe decays via electron capture to ^{55}Mn and this process has a half-life of 2.737 years (138). The ^{55}Fe decay process consists mainly of Auger electrons (5.19 keV with a probability of 60.7%) and X-Rays (5.89 keV with a probability of 27.8%). The first one is generated when the photons produced by the transition of the internal electrons ionize the external shells electrons while the second when the photons escape from the atom as X-ray radiation. It is important to notice that because of the penetration power of X-ray, those 5.89 keV photons are the ones expected to interact inside the detector. Figure 10 shows examples of electrons recoils produced by a ^{55}Fe source placed near to the LEMOn detector.

Figura 10 – Examples of electrons recoils due to ^{55}Fe source inside LEMOn detector



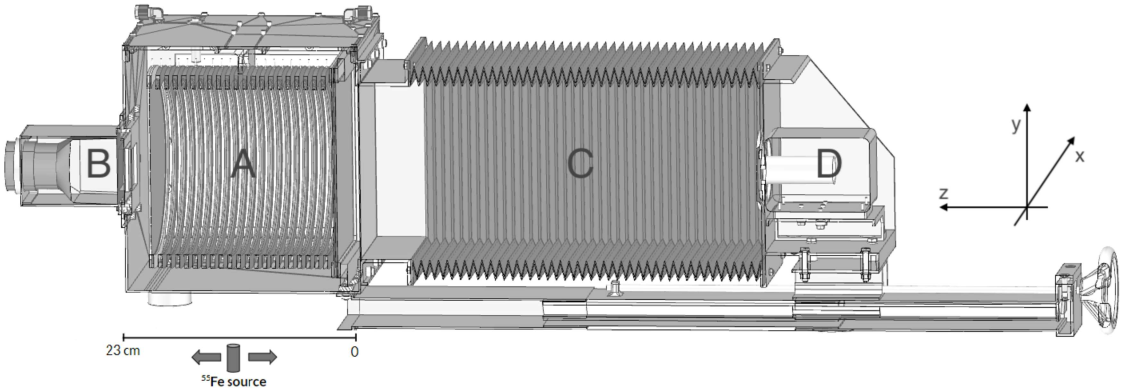
Source: Prepared by the author (2020).

4 EXPERIMENTAL SETUP

4.1 LEMON DETECTOR

One of the most recent CYGNO Experiment's prototype is Large Elliptical MOdule (LEMOn) and it was used to take all the data used in this thesis. The LEMOn detector, as shown in Figure 11, is composed of an elliptical field cage ($20 \times 20 \times 24$ cm^3) inside a 7 liter active drift volume and closed by a 20×24 cm^2 Triple GEM structure that amplifies the signal coming from the sensitive volume. Then, the photons produced in the GEM are readout by an Orca Flash 4 CMOS-based camera[1] placed at a distance of $52.5cm$ (i.e. 21 Focal Length, FL). In order to operate the detector a He/CF_4 gas mixture in the proportion of 60/40 was used to fill the LEMOn drift chamber and electric fields are applied to the TPC drift volume and between the GEMs. They are called drift field (E_d) and transfer field (E_t) respectively. More details are described in (139, 11, 131, 140). The regular running settings of the detector, as used in this work, are: $E_d = 500$ V/cm, $E_t =$ 2.5 kV/cm, and a voltage difference across the GEM sides (V_{GEM}) of 460V.

Figura 11 – Drawing of the experimental setup. In particular, the elliptical field cage close on one side by the triple-GEM structure and on the other side by the semitransparent cathode (A), the PMT (B), the adaptable bellow (C) and the CMOS camera with its lens (D) are visible



Source: Extracted from (139).

4.2 DATASETS

All the data used in this work were collected using auto-trigger mode. Three different datasets were acquired and used in order to conduce the intended study, as listed below:

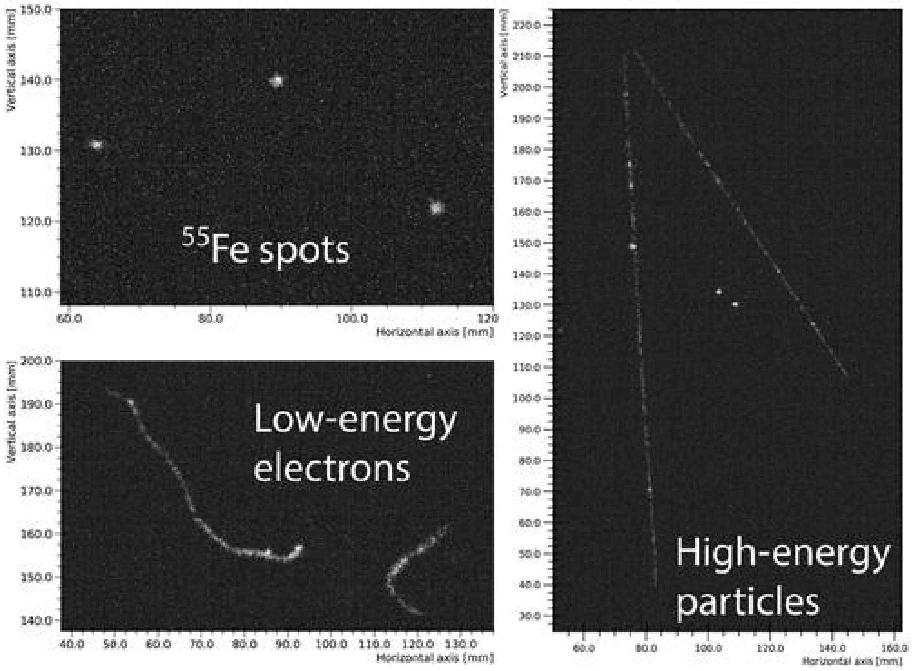
[1] For more details visit the site www.hamamatsu.com

- Electronic noise (EN) dataset: formed by lowering down V_{GEM} to a value where there is no multiplication process, in order to recorded only electronic noise (6478 images registered);
- Natural radioactivity (NRAD) dataset (composed of cosmic rays and environmental radioactivity): produced by turning on the detector to the regular running settings (see Sec. 4.1) allowing charge multiplication and secondary light emission during this process (864 images registered);
- Electron Recoils (ER) dataset: equal to the anterior item but placing a ^{55}Fe source next to the detector drift volume, as shown in Figure 11 (864 images registered).

4.3 EXPECTED DETECTOR'S SIGNALS

The expected detector's signals for the acquisition datasets defined in section 4.2 are shown in Fig. 12. In the left top image it is possible to observe three interactions of ^{55}Fe photons that are coming from the ^{55}Fe source placed near the detector, which releases 5.9 keV round spots on the image. And the other particles interaction expected in the detector are due to the natural radioactivity and cosmic rays muons, as it can be see in the examples: two low-energy electrons in the left bottom image and two high-energy particles in the right image.

Figura 12 – Examples of signals that can occur using the described configuration

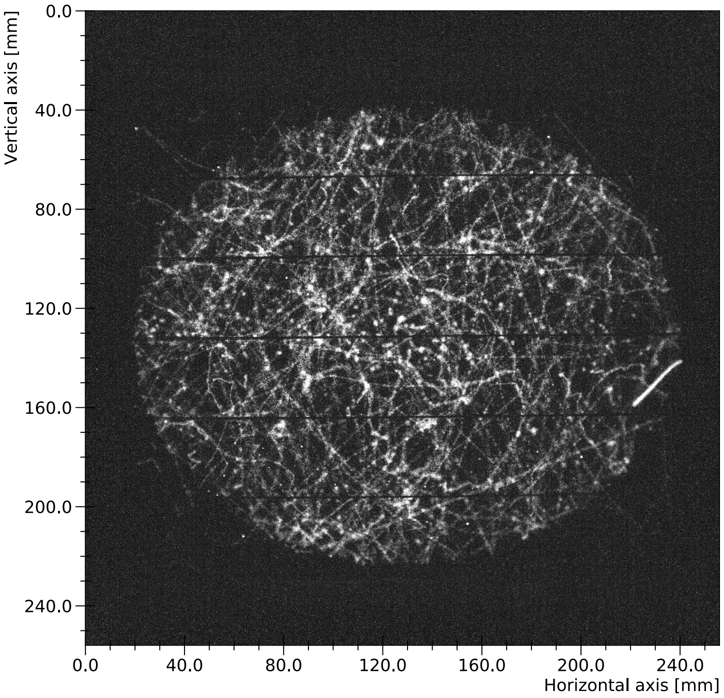


Source: Prepared by the author (2020).

The signals of interest for this thesis are the spots generated by the ^{55}Fe source, which can be used to calibrate the conversion factor between the measured photons and the energy in keV, as it is known that the ^{55}Fe produces monochromatic tracks at about 5.9 keV (see section **3.3.2**). And also they are in the energy rang of few keV, which is the expected energy region of the DM particles. Therefore they are used to evaluate the impact of the considered clustering algorithms on the detector characteristics, concentrating mostly on its energy resolution and background-events rejection performance.

Figure 13 was taken using an exposure time of 10s in the sCMOS sensor in order to show the region within the field cage, region where the detector are sensible to tracks. Also, it is possible to observe the high NRAD signals occupancy when the detector is placed at the surface.

Figura 13 – Example of an image taken with 10 seconds of exposure time



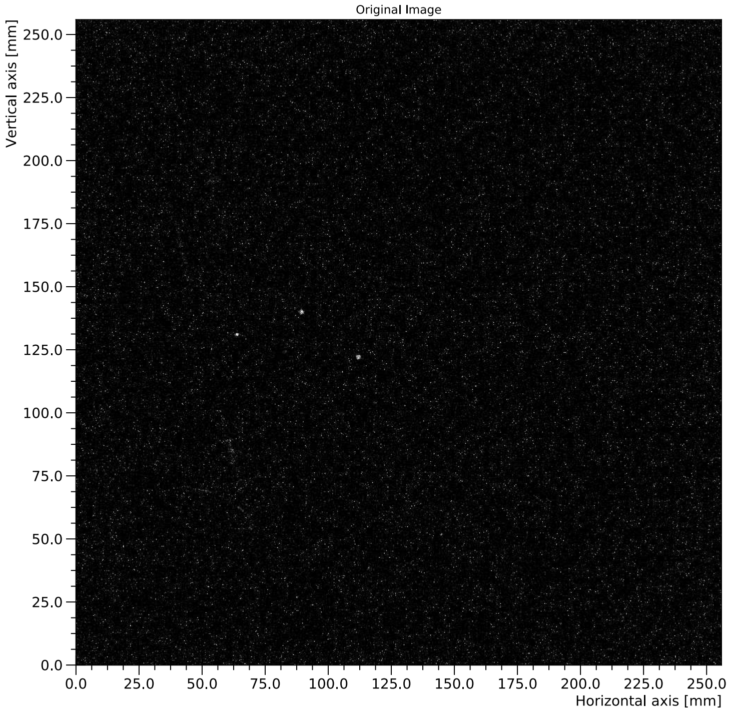
Source: Prepared by the author (2020).

5 DATA ANALYSIS ALGORITHM

5.1 DATA STRUCTURE

The output of readout system are images with 2048 × 2048 pixels captured by the Orca Flash 4 CMOS sensor, as illustrated by Figure 14, that shows the full resolution image where it is possible to see some very fainted tracks and 3 high density spots.

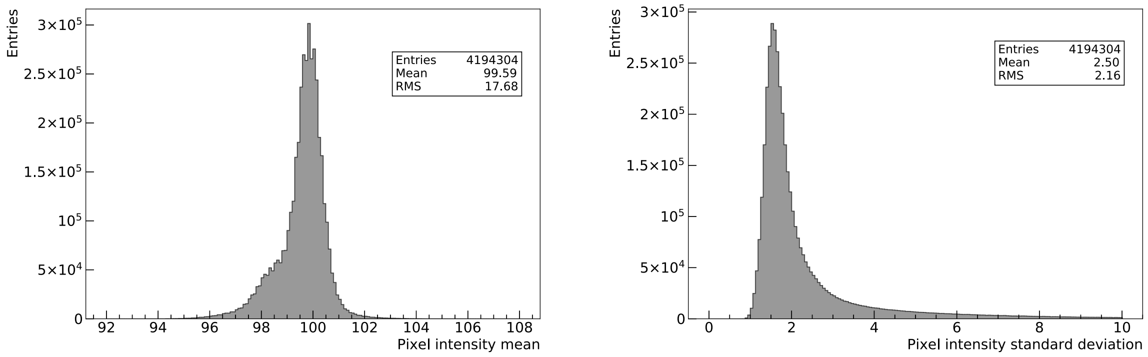
Figura 14 – The original image in full resolution



Source: Prepared by the author (2020).

The photo sensor has an sensitive area of 13312 μm^2 and each pixel has a size of $6.5\mu m \times 6.5\mu m$. The camera comprises an area of 26 × 26 cm^2 in relation to the plane of the last layer of the GEM detector and its exposure time was set to 40ms. Each one of the camera's pixel gives a response, here called intensity, which is proportional to the number of collected photons (129) combined with a baseline. The latter is also known as pedestal and can be interpreted as the intensity value corresponding to zero photons. Another relevant parameter to take into account is the pixel noise level that can vary from one pixel to another. Figure 15 shows the mean and standard deviation distributions of the noise for each pixel, as it was computed with the EN dataset, which illustrate that the pedestal average value of the sensor is about 99 counts, however it can differ from 90 to 110 for each pixel and the noise level average value is about 2.5, but as the mean, it also can fluctuate between 0 to about 10. Consequently, to run the event reconstruction procedure and measure the number of photons that were collected by the camera coming from the particles interactions it is required to calculate beforehand the pixel baseline (μ_i) and its average noise (σ_i).

Figura 15 – Mean and standard deviation distributions of the sensor's pixels noise

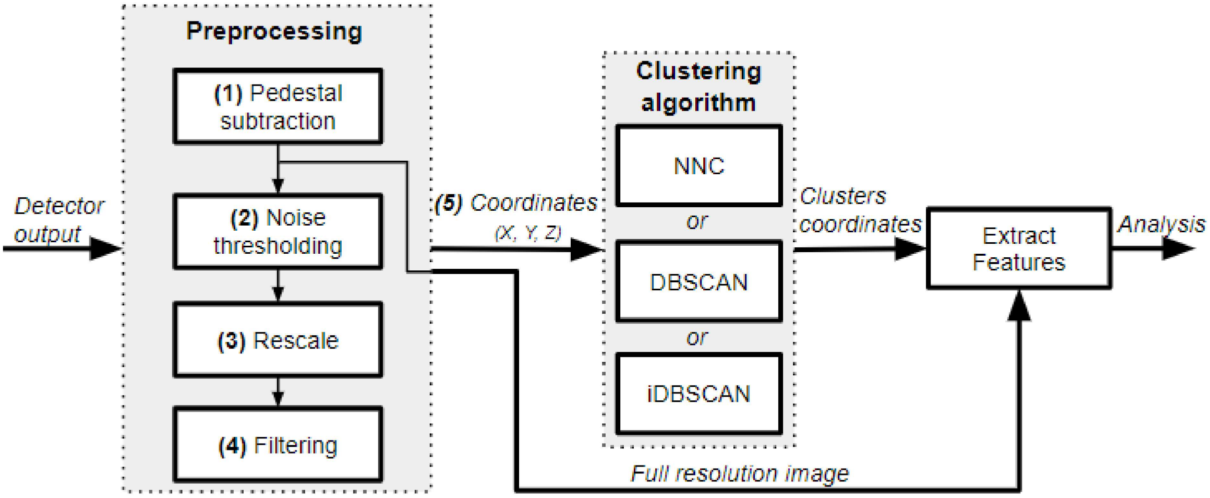


Source: Prepared by the author (2020).

5.2 OVERVIEW OF THE EVENT RECONSTRUCTION PROCEDURE

The current CYGNO's event-reconstruction algorithm is pictured in the flowchart shown in Figure 16 and its description is enumerated right below.

Figura 16 – Flowchart of the CYGNO's event-reconstruction algorithm

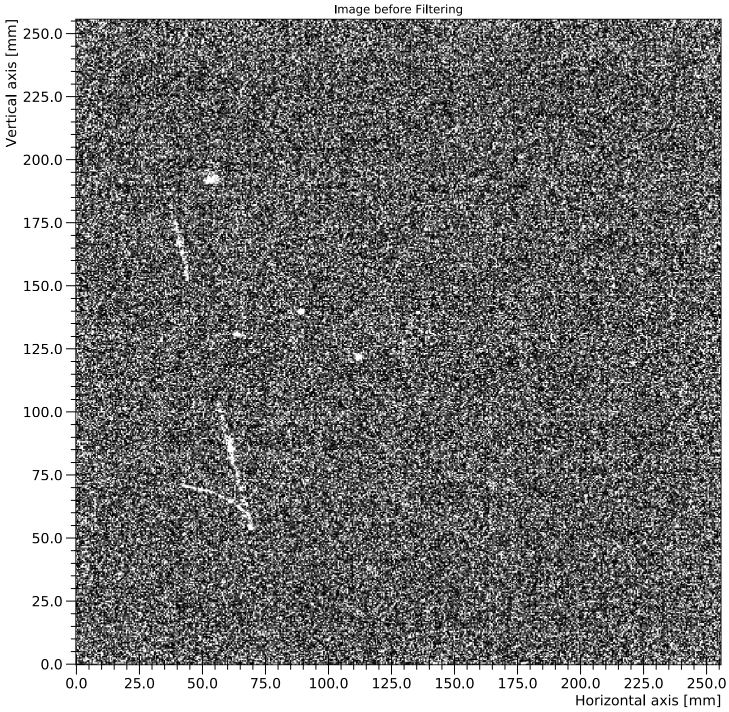


Source: Prepared by the author (2020).

1. First, each pixel original intensity value is subtracted from its previous calculated pedestal (μ_i), producing new intensity values defined as I_i, this process is called here Pedestal subtraction.

2. Then, in the Noise thresholding phase the upper limit is set to 100 counts, while the lower limit is set to 1.3 times σ_i, both of them are applied to I_i. The pixel intensity for the ones that are outside those limits are then reset to zero. The upper limit allows to discard pixels with a intensity much higher than the expected from particles, those intensities could be produced by leakage currents that go into sensor wells - also known as hot-pixels. On the other hand the lower limit was optimized to remove the fluctuation due to electronic noise and set to be just above it in order to provide a good detection efficiency, but at the same time reducing the amount of noise pixels that go to the event-reconstruction algorithm. Figure 17 illustrates the image after passing through the described process and now the tracks are easier to find by eye but it is also possible to note the noise environment where the signals of interest are immerse.

Figura 17 – Rebinned image after passing through the pedestal subtraction and noise thresholding

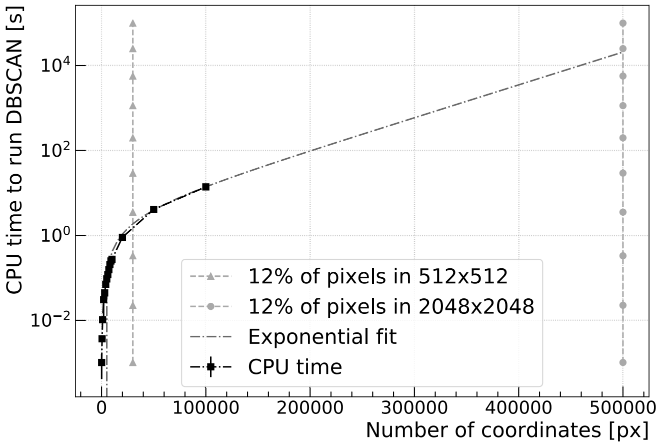


Source: Prepared by the author (2020).

3. Images are then rescaled to 512 × 512 pixels, for CPU reasons, so that each 4 × 4 matrix, called macro-pixel, is assigned an intensity value corresponding to the average of the intensities I_i of the 16 pixels occupying the same area of the sensor. The rescale process is necessary because the CPU time needed to run clustering algorithms increases when the number of pixels that are sent to them grows. As the example in Figure 18 shows, in the hypothesis that the preprocessing step removes

approximately 88% of the pixels from the image (rebined or not), the full resolution image would take about 10^4 times longer to be analyzed.

Figura 18 – DBSCAN CPU time consumption in relation of the number of pixels to be analysed



Source: Prepared by the author (2020).

4. A 4 × 4 median filter is applied to the rescaled image, replacing a given macro-pixel intensity by the median of all macro-pixels in its neighborhood w, $g(x, y)$, as given by Equation 5.1 (141), where $f(x, y)$ is the intensity of the macro-pixel (x, y).

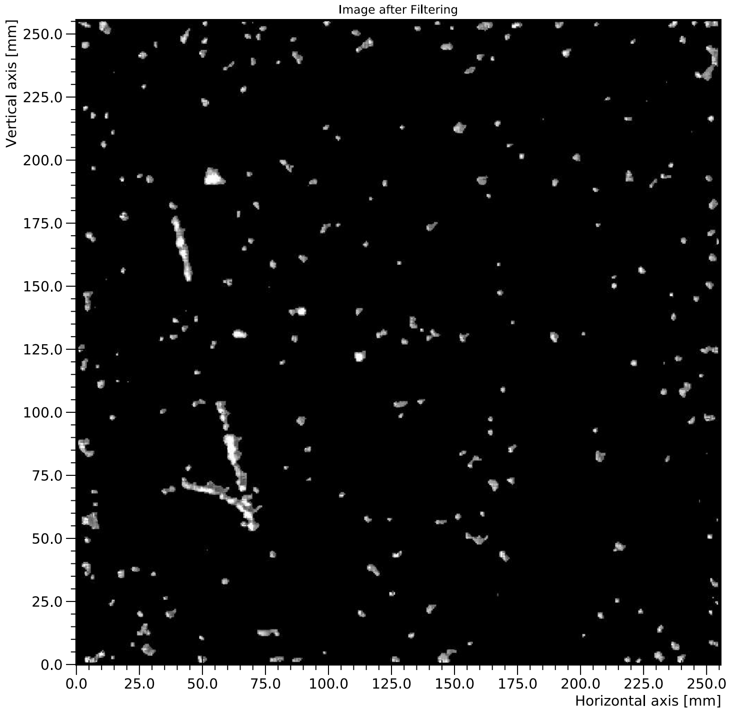
$$g(x, y) = median\{f(x, y), (x, y) \in w\} \tag{5.1}$$

The choice by such filter is justified by its effective noise suppression capability and high computational efficiency (142), which makes it very attractive for various applications. Tests performed on the EN dataset (see section 4.2) showed that this filter is able to reduce the number of noise pixels sent to the clustering algorithm by a factor of 3.07 ± 0.02, which is illustrated qualitatively in Figure 19.

5. Lastly, the clustering algorithm receives as input the coordinates (X, Y) and respective intensities (Z) of the pixels with non-zero I_i values, then the output of those algorithms is used to extract clusters' features such as integrated light, length and width. All the features are computed over the full-resolution image after the Pedestal subtraction, as indicated by the flowchart.

The event-reconstruction algorithm computes more than 20 different features from each cluster, however, for the scope of this work, it will be important to know only three of them, described as follow:

Figura 19 – Image after applying the median filter



Source: Prepared by the author (2020).

Length and width: defined as the full length of the major and minor axes along the two eigenvectors of the (X,Y) pixel matrix in the context of Principal Component Analysis (PCA) (143) are assigned as the length and width of the cluster, respectively.

Cluster light: calculated as the sum of all the pixel I_i intensities belonging to the cluster.

The CYGNO collaboration has presented in (139) a detector performance study using a clustering algorithm based on the widely employed NNC method. In order to observe the advantages of using iDBSCAN, it was compared to NNC and the DBSCAN one, all of them passing through the same preprocessing steps to guarantee a fair comparison between them.

5.3 THE CYGNO'S DENSITY-BASED CLUSTERING ALGORITHM

5.3.1 iDBSCAN

The search to improve and optimize most of the technological tasks in physics experiments, in order to achieve even better results, opens space to try different approaches. In the particle physics area, like in several others, a priori knowledge about the detection system and its data can be used to improve the performance of the clustering task (9). Hereupon, to fit in a finer way the experimental conditions and data of the LEMOn detector, an adaptation of DBSCAN (41) clustering algorithm was carried out.

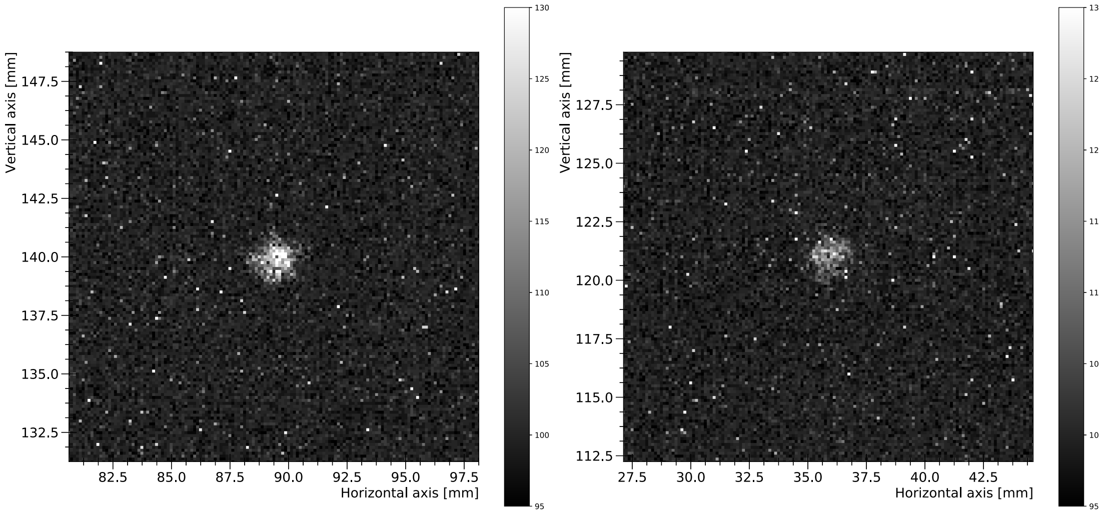
As explained in section *2.1.2.1*, DBSCAN has only two parameters: ϵ and N_{min}. The search for clusters starts from a random pixel, where it is open a hyper-sphere of radius ϵ, if the number of neighboring components inside it reaches the N_{min} value that pixel is set as a core point. Then, the same process is repeated to all the neighboring components trying to gather together more points in order to form the final cluster. After closed the first clusters the algorithm keep looking for more clusters until all the data components are analyzed.

However, to better match the CYGNO conditions, rather than just counting the number of components inside a hyper-sphere to decide if it will be member of a cluster, the intensity value of each pixel are also taking into account. In other words, the iDBSCAN approach look to the density of pixels and the N_{min} turn into a parameter associated to the total intensity within a hyper-sphere instead of to the number of elements. Consequently, whenever the total intensity inside a hyper-sphere reaches the N_{min} value, they are treated as belonging to a cluster. In the course of the development of the iDBSCAN algorithm, an effort was made in order to test many ϵ and N_{min} values, which has pointed to values around 5.8 and 30, respectively, however a validation of these parameters will be presented in section **5.3.2**.

Also, for all the three clustering methods analyzed in this work, it is required for a cluster to have more than two macro-pixels, otherwise it is discarded. In this way it is possible to increase their abilities to reject electronic noise and intensity spikes.

Figure 20 shows an example of two ^{55}Fe spots with different values of light, 3000 (left) and 1865 (right) photons. They have almost the same number of pixels, length and width, but a quiet difference on the mean intensity of each pixel. Therefore as DBSCAN inputs are only the coordinates (X,Y) that passed through the preprocessing phase, the algorithm sees barely no difference between these two clusters and it also could happen with an even fainter track. This behavior could lead to an increase in the number of fake clusters that the algorithm finds. In the other hand, as iDBSCAN also takes in to account the intensity of each pixel, it is able to better reject fake clusters, begin one of the reasons why the CYGNO Collaboration is currently using iDBSCAN for the clustering method in its event-reconstruction.

Figura 20 – Example of two ^{55}Fe spots with 3000 (left) and 1865 (right) photons



Source: Prepared by the author (2020).

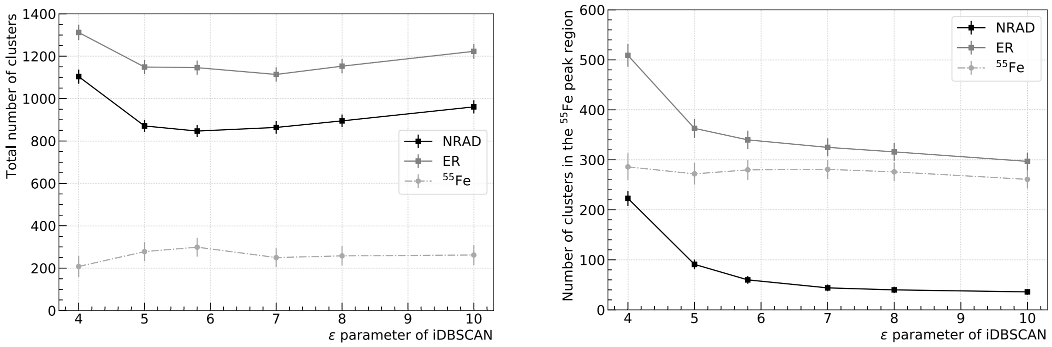
5.3.2 Validation of the iDBSCAN parameters

The iDBSCAN performance for signals originated by the interactions of photons from ^{55}Fe has been investigated as a function of different values of its parameters: ϵ and N_{min}. A test on the detector efficiency and background rejection was realized to inspect those values: a scan over the two iDBSCAN parameters. Although the ϵ (N_{min}) parameter will be fixed to a value of 5.8 (30), the other parameter's value will be swept from 5 to 50 (4 to 10). Figure 21 (left) shows the total number of clusters encountered as a function of ϵ for two different datasets: ER and NRAD. For low ϵ values, the number of NRAD clusters leans to rise, which could suggest an increase of background contamination. Nevertheless, for ϵ values between 5 and 7, this contamination rate tends to keep in steady around a minimum value. Figure 21 (right) shows a similar behavior, while counting only clusters with an integral in the range 2000–4000 photons, characteristic of ^{55}Fe deposits. This region are analyzed because it refers to the energy region of the ^{55}Fe produced electron recoils (see Fig. 27).

Correspondingly, a scan over the N_{min} parameter has been performed as shown in Fig. 22. Using the same logic for the ϵ parameter, the plot on the left indicates a low contamination region for N_{min} values between 20 and 40, and the right plot to a region for $N_{min} \leq 30$. In both of the cases, when stable, its possible to estimate the number of ^{55}Fe clusters by the difference between the results, which gives a total number of about 280.

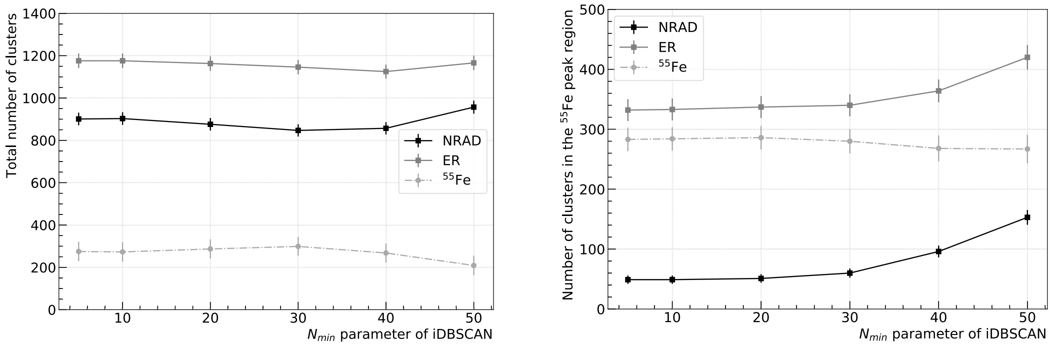
Lastly, energy resolution for all the tested iDBSCAN parameters has also been measured. The results shows a negligible variation in the energy resolution as a function

Figura 21 – Total number of reconstructed clusters using iDBSCAN (left) and Number of clusters in the ^{55}Fe peak region (right) as a function of ϵ for ER and NRAD runs and also a line for the ^{55}Fe, which means ER-NRAD



Source: Prepared by the author (2020).

Figura 22 – Total number of reconstructed clusters using iDBSCAN (left) and Number of clusters in the ^{55}Fe peak region (right) as a function of N_{min} for ER and NRAD runs and also a line for the ^{55}Fe, which means ER-NRAD



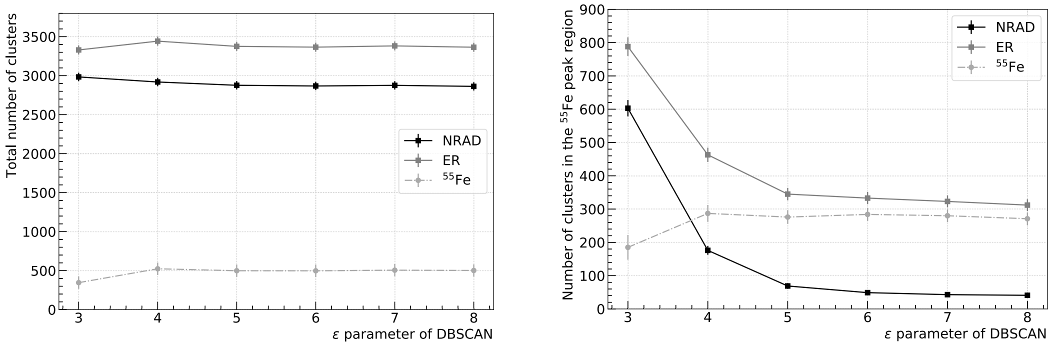
Source: Prepared by the author (2020).

of ϵ and N_{min}, while the mean value was found to be around 12.2%. More details about the energy resolution measurement are provided in Section 6.3.

5.3.3 Validation of the DBSCAN parameters

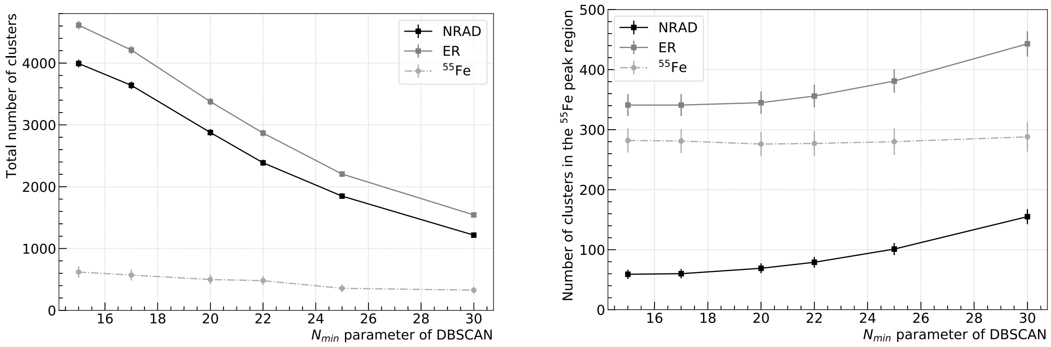
The same method performed to choose iDBSCAN parameters was also utilized for DBSCAN. The resulting values for the DBSCAN parameters were 6 for ϵ and 20 for N_{min}. It is evident that the value of ϵ for DBSCAN is very close to the 5.8 found by iDBSCAN, and that shows consistency, since the two-dimensional space is the same for both algorithms. The DBSCAN graphs are shown in the Figs. 23 and 24 and it is possible to observe that they have characteristics similar to those presented in Figs. 21 and 22.

Figura 23 – Total number of reconstructed clusters using DBSCAN (left) and Number of clusters in the ^{55}Fe peak region (right) as a function of ϵ for ER and NRAD runs and also a line for the ^{55}Fe, which means ER-NRAD



Source: Prepared by the author (2020).

Figura 24 – Total number of reconstructed clusters using DBSCAN (left) and Number of clusters in the ^{55}Fe peak region (right) as a function of N_{min} for ER and NRAD runs and also a line for the ^{55}Fe, which means ER-NRAD



Source: Prepared by the author (2020).

6 IDBSCAN COMPARED TO DBSCAN AND NNC

In this chapter, a comparison will be shown between the two clustering algorithms discussed in this work on three different data sets and also in terms of slimness cut and energy resolution.

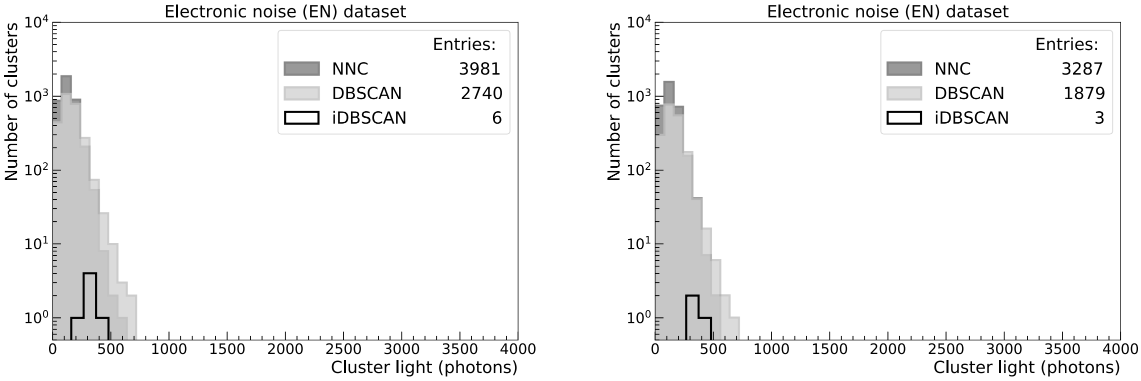
6.1 ELECTRONIC NOISE, NATURAL RADIOACTIVITY AND ^{55}FE ENERGY SPECTRA

In order to evaluate the detection efficiency and background rejection of both methods, the well-known energy deposition signature of 5.9 keV photons coming out from the ^{55}Fe source is exploited. The ER dataset will be used for signal characterization, while the background rejection measurements will be done by analyzing the EN and NRAD datasets. The EN acquired data provides low energy clusters with a distribution compressed in the region below 500 photons as shown in Fig. 25, NRAD provides an energy distribution widely spread by a heavy tail component as shown in Fig. 26 while ER forms an additional narrow distribution centered at around 3000 photons as shown in Fig. 27. In this last case, the energy spectrum is composed of background and ^{55}Fe induced deposits. Consequently, to remake the ^{55}Fe energy distribution, the background distribution should be subtracted. All the distributions were generated with 864 images, the same amount for each one of them, except for the iDBSCAN distributions of Fig. 25 which used 6478 images to collect enough EN-clusters, which occurs at a low rate. Additionally, the cluster aspect ratio, called slimness, enhances the signal purity. Slimness is defined here as the ratio between the minor axis (width) and major axis (length) of each cluster.

Figure 25 compares the energy spectrum of clusters generated by NNC and DBSCAN with those generated by iDBSCAN for EN events with and without a selection based on the slimness parameter, considering only clusters with slimness greater than 0.4 for the latter case. The computed numbers of EN-clusters per image for NNC, DBSCAN and iDBSCAN were 4.61 ± 0.17, 3.17 ± 0.12 and $(9 \pm 4) \times 10^{-4}$, respectively. Regarding NNC and DBSCAN, EN-clusters dominate the background rate for energies below 500 photons which can be noticed by comparing the EN energy distribution of Fig. 25 with that of the NRAD shown in Fig. 26. Selection on slimness variable decreases the number of clusters per image to 3.80 ± 0.14, 2.17 ± 0.09 and $(5 \pm 3) \times 10^{-4}$ for NNC, DBSCAN and iDBSCAN, respectively. Thus, when compared to NNC and DBSCAN, iDBSCAN is able to reduce the number of EN-clusters per image by a factor of $(3 \div 7) \times 10^{3}$.

Figure 26 shows the energy distributions for the NNC, DBSCAN and iDBSCAN clusters using the NRAD dataset without (left) and with (right) a selection on slimness. iDBSCAN presents a clear peak evolution of around 300 photons. At the same time, NNC

Figura 25 – Clusters energy distribution for NNC and iDSBSCAN applied to the EN dataset, without (left) and with (right) a cut on the slimness



Source: Prepared by the author (2020).

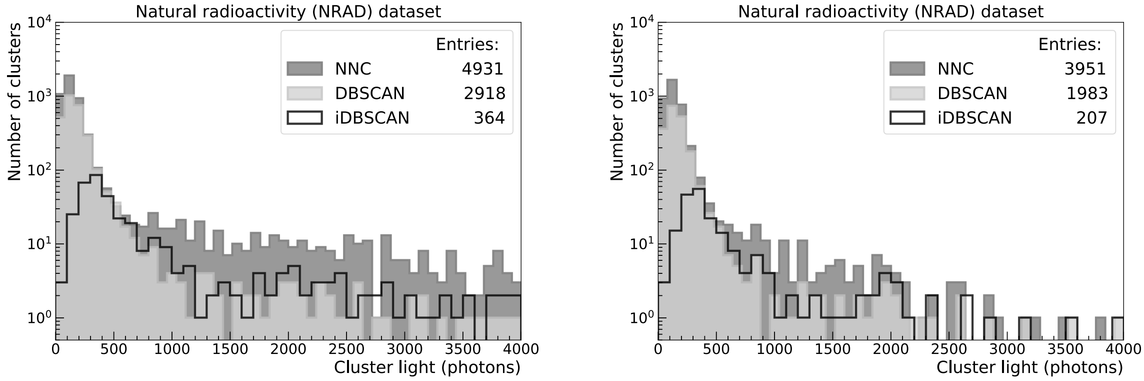
and DBSCAN accumulate clusters with lower energies due to EN-clusters. iDBSCAN and DBSCAN reduce the number of background events in the region between 2000 and 4000 photons when compared to NNC, which is the region where the ^{55}Fe events are expected to be, as mentioned before, providing better background rejection for low energy events as for the 5.9 keV photons. On the right of Fig. 26, the distribution of light, only considering clusters with slimness greater than 0.4 is shown. This selection reduces, even more, the number of background events in the ^{55}Fe region. It brings NNC closer to the other methods. Nonetheless, the number of fake clusters is only reduced on a small scale for the lower energy region, and that causes iDBSCAN to maintain a better background rejection efficiency when compared to NNC and DBSCAN.

Figure 27 shows the results of the same analysis performed on the ER dataset. The sum of the distribution obtained in the NRAD sample and the one from ^{55}Fe interactions are expected in this case. As explained before, all three clustering algorithms are sensitive to the 5.9 keV photon events. Yet a higher purity level is achieved using iDBSCAN. As shown in the right plot of Fig. 27, after applying the slimness threshold, the distributions around the ^{55}Fe peak of NNC, DBSCAN, and iDBSCAN get closer, indicating that the three methods have similar detection efficiency considering that the number of ^{55}Fe spots found by each method is basically the same.

6.2 SLIMNESS SELECTION OPTIMIZATION

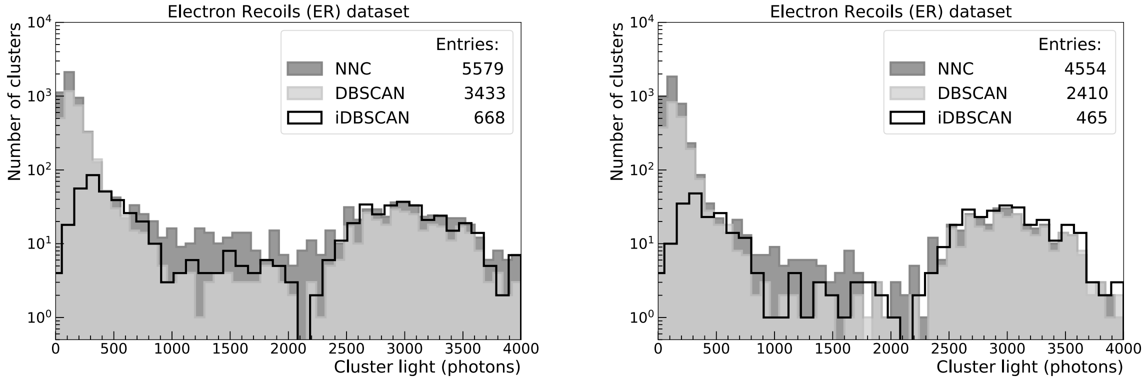
Figure 28 shows the slimness cumulative distribution of clusters for an interval between 0 and 1, applied to the NRAD and ER datasets for NNC, DBSCAN, and iDBSCAN. As shown, in all cases ^{55}Fe spots tend to have slimness higher than about 0.4.

Figura 26 – Clusters energy distribution for NNC and iDSBSCAN applied to the NR dataset, without (left) and with (right) a cut on the slimness



Source: Prepared by the author (2020).

Figura 27 – Clusters energy distribution for NNC and iDSBSCAN applied to the ^{55}Fe dataset, without (left) and with (right) a cut on the slimness

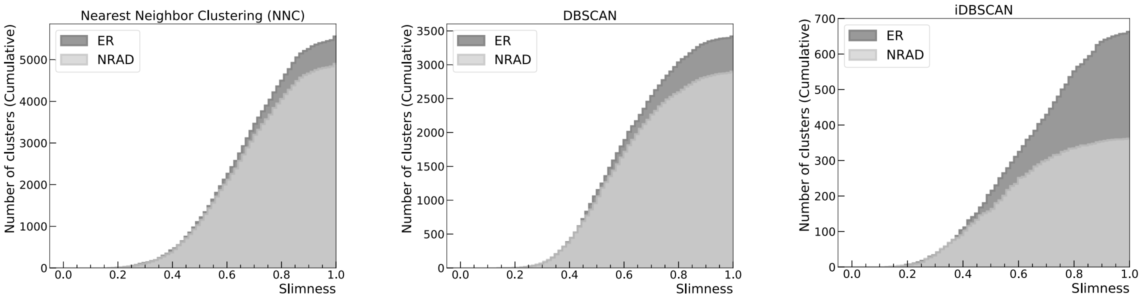


Source: Prepared by the author (2020).

This variable can be used in conjunction with energy measurement to segregate ^{55}Fe spots from background clusters. The value of slimness will be swept in this section so that it will be possible to determine the most relevant value for its use as an event selection parameter as well as to evaluate its impact when applied together with the energy measurement.

The number of clusters within the selected ^{55}Fe energy region (from 1500 to 4500 photons) was measured for various slimness threshold values ($X \geqslant x$) as shown in Fig. 29 for the NNC, DBSCAN, and iDBSCAN algorithms to evaluate the signal efficiency and purity as a function of the slimness selection for the two algorithms. This figure also shows that DBSCAN and iDBSCAN can find a similar number of clusters in the ^{55}Fe region

Figura 28 – Cumulative distribution of the slimness for NRAD and ER data, for NNC, DBSCAN and iDSBSCAN

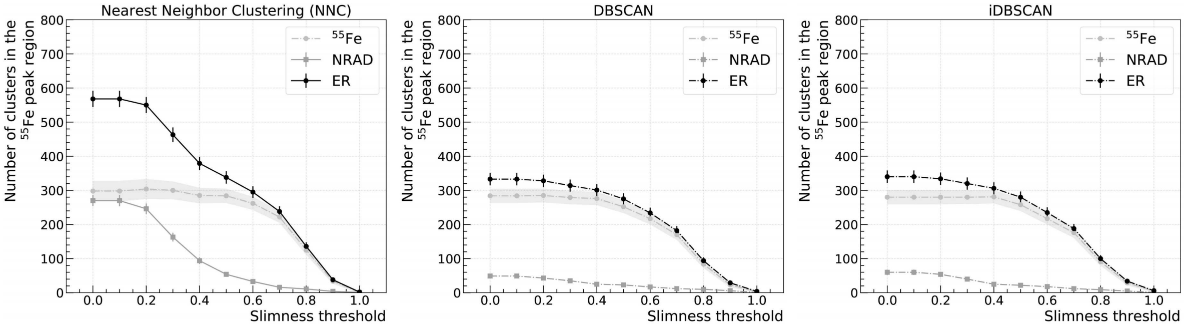


Source: Prepared by the author (2020).

when compared to NNC for slimness below 0.4, given by the difference between the ER and NRAD curves, but with lesser contamination (NRAD curves).

Considering that the ^{55}Fe clusters develop an intensity that follows a Gaussian distribution with an average value of about 3000 photons and standard deviations of 550, 385 and 371, for NNC, DBSCAN, and iDBSCAN respectively (see Fig. 30), then more than 99% of the ^{55}Fe clusters are selected between 1500 and 4500 photons. Differently, for the same region, the subtraction of the natural radioactivity events between the ER and NRAD acquisition runs has a mean value equal to zero but a fluctuation of about 23 (14), 10 (7) and 11 (7) clusters for slimness equal to 0.0 (0.4), for NNC, DBSCAN and iDBSCAN respectively. Consequently, the dashed line of Fig. 29 is made primarily of ^{55}Fe events plus a few background events produced by the statistical fluctuation that occurs in the process of subtracting natural radioactivity. As can be noticed by observing Fig. 26, DBSCAN and iDBSCAN tend to have less background contamination than NNC, which reduces the statistical uncertainty related to the background subtraction. This effect can also be shown by the shaded band drawn around the dashed lines of Fig. 29.

Figura 29 – Scan in the number of clusters on the ^{55}Fe peak region (between 1500 and 4500 photons) when changing the threshold on the slimness for NRAD and ER data, for NNC, DBSCAN and iDSBSCAN



Source: Prepared by the author (2020).

The impact of the slimness parameter, based on the measurements of Fig. 29, can be assessed by measuring the relative efficiency (ε_{sel}) concerning the bin with the highest content in the ^{55}Fe curve (so that for such a bin, $\varepsilon_{sel} = 100\%$), and fake events ($F_{evts}$), as defined below:

- ε_{sel}: number of clusters found in the ER dataset (nFe) subtracted by the number of clusters found in the NRAD dataset (nRd) divided by the maximum value of the $nFe - nRd$ subtraction among all slimness values (see Equation 6.1);

$$\varepsilon_{sel} = \left(\frac{nFe - nRd}{\max\left(nFe - nRd\right)} \right) \tag{6.1}$$

- F_{evts}: ratio between the number of clusters found in the NRAD dataset (nRd) and the number of clusters found in the ER dataset (nFe) (see Equation 6.2a). This measure can also be understood in terms of background rejection (B_{rj}) as shown by Equation 6.2b;

$$F_{evts} = \left(\tfrac{nRd}{nFe}\right) \quad (a) \quad , \qquad B_{rj} = 1 - F_{evts} \quad (b) \tag{6.2}$$

Figure 28 shows that the efficiency for background events is low for slimness below 0.4, while most of the ^{55}Fe events are retained. Tables 2 and 3 shows, respectively, the computed ε_{sel} and F_{evts} for both clustering methods and different thresholds on the slimness variable ranging from 0.0 to 0.8. The errors presented in these tables were computed considering a confidence interval of 95% for a binomial proportion (144). For the high-efficiency region (≥ 0.94), occurring for slimness values from 0.0 to 0.4, iDBSCAN and DBSCAN achieved a lower fake event probability, always about 3 times less than NNC. For slimness greater than or equal to 0.6 all methods begin to lose efficiency. More specifically, for a slimness threshold of 0.4, the efficiency is still close to 100% compared to not using slimness, but the number of fake events is reduced by a factor of about 2 for all the methods.

Tabela 2 – ε_{sel} comparison between iDBSCAN, DBSCAN and NNC.

Slimness	$\varepsilon_{\mathbf{sel}}$		
(width/length)	iDBSCAN	DBSCAN	NNC
0.0	$1.00^{+0.00}_{-0.02}$	$1.00^{+0.00}_{-0.02}$	$0.98^{+0.01}_{-0.02}$
0.2	$1.00^{+0.00}_{-0.02}$	$1.00^{+0.00}_{-0.01}$	$1.00^{+0.00}_{-0.01}$
0.4	$1.00^{+0.00}_{-0.01}$	$0.97^{+0.02}_{-0.03}$	$0.94^{+0.02}_{-0.03}$
0.6	$0.77^{+0.05}_{-0.05}$	$0.76^{+0.05}_{-0.05}$	$0.86^{+0.03}_{-0.04}$
0.8	$0.32^{+0.06}_{-0.05}$	$0.29^{+0.05}_{-0.05}$	$0.41^{+0.05}_{-0.06}$

Source: Prepared by the author (2020).

Tabela 3 – F_{evts} comparison between iDBSCAN, DBSCAN and NNC.

Slimness	F_{evts}			iDBSCAN B_{rj} variation (%)	
(width/length)	iDBSCAN	DBSCAN	NNC	DBSCAN	NNC
0.0	$0.18^{+0.04}_{-0.04}$	$0.15^{+0.04}_{-0.04}$	$0.48^{+0.04}_{-0.04}$	-3.4 $^{+7.1}_{-6.5}$	57.0 $^{+11.5}_{-12.3}$
0.2	$0.16^{+0.04}_{-0.04}$	$0.13^{+0.04}_{-0.03}$	$0.45^{+0.04}_{-0.04}$	-3.5 $^{+6.9}_{-6.4}$	51.7 $^{+10.7}_{-11.6}$
0.4	$0.08^{+0.04}_{-0.03}$	$0.08^{+0.04}_{-0.03}$	$0.25^{+0.05}_{-0.04}$	0.1 $^{+5.3}_{-6.4}$	22.6 $^{+6.6}_{-8.0}$
0.6	$0.08^{+0.04}_{-0.03}$	$0.07^{+0.04}_{-0.03}$	$0.11^{+0.04}_{-0.03}$	-0.4 $^{+6.4}_{-4.7}$	4.0 $^{+4.9}_{-6.7}$
0.8	$0.09^{+0.07}_{-0.04}$	$0.11^{+0.08}_{-0.05}$	$0.08^{+0.06}_{-0.04}$	1.8 $^{+6.5}_{-10.5}$	-1.0 $^{+10.2}_{-6.4}$

Source: Prepared by the author (2020).

The last column of Table 3 shows the iDBSCAN background-rejection improvement compared to NNC. For slimness equal to 0.4, for example, iDBSCAN has 92% of background rejection efficiency while NNC has 75%, leading to a relative improvement of (92-75)/75 ≈ 23%. Finally, The second-last column of this same table shows that iDBSCAN and DBSCAN present similar background-rejection performances.

6.3 LIGHT YIELD RESOLUTION

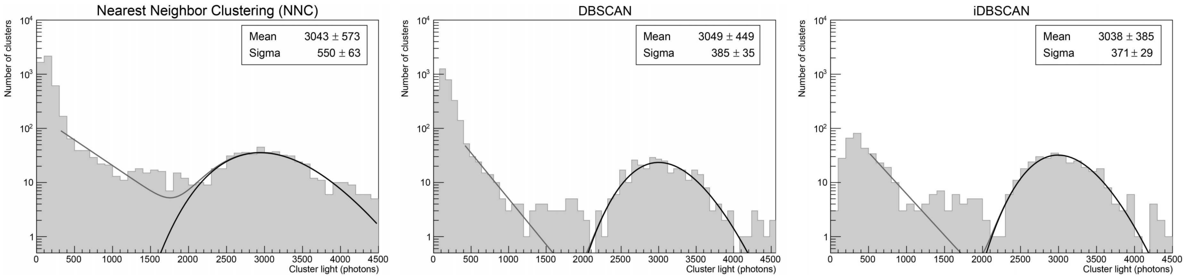
The detector energy resolution was estimated by a fit to the clusters energy distributions accounting for natural radioactivity and the ^{55}Fe events. The former was modeled by an exponential function and the latter by a Polya function (145):

$$P(n) = \frac{1}{b\overline{n}}\frac{1}{k!}\left(\frac{n}{b\overline{n}}\right)^{k} \cdot e^{-n/b\overline{n}} \tag{6.3}$$

where b is a free parameter and $k = 1/b - 1$. The distribution has $\overline{n}$ as expected value, while the variance is governed by $\overline{n}$ and the b parameter, as follows: $\sigma^2 = \overline{n}(1 + b\overline{n})$. The total likelihood is given by the sum of the two functions.

Figure 30 shows the fit results for NCC, DBSCAN and iDBSCAN clusters without applying any selection on the slimness parameter. Based on the computed values, energy resolution were measured to be (18.1 ± 3.9)%, (12.6 ± 2.2)% and (12.2 ± 1.8)% for NNC, DBSCAN and iDBSCAN respectively, and the energy conversion factor approximately 515 ADC units per keV for all of them. Conversion factor and energy resolution are computed using the *mean* and *sigma* parameters shown in Fig. 30. The former is the *mean* divided by 5.9 keV (ER energy), while the latter is given by dividing the *sigma* by the *mean.*

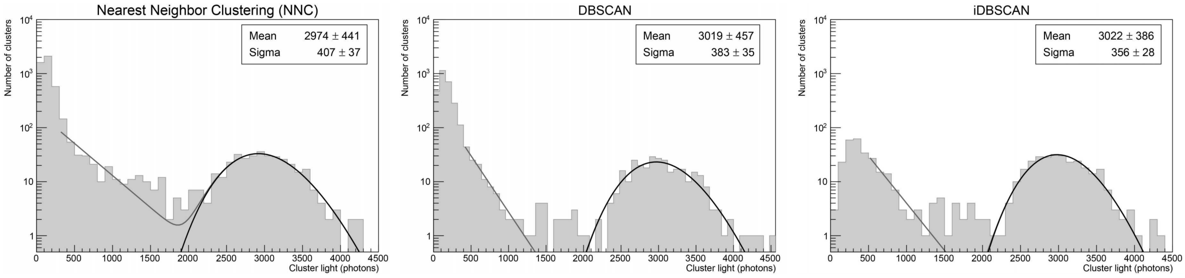
Figura 30 – Results of the fit applied to the NNC, DBSCAN and iDBSCAN energy distributions



Source: Prepared by the author (2020).

Figure 31 shows the fit results when considering only clusters with slimness greater than 0.4. The estimated energy resolutions are 13.7 ± 2.4%, 12.7 ± 2.3% and 11.8 ± 1.7% for NNC, DBSCAN and iDBSCAN, respectively, with a conversion factor of about 510 ADC units per keV.

Figura 31 – Results of the fit applied to the NNC, DBSCAN and iDBSCAN energy distributions for clusters with slimness higher than 0.4



Source: Prepared by the author (2020).

Lastly, Table 4 shows the resulting energy resolution for NNC, DBSCAN and iDBSCAN in correspondence of the different thresholds applied to the slimness. The energy resolution obtained with NNC, due to its higher background contamination, decreases as the slimness threshold value increases, reaching eventually the energy resolution obtained with iDBSCAN. The energy resolutions obtained with DBSCAN and iDBSCAN are similar and much less dependent on the slimness parameter when compared to NNC, indicating a greater purity in the selection of ^{55}Fe clusters for these two methods.

Tabela 4 – Detector resolution comparison between NNC, DBSCAN and iDBSCAN as a function of slimness.

Slimness	Resolution (%)					
(width/length)	iDBSCAN		DBSCAN		NNC	
0.0	12.2	± 1.8	12.6	± 2.2	18.1	± 4.0
0.2	12.0	± 1.7	12.6	± 2.2	17.3	± 3.7
0.4	11.8	± 1.8	12.7	± 2.3	13.7	± 2.4
0.6	12.0	± 2.0	12.9	± 2.8	11.8	± 1.8
0.8	12.3	± 3.8	10.4	± 3.1	11.1	± 2.8

Source: Prepared by the author (2020).

7 CONCLUSIONS

A fundamental task of the CYGNO experiment occurs in the clustering stage applied to the signals collected by the optical readout of its detector. In this context, this thesis offered a first study of the impact of two of the most commonly mentioned clustering algorithms in the scientific literature on the CYGNO experiment, known as NNC and DBSCAN. This study led to a modified version of the DBSCAN, called intensity-DBSCAN (iDBSCAN). The impact of this new algorithm has been examined using 5.9 keV photons from a ^{55}Fe radioactive source and compared with an outcome obtained with the standard DBSCAN and NNC algorithms. iDBSCAN has shown to be able to improve the energy resolution and background rejection of the experiment.

The achieved results showed that the clustering process of the CYGNO's event-reconstruction algorithm can achieve, with iDBSCAN and without any other event-selection routine, a natural radioactivity background rejection in the energy region around 5.9 keV (from 3.0 keV to 8.8 keV) of $0.82^{+0.04}_{-0.04}$ and a number of electronic-noise clusters per image of $(9 \pm 4) \times 10^{-4}$, occurring predominantly in the region below 1 keV ($\approx$ 500 photons). These results represent an enhancement of 57% for the former in comparison to NCC, and, for the latter, a advancement by a factor of a few thousand. In comparison to DBSCAN, iDBSCAN obtained comparable performance regarding background rejection in the ^{55}Fe energy region; yet, iDBSCAN has managed to considerably cut down the number of electronic noise clusters in comparison to DBSCAN. As a result, DBSCAN was not as efficient as iDBSCAN in reducing the effects of electronic noise, even though it accomplishes similar performance concerning iDBSCAN in the rejection of background radiation.

Lastly, the detector energy resolution using iDBSCAN was measured to be (12.2 ± 1.8)% for 5.9 keV electron recoil events. By requiring spots with slimness larger than 0.4, a rate of electronic-noise clusters per image of $(5 \pm 3) \times 10^{-4}$, a natural radioactive background rejection of $0.92^{+0.03}_{-0.04}$ and an energy resolution of (11.8 ± 1.7)% were achieved.

REFERENCES

1 XU, R.; WUNSCH, D. C. **Survey of clustering algorithms**. Institute of Electrical and Electronics Engineers (IEEE), 2005.

2 SHEN, J. et al. **Real-time superpixel segmentation by DBSCAN clustering algorithm**. *IEEE Transactions on Image Processing*, IEEE, v. 25, n. 12, p. 5933–5942, 2016.

3 VISWANATH, P.; PINKESH, R. **l-dbscan: A fast hybrid density based clustering method**. In: IEEE. *18th International Conference on Pattern Recognition (ICPR'06).* [S.l.], 2006. v. 1, p. 912–915.

4 TRAN, T. N.; DRAB, K.; DASZYKOWSKI, M. **Revised DBSCAN algorithm to cluster data with dense adjacent clusters**. *Chemometrics and Intelligent Laboratory Systems*, Elsevier, v. 120, p. 92–96, 2013.

5 RUIZ, C.; SPILIOPOULOU, M.; MENASALVAS, E. **C-dbscan: Density-based clustering with constraints**. In: SPRINGER. *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing.* [S.l.], 2007. p. 216–223.

6 CHEESEMAN, P. et al. **Autoclass: A Bayesian classification system**. In: *Machine learning proceedings 1988.* [S.l.]: Elsevier, 1988. p. 54–64.

7 FISHER, D. H. **Knowledge acquisition via incremental conceptual clustering**. *Machine learning*, Springer, v. 2, n. 2, p. 139–172, 1987.

8 DAVIDSON, I.; BASU, S. **A survey of clustering with instance level constraints**. *ACM Transactions on Knowledge Discovery from data*, ACM, v. 1, p. 1–41, 2007.

9 WAGSTAFF, K.; CARDIE, C. **Clustering with instance-level constraints**. *AAAI/IAAI*, v. 1097, p. 577–584, 2000.

10 ESTER, M. et al. **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.* AAAI Press, 1996. (KDD'96), p. 226–231. Disponível em: <http://dl.acm.org/citation.cfm?id=3001460.3001507>.

11 PINCI, D. et al. **Cygnus: development of a high resolution TPC for rare events**. *PoS*, EPS-HEP2017, p. 077, 2017.

12 SØRENSEN, T. J. ***A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons***. [S.l.]: I kommission hos E. Munksgaard, 1948.

13 AGGARWAL, C. C.; REDDY, C. K. **Data clustering**. *Algorithms and applications. Chapman&Hall/CRC Data mining and Knowledge Discovery series, Londra*, Citeseer, 2014.

14 HAN, J.; PEI, J.; KAMBER, M. ***Data mining: concepts and techniques***. [S.l.]: Elsevier, 2011.

15 JAIN, A. K. **Data clustering: 50 years beyond K-means**. *Pattern recognition letters*, Elsevier, v. 31, n. 8, p. 651–666, 2010.

16 LUXBURG, U. V. **A tutorial on spectral clustering**. *Statistics and Computing*, v. 17, n. 4, p. 395–416, 2007.

17 FREY, B. J.; DUECK, D. **Clustering by passing messages between data points**. *science*, American Association for the Advancement of Science, v. 315, n. 5814, p. 972–976, 2007.

18 NG, R. T.; HAN, J. **Efficient and effective clustering methods for spatial data mining**. In: *Proceedings of VLDB*. [S.l.: s.n.], 1994. p. 144–155.

19 HAN, J.; KAMBER, M. et al. **Data mining concept and technology**. In: *Th Annual International Symposium on Supply Chain Management*. [S.l.: s.n.], 2001. v. 132, p. 70–72.

20 ESTER, M. et al. **A density-based algorithm for discovering clusters in large spatial databases with noise.** In: *Kdd*. [S.l.: s.n.], 1996. v. 96, n. 34, p. 226–231.

21 HINNEBURG, A.; KEIM, D. A. et al. **An efficient approach to clustering in large multimedia databases with noise**. Bibliothek der Universität Konstanz, 1998.

22 KAILING, K.; KRIEGEL, H.-P.; KRÖGER, P. **Density-connected subspace clustering for high-dimensional data**. In: SIAM. *Proceedings of the 2004 SIAM international conference on data mining*. [S.l.], 2004. p. 246–256.

23 XU, X. et al. **Scan: a structural clustering algorithm for networks**. In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2007. p. 824–833.

24 CHEN, Y.; TU, L. **Density-based clustering for real-time stream data**. In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2007. p. 133–142.

25 ESTER, M. et al. **Incremental Clustering for Mining in a Data Warehousing Environment**. In: *Proceedings of the 24rd International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998. (VLDB '98), p. 323–333. ISBN 1558605665.

26 SANDER, J. et al. **Density-based clustering in spatial databases: The algorithm gdbscan and its applications**. *Data mining and knowledge discovery*, Springer, v. 2, n. 2, p. 169–194, 1998.

27 BRECHEISEN, S.; KRIEGEL, H.-P.; PFEIFLE, M. **Efficient density-based clustering of complex objects**. In: IEEE. *Fourth IEEE International Conference on Data Mining (ICDM'04)*. [S.l.], 2004. p. 43–50.

28 SARMAH, S.; SARMAH, R. D.; BHATTACHARYYA, D. K. **An effective density-based hierarchical clustering technique to identify coherent patterns from gene expression data**. In: SPRINGER. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. [S.l.], 2011. p. 225–236.

29 KRIEGEL, H.-P.; PFEIFLE, M. **Density-based clustering of uncertain data**. In: *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. [S.l.: s.n.], 2005. p. 672–677.

30 ANKERST, M. et al. **OPTICS: ordering points to identify the clustering structure**. *ACM Sigmod record*, ACM New York, NY, USA, v. 28, n. 2, p. 49–60, 1999.

31 WANG, X.; HAMILTON, H. J. **DBRS: a density-based spatial clustering method with random sampling**. In: SPRINGER. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. [S.l.], 2003. p. 563–575.

32 BIRANT, D.; KUT, A. **ST-DBSCAN: An algorithm for clustering spatial–temporal data**. *Data & knowledge engineering*, Elsevier, v. 60, n. 1, p. 208–221, 2007.

33 ESTER, M. et al. **Density-Connected Sets and their Application for Trend Detection in Spatial Databases.** In: *KDD*. [S.l.: s.n.], 1997. v. 97, p. 10–15.

34 FALKOWSKI, T.; BARTH, A.; SPILIOPOULOU, M. **Dengraph: A density-based community detection algorithm**. In: IEEE. *IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)*. [S.l.], 2007. p. 112–115.

35 NTOUTSI, I. et al. **Density-based projected clustering over high dimensional data streams**. In: SIAM. *Proceedings of the 2012 SIAM international conference on data mining*. [S.l.], 2012. p. 987–998.

36 KRIEGEL, H.-P.; PFEIFLE, M. **Hierarchical density-based clustering of uncertain data**. In: IEEE. *Fifth IEEE International Conference on Data Mining (ICDM'05)*. [S.l.], 2005. p. 4–pp.

37 SEO, J.; SHNEIDERMAN, B. **Interactively exploring hierarchical clustering results [gene identification]**. *Computer*, IEEE, v. 35, n. 7, p. 80–86, 2002.

38 BISWAS, A.; JACOBS, D. **Active image clustering: Seeking constraints from humans to complement algorithms**. In: IEEE. *2012 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.], 2012. p. 2152–2159.

39 CATANI, M. et al. **Virtual in vivo interactive dissection of white matter fasciculi in the human brain**. *Neuroimage*, Academic Press, v. 17, n. 1, p. 77–94, 2002.

40 KOBAYASHI, T. et al. **An anytime algorithm for camera-based character recognition**. In: IEEE. *2013 12th International Conference on Document Analysis and Recognition*. [S.l.], 2013. p. 1140–1144.

41 PEDREGOSA, F. et al. **Scikit-learn: Machine Learning in Python**. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

42 SCULLEY, D. **Web-scale k-means clustering**. In: *Proceedings of the 19th international conference on World wide web*. [S.l.: s.n.], 2010. p. 1177–1178.

43 COMANICIU, D.; MEER, P. **Mean shift: A robust approach toward feature space analysis**. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, v. 24, n. 5, p. 603–619, 2002.

44 ANKERST, M. et al. **OPTICS: ordering points to identify the clustering structure**. *ACM Sigmod record*, ACM New York, NY, USA, v. 28, n. 2, p. 49–60, 1999.

45 ZHANG, T.; RAMAKRISHNAN, R.; LIVNY, M. **BIRCH: an efficient data clustering method for very large databases**. *ACM sigmod record*, ACM New York, NY, USA, v. 25, n. 2, p. 103–114, 1996.

46 CERN. ***Physics***. 2015. Accessed: 2015-12-21. Disponível em: <http://home.cern/about/physics>.

47 HINSHAW, G. et al. **Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: cosmological parameter results**. *The Astrophysical Journal Supplement Series*, IOP Publishing, v. 208, n. 2, p. 19, 2013.

48 ADRIANI, O. et al. **Cosmic-ray positron energy spectrum measured by PAMELA**. *Physical review letters*, APS, v. 111, n. 8, p. 081102, 2013.

49 BOND, J. R.; EFSTATHIOU, G.; SILK, J. **Massive neutrinos and the large-scale structure of the universe**. *Physical Review Letters*, APS, v. 45, n. 24, 1980.

50 ELLIS, J. et al. ***Supersymmetric relics from the big bang***. 1984.

51 SIKIVIE, P. **Dark matter axions**. *International Journal of Modern Physics A*, World Scientific, v. 25, n. 02n03, p. 554–563, 2010.

52 PRESKILL, J.; WISE, M. B.; WILCZEK, F. **Cosmology of the invisible axion**. *Physics Letters B*, Elsevier, v. 120, n. 1-3, p. 127–132, 1983.

53 DODELSON, S.; WIDROW, L. M. **Sterile neutrinos as dark matter**. *Physical Review Letters*, APS, v. 72, n. 1, p. 17, 1994.

54 ELLIS, J.; OLIVE, K. A. **Supersymmetric dark matter candidates**. *arXiv preprint arXiv:1001.3651*, 2010.

55 SHAPOSHNIKOV, M. **Sterile neutrinos in cosmology and how to find them in the lab**. In: IOP PUBLISHING. *Journal of Physics: Conference Series*. [S.l.], 2008. v. 136, n. 2, p. 022045.

56 SCHUMANN, M. **Direct detection of WIMP dark matter: concepts and status**. *Journal of Physics G: Nuclear and Particle Physics*, IOP Publishing, v. 46, n. 10, p. 103003, 2019.

57 HARVEY, D. et al. **The nongravitational interactions of dark matter in colliding galaxy clusters**. *Science*, American Association for the Advancement of Science, v. 347, n. 6229, p. 1462–1465, 2015.

58 CLOWE, D. et al. **A direct empirical proof of the existence of dark matter**. *The Astrophysical Journal Letters*, IOP Publishing, v. 648, n. 2, p. L109, 2006.

59 FRENK, C. S.; WHITE, S. D. **Dark matter and cosmic structure**. *Annalen der Physik*, Wiley Online Library, v. 524, n. 9-10, p. 507–534, 2012.

60 LOPEZ-HONOREZ, L. et al. **Warm dark matter and the ionization history of the Universe**. *Physical Review D*, APS, v. 96, n. 10, p. 103539, 2017.

61 KAHLHOEFER, F. **Review of LHC dark matter searches**. *International Journal of Modern Physics A*, World Scientific, v. 32, n. 13, p. 1730006, 2017.

62 GASKINS, J. M. **A review of indirect searches for particle dark matter**. *Contemporary Physics*, Taylor & Francis, v. 57, n. 4, p. 496–525, 2016.

63 BARACCHINI, E. et al. **CYGNO: a CYGNUs Collaboration 1 mˆ 3 Module with Optical Readout for Directional Dark Matter Search**. *arXiv preprint arXiv:1901.04190*, 2019.

64 GOODMAN, M. W.; WITTEN, E. **Detectability of certain dark-matter candidates**. *Physical Review D*, APS, v. 31, n. 12, p. 3059, 1985.

65 LEWIN, J.; SMITH, P. ***Review of mathematics, numerical factors, and corrections for dark matter experiments based on elastic nuclear recoil***. [S.l.], 1996.

66 APRILE, E. et al. **Dark matter results from 225 live days of XENON100 data**. *Physical review letters*, APS, v. 109, n. 18, p. 181301, 2012.

67 BERNABEI, R. et al. **First results from DAMA/LIBRA and the combined results with DAMA/NaI**. *The European Physical Journal C*, Springer, v. 56, n. 3, p. 333–355, 2008.

68 AALSETH, C. et al. **Search for an annual modulation in three years of CoGeNT dark matter detector data**. *arXiv preprint arXiv:1401.3295*, 2014.

69 AHLEN, S. et al. **Limits on cold dark matter candidates from an ultralow background germanium spectrometer**. *Physics Letters B*, Elsevier, v. 195, n. 4, p. 603–608, 1987.

70 AGUILAR-AREVALO, A. et al. **Search for low-mass WIMPs in a 0.6 kg day exposure of the DAMIC experiment at SNOLAB**. *Physical Review D*, APS, v. 94, n. 8, p. 082006, 2016.

71 CRISLER, M. et al. **SENSEI: first direct-detection constraints on sub-GeV dark matter from a surface run**. *Physical Review Letters*, APS, v. 121, n. 6, p. 061803, 2018.

72 BERNABEI, R. et al. **First model independent results from DAMA/LIBRA–phase2**. *Universe*, Multidisciplinary Digital Publishing Institute, v. 4, n. 11, p. 116, 2018.

73 PARK, J. **Status of the COSINE Experiment**. In: *Proceedings, 12th Patras Workshop on Axions, WIMPs and WISPs (PATRAS 2016): Jeju Island, South Korea*. [S.l.: s.n.], 2016. p. 125–128.

74 BRONIATOWSKI, A. et al. **Cryogenic Ge detectors with interleaved electrodes: Design and modeling**. *Journal of Low Temperature Physics*, Springer, v. 151, n. 3-4, p. 830–834, 2008.

75 AKERIB, D. et al. **Surface Event Rejection Using Phonon Information in CDMS**. *Nuclear Physics B-Proceedings Supplements*, Elsevier, v. 173, p. 137–140, 2007.

76 AGNES, P. et al. **Results from the first use of low radioactivity argon in a dark matter search**. *Physical Review D*, APS, v. 93, n. 8, p. 081101, 2016.

77 AMAUDRUZ, P.-A. 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*, Elsevier, v. 85, p. 1–23, 2016.

78 AKERIB, D. S. et al. **First results from the LUX dark matter experiment at the Sanford Underground Research Facility**. *Physical review letters*, APS, v. 112, n. 9, p. 091303, 2014.

79 ANGLE, J. et al. **Search for light dark matter in XENON10 data**. *Physical Review Letters*, APS, v. 107, n. 5, p. 051301, 2011.

80 AUBIN, F. et al. **Discrimination of nuclear recoils from alpha particles with superheated liquids**. *New Journal of Physics*, IOP Publishing, v. 10, n. 10, p. 103017, 2008.

81 BARNABÉ-HEIDER, M. 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*, Elsevier, v. 555, n. 1-2, p. 184–204, 2005.

82 AMOLE, C. et al. **Improved dark matter search results from PICO-2L Run 2**. *Physical Review D*, APS, v. 93, n. 6, p. 061101, 2016.

83 AMOLE, C. et al. **Dark matter search results from the PICO-60 CF 3 I bubble chamber**. *Physical Review D*, APS, v. 93, n. 5, p. 052014, 2016.

84 BOLTE, W. et al. **A bubble chamber for dark matter detection (the COUPP project status)**. In: IOP PUBLISHING. *Journal of Physics: Conference Series*. [S.l.], 2006. v. 39, n. 1, p. 126.

85 COPI, C. J.; HEO, J.; KRAUSS, L. M. **Directional sensitivity, WIMP detection, and the galactic halo**. *Physics Letters B*, Elsevier, v. 461, n. 1-2, p. 43–48, 1999.

86 SANTOS, D. et al. **MIMAC: MIcro-tpc MAtrix of Chambers for dark matter directional detection**. In: *J. Phys. Conf. Ser.* [S.l.: s.n.], 2013. v. 469, n. 012002, p. 1311–0616.

87 BATTAT, J. et al. **Low threshold results and limits from the DRIFT directional dark matter detector**. *Astroparticle Physics*, Elsevier, v. 91, p. 65–74, 2017.

88 AMARÉ, J. et al. **First results on dark matter annual modulation from the ANAIS-112 experiment**. *Physical review letters*, APS, v. 123, n. 3, p. 031301, 2019.

89 JIANG, H. et al. **Limits on light weakly interacting massive particles from the first 102.8 kg× day data of the CDEX-10 experiment**. *Physical review letters*, APS, v. 120, n. 24, p. 241301, 2018.

90 AGNESE, R. et al. **New results from the search for low-mass weakly interacting massive particles with the CDMS low ionization threshold experiment**. *Physical review letters*, APS, v. 116, n. 7, p. 071301, 2016.

91 ADHIKARI, G. et al. **An experiment to search for dark matter interactions using sodium iodide detectors**. *arXiv preprint arXiv:1906.01791*, 2019.

92 ANGLOHER, G. et al. **Results on light dark matter particles with a low-threshold CRESST-II detector**. *The European Physical Journal C*, Springer, v. 76, n. 1, p. 1–8, 2016.

93 ABDELHAMEED, A. et al. **First results from the CRESST-III low-mass dark matter program**. *Physical Review D*, APS, v. 100, n. 10, p. 102002, 2019.

94 AGNES, P. et al. **Low-mass dark matter search with the DarkSide-50 experiment**. *Physical review letters*, APS, v. 121, n. 8, p. 081307, 2018.

95 COLLABORATION, D. et al. **Search for dark matter with a 231-day exposure of liquid argon using DEAP-3600 at SNOLAB**. *Physical Review D*, APS, v. 100, n. 2, p. 022004, 2019.

96 HEHN, L. et al. **Improved EDELWEISS-III sensitivity for low-mass WIMPs using a profile likelihood approach**. *The European Physical Journal C*, Springer, v. 76, n. 10, p. 548, 2016.

97 AKERIB, D. et al. **Results from a search for dark matter in the complete LUX exposure**. *Physical review letters*, APS, v. 118, n. 2, p. 021303, 2017.

98 ARNAUD, Q. et al. **First results from the NEWS-G direct dark matter search experiment at the LSM**. *Astroparticle Physics*, Elsevier, v. 97, p. 54–62, 2018.

99 CUI, X. et al. **Dark matter results from 54-ton-day exposure of PandaX-II experiment**. *Physical review letters*, APS, v. 119, n. 18, p. 181302, 2017.

100 BEHNKE, E. et al. **Final results of the PICASSO dark matter search experiment**. *Astroparticle Physics*, Elsevier, v. 90, p. 85–92, 2017.

101 AMOLE, C. et al. **Dark matter search results from the complete exposure of the PICO-60 C 3 F 8 bubble chamber**. *Physical Review D*, APS, v. 100, n. 2, p. 022001, 2019.

102 ABRAMOFF, O. et al. **SENSEI: Direct-detection constraints on sub-GeV dark matter from a shallow underground run using a prototype skipper CCD**. *Physical review letters*, APS, v. 122, n. 16, p. 161801, 2019.

103 AGNESE, R. et al. **First dark matter constraints from a SuperCDMS single-charge sensitive detector**. *Physical review letters*, APS, v. 121, n. 5, p. 051301, 2018.

104 APRILE, E. et al. **XENON100 dark matter results from a combination of 477 live days**. *Physical Review D*, APS, v. 94, n. 12, p. 122001, 2016.

105 COLLABORATION, X. et al. **Dark matter search results from a one ton-year exposure of xenon1t**. *Physical review letters*, APS, v. 121, n. 11, p. 111302, 2018.

106 ABE, K. et al. **A direct dark matter search in XMASS-I**. *Physics Letters B*, Elsevier, v. 789, p. 45–53, 2019.

107 PATRIGNANI, C. et al. **Particle data group**. *Chin. Phys. C*, v. 40, n. 100001, p. 253, 2016.

108 NYGREN, D. ***Time-projection chamber-1975***. [S.l.], 1975.

109 NYGREN, D. R. **The time projection chamber**. 1978.

110 SAULI, F. **Micro-pattern gas detectors**. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, Elsevier, v. 477, n. 1-3, p. 1–7, 2002.

111 TITOV, M.; ROPELEWSKI, L. **Micro-pattern gaseous detector technologies and RD51 collaboration**. *Modern Physics Letters A*, World Scientific, v. 28, n. 13, p. 1340022, 2013.

112 PINTO, S. D. **Micropattern gas detector technologies and applications the work of the RD51 collaboration**. In: IEEE. *IEEE Nuclear Science Symposuim & Medical Imaging Conference*. [S.l.], 2010. p. 802–807.

113 OED, A. **Position-sensitive detector with microstrip anode for electron multiplication with gases**. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, Elsevier, v. 263, n. 2-3, p. 351–359, 1988.

114 GEM, F. S. **A new concept for electron amplification in gas detectors Nucl**. *Instr. and Meth. A*, v. 386, p. 531, 1997.

115 BACHMANN, S. et al. **Development and applications of the gas electron multiplier**. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, Elsevier, v. 471, n. 1-2, p. 115–119, 2001.

116 MARINHO, P. R. B. ***Desenvolvimento de detectores Sensíveis à posição Multifilares e Multi-GEM para Obtenção de Imagens de Raios-X***. Tese (Doutorado) — Tese de Doutorado, CBPF. Rio de Janeiro, 2006.

117 ALTUNBAS, C. et al. **Construction, test and commissioning of the triple-GEM tracking detector for COMPASS**. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, Elsevier, v. 490, n. 1-2, p. 177–203, 2002.

118 ALFONSI, M. et al. **High-rate particle triggering with triple-GEM detector**. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, Elsevier, v. 518, n. 1-2, p. 106–112, 2004.

119 LAMI, S. et al. **A triple-GEM telescope for the TOTEM experiment**. *Nuclear Physics B-Proceedings Supplements*, Elsevier, v. 172, p. 231–233, 2007.

120 BAUDIS, L. **Direct dark matter detection: the next decade**. *Physics of the Dark Universe*, Elsevier, v. 1, n. 1-2, p. 94–108, 2012.

121 SCHNEIDER, P. ***Extragalactic astronomy and cosmology: an introduction***. [S.l.]: Springer, 2014.

122 PASACHOFF, J. M.; FILIPPENKO, A. ***The cosmos: Astronomy in the new millennium***. [S.l.]: Cambridge University Press, 2013.

123 DEROCCO, W. et al. **Supernova signals of light dark matter**. *Physical Review D*, APS, v. 100, n. 7, p. 075018, 2019.

124 KNIRCK, S. et al. **Directional axion detection**. *Journal of Cosmology and Astroparticle Physics*, IOP Publishing, v. 2018, n. 11, p. 051, 2018.

125 IRASTORZA, I. G.; GARCÍA, J. A. **Direct detection of dark matter axions with directional sensitivity**. *Journal of Cosmology and Astroparticle Physics*, IOP Publishing, v. 2012, n. 10, p. 022, 2012.

126 SÉGUINOT, J.; ZICHICHI, A.; YPSILANTIS, T. ***A high rate solar neutrino detector with energy determination***. [S.l.], 1992.

127 ARPESELLA, C.; BROGGINI, C.; CATTADORI, C. **A possible gas for solar neutrino spectroscopy**. *Astroparticle Physics*, Elsevier, v. 4, n. 4, p. 333–341, 1996.

128 BARACCHINI, E. et al. **Negative Ion Time Projection Chamber operation with $\mathbf{SF_6}$ at nearly atmospheric pressure**. *JINST*, v. 13, n. 04, p. P04022, 2018.

129 MARAFINI, M. et al. **ORANGE: A high sensitivity particle tracker based on optically read out GEM**. *Nucl. Instrum. Meth.*, A845, p. 285–288, 2017.

130 ANTOCHI, V. C. et al. **Combined readout of a triple-GEM detector**. *JINST*, v. 13, n. 05, p. P05001, 2018.

131 Mazzitelli, G. et al. **A high resolution TPC based on GEM optical readout**. In: *2017 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*. [S.l.: s.n.], 2017. p. 1–4. ISSN 2577-0829.

132 PINCI, D. et al. **High resolution TPC based on optically readout GEM**. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 2018. ISSN 0168-9002. Disponível em: <http://www.sciencedirect.com/science/article/pii/S016890021831711X>.

133 ANTOCHI, V. et al. **A GEM-based Optically Readout Time Projection Chamber for charged particle tracking**. *arXiv preprint arXiv:2005.12272*, 2020.

134 SAULI, F. **The gas electron multiplier (GEM): Operating principles and applications**. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, Elsevier, v. 805, p. 2–24, 2016.

135 CLEMENTE, F. D. ***Simulations of the CYGNO detector for Dark Matter direct search***. Dissertação (Mestrado) — Sapieza Università di Roma, 10 2020.

136 PHAN, N.; LEE, E.; LOOMBA, D. **Imaging 55Fe electron tracks in a GEM-based TPC using a CCD readout**. *Journal of Instrumentation*, IOP Publishing, v. 15, n. 05, p. P05012, 2020.

137 SANGIORGIO, S. et al. **First demonstration of a sub-keV electron recoil energy threshold in a liquid argon ionization chamber**. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, Elsevier, v. 728, p. 69–72, 2013.

138 AUDI, G. et al. **The Nubase evaluation of nuclear and decay properties**. *Nuclear Physics A*, v. 729, n. 1, p. 3 – 128, 2003. ISSN 0375-9474. The 2003 NUBASE and Atomic Mass Evaluations. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0375947403018074>.

139 COSTA, I. A. et al. **Performance of optically readout GEM-based TPC with a 55Fe source**. *Journal of Instrumentation*, IOP Publishing, v. 14, n. 07, p. P07011–P07011, jul 2019.

140 Mazzitelli, G. et al. **A high resolution TPC based on GEM optical readout**. In: *2018 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*. [S.l.: s.n.], 2018. Under publication in IEEE Nuclear Science Symposium Medical Imaging Conference, 2018.

141 LOPES, G. et al. **Study of the Impact of Pre-processing Applied to Images Acquired by the Cygno Experiment**. In: _____. *Pattern Recognition and Image Analysis*. [S.l.]: Springer International Publishing, 2019. p. 520–530. ISBN 978-3-030-31320-3.

142 GONZALEZ, R. C.; WOODS, R. E. et al. **Digital Image Processing**, vol. 141, no. 7. *Publishing House of Electronics Industry*, 2002.

143 JOLLIFFE, I. T. **Springer series in statistics**. *Principal component analysis*, Springer New York, NY, v. 29, 2002.

144 BROWN, L. D.; CAI, T. T.; DASGUPTA, A. **Interval estimation for a binomial proportion**. *Statistical science*, JSTOR, p. 101–117, 2001.

145 BLUM, W.; ROLANDI, L.; RIEGLER, W. ***Particle detection with drift chambers***. Springer Science & Business Media, 2008. (Particle Acceleration and Detection, ISBN = 9783540766834). ISBN 9783540766834, 9783540766841. Disponível em: <http://www.springer.com/physics/elementary/book/978-3-540-76683-4>.

APPENDIX A – PUBLICATION LIST

1. Book Chapter Publication

   Lopes G.S.P. et al. (2019) **Study of the Impact of Pre-processing Applied to Images Acquired by the Cygno Experiment**. In: Morales A., Fierrez J., Sánchez J., Ribeiro B. (eds) Pattern Recognition and Image Analysis. IbPRIA 2019. Lecture Notes in Computer Science, vol 11868. Springer, Cham.

   This work proposes to evaluate the effect of digital filters when applied to images acquired by the ORANGE prototype of the Cygno experiment. A preliminary analysis is presented in order to understand if filtering techniques can produce results that justify investing efforts in the pre-processing stage of those images. Such images come from a camera sensor based on CMOS technology installed in an appropriate gas detector. To perform the proposed work, a simulation environment was created and used to evaluate some of the classical filtering techniques known in the literature. The results showed that the signal-to-noise ratio of the images can be considerably improved, which may help in subsequent processing steps such as clustering and particles identification.

2. Journal Publication

   Costa, I. Abritta, et al. **Performance of optically readout GEM-based TPC with a 55Fe source**. Journal of Instrumentation 14.07 (2019): P07011.

   Optical readout of large Time Projection Chambers (TPCs) with multiple Gas Electron Multipliers (GEMs) amplification stages has shown to provide very interesting performances for high energy particle tracking. Proposed applications for low-energy and rare event studies, such as Dark Matter search, ask for demanding performance in the keV energy range. The performance of such a readout was studied in details as a function of the electric field configuration and GEM gain by using a 55Fe source within a 7 litre sensitive volume detector developed as a part of the R&D for the CYGNUS project. Results reported in this paper show that the low noise level of the sensor allows to operate with a 2 keV threshold while keeping a rate of fake-events lesser than 10 per year. In this configuration, a detection efficiency well above 95% along with an energy resolution (σ) of 18% is obtained for the 5.9 keV photons demonstrating the very promising capabilities of this technique.

3. Accepted Journal Article

   Baracchini, Elisabetta, et al. **A density-based clustering algorithm for the CYGNO data analysis**. Journal of Instrumentation (2020).

   Time Projection Chambers (TPCs) working in combination with Gas Electron Multipliers (GEMs) produces a very sensitive detector capable of detecting low energy events by capturing photons generated during the GEM electron multiplication process by means of a high-resolution photo camera. The CYGNO Experiment has recently developed a TPC-Triple GEM detector coupled to a low noise and high spatial resolution CMOS sensor. For the image analysis, an algorithm based on an adapted version of the well-known DBSCAN was implemented. In this paper a description of the CYGNO's DBSCAN-based algorithm will be given, including test and validation of its parameters, and a comparison with a widely used algorithm known as Nearest Neighbor Clustering (NNC). The results will show that the adapted version of DBSCAN is capable of providing full signal detection efficiency and very good energy resolution while improving the detector background rejection.