UCLouvain

FACULTE DES SCIENCES

Faculté des sciences

École de Physique

CP3

Optimisation of passing condition variables of ELOWEN particles with unsupervised machine learning tools

Auteur : Maxime Harvengt

Promoteur : Gwenhaël de Wasseige

Lecteurs : Eduardo Cortina Gil

Vincent Lemaitre

Année académique 2023-2024

Mémoire réalisé en vue de l'obtention du grade académique de Master [120] en sciences physiques, finalité approfondie.

Acknowledgement

I would like to thank my family and Laurine for their support.

Special thanks to my friends, especially Sébastien and Aurélien for their encouragement and motivation.

I would like to thank Karlijn from the CP3 for her guidance and coaching.

And finally I would like to thank my supervisor, Gwenhaël de Wasseige, for her dedicated time, expertise, and advice.

Contents

1 Neutrino Astronomy **5**
- 1.1 Solar neutrinos . . . 5
- 1.2 Supernova neutrinos . . . 6
- 1.3 Atmospheric origin . . . 8
- 1.4 Other sources . . . 10
- 1.5 Solutions . . . 10

2 The IceCube Neutrino Observatory **13**
- 2.1 Description of IceCube . . . 13
- 2.2 Event signatures . . . 14
- 2.3 Data process . . . 16
- 2.4 Background . . . 17
- 2.5 Results . . . 18
- 2.6 Decrease energy threshold . . . 18

3 New approach **21**
- 3.1 Supervised versus Unsupervised learning . . . 21
- 3.2 Unsupervised learning . . . 22
  - 3.2.1 PCA . . . 22
  - 3.2.2 Isomap . . . 23
  - 3.2.3 t-SNE . . . 24
  - 3.2.4 UMAP . . . 26
- 3.3 Application to ELOWEN . . . 27

4 A Novel Methodology **29**
- 4.1 Data processing . . . 29
- 4.2 Cuts . . . 31
- 4.3 Additional information . . . 32

5 Results **37**
- 5.1 Background . . . 37
- 5.2 Low energy neutrino . . . 38

5.3 PCA analysis 40
5.4 Neighbour graph based algorithms analysis 42
5.5 Comparison of unsupervised learning algorithms 44
5.5.1 2D histograms and cuts analysis 44
5.5.2 Quality of the cuts 45
5.6 Conclusion 46

6 My contribution **51**

A Full results **53**

Chapter 1

Neutrino astronomy

Neutrino astronomy is a fascinating branch of astrophysics that studies neutrinos, light elementary particles difficult to detect, coming from various cosmic sources. Due to their weak interactions and neutral charge, neutrinos can traverse vast distances without being deflected or absorbed, making them essential for probing any energy phenomena in the Universe. Moreover, detecting high-energy neutrinos, those beyond the energies produced naturally on Earth or by human activities, is important to confront our models and understand the fundamental properties of neutrinos. This chapter discusses the potential sources of neutrinos and the solutions that could enhance our understanding of the neutrino energy spectrum.

1.1 Solar neutrinos

The Sun has been studied for a long time to understand the reason for its energy production. The mass lost to the sun is converted into energy through several sequences of nuclear reactions. Among these reactions, some give rise to disappearance of electrons compensated by production of neutrinos and associated with a characteristic spectrum. As these neutrinos come from the sun it is appropriate to call them "solar neutrinos". Solar neutrinos have been the driving force behind the advances in solar physics and neutrino physics for over 50 years. The Standard Solar Model (SSM) has made an enormous contribution to the study of solar neutrinos. This model is based on a description of the composition of the Sun and the processes that occur within it, in order to accurately predict the profile of the Sun's density, pressure, temperature and neutrino fluxes [1] and [2]. Solar neutrinos are classified according to two fusion cycles. The reactions of these cycles are the following :

$$\mathrm{p} + \mathrm{p} \longrightarrow {}^{2}\mathrm{H} + \mathrm{e}^{+} + \nu_e \tag{1.1}$$
$$\mathrm{p} + \mathrm{e}^{-} + \mathrm{p} \longrightarrow {}^{2}\mathrm{H} + \nu_e \tag{1.2}$$
$${}^{3}\mathrm{He} + \mathrm{p} \longrightarrow {}^{4}\mathrm{He} + \mathrm{e}^{+} + \nu_e \tag{1.3}$$
$${}^{7}\mathrm{Be} + \mathrm{e}^{-} \longrightarrow {}^{7}\mathrm{Li} + \nu_e \tag{1.4}$$
$${}^{8}\mathrm{B} \longrightarrow 2\,({}^{4}\mathrm{He}) + \mathrm{e}^{+} + \nu_e, \tag{1.5}$$

for the pp-chain and

$${}^{13}\mathrm{N} \longrightarrow {}^{13}\mathrm{C} + \mathrm{e}^{+} + \nu_e \tag{1.6}$$
$${}^{15}\mathrm{O} \longrightarrow {}^{15}\mathrm{N} + \mathrm{e}^{+} + \nu_e \tag{1.7}$$
$${}^{17}\mathrm{F} \longrightarrow {}^{17}\mathrm{O} + \mathrm{e}^{+} + \nu_e, \tag{1.8}$$

for the CNO cycle [2].

The pp-chain contributes to a large amount of solar neutrino and is also the primary source of solar power (99%). Equations (1.1), (1.2), (1.3), (1.4) and (1.5) are commonly known as pp, pep, hep, ^{7}Be and ^{8}B, respectively. Figure 1.1 shows the flux of these reactions as a function of the energy as predicted by the SSM. According to SSM, pp represents more than 90% of the solar neutrino flux coming to Earth, whereas hep contributes to less than 0.0001%, which makes it an extremely challenging flow to measure [1] and [3].

Concerning the CNO cycle, despite its dominant energy production for stars greater than $1.44M_{\odot}$, CNO is predicted from SSM to yield 1% of the solar luminosity. However, to date no neutrino derived from this cycle has been detected without large uncertainties. As metals affect the CNO cycle, CNO tells us about the metal component of the sun's core, the temperature and density profile. This makes it one of the most exciting open questions [3].

Solar neutrinos are physical interest for many reasons. They are used for research into flavour-oscillation and neutrino mass, especially helpful in understanding physics beyond the standard model of particle physics. They also tell us in detail about the behaviour of the sun, i.e. they can give us information about the reactions taking place in our nearest star in real time. Unlike photons, which escape the sun in 100,000 years, neutrinos take only 2 s to emerge from the last layers of the sun [1] and [3].

1.2 Supernova neutrinos

Core-collapse supernova is the cosmic source the most powerful of neutrino energy in the range of MeV. This astronomical phenomenon is connected to the

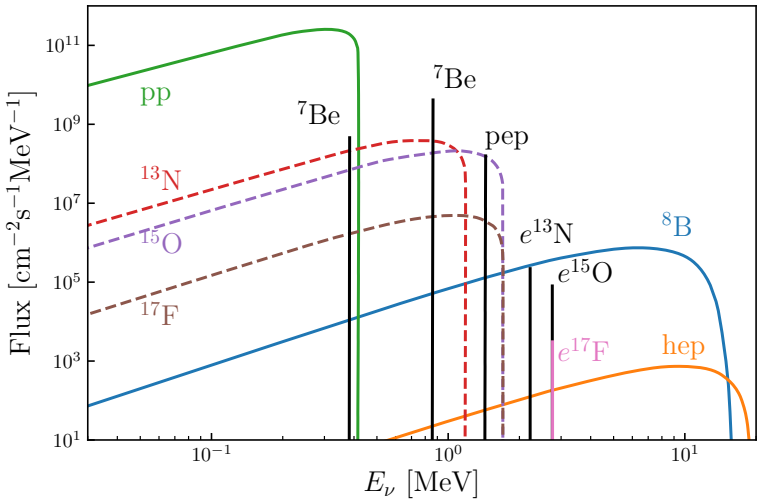

Figure 1.1: Neutrino spectrum predicted by Standard Solar Model.

formation of a large density object composed of neutrons, the neutron star. At the center of massive stars, nuclear reactions involving heavy elements take place, leading to the formation of a core mainly composed of iron. When the core exceeds the Chandrasekhar mass of $1.44M_{\odot}$, the force of gravity overcomes electron degeneracy pressure and the core begins to collapse slowly [1] and [4]. The core-collapse supernova occurs according to [4] in different phases:

- **Accelerating collapse** The iron core subjected to a slow contraction, becomes extremely hot such that photons have sufficient energy to separate iron-group nuclei into α-particles and free nucleons. Then, the pressure increases with density and the electron capture becomes possible. The drop of electrons leads to accelerating the collapse close to free-fall speeds. The electron capture allows the production of neutrinos which can initially escape unhindered and will later be trapped when the density of the core exceeds $10^{11}\ g\ cm^{-3}$.

- **Core bounce** Fraction of seconds after the acceleration, a proto-neutron star is formed with a nuclear-matter density. The nuclear force of this newborn acts like a wall and sends back the remaining matter of the core which until now has fallen freely, triggering a bounce shock opposing the ongoing collapse.

- **Neutrino burst and the shock propagation** As the shock propagates, its density decreases until it becomes transparent to neutrinos. It is during this period that the supernova's greatest neutrino intensity occurs, the breakout

burst. Furthermore, as the kinetic energy of the shock is lost in thermal energy and disintegration of heavy nuclei, the expansion of the bounce shock is weakening. However, the shock doesn't stop immediately and speeds up again with neutrino heating (most promising explanation).

The resulting neutron star contracts and accretes more matter. No more than one second after the outward acceleration of the shock wave, the accretion stops. The star cools and emits neutrinos of all flavours. These neutrinos are absorbed, re-emitted and down scattered billions of times before reaching the neutrinosphere from which neutrinos can be released [4].

On 24 February 1987 the supernova, SN1987A in the Large Magellanic Cloud about 50 kpc away, was observed by three underground experiments (Kamiokande II, Irvine-Michigan-Brookhaven and Baksan). A total of 24 neutrinos were detected at energies ranging from 10 MeV to 50 MeV, in the interval of 12 seconds. Most of the neutrino events happened in the two first seconds. Then, this supernova involved a neutrino burst as in the core-collapse model discussed previously. However, neutrino bursts are exceptional events that occur only a few times a century and SN1987A was the only event of its kind detected to date [5].

1.3 Atmospheric origin

Earth's atmosphere is continuously bombarded by cosmic rays, high energy protons and atomic nuclei originating from outer space. When these high-energy particles collide with nucleons in the atmosphere, they generate a cascade of interactions that produces a variety of secondary particles, such mesons like pions and kaons, as well as protons and neutrons. Figure 1.2 illustrates the secondary particles generated in the atmosphere and their subsequent decay into fundamental particles. The figure highlights electronic showers triggered by π^0 decay and muon decay, as well as the muonic components and neutrinos produced by kaon and $\pi^{\pm}$ decay.

Pions mainly contribute to the production of atmospheric neutrinos. At low energies, pions decay before they can interact, offering a relatively constant flux of neutrinos. Above 100 GeV, the mean free path of the pion is smaller than the decay length. Therefore, for higher energy, pions are more likely to lose energy before decaying. This has consequences for the neutrino spectrum in the range 100 GeV - 100 TeV, whose slope becomes steeper and steeper with a distribution $\sim E_{\nu}^{-3.7}$.

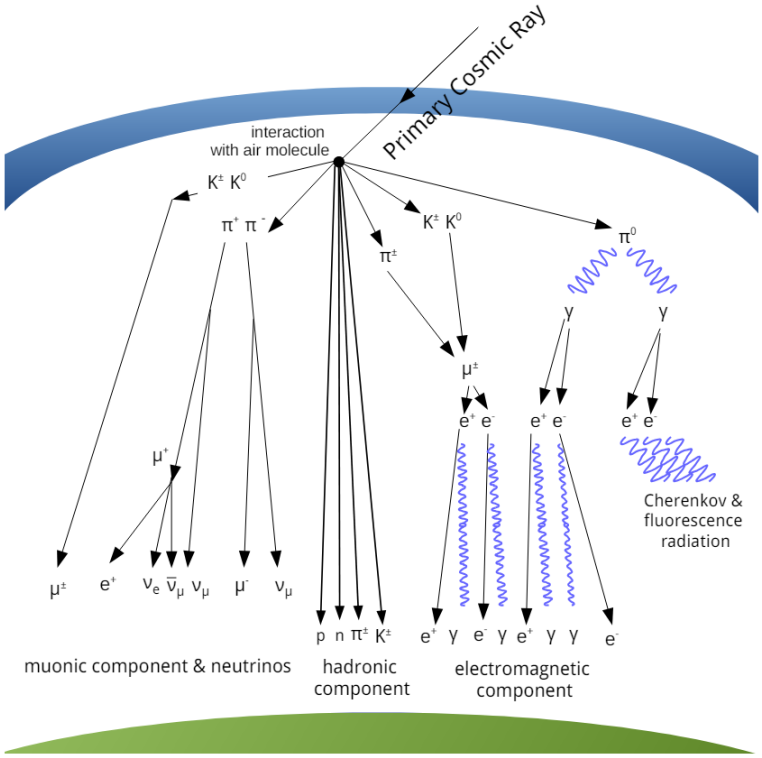

Figure 1.2: Various interactions resulting from primary cosmic rays.

The neutrinos (and muons) produced in the atmosphere form an unavoidable background that makes it laborious to observe the galactic sources associated with neutrinos of similar energy [1].

Additionally, charmed particles are produced when cosmic rays interact with the atmosphere. Their decay length is shorter than their interacting length such that they decay almost instantaneously into neutrinos. Despite the unstable estimations of the theory, those prompt neutrinos are predicted to exceed the contribution of pion and kaon around 100 TeV. However, no prompt neutrino has been detected so far [1].

1.4 Other sources

Looking at the Figure 1.3, aside from terrestrial neutrinos and reactor neutrinos, there are some sources yet to be discussed. Below is a brief overview of these sources.

Cosmological neutrino refers to neutrinos that were produced in the early universe, such as Cosmic Neutrino Background (CNB) or Big Bang Nucleosynthesis (BBN) Neutrino. They have been traveling through space ever since and provide important information about the processes that occurred shortly after the Big Bang.

Blazars are a type of Active Galactic Nucleus (AGN). They emit relativistic jets of particles able to accelerate protons to very high energies. The interactions of these protons with the surrounding matter can produce neutrinos with energies ranging from a TeV to a PeV. Similar energy ranges could also be reached by Star Forming Galaxies (SFGs) or Gamma-ray Bursts (GRBs).

Cosmogenic neutrinos are very high-energy neutrinos produced by the interaction of ultra-high-energy cosmic rays (UHECRs) with the cosmic microwave background (CMB). These neutrinos have not been detected as their energy range exceeds the scope of the current detectors. Moreover, the sources and mechanisms of UHECRs, which could have energies up to 10^{20}, remain unknown [6].

1.5 Solutions

Since the inception of neutrino astronomy, only a few sources have been identified. IceCube has opened a window on our universe with its sensitivity ranging

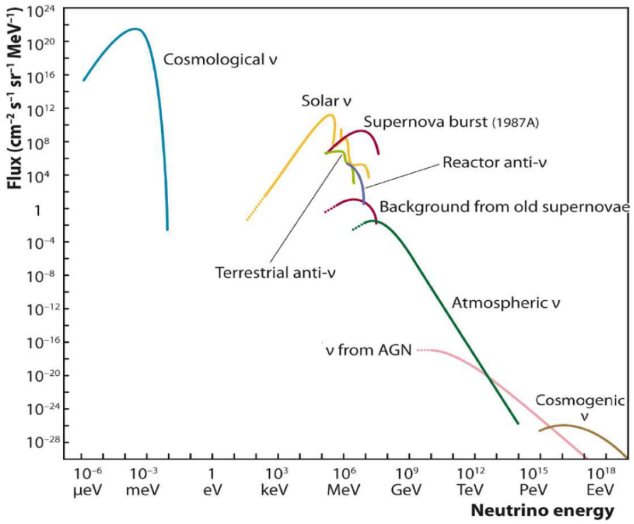

Figure 1.3: Neutrino flux as a function of energy from different astrophysical sources and from the Earth.

from TeV to PeV energies. However, additional information is needed to fully characterize the entire neutrino flux. Several future scientific projects aim to extend the observable energy range of neutrinos.

IceCube-Gen2 is an extension of IceCube. It is designed to reach energies from TeV to EeV, to increase the annual rate of detected cosmic neutrinos by a factor of ten, and to improve sensitivity to high-energy neutrinos [7]. Similarly, the Radio Neutrino Observatory in Greenland (RNO-G) pursues the same goal. It would utilize the Askaryan effect, which occurs in neutrino-induced particle cascades within ice, to detect astrophysical neutrinos with energies ranging from several PeV to EeV[8].

Another approach is to fully exploit the existing data from IceCube, which has accumulated nearly 15 years of observations. By refining data selection techniques, it may be possible to lower the energy threshold for detected neutrinos. That will be the focus of this master's thesis.

Chapter 2

The IceCube Neutrino Observatory

In this chapter, we will first explain the operation of the IceCube Neutrino Observatory. Following that, Section 2.2 will detail the two main neutrino event signatures that can be recorded. Section 2.3 will then describe the data processing of IceCube, as well as the various sources of possible background in Section 2.4. Finally, Section 2.6 will present a significant result from IceCube, leading to the motivation for this master thesis.

2.1 Description of IceCube

IceCube is a neutrino telescope located on the South Pole with a volume of a km3. It was completely built in 2010 and it became fully functional in May 2011. As shown in Figure 2.1, IceCube is made up of 86 strings buried on the ice sheet until the depth of 2450 m. At each string, are attached 60 sensors called digital optical modules, or DOMs, spread along the strings between 1450 m and 2450 m, for a total of 5160 sensors. In most of the detector, the strings are spaced approximately by 125 m and the sensors are vertically separated by 17 m. The region of DeepCore is an exception where the space dimensions are 70 m and 7 m, respectively, allowing to probe low energy ranges. The IceTop is located in the ice surface and consists of 81 stations. Each station has two frozen water tanks containing two DOMs. IceCube is also composed of its precursor Amanda II which was used as a proof of concept before the total deployment of IceCube [9], [10] and [11].

DOMs detect the neutrino indirectly via the Cerenkov effect of charged particles produced by the interaction between the neutrino and ice. When a high energy charged particle passes through a medium, it excites the electrons in the outer layers of the atoms within that medium. When electrons return to their lower energy states, they release photons at optical wavelengths. Whether the charged

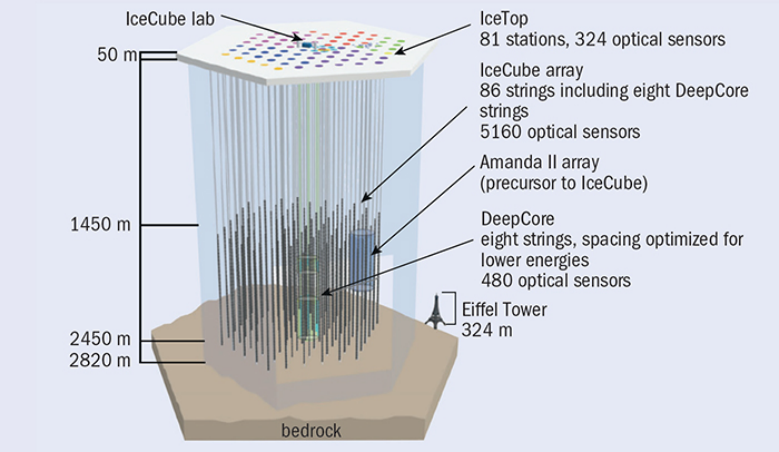

Figure 2.1: IceCube telescope. Each region of the detector is depicted, including the precursor of IceCube, Amanda II.

particle travels in the medium at a velocity higher than the speed of light, light surroundings the particle takes the shape of a cone generating an electromagnetic shock wave. This phenomenon is referred as the Cerenkov effect [9].

IceCube is deployed in ice, unlike many other neutrino detectors built in water, such as KM3NeT. Ice has both advantages and disadvantages compared with water. The optical characteristics of IceCube rely on the quality of the ice, i.e., on the dust concentration. Indeed, dust concentration varies as a function of depth with a maximum at 1300m, associated to the last glacial peak. This adds complexity to the analysis. Nonetheless, ice provides a lower noise than water, 560 Hz against 20-40 kHz for water. These last values are primarily due to bioluminescence and against ^{40}K decay, processes that are missing in the ice sheet of the South Pole. An additional distinction concerns scattering and absorption. The shorter scattering length of ice leads to a bad angular resolution caused by the delay of photons while the weaker absorption offers larger quantities of photons reaching the DOMs [9].

2.2 Event signatures

As mentioned previously, the neutrino is observed via Cerenkov radiation of a subsequent particle. For instance, a neutrino of 10 GeV could transfer 50% of its energy (and it may achieve 80% for higher energy) to a lepton through charged current (CC) interaction. The remaining energy is sent to the nucleus of the target causing a hadronic shower. Cerenkov light is generated by both, the shower and

the lepton. As shown in Figure 2.2, different topologies occur depending on the lepton type.

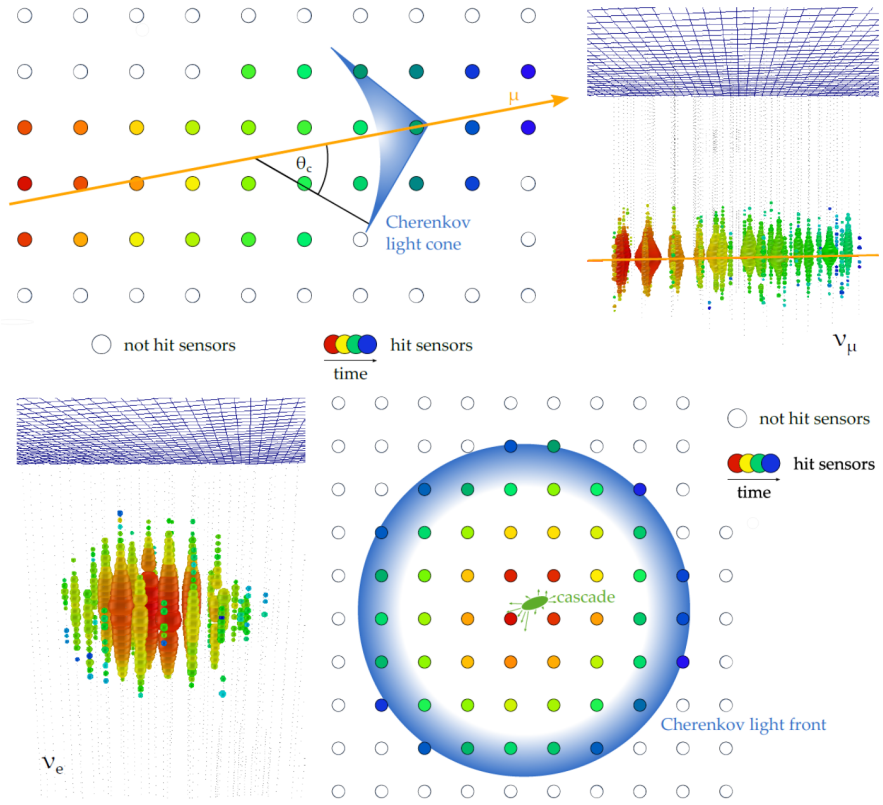

Figure 2.2: The two different topologies of an event as seen by IceCube are illustrated here. The left side represents a track-like event, while the right side shows a cascade-like event. In each event, time progresses from red to blue. The size of the DOMs indicates the number of photons they have detected: the larger the DOM, the greater the number of photons received.

Firstly, the track-like event on the left of Figure 2.2 is triggered by a muon traveling faster than the speed of light in the ice. Muons lose energy through mechanisms such as ionization, pair production, bremsstrahlung, and photo-nuclear interactions. A muon with an energy of 10 TeV can travel 10 km before coming to a stop. The long track left by the muon can precisely indicate the direction of the originating neutrino. However, accurately estimating the neutrino energy is challenging, as the entire track is not always contained within the detector.

Secondly, cascade-like events can occur from CC interactions of ν_e and ν_τ , as

well as neutral current (NC) processes from neutrino of any flavor. A ν_e transfers its energy to an electron and a nucleus, resulting in undifferentiated electromagnetic and hadronic showers, respectively. The shape of the cascade is spherical but non-isotropic as illustrated in Figure 2.2. A ν_τ interaction produces a similar shape due to the tau lepton's short lifetime, except at energies above the PeV range where two isolated showers may appear. Unlike the track-like event determining the direction of ν_e and ν_τ is challenging. However, reconstructing their energy is easier because the entire cascade is included within the detector [12].

2.3 Data process

When a photon enters a DOM in IceCube, it passes into the photo multiplier tube (PMT) and is converted into a photo electron by the photo cathode. This photo electron is then accelerated by the dynodes, causing an avalanche that amplifies the electric signal. If this signal exceeds 25% of the anode voltage level produced by a single photon, digitization begins. Multiple photons may hit the DOM within the signal window due to the finite time (2 ns) and voltage resolution of the DOM.

IceCube, located in an isolated position, relies on satellite connections to transfer data. However, this connection is available only for a few hours per day, allowing for the export of just 800 GB daily. To manage this limitation, IceCube uses a triggering system that limits the digitization of noise and focuses on interesting events. The primary trigger mechanism involves the Hard Local Coincidence (HLC), which requires two neighboring DOMs (either nearest or second nearest neighbors) to exceed the 25% threshold within 1 μs.

The most significant trigger is the Simple Multiplicity Trigger (SMT), especially SMT-3, which is used in this work. SMT-3 requires three HLC hits within a 2.5 μs time interval. Once the raw data is compressed, it is sent to the Northern Hemisphere and converted into a specialized IceCube data format using a modular software framework named IceTray. This system allows researchers to access detailed information about the triggered events, such as the timing of the hit, the location of the DOM that recorded the hit, and the amount of charge generated by the hit [13] and [9].

2.4 Background

The challenge of background within the IceCube detector builds a significant barrier to neutrino research. This encompasses not only events originating from the sky but also noise generated within the detector itself. The following discussion explains the undesired events observed within IceCube.

As mentioned in Section 1.3, the interaction of cosmic rays with the upper atmosphere produces pions and kaons. Most muons are generated from the decay of these pions and kaons. Due to the high energy of cosmic rays, the resulting muons reach relativistic speeds, allowing them, through time dilation, to travel long distances and attain the IceCube detector. Depending on their energy, these muons can penetrate the ice and generate noise, even though the DOMs are deeply embedded. They create track-like events that imitate signals from muons originating from muon neutrinos. Atmospheric muons account for approximately 2.6 kHz to 3.0 kHz of the total event rate.

Besides atmospheric muons, there's a noise originating from processes within the PMT that generates electrons independently of incident photons from the Cherenkov effect. This is known as dark noise.

In the PMT, if a photon hits a dynode without first striking the photocathode, it can prematurely eject a photoelectron, creating what is known as a pre-pulse. This occurrence is rare and generally negligible. Another source of noise is the afterpulse. Electrons accelerated by the dynode can ionize residual gases, and these ions, attracted to the photocathode, can cause the release of additional electrons. This process also generates an avalanche, producing a secondary signal. Pre-pulses and afterpulses contribute about 30 Hz of noise in the DOM.

Thermal emissions occur in the photocathode. They are time-dependent electron emissions that follow Richardson's law and produce a noise at a rate of 10 Hz. Furthermore, radioactive elements present in the PMT and the pressure sphere glass contribute to the noise. Both thermal emission and radioactive decay create a Poissonian noise in the time interval distribution, with an event rate ranging from 230 to 250 Hz.

There is also time-dependent correlated noise. Its exact process is not fully understood. It may be caused by luminescence in the glass sphere triggered by the radioactive decay of potassium and uranium, but it is uncertain. This type of noise occurs at a rate of approximately 280 Hz to 340 Hz [13].

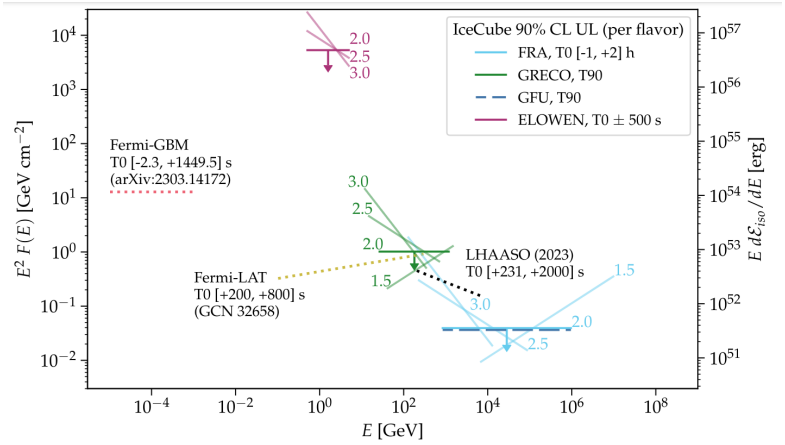

Figure 2.3: IceCube telescope.

2.5 Results

To ensure that neutrinos fully fulfill their role as multi-messenger particles and to match observations with predictions, many astronomy researchers aim to establish a link between neutrino and electromagnetic emissions. Some of these astronomers target the GRB which may be key to this connection. The study [14] focuses on the event of GRB221009A, which recorded the highest light intensity in history over 325 seconds. The study seeks to detect neutrinos within a time window encompassing T_0, the moment when the Gamma-ray Burst Monitor on the Fermi satellite was triggered by the GRB. This is achieved, as described in [14], by combining four analyses that probe potential neutrino emissions from GRB221009A across a range from MeV to PeV. Although no signal was observed, an upper limit on the time-integrated neutrino flux has been established for each analysis, below which the background prevents detection. The analysis of particular interest is the ELOWEN selection, focusing on 0.5-5 GeV neutrinos. It utilizes the eight string of DeepCore to expand sensitivity to lower energy thresholds. As shown in Figure 2.3, the ELOWEN selection has the highest threshold, 3 orders of magnitude higher than the extremely bright burst detected by the satellite Fermi, making it the best candidate for observing a neutrino excess. This highlights a strong motivation for further work in this area.

2.6 Decrease energy threshold

Initially, IceCube was constructed to study TeV neutrinos. As discussed in the Section, it is achievable to expand the sensitivity to lower energy levels using Deep-

Core. This work centers on Elowen (Extremely Low Energy) neutrinos. Section 2.4 highlighted the various backgrounds encountered when examining low-energy events. The ELOWEN selection was designed to address these backgrounds by applying passing conditions on variables based on the position, time, and charge of each hit detected by the DOMs. However, these criteria were the best obtained manually. Consequently, to drop energy threshold down GeV poses a significant challenge. Despite these challenges, more efficient approaches exist. This master thesis proposes the use of machine learning algorithms.

Chapter 3

New approach for low-energy neutrino searches

This chapter describes the machine learning tools employed in the methodology. Section 3.1 explains the difference between supervised and unsupervised learning, and why using supervised tools on our dataset is not advisable. Later, four dimensional reduction techniques will be discussed: PCA, which involves an eigenvector search, and Isomap, t-SNE, and UMAP, which are based on neighbourhood graph methods. Finally, Section 3.3 demonstrates how to apply these dimensional reduction techniques to the ELOWEN selection previously mentioned in Section 2.6.

3.1 Supervised versus Unsupervised learning

Supervised and unsupervised learning are two approaches in artificial intelligence. They differ in how they handle data. Supervised learning uses labeled data to train an algorithm, requiring human influence for labeling. During training, the algorithm iteratively adjusts to make accurate predictions and classifications. Unlike supervised learning, unsupervised learning clusters and analyzes unlabeled data to find hidden patterns. While it operates independently, human validation is needed to validate the discovered patterns [15].

Later, this study will demonstrate that the background simulation in the ELOWEN energy range does not match the IceCube data. Consequently, the data will be approximated as background, since low energy neutrino events are rarely detected. Employing supervised learning tools would label the data as background during training, causing neutrino events to be misclassified as background. This misunderstanding could bias the training process. Therefore, only unsupervised

learning will be utilized.

3.2 Unsupervised learning

Unsupervised learning models allow for the discovery of patterns within data without the need for explicit instructions or labeled data. These models can be classified into clustering, association rules, or dimensional reduction. The latter, in particular, is of interest to this work. It plays a vital role in data science, serving as practical technique for data visualization. Dimensional reduction techniques aim to compress the various features of the dataset into two or three dimensions. These algorithms either aim to preserve the pairwise distance structure between all data samples (maintain global distance), such as PCA, or prioritize preserving local distances over global distances, as seen in methods like t-SNE, Isomap, and UMAP [16]. Throughout this section, the input space or high-dimensional space, will refer to the place where the original data set X resides, the embedding space or low-dimensional space will point to the region of the projected data points Y. The data are denoted individually as x_1, ..., x_n and y_1, ..., y_n, respectively, where n is the size of the sample. As d will designate the dimension of the embedding space and D the dimension of the input space, it is suitable to consider X as $n \times D$ matrix and y as $n \times d$ matrix.

3.2.1 PCA

The method of PCA, or Principal Component Analysis, aims to enhance interpretability while minimizing loss of information. In other words, it transforms correlated variables into new, uncorrelated variables. Let $x'_1, ..., x'_D$ be the columns of X, each one representing a variable, and let $a = (a_1, ..., a_D)$ be a vector. PCA finds the linear combination $\sum_{j=1}^{D} a_j x'_j = Xa$, which maximizes the variance $V_a(Xa) = a^T Sa$, where S is the sample covariance matrix of X which is a $D \times D$ matrix. Therefore, finding this linear combination is equivalent to determine a. It can be shown [17] that a is determined by the eigenvector equation

$$Sa = \lambda a, \tag{3.1}$$

where λ is the eigenvalue of S. Obviously, there is more than one solution to this equation. Actually, in most cases, there are D vectors a and hence, D principal components. PCA can find all the vectors a by centering the data and using singular value decomposition (SVD). Let's define our centered data by $X_c = X - \bar{X}$ where each column i of $\bar{X}$ represents the mean of the components of x'_i. If the centered matrix is assumed to be of rank r, it can then be decomposed as follows

$$X_c = BLA, \tag{3.2}$$

with B an orthonormal matrix $n \times r$, A an orthonormal matrix $r \times D$, and L a diagonal matrix $r \times r$. By definition,

$$(n-1)S = X_c^T X_c \tag{3.3}$$

and by orthogonality of matrices A and B, it can be easily demonstrated that

$$(n-1)S = A^T L^2 A. \tag{3.4}$$

The singular vectors (i.e., the columns of A^T) are therefore the various solutions for a and the singular values (i.e., the diagonal components of L^2) are thus the eigenvalues of $(n-1)S$. The singular vectors a are principal axes with the largest part of variance, the singular values. Consequently, the set of principal components (centralized) is stored in the matrix $X_c A^T$. It represents all the data projected in the D principal axis. PCA sorts itself the axis in descending order of singular value. It then retains the first d axis and projects the centralized data set on them creating a new embedding space :

$$Y = X_c A_d, \tag{3.5}$$

where the matrix A_d is the first d column of the sorted matrix of A. Note that the new variables are derived from the dataset itself, making PCA an adaptive data analysis technique. While this adaptability can be advantageous, it can also introduce issues in the reproducibility of cuts. Additionally, it is noteworthy that the singular vectors indicate how much each variable contributes to form the principal axis, i.e., each singular vector is a linear combination of the variables in the input space. The singular value is also important as it indicates how much the singular vector explained dataset [17].

3.2.2 Isomap

The methodology of Isomap, known as Isometric Mapping, is a fusion between the major algorithmic attributes of PCA that allow computational efficiency and the k-neighbour graphs that enable to preserve intrinsic geometry. Let M be the manifold representing the geometric structure of the data point of the input space X. Isomap typically involves three main stages. Firstly, a graph G is constructed overall data from X by connecting points i and j if i is one of the k-nearest neighbours of j, and associating each pair an edge lengths $d_X(i,j)$. For n points, the matrix $D_G = (d_G(i,j))_{i,j=1,\ldots,n}$ is initialized as

$$\begin{cases} d_G(i,j) = d_X(i,j), & \text{i is a k-nearest neighbour of j} \\ d_G(i,j) = \infty, & \text{otherwise.} \end{cases} \tag{3.6}$$

Then, the algorithm computes the geodesic distances (or shortage distances) between all the points on the manifold M. Geodesic distances are found using only the distance of the k-neighbour graphs defined previously. In other word, the geodesic distance between two points i and j is either $d_X(i,j)$ if they are nearest neighbour, or $d_X(i,k_1) + d_X(k_1,k_2) + ... + d_X(k_n,j)$ if they are far away points. Mathematically, all the components of D_G become, for $k = 1, ..., n$

$$d_G(i,j) = min\{d_G(i,j), d_G(i,k) + d_G(k,j)\}. \tag{3.7}$$

In order to generate a data mapping into a d-dimensional output space Y while preserving the estimated intrinsic geometry of the manifold, a method similar to PCA is used. The goal of this step is to minimize $\|\tau(D_G) - \tau(D_Y)\|_{L^2}$. Here, $\|A\|_{L^2}$ denote the L^2 matrix norm $\sqrt{\sum_{i,j} A_{ij}^2}$ and $D_Y = (\|y_i - y_j\|)_{i,j=1,...,n}$ is the matrix of distance in the embedding space. Let's introduce τ such that

$$\tau(D_G) = \frac{-HSH}{2}. \tag{3.8}$$

where S is the matrix of squared distances $S = (d_G^2(i,j))_{i,j=1,...,n}$ and H is the centering matrix $H = I_n - \frac{1}{n}J_n$, with I_n the $n \times n$ identity matrix and J_n an $n \times n$ matrix of all ones. The τ operator transforms distances to facilitate efficient optimization processes. The problem of minimization becomes an eigenvalue search of the matrix $\tau(D_G)$. Let λ_p be the p-th eigenvalue and v_p^i be the i-th component of the p-th eigenvector [18]. Therefore, the points of the output space are

$$y_i = (\sqrt{\lambda_1}v_1^i, ..., \sqrt{\lambda_d}v_d^i). \tag{3.9}$$

3.2.3 t-SNE

t-SNE combines elements of the Student t-distribution with Stochastic neighbour Embedding (SNE). It was developed as an enhancement to SNE, which faces challenges with its optimization cost function and a phenomenon known as the "crowding problem". t-SNE is much easier to optimize compared to its predecessor, and it provides better visualizations. t-SNE operates by establishing pairwise probability between the points in both the high-dimensional space and the low-dimensional space. Let's start by defining these probabilities. Let $v_{j|i}$ denote the Gaussian distribution based on the Euclidean distance between x_i and x_j :

$$v_{j|i} = \exp\left(-\|x_i - x_j\|_2^2 / 2\sigma_i^2\right). \tag{3.10}$$

Here, σ_i^2 is defined according to a value called perplexity, which corresponds to the number of neighbours around each point i. This value is set by the user in advance. The greater the perplexity, the greater the variance. More details about the calculations of σ_i can be found in [18]. Generally, the perplexity value is between 5 and 50. Above 50, calculations become unnecessarily hard and below 5, information is missing. Moreover, the value must be lower than the sample size. Varying perplexity yields different outcomes.

Note that $j|i$ refers to asymmetric quantities, while ij designates the symmetric one's. $v_{j|i}$ is converted into conditional probability distributions through normalization:

$$p_{j|i} = \frac{v_{j|i}}{\sum_{k \neq i} v_{k|i}}. \tag{3.11}$$

The probability distribution is symmetrized to form the pair-wise probability distribution of the input space:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}. \tag{3.12}$$

These probabilities, which represent the similarities between all points in high-dimensional space, are computed only once in the early stages of the process remaining constant thereafter. Now, let's replicate the procedure for the embedding space. Here, we define w_{ij} using a Student t-distribution with one degree of freedom:

$$w_{ij} = (1 + \|y_i - y_j\|_2^2)^{-1} \tag{3.13}$$

and its corresponding joint probability

$$q_{ij} = \frac{w_{ij}}{\sum_{k \neq l} w_{kl}}. \tag{3.14}$$

Like most algorithms discussed here, t-SNE employs the gradient descent method. Consequently, the positions of points y_1 to y_n and then q_{ij} will be updated in each iteration. t-SNE aims to minimize a cost function that quantifies the Kullback-Leibler divergence, decomposed into constant and non-constant components as following :

$$C = \sum_{i \neq j} p_{ij} \log p_{ij} - p_{ij} \log q_{ij}. \tag{3.15}$$

t-SNE has a useful parameter called random state, which allows to set the seed for the random number generator employed during the initial projection on the low-dimensional space. By setting this random state, the initial positions of the points will always be the same, ensuring the results reproducible. Similarly, the following dimensional reduction, UMAP, also utilizes this parameter.

3.2.4 UMAP

UMAP (Uniform Manifold Approximation and Projection) is built upon a theoretical framework based on Riemannian geometry and algebraic topology. In this dimensional reduction technique, it is assumed that a manifold exists in the input space and the data are uniformly distributed on it. UMAP aims to preserve the topological structure of this manifold during the mapping process. The theory underlying the algorithm can be somewhat complex for readers without a background in topological data analysis and category theory. For the following discussion, UMAP will be explained in a more computational description to facilitate comparison with t-SNE. [16] is recommended for a rigorous mathematical description. As t-SNE, UMAP initialy computes a weighted graph of v_{ij}, except that UMAP chooses k-nearest neighbours to each point that will be considered, instead of an abstract parameter as perplexity. For instance, if $k = 6$ and the point of focus is x_i, only the p_{ij} of the nearest neighbour x_{i_1}, ..., x_{i_6} are computed. UMAP uses non-normalize probabilities

$$v_{j|i} = \exp\left[(-d(x_i, x_j) - \rho_i)/\sigma_i\right], \tag{3.16}$$

with their symmetrical counterparts

$$v_{ij} = (v_{j|i} + v_{i|j}) - v_{j|i}v_{i|j}. \tag{3.17}$$

$d(x_i, x_j)$ is the distance between x_i and x_j. The distance does not necessarily have to be Euclidean, but in our application of UMAP, only the Euclidean distance will be employed. ρ_i is the distance to the nearest neighbour and σ depends on the number of nearest neighbour as well as σ^2 depends on the perplexity with t-SNE. UMAP also uses non-normalize probabilities for low-dimension points

$$w_{ij} = (1 + a\|y_i - y_j\|_2^{2b})^{-1}. \tag{3.18}$$

The optimal values for the parameters are $a \approx 1.929$ and $b \approx 0.7915$. Notably, for $a = 1$ and $b = 1$, w_{ij} for t-SNE is founded. The low dimensional arrangement of points is computed via stochastic gradient descent minimizing the cost function

$$C = \sum_{i \neq j} v_{ij} \log v_{ij} + (1 - v_{ij}) \log(1 - v_{ij}) \tag{3.19}$$

$$-v_{ij} \log w_{ij} - (1 - v_{ij}) \log(1 - w_{ij}). \tag{3.20}$$

Unlike t-SNE, UMAP is faster due to its optimization and has strong mathematical foundations. It has a superior balance between locality and globality in clustering. UMAP excels at reducing a lot of dimensions, as it imposes no computational restrictions on embedding dimensions. Furthermore the algorithm scales remarkably well to larger datasets compared to t-SNE.

3.3 Application to ELOWEN

The ELOWEN selection method allows for the extraction of low-energy neutrinos by applying various cuts on variables characterizing luminosity (number of photons) and causality between hits. However, only sequential cuts have been performed so far. Utilizing PCA provides a combination of variables on which no cuts have yet been applied. This master thesis aims to explore how effective the variable combinations are. Additionally, Isomap, t-SNE and UMAP preserve structure and cluster low-energy neutrinos and background events in different locations, offering an alternative and potentially more effective approach.

Chapter 4

A Novel Methodology

This chapter encompasses the entire methodology used to achieve the results presented in the following chapter. Firstly, the filtering variables and steps before the dataset projections are discussed in Section 4.1. Later, the procedure of the cuts and strategies for achieving optimal cut quality will be explained. Finally, Section 4.3 will detail the approach for encoding geometric parameters and ensuring reproducibility of the results.

4.1 Data processing

The main goal of this master thesis was to distinguish neutrinos in the energy range 0.5 to 5 GeV from the background recorded by IceCube telescope. To achieve this, I had access to simulated data for the neutrinos of all flavours from the GENIE Neutrino Monte Carlo Generator [9], for the dark noise from Monte Carlo that follows an empirical distribution fitted to the data (called Vuvuzela) and for atmospheric muons from COsmic Ray SImulations for KAscade (CORSIKA), and data from the IceCube telescope for the background, available on IceTray. As far as the noise is concerned, since the neutrino rate contained in the real data is negligible, it is appropriate to use the IceCube data as background. The raw data for each event are transformed according to the schema of the different variables. These variables characterize the physics of low-energy neutrinos, and depending on the transformed data, the variables can be used to separate these neutrinos from the rest of the particles detected.

The main distinction between low- and high-energy events lies in the number of photons generated in the ice and therefore, absorbed by the DOMs. Applying passing conditions on these quantities below would eliminate high-energy events, including muons and neutrinos in excess of 5 GeV.

- **Number of Hard Local Coincidence in the DeepCore (#HLC in DC) or outside the DeepCore (#HLC w/o DC)** HLC occurs when two nearest neighbours DOMs or second nearest neighbours DOMs on the same string detect a signal surpassing the threshold within a 1 μs time window, as already defined in Section 2.3.

- **Number of causal connections (#SRT)** The number of SRT is the number of causally correlated hits within an event. It can be considered as the amount of light emitted during the event.

- **Passing filter** Filters are already built by the software (Icetray) of the program to target specific kind of events. IceCube filters identify low-energy neutrinos whereas many filters are designed to recognize high energy muons or cascades. In addition, the Low-up filter tags low-energy neutrinos coming from the Northern sky and the Full Sky Starting filter prevents contamination from incoming muon events, allowing to search for low-energy neutrinos from all directions.

The previous variables enable to remove a large part of high energy event. The following ones allow us to increase the purity of the sample by focusing and the morphology and by normalizing the data and the signal distributions.

- **Total charge** The total charge in an event recorded by the PM of each DOM.

- **Charge ratio** For one event, it is the ratio between the charge collected in the DeepCore 600 ns after the first HLC hit and the total charge deposited in the whole detector.

- **Depth** Depth of the first HLC hit in the DeepCore

- **Distance** Distance between the first and the second HLC hit in the DeepCore.

- **Delay** Time between the first and the second HLC hit in the DeepCore.

For simplicity, before harnessing machine learning tools, the initial step is to establish a passing condition on the variable "Passing filter". As outlined in [9], the methodology exclusively considers events that pass the filter combinations in Table 4.1 and fail to all other filters. The remaining raw data are then transformed to match the variables defined above. The new data is then scaled to ensure that each variable contributes equally to the results produced by unsupervised tools. The data are categorized into "neutrino" and "background" to distinguish events and to quantify parameters defined in the next subsection.

Variable	Combination
Passing filter	DeepCore
	DeepCore + LowUp
	DeepCore + FSS

Table 4.1: Filtering criteria for the passing filter variable. The condition is the success in the combination and the fail in all other filters.

4.2 Cuts

In a conventional data selection approach, it is generally considered judicious to set up filtering criteria based on the variables stated in the previous subsection. The researcher then has to manually select the most appropriate condition. Rather than opting for the conventional method, this subsection presents a new approach for data selection using unsupervised learning techniques. Nevertheless, the ELOWEN selection (classical approach) was still applied to the new dataset for comparing with dimensional reduction techniques. The optimal cut will be determined by the value of the following quantities :

- **Signal to noise ratio (S/NR)** $= \frac{Signal}{\sqrt{Signal+Background}}$
- **Purity** $= \frac{Signal}{Signal+Background}$
- **Purity ratio** $= \frac{Purity\ of\ the\ new\ approach}{Purity\ of\ the\ ELOWEN\ selection}$
- **Low energy neutrino ratio (LENR)** $= \frac{\#low\ energy\ neutrinos\ before\ the\ cut}{\#low\ energy\ neutrino\ after\ the\ cut}$
- **Background ratio (BR)** $= \frac{\#background\ before\ the\ cut}{\#background\ after\ the\ cut}$.

Efficient optimization of these variables involves maximizing for the first four and minimizing for the last one. In practice, two rules were followed during the cut process: maximize the purity while ensuring the LENR remains above the level obtained in the classical approach.

The data is compressed into a set of points in a two-dimensional space where, ideally, neutrinos and noise are optimally separated. The objective is to draw geometric shapes that effectively isolate neutrinos from the background enhancing the value of the five quantities. Figure 4.1 and Figure 4.2 shows two geometric shapes used in this study : the linear shape and the elliptical shape. In both figures, it can be seen that the cluster below the two segments and the cluster outside the ellipse has been eliminated. However, it's important to note that the reverse could also have been possible.

The next step is to assign consistent weights to each low-energy neutrino and to each event of the background. The weight represents the rate of occurrence of the event. Initially, simulated neutrinos are weighted relative to each other, but not adjusted for the background, where each event initially receives a weight of 1. A study has been set up to evaluate the appearance rate of the neutrino compared with background. It has been decided to assign an appearance rate of 0.00181 % after the "Passing filter" selection. The weights are then normalized so that the sum of all the weights of the background events and the simulated neutrinos reach 15 and 0.002715, respectively. After the cut, the weights of the remaining low-energy neutrinos and the background is summed and then the four quantities defined above are computed.

4.3 Additional information

This Section provides more detail about the process and enables the reader to replicate the experiment with the same geometric cuts. As explained in Subsection 4.2, the target of the procedure was to maximize S/NR, purity, purity ratio and LENR, while minimizing BR. However, altering one quantity can affect another. For instance, increasing the purity, is often followed by a drop in the LENR because background data and low-energy neutrinos are difficult to separate. To establish the rules, the former set up passing conditions on the filter variables as described in [9] are used. I applied these passing conditions on the various configurations of data employed during the experiment. Table 4.2 shows the composition of the data and the effectiveness of the passing conditions. For simplicity, each sample S_i consists of the same events. As mentioned in the section above, the objective was to fix a threshold on LENR determined by the value considered in Table 4.2, whereas maximizing the purity.

Another important aspect of the procedure is the construction of the geometric cuts. Some features characterize cuts such as elliptic shape of the geometry or if the data kept are in or out of the ellipse. This information is essential to replicate the results of this research. The following terms describe these geometric cuts :

- **Linear shape :**
  - UP or DOWN if the events kept during the cut are upside or downside from the two lines.
  - x_L, y_L; x_C, y_C; x_R, y_R are the abs and ordinates of the left-side, middle-side and right-side, respectively.
- **elliptical shape :**

Sample	S_1	S_2	S_3
S/NR	2.493×10^{-3}	2.428×10^{-3}	2.342×10^{-3}
Purity	3.201×10^{-3}	3.084×10^{-3}	2.878×10^{-3}
LENR	0.7150	0.7041	0.7018
BR	0.0403	0.0412	0.0440
$\#\nu_e$ 0.5-1 GeV	500	667	1000
$\#\nu_e$ 1-5 GeV	500	667	1000
$\#\nu_\mu$ 0.5-1 GeV	500	667	1000
$\#\nu_\mu$ 1-5 GeV	500	667	1000
$\#\nu_\tau$ 1-5 GeV	1000	1334	2000
# background	3000	4000	6000

Sample	S_4	S_5	S_6
S/NR	2.243×10^{-3}	2.237×10^{-3}	2.223×10^{-3}
Purity	2.935×10^{-3}	2.940×10^{-3}	2.929×10^{-3}
LENR	0.6311	0.6271	0.6215
BR	0.0388	0.0385	0.0383
$\#\nu_e$ 0.5-1 GeV	3000	3000	3000
$\#\nu_e$ 1-5 GeV	30000	50000	210000
$\#\nu_\mu$ 0.5-1 GeV	3000	3000	3000
$\#\nu_\mu$ 1-5 GeV	30000	50000	210000
$\#\nu_\tau$ 1-5 GeV	30000	50000	210000
# background	100000	150000	200000

Table 4.2: The upper part of the table includes the results for the handmade passing conditions of the filter variables. The lower part shows the composition of each sample.

- IN or OUT if the events kept are located inside or outside of the ellipse.
- a and b are respectively the horizontal and vertical semi-axis of the ellipse, while $O = (O_x, O_y)$ represents the center coordinates of the ellipse.

The two lines forming the linear shape must depend on three input points provided by the user, i.e., $\{(x_L, y_L), (x_C, y_C), (x_R, y_R)\}$. The two lines are defined for the slopes

$$\begin{cases} m_1 = \frac{y_C - y_L}{x_C - x_L} \\ m_2 = \frac{y_R - y_C}{x_R - x_C} \end{cases} \tag{4.1}$$

and for $x \in \{x_L, x_R\}$, as

$$\begin{cases} D_1: & y = (x - x_L)m_1 + y_L \\ D_2: & y = (x - x_c)m_2 + y_c. \end{cases} . \tag{4.2}$$

Regarding the elliptical shape, the input parameters are a, b and O. The upper and lower part of the ellipse are respectively defined for $x \in [O_x - a, O_x + a]$ as

$$\begin{cases} E_1: & y = b(1 - (\frac{x - O_x}{a})^2)^{\frac{1}{2}} + O_y \\ E_2: & y = -b(1 - (\frac{x - O_x}{a})^2)^{\frac{1}{2}} + O_y \end{cases} . \tag{4.3}$$

.

The last parameters of the cut for the data points (x_i, y_i), where $i = 1, ..., n$, are simply the condition described by

$$\text{UP} = \begin{cases} \ominus & \text{if } y_i < D_1 \\ \ominus & \text{if } y_i < D_2 \\ \oplus & \text{otherwise} \end{cases} , \tag{4.4}$$

$$\text{DOWN} = \begin{cases} \ominus & \text{if } y_i > D_1 \\ \ominus & \text{if } y_i > D_2 \\ \oplus & \text{otherwise} \end{cases} , \tag{4.5}$$

$$\text{IN} = \begin{cases} \ominus & \text{if } (y_i > E_1) \ \& \ (x_i \in [O_x - a, O_x + a]) \\ \ominus & \text{if } (y_i < E_2) \ \& \ (x_i \in [O_x - a, O_x + a]) \\ \ominus & \text{if } x_i < O_x - a \\ \ominus & \text{if } x_i > O_x + a \\ \oplus & \text{otherwise} \end{cases} , \tag{4.6}$$

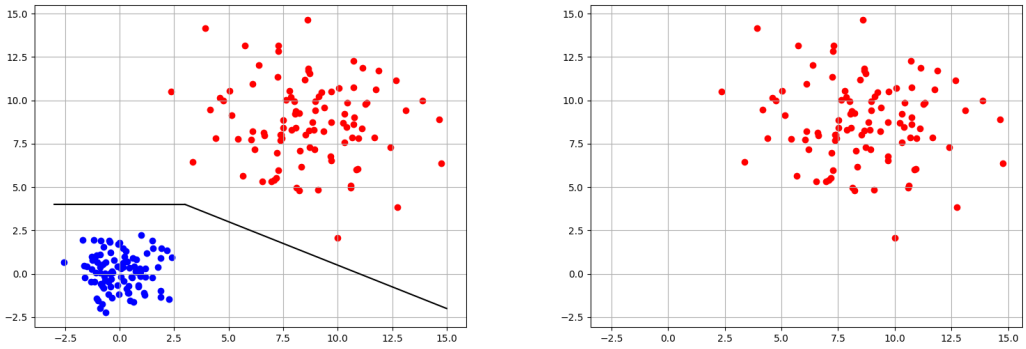

Figure 4.1: Illustration of a cut whose boundary shape is linear. Three points are required as input parameters to define the two line segments shown in the figure on the left. Everything below the two segments is then eliminated in the figure on the right.

$$\text{OUT} = \begin{cases} \oplus & \text{if } (y_i > E_1) \ \& \ (x_i \in [O_x - a, O_x + a]) \\ \oplus & \text{if } (y_i < E_2) \ \& \ (x_i \in [O_x - a, O_x + a]) \\ \oplus & \text{if } x_i < O_x - a \\ \oplus & \text{if } x_i > O_x + a \\ \ominus & \text{otherwise} \end{cases} , \tag{4.7}$$

where $\oplus$ is the symbol to retain the data and $\ominus$ to remove the data. These parameters were encoded in the appendix for each cut performed in this study. Obviously, machine learning tools and their parameter perplexity, number of neighbors and random state, explained in the previous section, are also required to clarify the cut.

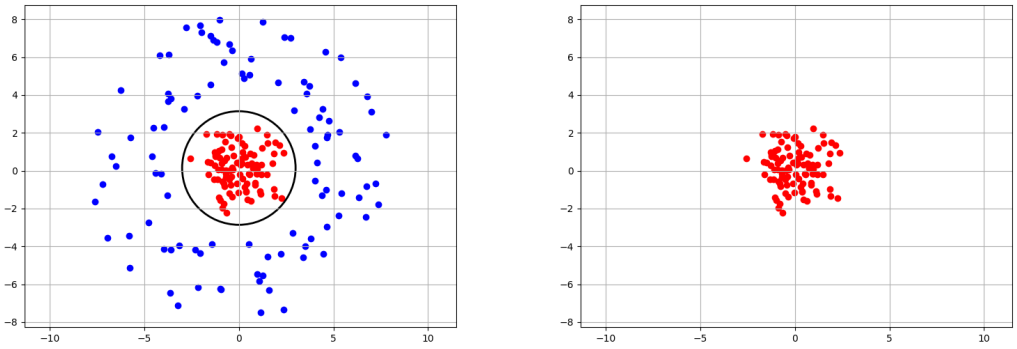

Figure 4.2: Illustration of a cut whose boundary shape is elliptical. The point of origin, the horizontal axis and the vertical axis of the ellipse are required as input parameters to define the ellipse shown in the figure on the left. Everything outside the ellipse is then eliminated in the figure on the right.

Chapter 5

Results

In this study, six types of samples were used as shown in Table 4.2. Each cut associated with an unsupervised learning tool utilized both a large-size and a low-size sample. The sample size depends on the algorithm's capacity to handle a large volume of data. Concerning the k-neighbours graph algorithm (UMAP, Isomap and t-SNE), dataset was trained with varying numbers of neighbours or different perplexity levels in the case of t-SNE.

5.1 Background

Section 4.1 explained that IceCube data was labeled as background on this research. There is a specific reason for this approach. The Figure 5.1 presents the distributions the actual data and the simulated backgrounds: Vuvuzela for DOM noise and Corsika for atmospheric muons. Let's analyze each histogram in detail.

- #**HLC w/o DC** Many Vuvuzela events produce only a single HLC hit outside DeepCore, whereas Corsika events can generate a number of HLC hits outside DeepCore, in a wide range of values.
- #**HLC in DC** The distributions of Corsika and Vuvuzela are almost similar, except that Vuvuzela's distribution is higher in the DeepCore until approximately 11 HLC hits and then, the distribution of Corsica surpasses Vuvuzela. Note that both distributions begin at 3 HLC hits in the DeepCore due to the application of the DeepCore filter to the data.
- #**SRT hits** This graph clearly illustrates that showers generated by atmospheric muons produce events that are interconnected, resulting in a flat distribution that is concentrated on the right. Contrary to DOM noise, which is independent for each DOM, has a distribution with an expectation value of five SRT hits and a steeper slope.

- **Q ratio in DC** On average, a large portion of muon energy is dissipated in the DeepCore, unlike the DOM noise, which is scattered throughout the entire detector.

- **Q total of the event** The simulations appear to match the IceCube data perfectly. This variable depends on the energy of the event, which correlates with the number of SRT hits, as the distributions strongly look alike each other. The Gaussian Vuvuzela distribution and the Poisson Corsika distribution are clearly visible, with expectation values of 45 V and 75 V, respectively. Consequently, it can be concluded that the deposited energy of Vuvuzela is lower than the one of Corsika.

- **Distance btw 1st 2 hits** In both simulations, the distance between the first two hits in the DeepCore is highly concentrated at short distances and then decreases rapidly.

- **Time btw 1st 2 hits** The time between the first two hits in the DeepCore is short for muons, while the DOM noise is almost uniformly distributed in this histogram.

- **Depth 1st hit** The depth of the first hit in the DeepCore is interesting, since no photons are observed between -150 and 100 m, due to the layer of dust in this area. Noise tends to occur below this layer of dust, but muons emit photons in both the upper and lower regions.

It is hard to overlay the two simulations correctly despite properly taken into account weighting when needed. However, it is obvious that the Vuvuzela and Corsika distributions do not match perfectly the background distribution on the various histograms. This strongly reinforces our idea that using the simulations are not able to perfectly reproduce data and should therefore not be used as background for our study.

5.2 Low energy neutrino

This section aims to provide a better understanding of low-energy neutrinos. Figure 5.2 describes the characteristics of low energy neutrino Elowen (0.5-1 GeV) and Lowen (1-5 GeV) across the filter variables. These samples of neutrino were not submitted by any filter selection. Firstly, it is important to note that low-energy neutrinos naturally have more than 2 HLC hits in the DeepCore, supporting the motivation to select SMT3 triggered data in the DeepCore. Furthermore, most events have no HLC hits outside of the DeepCore, as the DOMs are less

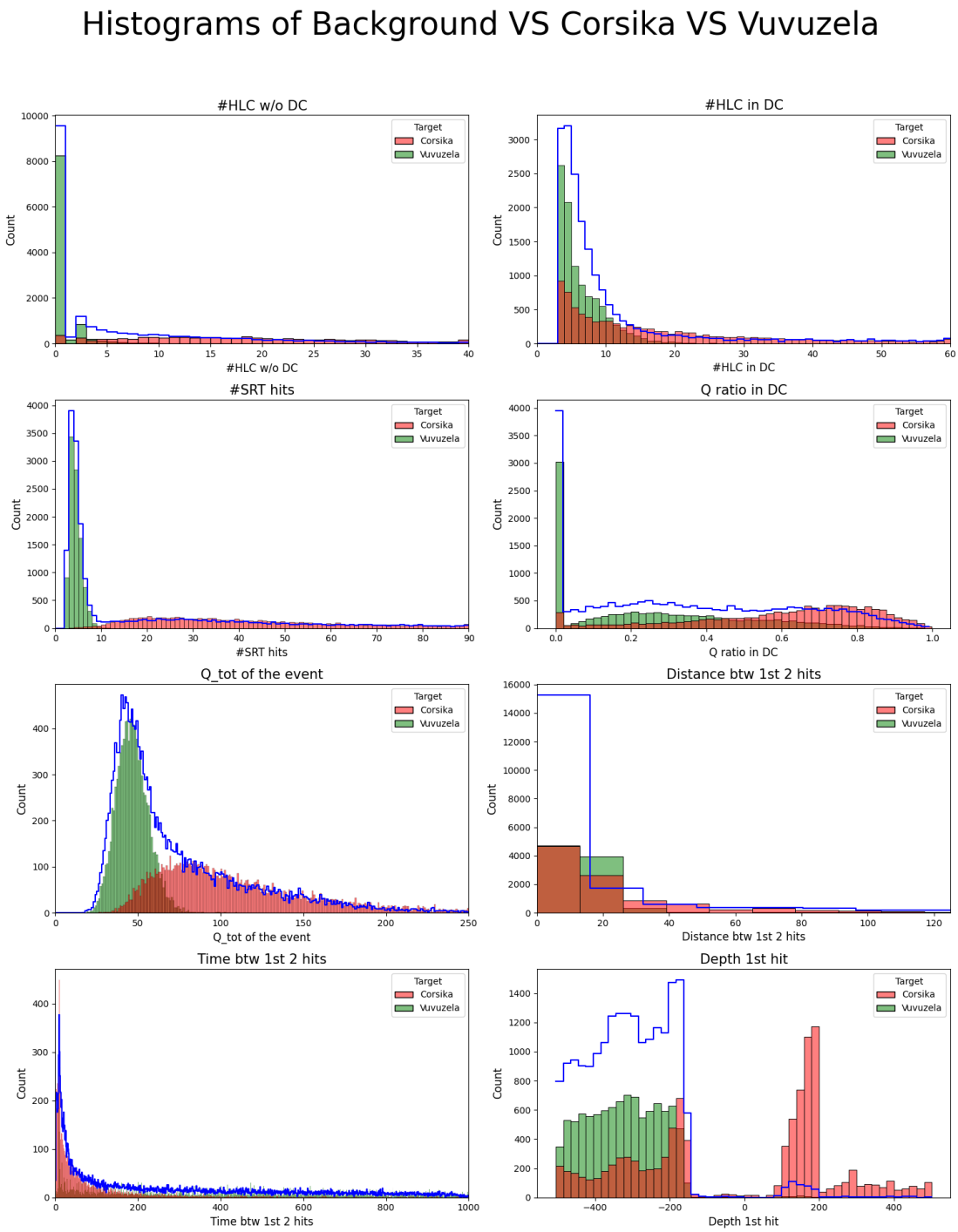

Figure 5.1: Histograms of the 8 filter variables used in the study. The background represents IceCube data, while Vuvuzela and Corsika are simulations representing DOM noise and atmospheric muons, respectively.

concentrated in this region. Additionally, the high q ratio, the short time and the distance between the two first hits in the Deepcore, indicate that event morphology is very concentrated. The low total charge and number of SRT hits further confirm the low energy of these events. Finally, the Elowen et Lowen neutrinos behave quite similarly, although their energy differences.

5.3 PCA analysis

As shown in Table A.1, PCA was applied to two different samples. Comparing the purity ratio, it is evident that the cut quality on S_6 is superior, as unsupervised learning techniques scale better with larger sample sizes. Additionally, PCA provides extra information about its variables in its projections. As introduced in Section 3.2.1, the two principal components are linear combinations of the filter variables. The coefficients of these linear combinations are listed in Table 5.2. It is observed that the principal components change from S_3 to S_6, indicating their dependence on the data set. The table also includes the variance ratio, which shows the contribution of each principal component to the sample's variance. In other words, it indicates how much the principal components explain the distribution of the dataset. The closer the variance ratio is to 1, the more our cut will increase the purity. Here, the two principal components together account for 53% of the variance, which is a good value but could be improved with a 3D cut.

Sample	S_3	S_6
S/NR	$2.221\ 10^{-3}$	$2.356\ \times 10^{-3}$
Purity	$2.587\ \times 10^{-3}$	$3.237\ \times 10^{-3}$
LENR	0.7021	0.6318
BR	0.049	0.0352

Table 5.1: The two results for the dimensional reduction technique PCA (the low size S_3 and the large size S_4) lie in this table. The quality of the cuts is evaluated based on four variables described in the methodology.

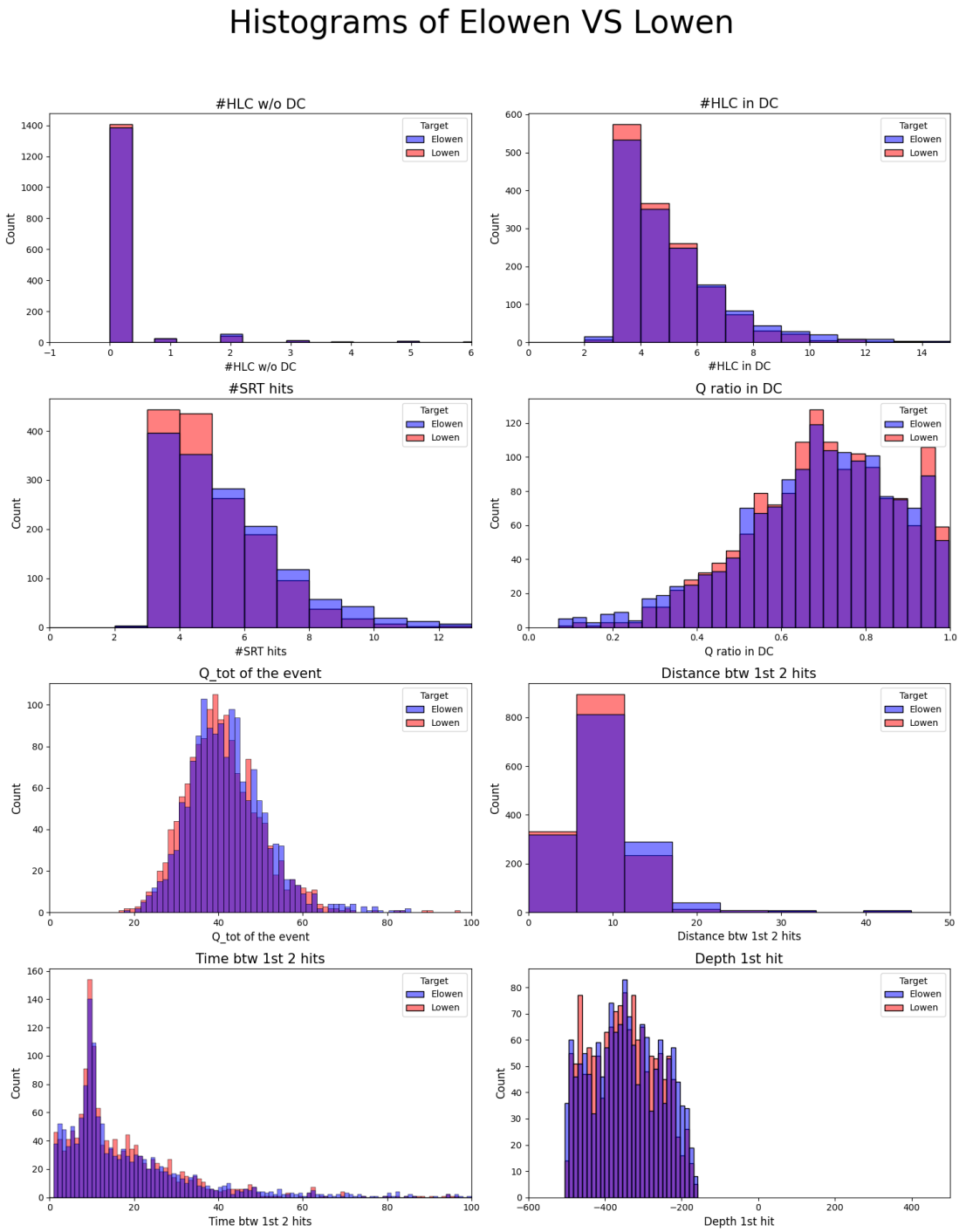

Figure 5.2: Histograms of the 8 filter variables used in the study. Elowen are neutrinos in energy range of 0.5-1 GeV, while Lowen have energy of 1-5 GeV.

Variable	$PC1_{S_3}$	$PC2_{S_3}$	$PC1_{S_6}$	$PC2_{S_6}$
#HLC w/o DC	0.3231	0.5297	0.3391	0.5189
#HLC in DC	0.5252	-0.2264	0.5190	-0.2534
#SRT hits	0.4695	0.2762	0.4670	-0.2833
Q ratio in DC	-0.0512	0.3017	-0.0993	-0.2675
Q_{tot} of the event	0.2702	0.5880	0.2834	0.5813
Distance btw 1st 2 hits	0.0724	0.1238	0.0940	0.1302
Time btw 1st 2 hits	0.0568	0.2941	0.0644	0.3056
Depth 1st hit	0.1974	-0.0062	0.1574	-0.0101
Variance ratio	0.3420	0.1880	0.3457	0.1820

Table 5.2: The 2 principal components of PCA for S_3 and S_6. Their contribution to the variance of sample is represented by the variance ratio.

5.4 Neighbour graph based algorithms analysis

The optimal number of neighbours for k-neighbours graph tools is not straightforward to determine and often requires several tests to identify the best value. In Tables A.2 and A.3, cuts were applied on maps resulting from training with k values of 5, 20, 60 and 100; and perplexity values of 5, 15, 30 and 50. To better visualize the quality of these cuts a a graph comparing the purity ratio is described in Figure 5.3. It is observed that for t-SNE and UMAP, the highest purity is achieved by promoting local structure, with the optimal results found at k = 20 and Perp = 5. Conversely, Isomap favors global structure, yielding the highest purity at k = 100. Additionally, larger sample sizes make it easier to maximize purity compared to smaller sample sizes. This can be attributed to two reasons. First, the size of the sample itself may influence the results, i.e., the three tools scale with the size of the dataset. Second, in larger datasets, neutrinos with energies of 0.5-1 GeV are proportionally fewer than in smaller datasets. As the behavior of 0.5-1 GeV neutrinos is less distinct from the background than 1-5 GeV neutrinos, optimal purity is harder to find.

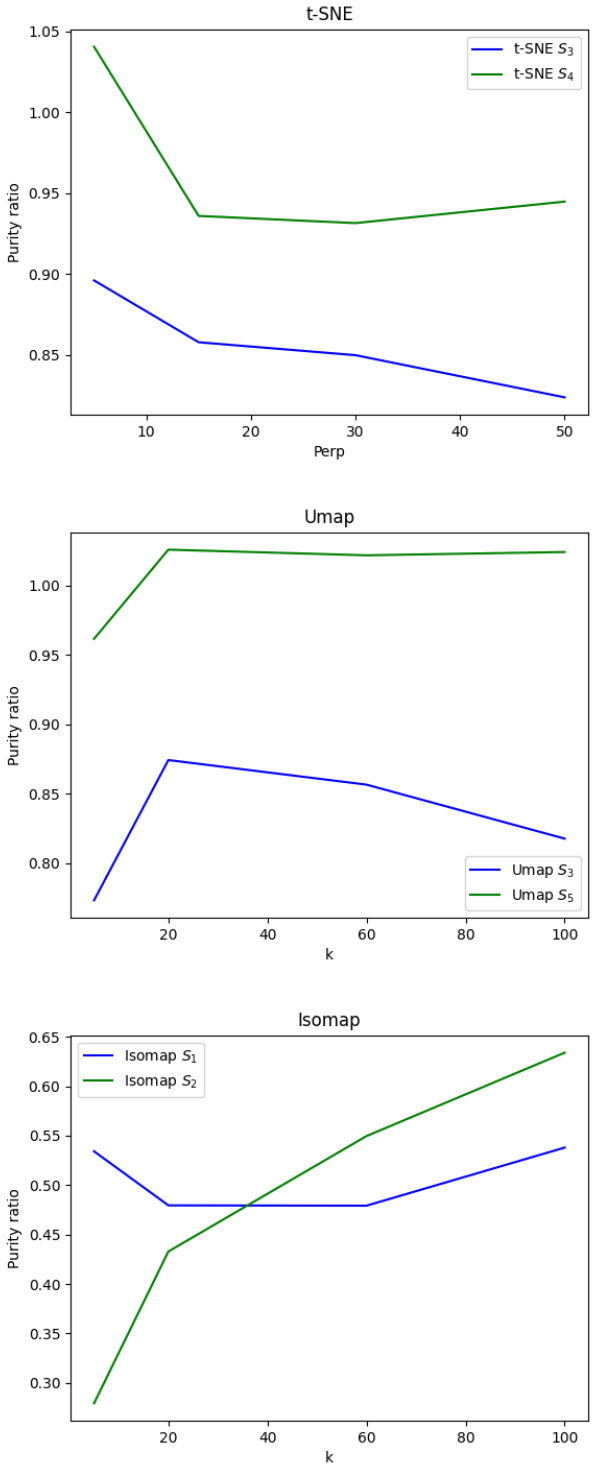

Figure 5.3: Comparison of purity ratio across different number of neighbours k or perplexity level Perp. The purity ratio was computed according to ELOWEN selection in the Table 4.2 and various cuts stored on the appendix. Each figure represents purity ratio for a low-size sample and a larger one.

5.5 Comparison of unsupervised learning algorithms

This section gathers the optimal cuts from each unsupervised learning algorithm to determine which tool best distinguishes low-energy neutrinos from the background. The cuts and the projection of the data in 2D will be analyzed, followed by a discussion on the quality of these cuts based on their purity results.

5.5.1 2D histograms and cuts analysis

The projection of PCA is illustrated in Figure 5.4. The two graphs at the top show the distribution of the background and low-energy neutrinos. The high-density regions of the background and low-energy neutrinos are separable, as indicated by the two dotted lines. In this triangular-shaped distribution, neutrinos are located in a region ranging from -1 to 2 on the x-axis and from -3 to 3 on the y-axis, while the background spreads over a much larger area. Despite the small area retained by the cut, a large number of neutrinos (63%) are preserved, demonstrating how efficiently the algorithm has concentrated them.

Concerning t-SNE in Figure 5.5, the distribution has the shape of a ball. The best result was obtained with a projection that least preserved the local structure, i.e., with a perplexity of 5. The background data is concentrated in the top right corner, while the neutrino data tends to cluster in the bottom left corner. The distribution of the background is wider than the distribution of neutrinos. Note that this is the most dispersed distribution among the four. During the gradient descent process, the distances between points in the 2D space follow a t-distribution, which is flatter than a Gaussian curve. This flatter distribution helps attenuate the Crowding Problem [19], which tends to concentrate data points toward the center.

UMAP was the algorithm that demonstrated optimal global structure preservation as shown in Figure 5.6. Using 20 nearest neighbours ($k = 20$), the distribution of the data follows a complex form that doesn't completely divide the data, but separates a significant part.

Isomap was the most challenging algorithm. As described in Figure 5.7, a high density of neutrinos was clustered at the center, whereas the background was scattered all around. Then, a circular approach was employed to explore different cuts, aiming to retain as many neutrinos as possible while minimizing background contamination. Despite the background and the neutrino could be disentangled as illustrated with the red marker, the results were poor.

5.5.2 Quality of the cuts

As previously mentioned in the methodology, the initial simulation weighted the neutrinos and background such that their respective sums of weights were 15 and 0.002715. This configuration yields a purity of 0.000181 which is interesting to compare with the purity resulting from the algorithms. The outcomes are presented in the Table 5.3. The goal was to maintain a level of retained neutrinos above a specific threshold, defined by the number of neutrinos retained using the Elowen selection for the same sample.

Isomap performed the poorest among the unsupervised learning methods. It struggled with processing large samples, as its kernel crashed when training on tabular data exceeding 10,000 rows, mainly due to its calculation of the geodesic distances between all the points. Isomap was the only tool requiring the implementation of elliptical shape to cut the projected data. Additionally, it achieved the lowest purity result, falling below the purity of ElOWEN selection and could not compete with the performance of the other algorithm employing the same order of sample size, i.e., 12 000 rows.

UMAP achieved better results than Isomap but still had the second poorest performance. Nevertheless, it surpassed the handmade method by 2.6%. UMAP can handle large sample sizes, processing nearly twice the number of rows as t-SNE. Moreover, it scales with the number of variables (columns) used, although it could not be used in this research due to the restriction to height filter variables. Despite, the algorithm is claimed to preserve global distances better than t-SNE, its performance did not exceed t-SNE's for larger data sets.

The purity of t-SNE increased by 4.1% compared to the ELOWEN selection, making it the best nearest neighbours graph algorithm in this study. Although t-SNE was not the best performer overall, it has the potential to be more effective with better adaption of the perplexity parameter. In this research, perplexity values were chosen randomly and the goal was to observe the evolution of purity with changes in perplexity. It is possible that an optimal perplexity value, potentially between 5 and 15, could yield even better results.

Finally, the best method was PCA, which achieved results 10.5% higher than the ELOWEN selection, despite being the simplest method utilizing matrix diagonalization. The purity reached 3.237 $\times 10^{-3}$, more than 17$\times$ the initial purity after the filter selection. However as explained in Section 5.4, this result could be biased by a significantly larger portion of neutrinos of 1-5 GeV than those of 0.5 to 1 GeV energy range neutrinos. Note that another approach could improve

	PCA	t-SNE	Isomap	UMAP
S/NR	2.356×10^{-3}	2.290×10^{-3}	1.939×10^{-3}	2.269×10^{-3}
Purity	3.237×10^{-3}	3.054×10^{-3}	1.955×10^{-3}	3.016×10^{-3}
LENR	0.6318	0.6326	0.7087	0.629
BR	0.0352	0.0374	0.0655	0.03763
Purity ratio	1.105	1.041	0.634	1.026

Table 5.3: Optimal results of each method. The various samples and algorithm parameters employed are as follows : PCA with S_6. t-SNE with S_4 and Perp = 5. Isomap with S_2 and k = 100. UMAP with S_5 and k = 20. The geometric parameters employed can be found in the appendix.

results: instead of limiting to 2 principal components, using 4 or 5 and applying an ordinary pass band filter could enhance performance.

The results of PCA may seem unsatisfactory at first glance, but the dimensional reduction algorithms offer hidden benefits. The cuts made on the projected data provide a potential neutrino signal and a purer background including less low energy neutrino-like events. A supervised algorithm can then train on this refined background and simulated neutrinos to improve the potential neutrino signal. Combining unsupervised and supervised learning is a common practice in the AI field and could lead to more significant improvements.

5.6 Conclusion

The study began with an analysis of data behavior using various filter variables, revealing that the simulated backgrounds (Corsika and Vuvuzela) did not perfectly mimic the actual background. Consequently, IceCube data was used as the background due to its low concentration of low-energy neutrinos. However, since neutrinos are present in this new background, training the data with supervised learning tools could bias the results. Therefore, four dimensional reduction techniques were employed in this research: PCA, t-SNE, Isomap, and UMAP. PCA searches for eigenvectors of the sample covariance matrix, while the other three are based on k-nearest neighbours graph algorithms, requiring adaptation of the number of nearest neighbours. Cuts were made on the 2D projected space of these methods, with PCA yielding the best results. The outcome was a 10.5% increase in purity which is a not significant. However, using unsupervised learning tools, creates an unbiased background and provides an initial signal of neutrinos that can be trained with supervised learning techniques for potentially better results.

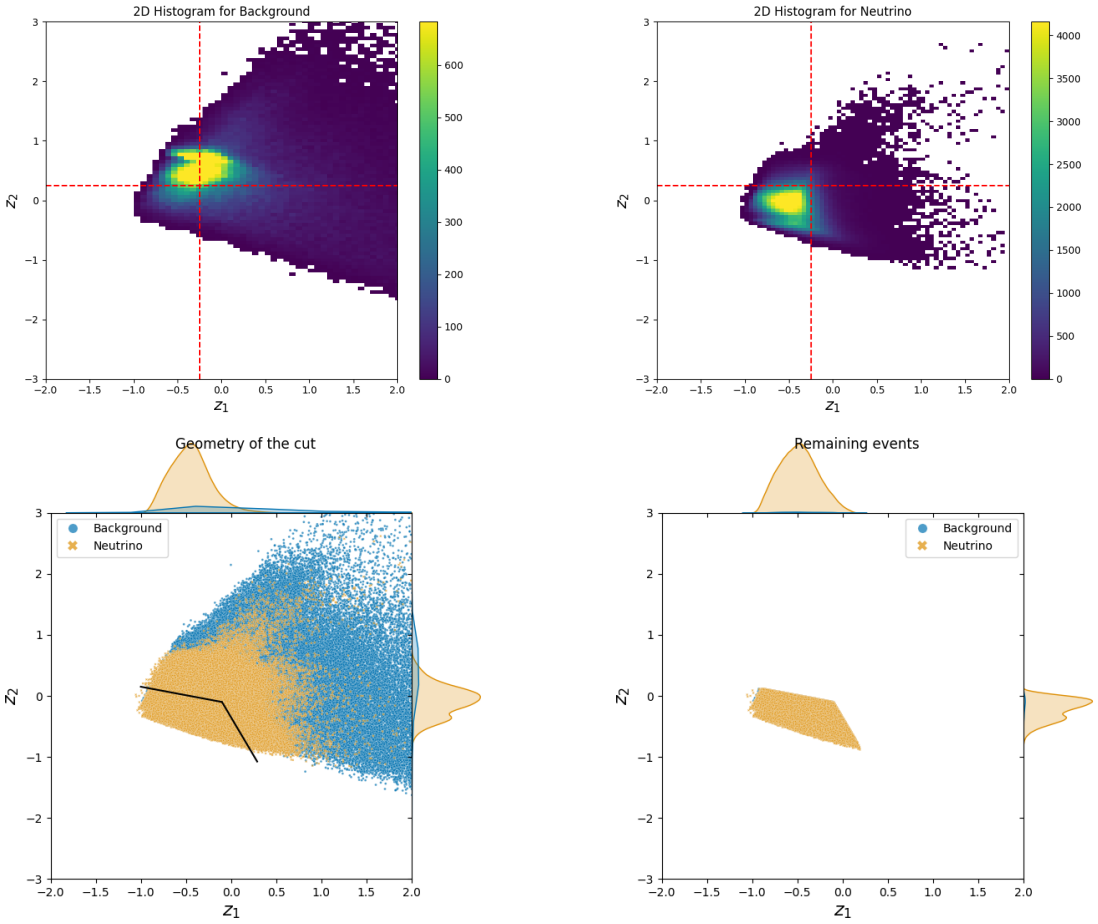

Figure 5.4: Representation of 2D dimensional reduction technique of PCA and its cuts as define in table 5.3. The two graphs on the top illustrate the dataset before and after applying the cuts, with the cutting method shown on the left and the remaining dataset on the right. The bottom graphs describe the density distributions of neutrinos and background data. In all graphs, the axes are labeled as $z_1 = PC1$ and $z_2 = PC2$, the principal components of PCA. The two dotted lines serve as a reference for comparing high neutrino and background densities.

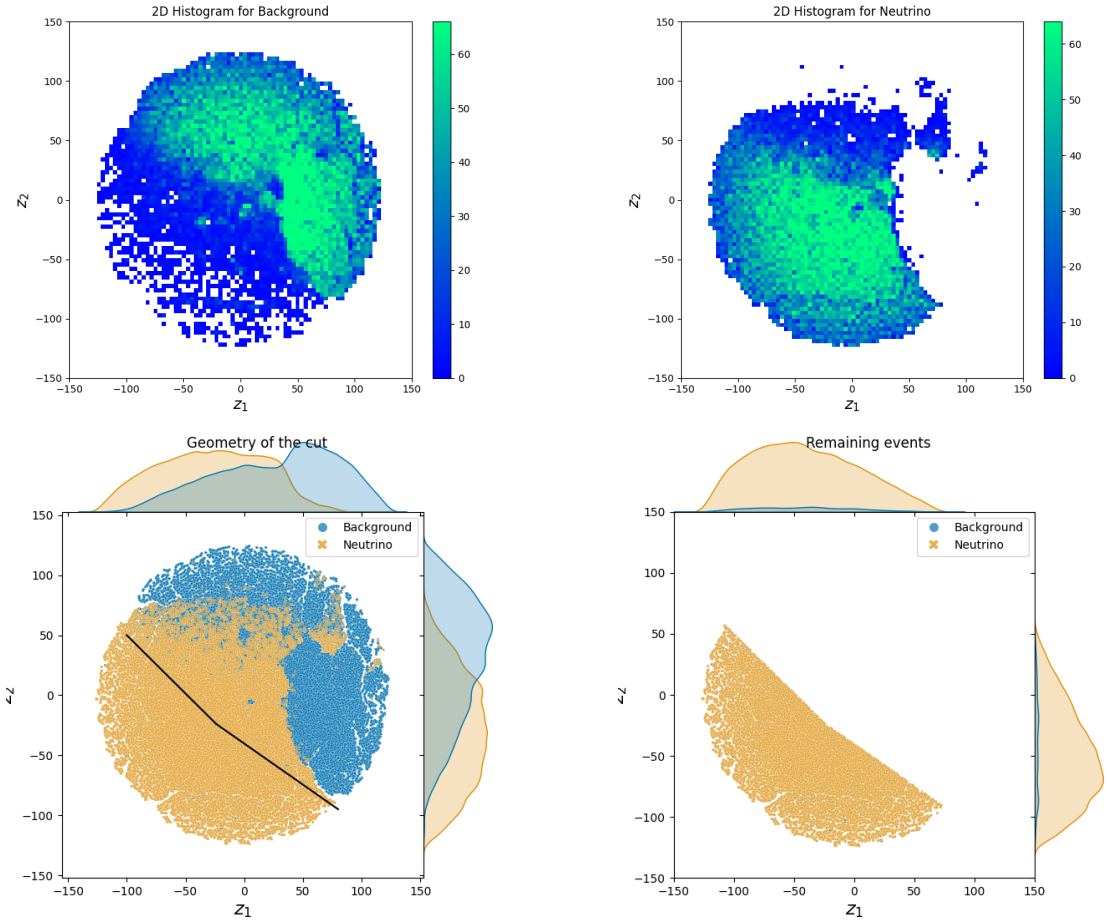

Figure 5.5: Representation of 2D dimensional reduction technique of t-SNE and its cuts as define in table 5.3. The two graphs on the top illustrate the dataset before and after applying the cuts, with the cutting method shown on the left and the remaining dataset on the right. The bottom graphs describe the density distributions of neutrinos and background data.

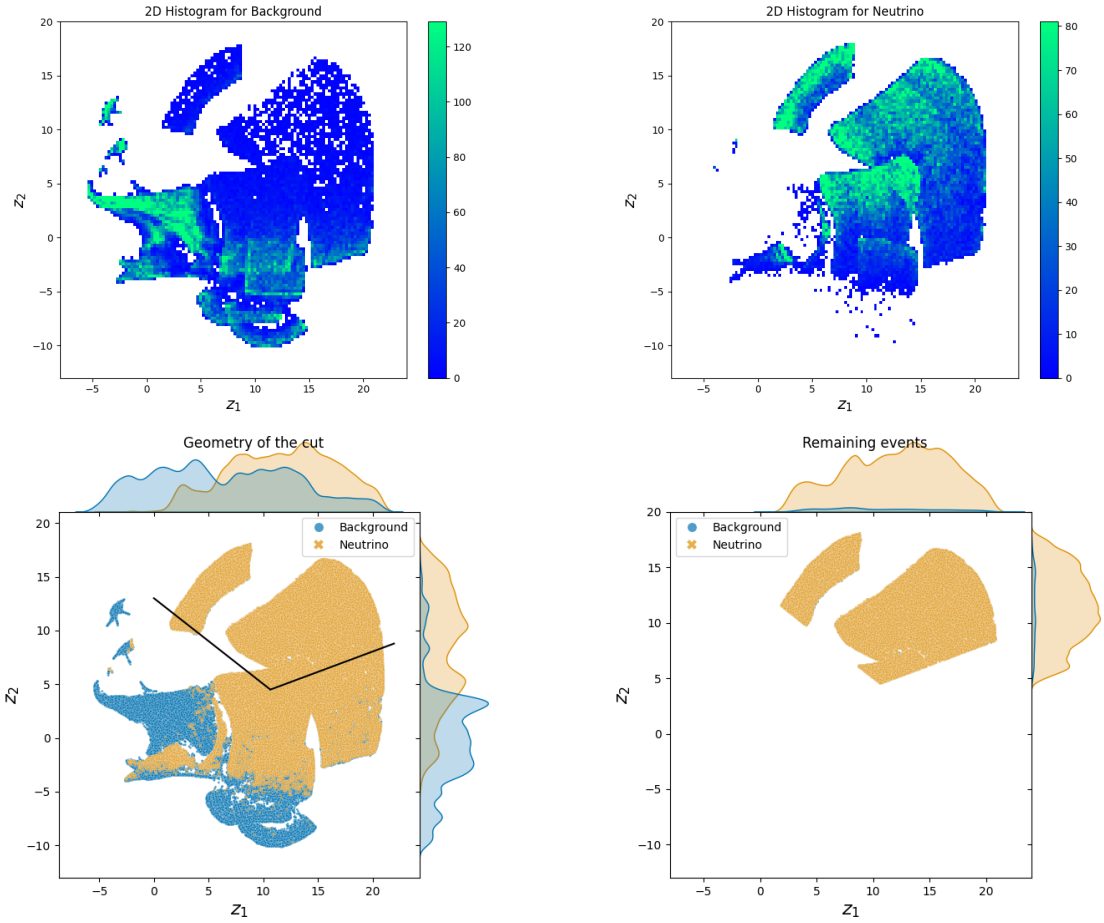

Figure 5.6: Representation of 2D dimensional reduction technique of UMAP and its cuts as define in table 5.3. The two graphs on the top illustrate the dataset before and after applying the cuts, with the cutting method shown on the left and the remaining dataset on the right. The bottom graphs describe the density distributions of neutrinos and background data.

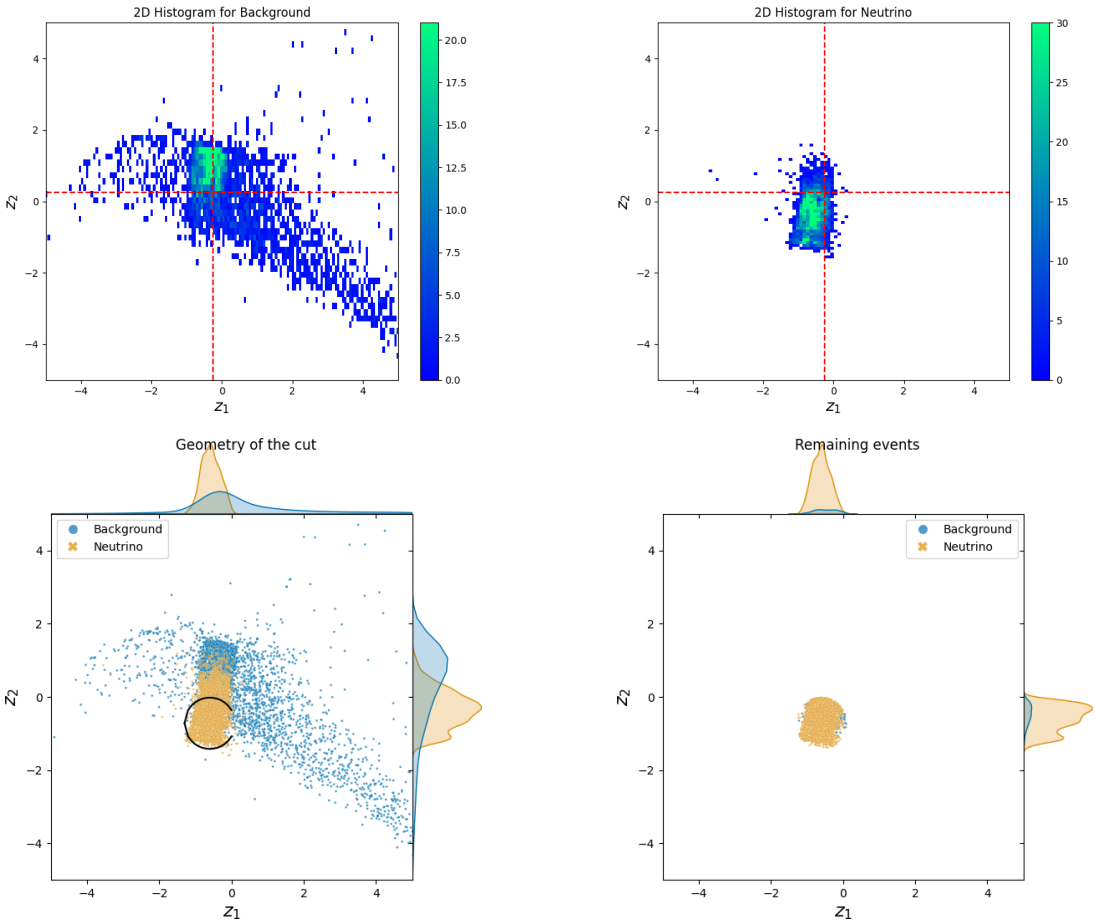

Figure 5.7: Representation of 2D dimensional reduction technique of Isomap and its cuts as define in table 5.3. The two graphs on the top illustrate the dataset before and after applying the cuts, with the cutting method shown on the left and the remaining dataset on the right. The bottom graphs describe the density distributions of neutrinos and background data. The two dotted lines serve as a reference for comparing high neutrino and background densities.

Chapter 6

My contribution

During the months I worked on my thesis I learned a lot. I started by discovering the IceTray software, which is the IceCube core software used to provide a uniform analysis framework for simulation, reconstruction and analysis modules to work within. My work consisted on selecting IceTray data and processing them on the Jupyter Notebook of IceCube server. I explored the potential of various mechanisms of unsupervised learning and used them in an innovative approach, which includes standardisation, dimensionality reduction and cut optimisation. This approach allows to identify promising paths to be used by the IceCube research team in CP3.

Appendix A

Full results

The appendix A store the cut quality variables and geometric parameters for all the cuts made during the research. The outcomes for PCA can be found in the Table A.1, for UMAP and Isomap in the Table A.2, and for t-SNE in Table A.3.

Sample	S_3	S_6
S/NR	$2.221\ 10^{-3}$	$2.356\ \times 10^{-3}$
Purity	$2.587\ \times 10^{-3}$	$3.237\ \times 10^{-3}$
LENR	0.7021	0.6318
BR	0.049	0.0352
(x_L, y_L)	$(-1.3, 0.24)$	$(-1, 0.15)$
(x_C, y_C)	$(-0.469, -0.165)$	$(-0.1, -0.1)$
(x_R, y_R)	$(0.13, -1)$	$(0.3, -1.1)$
Events kept	DOWN	DOWN

Table A.1: The two results for the dimensional reduction technique PCA (the low size S_3 and the large size S_6 define in the Table 4.2) lie in this table. The quality of the cuts is evaluated based on the four variables described in the methodology. The geometric parameters are also included in this table.

		k = 5	k = 20	k = 60	k = 100
Umap S_3	S/NR	2.080×10^{-3}	2.192×10^{-3}	2.162×10^{-3}	2.120×10^{-3}
	Purity	2.225×10^{-3}	2.516×10^{-3}	2.465×10^{-3}	2.353×10^{-3}
	LENR	0.7166	0.7036	0.6985	0.7037
	BR	0.0582	0.0505	0.0512	0.054
	(x_L, y_L)	$(0, 20)$	$(0, 20)$	$(-5, 2)$	$(4, -5)$
	(x_C, y_C)	$(7, 6)$	$(5, 6)$	$(5, 6)$	$(6.34, 5)$
	(x_R, y_R)	$(20, 6)$	$(20, 4.7)$	$(8, 15)$	$(15, 12)$
	Events kept	UP	UP	UP	DOWN
Umap S_5	S/NR	2.194×10^{-3}	2.269×10^{-3}	2.262×10^{-3}	2.268×10^{-3}
	Purity	2.827×10^{-3}	3.016×10^{-3}	3.004×10^{-3}	3.011×10^{-3}
	LENR	0.6272	0.629	0.6276	0.6292
	BR	0.04	0.03763	0.03771	0.0377
	(x_L, y_L)	$(-5, 2)$	$(0, 13)$	$(0, 13)$	$(0, 13)$
	(x_C, y_C)	$(12, 3.5)$	$(10.65, 4.5)$	$(10, 4.7)$	$(10.65, 4.4)$
	(x_R, y_R)	$(17, 1.5)$	$(22, 8.8)$	$(22, 8)$	$(22, 8.8)$
	Events kept	DOWN	UP	UP	UP
Isomap S_1	S/NR	1.811×10^{-3}	1.735×10^{-3}	1.717×10^{-3}	1.812×10^{-3}
	Purity	1.710×10^{-3}	1.535×10^{-3}	1.534×10^{-3}	1.722×10^{-3}
	LENR	0.7065	0.7221	0.7075	0.7021
	BR	0.0747	0.0850	0.0833	0.0737
	a	1	1	0.7	1
	b	0.7	1	0.8	1
	(O_x, O_y)	$(-1.15, 0)$	$(0, -1.2)$	$(0, -1)$	$(-1.28, 0)$
	Events kept	IN	IN	IN	IN
Isomap S_2	S/NR	1.286×10^{-3}	1.600×10^{-3}	1.814×10^{-3}	1.939×10^{-3}
	Purity	8.617×10^{-4}	1.335×10^{-3}	1.695×10^{-3}	1.955×10^{-3}
	LENR	0.7064	0.7069	0.7152	0.7087
	BR	0.1483	0.0957	0.0763	0.0655
	a	1.17	0.9	0.8	0.7
	b	0.55	0.55	0.55	0.7
	(O_x, O_y)	$(-1.2, 0)$	$(0, -0.61)$	$(0, -0.86)$	$(-0.6, -0.72)$
	Events kept	IN	IN	IN	IN

Table A.2: All results for the dimensional reduction techniques Umap and Isomap are summarized in this table. Various numbers of neighbors k and sample sizes, as defined in Table 4.2, were used. The quality of the cuts is evaluated based on the four variables described in the methodology. The geometric parameters are also included in this table. Umap operate with random state 57 for S_5 and 31 for S_3.

		Perp = 5	Perp = 15	Perp = 30	Perp = 50
t-SNE S_3	S/NR	2.218×10^{-3}	2.180×10^{-3}	2.156×10^{-3}	2.127×10^{-3}
	Purity	2.579×10^{-3}	2.469×10^{-3}	2.446×10^{-3}	2.371×10^{-3}
	LENR	0.7025	0.7088	0.7001	0.7026
	BR	0.0492	0.0518	0.0517	0.0535
	(x_L, y_L)	$(-80, 30)$	$(-75, 40)$	$(-90, 40)$	$(-83.5, 50)$
	(x_C, y_C)	$(-20, 0)$	$(-30, -15)$	$(-25, -10)$	$(-24.5, -9.5)$
	(x_R, y_R)	$(75, -100)$	$(75, -60)$	$(75, -65)$	$(75, -55)$
	Events kept	DOWN	DOWN	DOWN	DOWN
t-SNE S_4	S/NR	2.290×10^{-3}	2.174×10^{-3}	2.165×10^{-3}	2.179×10^{-3}
	Purity	3.054×10^{-3}	2.747×10^{-3}	2.734×10^{-3}	2.773×10^{-3}
	LENR	0.6326	0.6335	0.6317	0.6306
	BR	0.0374	0.0416	0.4171	0.04105
	(x_L, y_L)	$(-100, 50)$	$(-120, 60)$	$(-110, 10)$	$(-110, 11)$
	(x_C, y_C)	$(-23.899, -23.898)$	$(-27, -27)$	$(-87, 15)$	$(-87, 11)$
	(x_R, y_R)	$(80, -95)$	$(80, -85)$	$(80, -80)$	$(80, -80)$
	Events kept	DOWN	DOWN	DOWN	DOWN

Table A.3: All results for the dimensional reduction technique, t-SNE, are summarized in this table. Various perplexities Perp and two sample sizes, as defined in Table 4.2, were used. The quality of the cuts is evaluated based on the four variables described in the methodology. The geometric parameters are also included in this table. t-SNE operate with random state 68 for S_4 and 68 for S_3.

Bibliography

[1] Andrea Gallo Rosso, Carlo Mascaretti, Andrea Palladino, Francesco Vissani, *Introduction to neutrino astronomy*, arXiv:1806.06339, 26 Aug 2018.

[2] Xun-Jie Xu, Zhe Wang, Shaomin Chen, *Solar neutrino physics*, arXiv:2209.14832v2, 17 Mar 2023.

[3] Y.Koshio, G.D.Orebi Gann, E.O'Sullivan, and I.Tamborra, *Snowmass 2021 topical group report: Neutrinos from Natural Sources*, arXiv:2209.04298v2, 16 Sep 2022.

[4] Hans-Thomas Janka, *Neutrino Emission from Supernovae*, arXiv:1702.08713, 28 Feb 2017.

[5] D.Liu, J.Chang, S.Chen, H.Dai, C.Feng, B.Gao, G.Gong, M.Gu, F.Li, X.Ma, X.Wang, Z.Wang and X.Zuo on behalf of the LHAASO Collaboration, *Supernova Neutrino Detection with LHAASO-MD*, PoS(ICRC2021)1037.

[6] Edoardo Vitagliano, Irene Tamborra, Georg Raffelt, *Grand Unified Neutrino Spectrum at Earth:Sources and Spectral Components*, arXiv:1910.11878, 10 Dec 2020.

[7] The IceCube-Gen2 Collaboration, *IceCube-Gen2 : The Window to the Extreme Universe*, 10 Aug 2020, arXiv:2008.04323.

[8] J.A.Aguilar, P.Allison, J.J.Beatty, *Design and Sensitivity of the Radio Neutrino Observatory in Greenland (RNO-G)*, 31 Jan 2023, arXiv:2010.12279.

[9] Thesis from Gwenhaël de Wasseige, *Solar Flare Neutrinos in the Multi-Messenger Era : Flux Calculations and a Search with the IceCube Neutrino Observatory*, July 2018.

[10] Albrecht Karle, for the IceCube Collaboration, *IceCube*, PoS(Neutel 2013)053.

[11] D. Tosi, K. Jero, for the IceCube Collaboration, *IceTop as Veto for IceCube*, PoS(ICRC2015)1086.

[12] Thesis from Martin Casier, *Search for High-Energy Neutrino Production in Short Gamma Ray Bursts with the IceCube Neutrino Observatory*, September 2015.

[13] Thesis from David Freiherr Heereman von Zuydtwyck, *HitSpooling : An Improvement for the Supernova Neutrino Detection System in IceCube*, June 1st 2015.

[14] Karlijn Kruiswijk, Bennett Brinson, Rachel Procter-Murphy, Jessie Thwaites, Nora Valtonen-Mattila, for the IceCube Collaboration *IceCube search for neutrinos from GRB 221009A*, 17 Nov 2023, arXiv:2307.16354.

[15] https://www.ibm.com/think/topics/supervised-vs-unsupervised-learning

[16] Leland McInnes, John Healy, James Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, arXiv:1802.03426v3.

[17] Ian T.Jolliffe and Jorge Cadima, *Principal component analysis: a review and recent developments*, doi: 10.1098/rsta.2015.0202.

[18] Joshua B. Tenenbaum, Vin de Silva and John C. Langford, *A Global Geometric Framework for Nonlinear Dimensionality Reduction*, doi: 10.1126/science.290.5500.2319.

[19] Laurens van der Maaten and Geoffrey Hinton, *Visualizing Data using t-SNE*, Journal of Machine Learning Research 9 (2008) 2579-2605.

[20] Klas Hultqvist, *IceCube results*, PoS(NEUTEL2015)053.

[21] Christian Spiering, for the IceCube Collaboration, *IceCube : Astrophysics Results*, PoS(NEUTEL2017)083. —- si utilisé

[22] Dawn R. Williams, for the IceCube Collaboration, *Results from IceCube*, PoS(ICRC2019)016. —- si utilisé

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté des sciences
Place des sciences, 2 bte L6.06.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/sc**