Assessment of few-hits machine learning classification algorithms for low-energy physics in liquid argon detectors

Roberto Moretti [1,2*†], Marco Rossi[3,4,5†], Matteo Biassoni [1], Andrea Giachero [1,2], Michele Grossi [3], Daniele Guffanti [1,2], Danilo Labranca [1,2], Francesco Terranova [1,2], Sofia Vallecorsa [3]

[1*]Dipartimento di Fisica "G. Occhialini", Università di Milano - Bicocca, Milan, I-20126, Italy.
[2]Istituto Nazionale di Fisica Nucleare (INFN), Sezione di Milano-Bicocca, Milan, I-20126, Italy.
[3]European Organization for Nuclear Research (CERN), Geneva, CH-1211, Switzerland.
[4]Dipartimento di Fisica, Università degli Studi di Milano, Milan, I-20133, Italy.
[5]Istituto Nazionale di Fisica Nucleare (INFN) Sezione di Milano, Milan, I-20133, Italy.

*Corresponding author(s). E-mail(s): roberto.moretti@mib.infn.com;
†These authors contributed equally to this work.

Abstract

The physics potential of massive liquid argon TPCs in the low-energy regime is still to be fully reaped because few-hits events encode information that can hardly be exploited by conventional classification algorithms. Machine learning (ML) techniques give their best in these types of classification problems. In this paper, we evaluate their performance against conventional (deterministic) algorithms. We demonstrate that both Convolutional Neural Networks (CNN) and Transformer-Encoder methods outperform deterministic algorithms in one of the most challenging classification problems of low-energy physics (single- versus double-beta events). We discuss the advantages and pitfalls of Transformer-Encoder methods versus CNN and employ these methods to optimize the detector parameters, with an emphasis on the DUNE Phase II detectors ("Module of Opportunity").

Keywords: liquid argon TPC, machine learning, particle identification, Transformer-Encoder

1 Introduction

Liquid argon detectors play a prominent role in neutrino physics and - thanks to experiments like DUNE [1, 2] and DarkSide [3, 4] - will be one of the technologies of choice for the next generation of accelerator neutrino experiments and direct search of dark matter. Such prominence is grounded on scalability. DUNE, in particular, has brought the liquid argon TPC technique (LArTPC) to an unprecedented scale after the development of cryostats that do not need to be evacuated to reach the required purity in TPCs [5]. In turn, this finding allowed the use of commercial membrane cryostats for the DUNE modules [6–8]. Similarly, DarkSide is commissioning high-throughput facilities for depleted underground argon extraction, purification, and distillation [9, 10].

In the last few years and, notably, in the course of the 2021 Snowmass process, several collaborations have formed to fill the gap between LArTPC for beam neutrinos and detectors for rare event searches [11–16]. The first and second DUNE modules were engineered to achieve maximum performance for the observation of GeV-scale neutrinos but a wealth of DUNE physics resides in the observation of MeV events. Some of these channels are being actively pursued by DUNE because the event threshold E is located at $E > 10$ MeV. They are supernova neutrinos, sub-GeV atmospheric neutrinos, and the neutrinos originating from the sun from helium-proton fusion (”hep neutrinos”) [17, 18]. Other channels are currently outside the DUNE scope but may be addressed by the DUNE Phase II modules (third and fourth modules). Recent studies conducted in the framework of the Module of Opportunity (MoO) R&D effort [19] indicate that a module with improved radiopurity, light collection efficiency, and granularity can access a rich physics portfolio: enhance the sensitivity to solar neutrinos produced by the ^{8}B branch of the *pp*-chain, perform neutrinoless double-beta decay searches, boosted dark matter, and the direct detection of WIMP-like dark matter candidates [14, 20, 21].

The possibility of making DUNE the most sensitive neutrinoless double-beta decay experiment in the world is still speculative because of the need for a large ^{136}Xe mass to be dissolved in liquid argon with a few-percent concentration. Moreover, this search is hindered by the presence of radioactive isotopes of argon. Still, results on small prototypes that were doped at the level of 2% are very encouraging [22]. Further, the ProtoDUNE-SP detector at CERN (700 tons of high-purity liquid argon) was doped in 2021 with $\simeq$ 100 ppm and recorded no instability or performance deterioration [23]. The DUNE second module will be doped with (non-enriched) xenon even if the concentration of the double-beta decay isotope ^{136}Xe and the detector radiopurity and granularity cannot address the $^{136}\mathrm{Xe} \rightarrow^{136} \mathrm{Ba}^{*}\ 2e^{-}$ channel (Q-value: ~ 2.458 MeV) in a competitive manner. Several collaborations are addressing these challenges, which require high-throughput facilities for the extraction of underground argon depleted in both ^{39}Ar and ^{42}Ar [13] or the development of dedicated systems for the distillation of atmospheric argon to remove ^{42}Ar [24].

At the time of writing, the main obstacle toward large-mass LArTPCs in the few-MeV energy range is scalability. Charge readout in LArTPCs is carried out by wires or pixels, with pixels having been demonstrated to outperform the more traditional wire-based readout [25, 26]. Given the density of liquid argon, a few-MeV event will range out in ~ 1 cm and will produce just a few hits in the LArTPC. Bringing this number to a level comparable with supernova neutrinos would require a miniaturization of the charge readout system down to the diffusion scale (~ 1 mm) that is either impossible with wires or too expensive with pixels. In addition, the large increase in the number of readout channels impacts front-end electronics, data rate, and trigger complexity, and makes such a miniaturization approach cost-ineffective.

Machine learning (ML)-assisted algorithms have been proven to be the most effective choice to extract information in low-granularity detectors. Machine learning techniques were successfully applied to GeV neutrinos in LArTPC [27–30] and at lower energy in small-size devices [31, 32]. A wealth of novel techniques for supervised and unsupervised machine learning already found applications in particle physics [33] and they are particularly well suited to extract information when it is stored in variables that are non-trivially correlated, often surpassing the performances of non-ML approaches tailored for the task at hand. Few-hits low-energy events are thus an ideal target for these methods, which can compensate for the lack of detector granularity and relieve the burden of designing and operating a large number of channels in a cryogenic underground detector.

This paper addresses such a challenge by identifying a class of machine learning approaches that are suited for low-energy LArTPC. The physics benchmark we employed in our study is the separation capability of one versus two electrons emerging from single and double-beta decay. We chose this benchmark because it represents an important but challenging handle to suppress background events from radioactive argon isotopes in any low-energy physics channel of LArTPCs. β versus $\beta\beta$ separation is also useful to suppress ^{42}K beta decay signals originating from ^{42}Ar in the region-of-interest (ROI) for the neutrinoless double-beta decay of ^{136}Xe. After identifying the optimal machine learning approaches, we show that these techniques relieve the requirement to invest in pixel miniaturization thus reducing the cost and complexity of the next-generation LArTPCs and of MoO.

The main feature of LArTPCs of relevance for this study and the benchmark channels are introduced in Sec.2. Sec.3 presents the feature extraction and classification techniques for few-hits LArTPC events and the rationale of the different methods considered in this study. The performance of the methods against the benchmark channel (β-$\beta\beta$ separation) is discussed in Sec.4. In this section, we also discuss the impact on pixel miniaturization by comparing the overall effectiveness of deep learning algorithms with respect to a system that does not employ ML-assisted β-$\beta\beta$ separation techniques. We draw our conclusions in Sec.5.

2 Few-MeV events in LArTPC

A liquid argon TPC is a cryogenic device that provides the full reconstruction of neutrino interactions in a broad energy range. The detectors that have been developed so far at large scale (> 100 tons) are mostly based on wire anodes. Here, the electrons produced by ionization losses in LAr drift toward the anode driven by a constant electric field of ~ 500 V/cm. The drift velocity at such a field amounts to 1.6 mm/μs and the electrons travel for several meters. In this paper, we will mostly consider the electric field configuration of the first DUNE module (FD1-HD), where electrons drift horizontally ("Horizontal Drift" - HD) between the cathode and the anode for a maximum distance of 3.5 m. FD1-HD is a 65.8 m × 17.8 m × 18.9 m LArTPC segmented in four drift volumes for a total (fiducial) mass of 17 (10) kton. Electron recombination due to electronegative impurities has been addressed by decade-long R&D. The purity of the argon achieved in a LArTPC is usually expressed in terms of the electron lifetime τ. The lifetime is derived from the number of electrons that crosses a drift length $x = v_d t$ and it is given by:

$$N(t) = N_0 \exp -\frac{t}{\tau} = N_0 \exp -\frac{x}{v_d \tau} \tag{1}$$

where N_0 is the number of electrons (and ions) produced by a charged particle in a given volume of the detector, x is the distance from the anode, v_d is the drift velocity and t is the drift time, i.e. the time the electron travels in LAr before reaching the anode. The electron lifetime thus characterizes the electron survival probability against electronegative impurities. In 2014, ICARUS observed a record lifetime of > 15 ms, corresponding to 20 parts per trillion (ppt) of O_2-equivalent contamination using a vacuum-tight cryostat [34]. More recently, the DUNE Collaboration reached an even longer lifetime using a membrane cryostat. Data collected by the FD1-HD demonstrator (ProtoDUNE-SP) indicate a lifetime > 30 ms over a maximum drift length of 3.5 m [8]. Such a bold result boosted the DUNE "Vertical Drift" concept that will be employed for the second DUNE module (FD2-VD). FD2-VD is based on a 10 kton LAr volume whose maximum drift length corresponds to 6 m. In the following, we will test our classification algorithms in a drift volume equivalent to one of the drift volumes of FD1-HD, properly accounting for electron losses due to residual impurities. Thanks to the outstanding purity reached by ProtoDUNE-SP, we anticipate that results hold for a drift length comparable with the maximum drift length of FD2-VD, too.

The readout of ionization electrons at the cathode is generally performed by a set of wires that reconstruct the electron position in the anode plane, that is the plane perpendicular to the drift direction. In FD1-HD, charge reconstruction is performed by a set of 6 m × 2.3 m Anode Plane Assemblies (APAs). Each APA comprises four wire planes and the spatial resolution is dominated by the spacing of the wires inside the plane plus charge diffusion in LAr. FD1-HD employs 152 μm diameter copper-beryllium wires and the wire spacing on each layer is about 4.7 mm, corresponding to a spatial resolution of $\sim 4.7 \text{ mm}/\sqrt{12} = 1.36$ mm. FD2-VD will replace the APA wires with a pair of perforated PCBs, etched with readout strips. The collection strip corresponds to the strip where electrons are stopped and collected and it has a width of 5.1 mm. Induction strips crossed by the electrons before collection have a width of 7.65 mm. As a consequence, the space resolution of FD2-VD is comparable with FD1-HD. A novel readout based on pixels (pixel width: 4 mm) is employed in the DUNE near detector (NDLAr [35]) and is being considered for the third and fourth DUNE far detector modules [15]. Unlike early LArTPCs, all DUNE modules will be operated employing front-end electronics operated at liquid argon temperature (87 K) because cold electronics boards located next to the readout element (wire, strip, or pixel) offer unprecedented noise immunity. The cold electronics of FD1-HD operates with noise well below $800e^-$ per channel [36], corresponding to an energy threshold of < 50 keV. The front-end electronics for the pixelated system are under development and performance is expected to be comparable to or better than FD1-HD.

During the development and assessment of the ML-assisted event classifiers presented in this paper, the performance was estimated as a function of the spatial resolution and energy threshold within the range attainable by the technologies mentioned above.

Low-energy (1-10 MeV) events in liquid argon are recorded as a set of hits associated with an energy deposit per hit. The number of hits depends on the granularity of the LArTPC and, in particular, the size of the readout element (pixel or 2D hits reconstructed by the signals on wires). For a candidate neutrinoless double-beta decay of ^{136}Xe (Q-value: 2.458 MeV) it never exceeds 20 hits even in the most aggressive scenario (pixel size: 1 mm). Solar neutrinos offer a richer topology because the charged current interaction on ^{40}Ar is accompanied by a de-excitation photon from the ν_e $^{40}\mathrm{Ar} \rightarrow^{40} \mathrm{K}^*$ e^- reaction, while electron-neutrino scattering creates a single electron-like track only.

The ML-assisted identification techniques discussed below have been ranked against the most critical benchmark at the MeV scale: the identification of single versus double electrons in a given region-of-interest (ROI) when no other energy deposits are identified as a detached track (or hit) beyond the candidate electron track. The width of the ROI is determined by the energy resolution of the LArTPC, which exploits the total energy deposited estimated from the total collected charge and scintillation light. As a consequence, this information is not used to test the electron hypothesis. Further, we focused on backgrounds where the pulse-shape of the scintillation light cannot be exploited to identify the nature of the observed particle. This is a powerful technique in LAr for $\alpha - \beta$ and n/γ separation using the time profile of the scintillation light [37]. Still, for β particles the identification can only rely on the hit topology and the pattern of ionization (charge) losses per hit. This situation comprises the two most critical backgrounds at the MeV scale, both originating from the radioactive isotopes of natural argon: single beta decay of ^{42}Ar and pile-up events from ^{39}Ar.

The presence of radioactive isotopes in natural argon is considered the most serious drawback of LAr detectors in low-energy physics and, in particular, to search for rare events like WIMP interactions, the occurrence of neutrinoless double-beta decay, and electron-neutrino scattering. Moderate-size LArTPC can be filled with underground argon, which is depleted from radioactive isotopes, but the use of underground argon represents a challenge to the scalability of LArTPC to masses comparable with DUNE [38]. In particular, ^{39}Ar has a quite high natural abundance and contributes to an intrinsic activity of natural argon of 1.01 ± 0.08 Bq/kg [39]. This beta emitter has a lifetime of 269 y. Its Q-value (0.56 MeV) is immaterial for the physics processes considered in this paper (1-10 MeV) except in the occurrence of pile-up ($\beta\beta$ events). The same consideration holds for ^{42}Ar, which contributes with a modest activity (6×10^{-5} Bq/kg) and a sub-MeV Q-value (lifetime: 32.9 y). Unfortunately, the daughter isotope of ^{42}Ar (^{42}K) is a beta emitter in secular equilibrium with ^{42}Ar and has a Q-value of 3.525 MeV. It thus represents the leading background to search for neutrinoless double-beta decay in DUNE [21].

ML-assisted identification algorithms are thus requested to separate single β from double β events exploiting the hit information mentioned above within a given ROI. For the sake of concreteness, we defined the true hypothesis considering the search for neutrinoless double-beta decay in DUNE. The signal is, therefore, the occurrence of two electrons with an energy deposit within an ROI centred at the Q-value of ^{136}Xe. The hypothesis is tested against a single electron, whose energy is located inside the ROI. The comparison is performed among classification algorithms based on machine learning and compared with deterministic algorithms as the *blob* method developed by NEXT to address the same physics channel.

The performance of the classification techniques discussed in this paper is studied using two samples of simulated events: one representing the background (β), consisting of single electrons with energy equal to the ^{136}Xe Q-value, and a second made of ^{136}Xe neutrinoless double β decay events ($\beta\beta$) generated with energy and angular correlations as in [40]. The primary particles (single or double electrons) are then propagated inside a LAr volume using the Geant4 software package [41–43] and the ionization energy loss simulated at each step is used to compute the number of ionization electrons to be propagated to the anode. We account for electron diffusion and recombination in liquid argon as described in Sec.4. For simplicity, we consider in this work a pixel-based readout for it can highlight the impact of the system's spatial resolution on the classification performance more intuitively than a wire-based anode.

The pixel size and the lower limit on their threshold energy significantly affect the quality of β and $\beta\beta$ events spatial reconstruction in a LArTPC. We accounted for these experimental limitations by spatially downsampling the energy deposition profiles in order to match a specific pixel size, removing the ones that don't satisfy the energy threshold requirement.

3 Classification models and feature extraction

In this study, we considered three methods for classifying β and $\beta\beta$ decays by using three-dimensional tracking information extracted from a LArTPC. The first method does not employ machine learning techniques and relies on a physics-informed extraction of highly discriminating features based on *blob* detection, i.e. energy depositions in correspondence with the electron (or positron) trajectory endpoints. A variation of this technique, which we will refer to as "blob method", has already been applied by the NEXT Collaboration [44] for the β-$\beta\beta$ separation in a high-pressure ^{136}Xe gas TPC with satisfactory results for tracks of ~ 15 cm [45, 46]. The second and third methods are Deep Learning architectures called Convolutional Neural Network

(CNN) [47] and Transformer-Encoder, a variant of the Transformer [48]. CNNs are well-known models vastly applied in Computer Vision, while Transformers excel in Natural Language Processing problems thanks to the mechanism of self-attention. The NEXT Collaboration also developed CNN architectures [49] surpassing the blob method for background rejection in the $\beta\beta$ analysis. In the present work, CNN and Transformer solve the task of β-$\beta\beta$ binary classification with different feature processing strategies: the CNN analyses hit positions and energies as a set of pictures, while the Transformer treats hits energies and coordinates as sequences of correlated items. Moreover, the CNN specialises in learning from local features (i.e. pixel structures in a small neighbourhood) while the Transformer, due to its structure, captures both long and short-range dependencies equally [50].
Within this framework, the blob method has been implemented to set a performance benchmark in class separation with respect to the Deep Learning models, as well as to investigate its limits at different granularities of the LArTPC charge readout. In the following, we describe in more detail the characteristics of the blob method and the Neural Network architectures employed in our analysis.

3.1 The *blob* method

Due to the inverse-square velocity dependence of the average ionization energy loss per unit distance [51], ~ 1 MeV electrons release more of their energy when close to their trajectory endpoint, forming a blob. This implies that the double-beta ionization pattern will appear as a unique track in the LArTPC with two endpoint blobs, whereas single-betas only feature one endpoint. We detected blob candidates by finding an appropriate graph representation for each event and localizing the nodes corresponding to the blob position using a Breadth-First Search (BFS) [52]. For a n-hits track, the algorithm works as follows:

1. Assign every track hit to a graph node and connect every node pair corresponding to adjacent hits with edges of unitary weights. Two hits are considered adjacent if they share a surface, an edge or a vertex in the three-dimensional lattice.
2. Perform a BFS search to find the shortest path length between each pair of nodes in the graph i, j for i, $j = 1, \dots n$. Let D be the symmetric $n \times n$ matrix collecting the pairwise path length.
3. Find the indices $i'j'$ such that $D_{i'j'} \geq D_{ij} \; \forall \; i = 1, \dots n \; \wedge \; i < j < n$.
4. Associate the track endpoints (i.e. the candidate blobs centres) to the positions of hits i' and j'.
5. Sum the energies of all hits within a blob radius r from each centre, as depicted in Fig. 1, obtaining the new variables E_{b1} and E_{b2}, where b_1 is the more energetic blob candidate: $E_{b1} > E_{b2}$.

For $\beta\beta$ events, both candidates are expected to be true blobs, hence $E_{b1} \simeq E_{b2}$. On the contrary, for β decays E_{b2} should be significantly smaller than E_{b1}. These two variables allow for establishing a two-feature background rejection criterion. By leveraging the

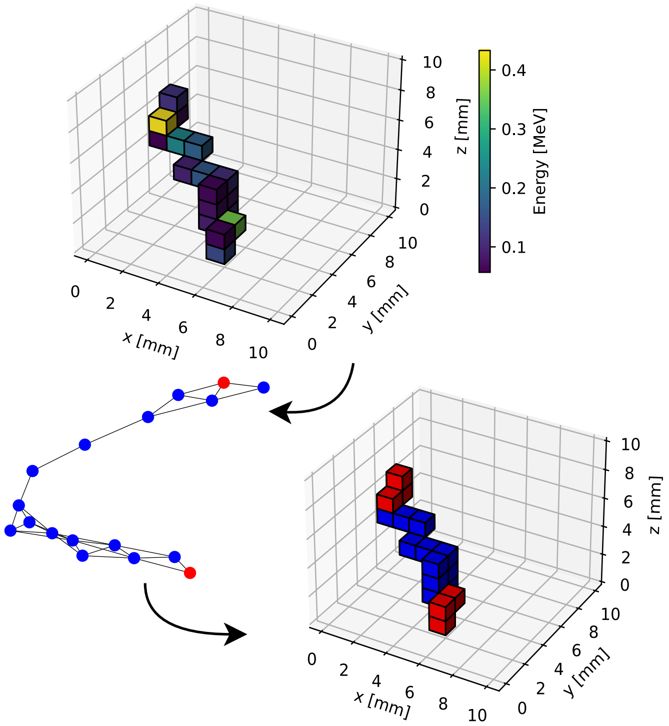


Fig. 1: Workflow of the blob detection algorithm. The three-dimensional profile reconstructed at the LArTPC (top) is transposed into the corresponding graph. The red nodes represent the track endpoints, i.e. the candidate blob position (bottom left). We then integrate hit energies within a radius $r = 2$ mm from the endpoints pair to determine E_{b1} and E_{b2} (bottom right).

Neyman-Pearson lemma [53], we employ the likelihood ratio as the test statistics:

$$q = \frac{P(E_{b1}, E_{b2} \,|\, \beta\beta)}{P(E_{b1}, E_{b2} \,|\, \beta)} \tag{2}$$

where $P(E_{b1}, E_{b2} \,|\, \beta\beta)$ and $P(E_{b1}, E_{b2} \,|\, \beta)$ are the probability distributions for an event with E_{b1}, E_{b2} under the $\beta\beta$ and β hypothesis, respectively. These probabilities are unknown *a priori* and we performed a data-driven estimation by sampling the E_{b1}, E_{b2} distributions in the training dataset. Fig. 2 shows an example of E_{b1} and E_{b2} feature distributions for the β and $\beta\beta$ classes.

Despite its simplicity, the blob method has two drawbacks:

- part of the track information is lost, i.e. only hits near the endpoints deliver feature information.

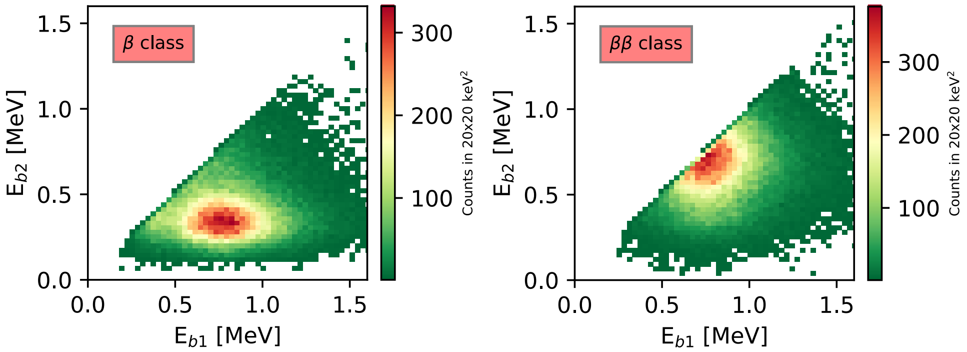


Fig. 2: Blob candidate energy distributions for the β class (top) and the $\beta\beta$ class (bottom). E_{b1} and E_{b2} are extracted from the three-dimensional LArTPC event reconstruction considering a pixel size of $1 \times 1 \times 1$ mm^3 and a hit energy threshold of 50 keV. As expected, the $\beta\beta$ distribution centroid appears closer to the bisector than the β's one, allowing for class separation.

- the BFS fails in determining blob positions if the track reconstructed by the LArTPC presents many gaps, for example due to inefficiencies or trigger requirements, like setting a lower limit energy threshold for the hits, as described in Sec.4.

In addition, this method cannot be generalized to background reduction for other physics channels.

3.2 Convolutional Neural Network

The fundamental building block of CNNs is the convolutional layer, which is a set of back-propagation learnable filters, i.e. tensors which perform a convolution operation on a fixed-size input. Thanks to a stack of convolutional layers, a CNN is able to process hierarchical features that are significant for the learning process [54]. In addition to convolutional layers, CNNs typically include pooling layers, which downsample the feature maps to reduce the computational training cost of the network and limit overfitting. Fully connected layers are then added to the network's end to map the high-level features to the desired output.

In order to define a scalable CNN architecture for the task of track classification in LArTPC when higher readout granularity rapidly increases the input dimension, we embedded the hit energy content into a three-dimensional tensor according to their position in cartesian coordinates, setting all the other entries to zero. We then integrated along the orthogonal axis X, Y, Z to get the three planar views (YZ, XZ and XY planes, respectively). At the cost of a dispensable loss of information, this allows the CNN to support a wide range of readout resolutions with fixed architecture and hyperparameters, with manageable computational costs and resources (only two-dimensional filters are needed), despite employing a large training dataset, containing

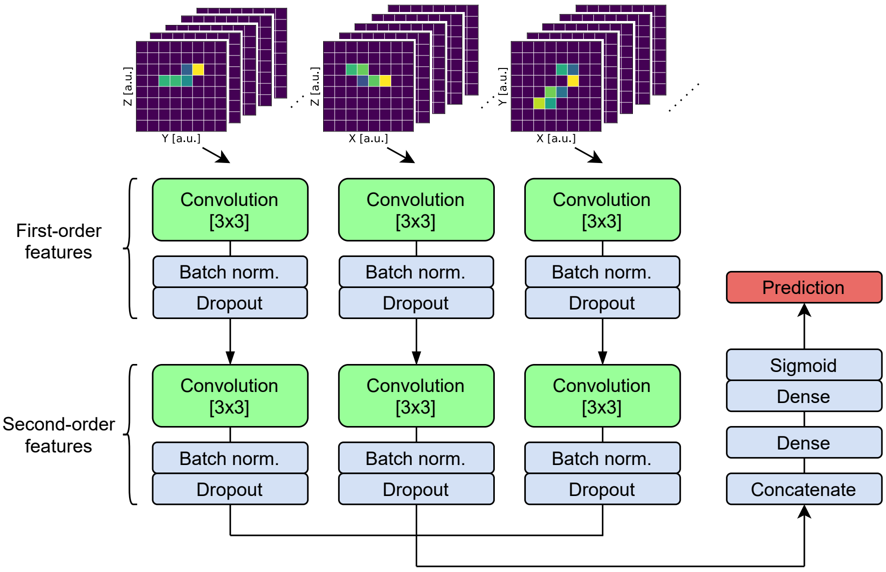


Fig. 3: Convolutional Neural Network scheme. A batch of LArTPC events split into three planar views are fed to two independent stacks of convolutional, batch normalization [55] and dropout [56] layers. The stack outputs merge into a single array, which passes through a fully connected layer. The output layer is a single neuron with a sigmoid activation function, which returns a [0, 1] bounded network predictive score. Each convolutional step comprises 25 filters with [3 × 3] dimensions, and all hidden layers are equipped with LeakyReLU activations [57]. Dropout layers were inserted to prevent overfitting of the model.

about 2×10^5 events. In this physics application, we were also able to remove the pooling layers, which compromise the performance for the low granularity configurations. Fig. 3 illustrates a schematic representation of the CNN architecture we designed for this study.

3.3 Transformer and self-attention

The attention mechanism is the core of Transformers [48], which allows the model to adaptively focus on specific parts of the input sequence, i.e. hits that are supposed to carry more information than others for the classification task. In particular, we refer to the Scaled Dot-Product attention [48] shown in equation 4. Given an ensemble of input sequences, the self-attention mechanism computes a set of query (Q), key (K), and value (V) matrices for each input sequence via fully-connected layers. These matrices are then used to compute a weighted sum of the values, where the weights are determined by the similarity between the query and key matrices. More specifically, the attention weights for a given query matrix are computed as a softmax function [58] over the dot products of the query and key matrices. The resulting weights are then used to compute a weighted sum of the value matrices, resulting in a context vector [59, 60] that captures the most relevant information for the query.

The self-attention mechanism is often used in a multi-head configuration, where multiple sets of query, key, and value matrices are computed in parallel. The resulting context vectors from each head are concatenated and passed through a linear layer, which combines the information from the different heads.

Transformers typically use an encoder-decoder architecture, which consists of an encoder network that processes the input sequence and a decoder network that generates an output sequence. The encoder and decoder both use stacked self-attention layers followed by feedforward layers and are connected by means of an additional attention layer that computes the context vector based on the encoded input sequence.

For binary classification tasks, the network's output simply consists of a single prediction value. For this reason, we employed a simplified architecture consisting only of the encoding part of the Transformer, with a stack of feed-forward layers mapping the encoded state to the output. In this architecture, only self-attention is needed. Self-attention is computed by:

$$\text{Attention(Q, K, V)} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{3}$$

$Q = K = V$ are matrices $\in \mathbf{R}^{d_{\text{model}} \times d_k}$, where d_{model} is the embedding size and d_k the sequence length. The Transformer-Encoder we developed for this work is depicted in Fig. 4. The Transformer-Encoder takes as input the full three-dimensional information. LArTPC events are thus treated as weighted point clouds in which every point corresponds to a hit position in space and the weight is determined by the hit energy. The hit energy and position are fed to the network as a unique array per event with four entries per hit (XYZ coordinates and its energy E). Zero-padding is required for every training batch in order to preserve dimensions through a forward-backwards pass. Compared to the traditional CNN, this Transformer implementation allows for more efficient memory management, with faster training at higher spatial resolutions. This is due to the fact that CNN needs to store progressively larger and sparser two-dimensional tensors and suffers from the increase in dimensionality. The Transformer-Encoder might emerge as a potential rival to tailored approaches such as sparse CNNs, which already overcome the memory issues of CNNs in the classification of GeV-scale events in LArTPCs [29].

4 Dataset and results

Particle ionization losses in liquid argon produce a number of charge carriers (electrons and ions) that is proportional to the deposited energy. The carriers drift across the medium thanks to the TPC electric field. LArTPCs enable three-dimensional tracking by recording the signal induced by ionization electrons as they approach the anode plane. The quality of the event reconstruction depends on the granularity of the read-out system, i.e. the anode wire pitch or pixel size, and the sampling rate of the induced signal. The latter determines the space precision along the drift coordinate Z. Few-MeV tracking is also sensitive to the minimum detectable charge by the front-end

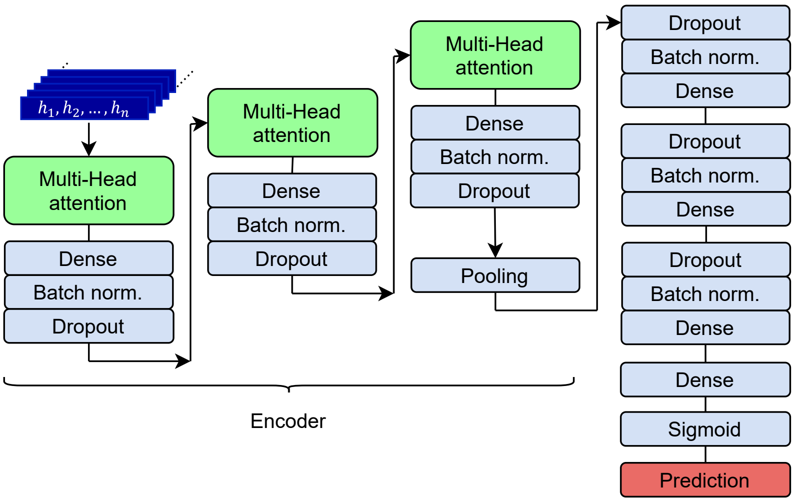


Fig. 4: Transformer-Encoder scheme. LArTPC events consist of a collection of hits ($h_1\, h_2, \dots h_n$), where n can change for different events. Every hit comes with four variables (three space coordinates and the hit energy), and the input vector size is $4m$, where m is the largest hit number for the events in the batch. The encoder consists of three stacks of multi-head attention layers and fully-connected layers followed by batch normalization and dropout. Each multi-head step comprises four parallel self-attention heads. The encoder output is mapped into the final prediction, i.e. a single-neuron layer with a sigmoid activation function. All hidden layers are equipped with LeakyReLU activations.

electronics as discussed in Sec.2. This limitation corresponds to an energy threshold of several tens of keV per hit.

In this work, we considered the same signal sampling rate as DUNE (2 MHz [1]), which corresponds to ~ 1 mm spatial resolution in the drift direction for an electric field of $E = 500$ kV/cm. We then trained the classification models introduced in Sec.3 by varying the pixel size w at different energy thresholds E_t.

The dataset consists of 2×10^5 events, equally split into the β and the $\beta\beta$ classes. Event distances from the readout plane are uniformly distributed between 0 m and the maximum drift length (3.5 m). Electron diffusion was taken into account by applying longitudinal and transversal dispersions according to:

$$\sigma_L = \sqrt{2D_L t} \tag{4}$$

$$\sigma_T = \sqrt{2D_T t} \tag{5}$$

where D_L and D_T are longitudinal and transversal diffusion coefficients estimated for liquid argon from empirical models [61] and t is the drift time. We estimated a maximum standard deviation of $\sigma_L \approx 1.7$ mm and $\sigma_T \approx 2.3$ mm, corresponding to an event occurring at 3.5 m distance from the readout (maximum drift time). As noted in Sec.2, we accounted for electron recombination by assuming an effective electron

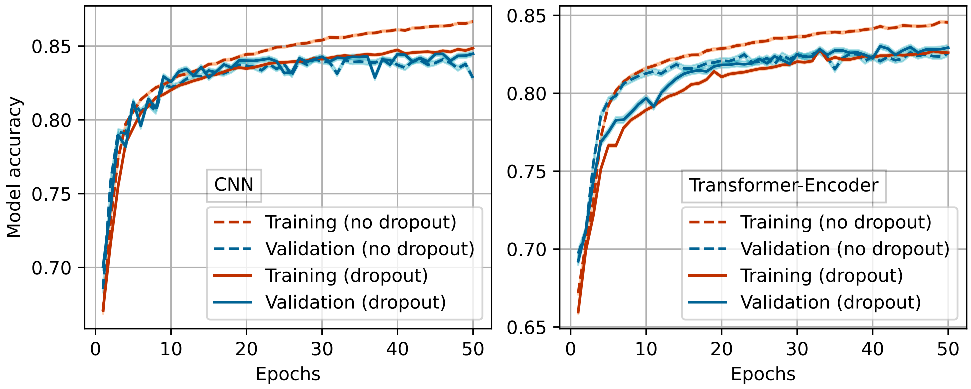
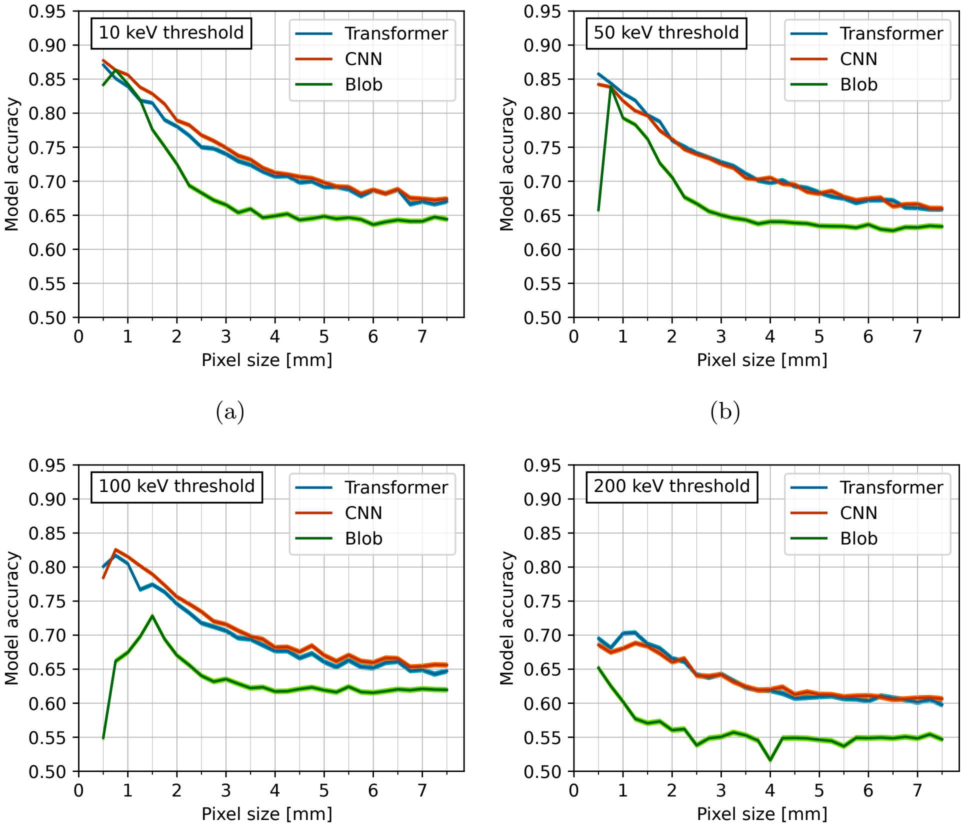


Fig. 5: Learning curves for the Convolutional Neural Network (left) and the Transformer-Encoder (right), considering $w = 1$ mm and $E_t = 50$ keV. The dashed lines show the performance in the absence of dropout layers, while the solid lines include the effect of dropout layers (inserted accordingly to Fig. 3 and 4), with a dropout rate of 0.15 and 0.02 for the CNN and the Transfomer-Encoder, respectively. We observe that in both cases the usage of dropout layers mitigates overfitting without compromising the asymptotical validation accuracy.

lifetime of $\tau = 30$ ms. Note that recombination plays a marginal role in our study since the maximum drift time (~ 2.2 ms) is much smaller than the electron lifetime. This implies that the majority of electrons will reach the anode before undergoing recombination.

For each training at a different pixel size and energy cutoff, we downsampled the MC simulation track information by integrating the energy depositions into hits of dimension $w \times w \times 1$ mm removing hits below the energy threshold. We trained each model by randomly partitioning the dataset into 140000 events for training (70%), 30000 for validation (15%) and 30000 for testing (15%). To establish the model performance, we chose the accuracy metric as the fraction of events correctly classified by the model. For a balanced dataset, i.e. with the same number of samples for each label, the accuracy value ranges between the rates of true $\beta\beta$ event acceptance and the true β event rejection. The CNN and the Trasformer-Encoder training was carried out with the Adam optimizer [62], a variant of the Stochastic Gradient Descent (SGD), with an initial learning rate of 10^{-3}. The learning rate halves every 20 consecutive stall epochs, i.e. epochs in which the validation accuracy does not increase with respect to the previous one. 50 epochs for the Transformer-Encoder and 40 for the CNN ensure a stable convergence of the learning curves.

We note that in the case of no dropout layers, both models exhibit significant overfitting, approximatively after 15-20 epochs, especially at small pixel size (w). Including dropout layers in our architecture is enough to minimise the occurrence of overfitting to negligible levels. We also observe that a dropout rate of 0.15 is needed for the CNN, while for the Transformer-Encoder, overfitting is prevented with a rate as small as 0.02. An example is shown in Fig. 5.



(a) (b)

(c) (d)

Fig. 6: Accuracy of the Transformer-Encoder, the Convolutional Neural Network and the blob method at different pixel sizes w with a 0.25 mm step applying a hit cutoff energy of E_t = 10 keV (a), 50 keV (b), 100 keV (c), 200 keV (d). The error bands account for $\pm 2\sigma$ statistical fluctuations.

Fig. 6 shows the classification accuracies of the blob method, the CNN, and the Transformer-Encoder for the test set at different pixel sizes and energy threshold values of 10, 50, 100, and 200 keV. It is important to notice that the average hit energy content scales with its volume. Since we keep one dimension fixed (as we only vary the pixel size and not the sampling frequency), the hit energies scale with w^2. As a consequence, the energy threshold has a bigger effect at low w.

Overall, the blob model is unable to reach Deep Learning-competitive performances except for a handful of configurations with low-energy cutoffs and small pixel size (w between 0.50 mm and 1.25 mm at 10 keV, w = 0.75 mm at 50 keV). As expected, the blob model performs at its best when tracks are fine-grained and most of the hits pass the energy threshold. The presence of gaps in the LArTPC track reconstruction severely affects the graph connections described in Sec. 3 and compromises the blob candidate localization through the BFS algorithm. ML algorithms, instead, are much more robust to gaps in traces, showing little to no accuracy losses where the blob model fails.

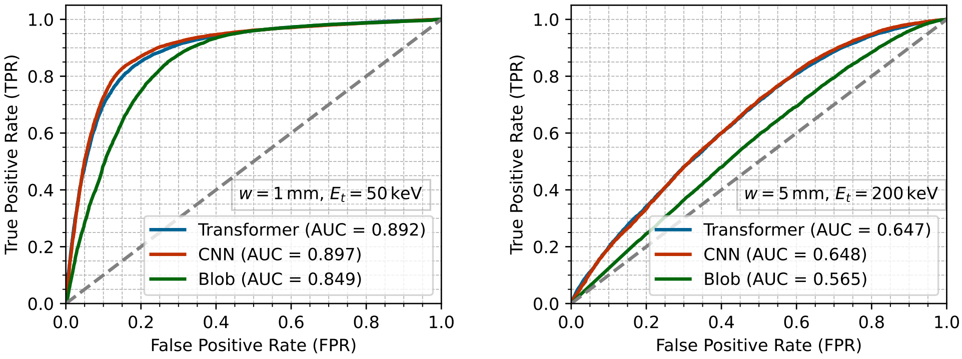
Transformer-Encoder and CNN present a similar trend, especially at intermediate to high pixel sizes. Their behaviour differentiates at the 1-2% level at $w \lesssim 2.5$ mm for $E_t = 10$ keV and $E_t = 100$ keV (CNN outperforms here) and $w \lesssim 1.5$ mm for $E_t = 50$ keV and $E_t = 200$ (Transformer-Encoder performs better). This trend is justified by the fact that more information is available when the number of hits increases and the effectiveness of the Transformer-Encoder learning method results in slightly better accuracy. We expect to gain further improvements in the classification accuracy for both models at small w, employing a larger training dataset and *ad hoc* architecture and hyperparameter optimization for each individual (w, E_t) configuration. For larger values of w, the analysis approaches the one-dimensional limit, as the time-axis alone carries most of the information, narrowing the improvement margin.

The results also emphasize how the accuracy dependence on w flattens as E_t increases. For $E_t = 200$ keV, reducing w from $w = 7.5$ mm to $w = 0.5$ mm – a substantial increase of the LArTPC complexity and cost – the rate of correctly classified events improves by just 10%, going from 60% to 70%.

Fig. 7 shows the Receiver Operating Characteristic (ROC) curves for each of the classification models and the corresponding Area Under Curve (AUC) evaluation metric [63], providing more complete information on two significant granularity-threshold combinations ($w = 5$ mm, $E_t = 200$ keV and $w = 1$ mm, $E_t = 50$ keV) in terms of tradeoffs between signal and background acceptances. The first one (Fig.7a) corresponds to a regime comparable to the ones expected for FD1 HD and FD2 VD module, while the second one (Fig.7b) to an optimal, yet achievable configuration (see Sec.2). In the first scenario, the ML techniques exhibit an overall superior performance with respect to the blob method. However, when setting a working point with high $\beta\beta$ selection efficiency (resulting, in turn, in a lower background rejection rate), all three models demonstrate comparable capabilities. Such a working point is the typical choice when searching for rare events in the presence of a dominant background, as would be the $0\nu\beta\beta$ process. Conversely, in the lower-resolution, high-threshold scenario, the ML models consistently outperform the blob method across all working points.

Machine learning techniques offer additional insight with respect to classic algorithms when detector optimization studies are considered. Both the CNN and Transformer-based classification algorithms point toward the prominence of readout electronics over granularity for sufficiently small values of w. This is an important finding since the increase of granularity in large-volume LArTPC is a major technical challenge. Such an increase does not represent a viable option for $w < 1$ mm due to the increase in the number of channels and the data throughput, while a reduction of E_t at the level of a few tens of keV is well within reach of current technologies.

At $E_t = 100$ keV and lower, the improvement in accuracy achievable by lowering w is prominent and drives the accuracy metric of the β vs $\beta\beta$ classification.



(a) $w = 1$ mm, $E_t = 50$ keV.

(b) $w = 5$ mm, $E_t = 200$ keV.

Fig. 7: ROC curves on a test dataset for Transformer-Encoder, CNN and blob models, displaying the tradeoff between the True Positive Ratio, defined as the fraction of correctly classified $\beta\beta$ events over the totality of them, and the False Positive Rate which corresponds to the fraction of β events erroneously classified as $\beta\beta$. The two panels refer to different resolution-threshold conditions.

5 Conclusions

In this paper, we discussed the performance of two major classes of machine learning algorithms for the identification of low-energy events in liquid argon and compared these findings with the performance of conventional techniques. In particular, we focused our attention on the most challenging classification problem, the discrimination of single β versus $\beta\beta$ events. We thus used as a benchmark the *blob* deterministic method developed by the NEXT Collaboration for the identification of neutrinoless double-beta decay modified to operate in a LArTPC. Both classes of machine learning algorithms - Convolutional Neural Networks and Transformer-Encoder - outperforms the blob algorithm. The CNN and transformer performance are comparable in most of the detector parameter space (see Fig. 6). Overfitting is mitigated in both cases by a dropout layer and is negligible even for small values of the pixel size w. Still, the Transformer-Encoder is more memory-efficient and robust against overfitting even with a dropout rate as small as 0.02. ML-assisted techniques are particularly effective for detector optimization studies since the redefinition of conventional algorithms in a broad detector parameter phase space is very cumbersome. The CNN and Transformer-based classification algorithms point toward the prominence of readout electronics over granularity for sufficiently small values of w. This is an important finding since the increase of granularity in large-volume LArTPC is a major technical challenge, while a reduction of E_t at the level of a few tens of keV is well within reach of current technologies.

Acknowledgments. We gratefully acknowledge the CIEMAT neutrino group and, in particular, C. Cuesta, I. Gil Botella, A. Alvarez Garrote and P. Barham Alzás, for many useful discussions on the DUNE low-energy physics.

Declarations

- Funding: MG and SV are supported by CERN through the CERN QTI. AG acknowledges support by the Horizon 2020 Marie Skłodowska-Curie actions (H2020-MSCA-IF GA No.101027746). This work is supported in part by the Italian MIUR Grant 2017-NAZ-0444 (PRIN2017).
- Conflict of interest/Competing interests: no conflicts to declare.
- Availability of data and materials: the dataset generated and analysed during this study is available upon reasonable request from the author.
- Code availability: the results presented in this manuscript have been produced with the DeepLAr software package available on the GitHub repository: https://github.com/CERN-IT-INNOVATION/DeepLAr

References

[1] Abi, B., *et al.*: Deep Underground Neutrino Experiment (DUNE), Far Detector Technical Design Report, Volume I Introduction to DUNE. JINST **15**(08), 08008 (2020) https://doi.org/10.1088/1748-0221/15/08/T08008 arXiv:2002.02967 [physics.ins-det]

[2] Abi, B., *et al.*: Deep Underground Neutrino Experiment (DUNE), Far Detector Technical Design Report, Volume IV: Far Detector Single-phase Technology. JINST **15**(08), 08010 (2020) https://doi.org/10.1088/1748-0221/15/08/T08010 arXiv:2002.03010 [physics.ins-det]

[3] Aalseth, C.E., *et al.*: DarkSide-20k: A 20 tonne two-phase LAr TPC for direct dark matter detection at LNGS. Eur. Phys. J. Plus **133**, 131 (2018) https://doi.org/10.1140/epjp/i2018-11973-4 arXiv:1707.08145 [physics.ins-det]

[4] Agnes, P., *et al.*: First Results from the DarkSide-50 Dark Matter Experiment at Laboratori Nazionali del Gran Sasso. Phys. Lett. B **743**, 456–466 (2015) https://doi.org/10.1016/j.physletb.2015.03.012 arXiv:1410.0653 [astro-ph.CO]

[5] Adamowski, M., *et al.*: The Liquid Argon Purity Demonstrator. JINST **9**, 07005 (2014) https://doi.org/10.1088/1748-0221/9/07/P07005 arXiv:1403.7236 [physics.ins-det]

[6] Montanari, D., *et al.*: First scientific application of the membrane cryostat technology. AIP Conf. Proc. **1573**(1), 1664–1671 (2015) https://doi.org/10.1063/1.4860907

[7] Montanari, D., Adamowski, M., Hahn, A., Norris, B., Reichenbacher, J., Rucinski, R., Stewart, J., Tope, T.: Performance and Results of the LBNE 35 Ton Membrane Cryostat Prototype. Phys. Procedia **67**, 308–313 (2015) https://doi.org/10.1016/j.phpro.2015.06.092

[8] Abi, B., *et al.*: First results on ProtoDUNE-SP liquid argon time projection chamber performance from a beam test at the CERN Neutrino Platform. JINST **15**(12), 12004 (2020) https://doi.org/10.1088/1748-0221/15/12/P12004 arXiv:2007.06722 [physics.ins-det]

[9] Agnes, P., *et al.*: Separating ^{39}Ar from ^{40}Ar by cryogenic distillation with Aria for dark-matter searches. Eur. Phys. J. C **81**(4), 359 (2021) https://doi.org/10.1140/epjc/s10052-021-09121-9 arXiv:2101.08686 [physics.ins-det]

[10] Alexander, T., et al.: The Low-Radioactivity Underground Argon Workshop: A workshop synopsis (2019). https://doi.org/10.48550/arXiv.1901.10108

[11] Abed Abud, A., et al.: Snowmass Neutrino Frontier: DUNE Physics Summary (2022). https://doi.org/10.48550/arXiv.2203.06100

[12] Borkum, A., et al.: Large Low Background kTon-Scale Liquid Argon Time Projection Chambers (2023). https://doi.org/10.48550/arXiv.2301.11878

[13] Back, H.O., *et al.*: A Facility for Low-Radioactivity Underground Argon. In: 2022 Snowmass Summer Study (2022). https://doi.org/10.48550/arXiv.2203.09734

[14] Avasthi, A., *et al.*: Low Background kTon-Scale Liquid Argon Time Projection Chambers. In: 2022 Snowmass Summer Study (2022). https://doi.org/10.48550/arXiv.2203.08821

[15] Parsa, S., *et al.*: SoLAr: Solar Neutrinos in Liquid Argon. In: 2022 Snowmass Summer Study (2022). https://doi.org/10.48550/arXiv.2203.07501

[16] Caratelli, D., Foreman, W., Friedland, A., Gardiner, S., Gil-Botella, I., al.: Low-Energy Physics in Neutrino LArTPCs (2022). https://doi.org/10.48550/arXiv.2203.00740

[17] Abi, B., *et al.*: Supernova neutrino burst detection with the Deep Underground Neutrino Experiment. Eur. Phys. J. C **81**(5), 423 (2021) https://doi.org/10.1140/epjc/s10052-021-09166-w arXiv:2008.06647 [hep-ex]

[18] Abi, B., et al.: Deep Underground Neutrino Experiment (DUNE), Far Detector Technical Design Report, Volume II: DUNE Physics (2020). https://doi.org/10.48550/arXiv.2002.03005

[19] Gil Botella, I., et al.: DUNE Module of Opportunity Workshop. Valencia, 2-4 Nov 2022, `https://congresos.adeituv.es/dune_science/`

[20] Capozzi, F., Li, S.W., Zhu, G., Beacom, J.F.: DUNE as the Next-Generation Solar Neutrino Experiment. Phys. Rev. Lett. **123**(13), 131803 (2019) https://doi.org/10.1103/PhysRevLett.123.131803 arXiv:1808.08232 [hep-ph]

[21] Mastbaum, A., Psihas, F., Zennamo, J.: Xenon-doped liquid argon TPCs as a neutrinoless double beta decay platform. Phys. Rev. D **106**(9), 092002 (2022) https://doi.org/10.1103/PhysRevD.106.092002 arXiv:2203.14700 [hep-ex]

[22] Campestrini, M., Stringari, P., Arpentinier, P.: Solid–liquid equilibrium prediction for binary mixtures of ar, o2, n2, kr, xe, and ch4 using the lj-slv-eos. Fluid Phase Equilibria **379**, 139–147 (2014)

[23] Gallice, N.: Xenon doping of liquid argon in ProtoDUNE single phase. JINST **17**(01), 01034 (2022) https://doi.org/10.1088/1748-0221/17/01/C01034 arXiv:2111.00347 [physics.ins-det]

[24] Guffanti, D., et al.: Depletion of atmospheric argon for neutrinoless double beta decay searches. In preparation

[25] Adams, C., Tutto, M.D., Asaadi, J., Bernstein, M., al.: Enhancing neutrino event reconstruction with pixel-based 3d readout for liquid argon time projection chambers. J. Instrum. **15**(04), 04009 (2020) https://doi.org/10.1088/1748-0221/15/04/P04009

[26] Kubota, S., Ho, J., McDonald, A.D., Tata, N., Asaadi, J., al.: Enhanced low-energy supernova burst detection in large liquid argon time projection chambers enabled by q-pix. Phys. Rev. D **106**, 032011 (2022) https://doi.org/10.1103/PhysRevD.106.032011

[27] Abi, B., *et al.*: Neutrino interaction classification with a convolutional neural network in the DUNE far detector. Phys. Rev. D **102**(9), 092003 (2020) https://doi.org/10.1103/PhysRevD.102.092003 arXiv:2006.15052 [physics.ins-det]

[28] Acciarri, R., Adams, C., Andreopoulos, C., Asaadi, J., al.: Cosmic ray background removal with deep neural networks in sbnd. Front. Artif. Intell. **4** (2021) https://doi.org/10.3389/frai.2021.649917

[29] Abratenko, P., Alrashed, M., An, R., Anthony, J., Asaadi, J., al.: Semantic segmentation with a sparse convolutional neural network for event reconstruction in microboone. Phys. Rev. D **103**, 052012 (2021) https://doi.org/10.1103/PhysRevD.103.052012

[30] Adams, C., Alrashed, M., An, R., Anthony, J., al.: Deep neural network for pixel-level electromagnetic particle identification in the microboone liquid argon time projection chamber. Phys. Rev. D **99**, 092001 (2019) https://doi.org/10.1103/PhysRevD.99.092001

[31] Buuck, M., Mishra, A., Charles, E., Di Lalla, N., Hitchcock, O.A., Monzani, M.E., Omodei, N., Shutt, T.: Low-energy Electron-track Imaging for a Liquid Argon Time-projection-chamber Telescope Concept Using Probabilistic Deep Learning. Astrophys. J. **942**(2), 77 (2023) https://doi.org/10.3847/1538-4357/aca329 arXiv:2207.07805 [astro-ph.IM]

[32] Acciarri, R., Adams, C., Asaadi, J., Baller, B., Bolton, T., al.: Demonstration of mev-scale physics in liquid argon time projection chambers using argoneut. Phys. Rev. D **99**, 012002 (2019) https://doi.org/10.1103/PhysRevD.99.012002

[33] Albertsson, K., et al.: Machine Learning in High Energy Physics Community White Paper. arXiv (2018). https://doi.org/10.48550/ARXIV.1807.02876

[34] Antonello, M., *et al.*: Experimental observation of an extremely high electron lifetime with the ICARUS-T600 LAr-TPC. JINST **9**(12), 12006 (2014) https://doi.org/10.1088/1748-0221/9/12/P12006 arXiv:1409.5592 [physics.ins-det]

[35] Hewes, V., *et al.*: Deep Underground Neutrino Experiment (DUNE) Near Detector Conceptual Design Report. Instruments **5**(4), 31 (2021) https://doi.org/10.3390/instruments5040031 arXiv:2103.13910 [physics.ins-det]

[36] Adams, D., *et al.*: The ProtoDUNE-SP LArTPC Electronics Production, Commissioning, and Performance. JINST **15**(06), 06017 (2020) https://doi.org/10.1088/1748-0221/15/06/P06017 arXiv:2002.01782 [physics.ins-det]

[37] Boulay, M.G., Hime, A.: Technique for direct detection of weakly interacting massive particles using scintillation time discrimination in liquid argon. Astropart. Phys. **25**, 179–182 (2006) https://doi.org/10.1016/j.astropartphys.2005.12.009

[38] Andringa, S., *et al.*: Low-energy physics in neutrino LArTPCs. J. Phys. G **50**(3), 033001 (2023) https://doi.org/10.1088/1361-6471/acad17

[39] Benetti, P., Calaprice, F., Calligarich, E., Cambiaghi, M., Carbonara, F., al.: Measurement of the specific activity of 39ar in natural argon. NIM-A **574**(1), 83–88 (2007) https://doi.org/10.1016/j.nima.2007.01.106

[40] Ponkratenko, O.A., Tretyak, V.I., Zdesenko, Y.G.: Event generator DECAY4 for simulating double-beta processes and decays of radioactive nuclei. Physics of Atomic Nuclei **63**(7), 1282–1287 (2000) https://doi.org/10.1134/1.855784

[41] Agostinelli, S., *et al.*: Geant4—a simulation toolkit. Nucl. Instrum. Methods Phys. Res. A **506**(3), 250–303 (2003) https://doi.org/10.1016/S0168-9002(03)01368-8

[42] Allison, J., *et al.*: Geant4 developments and applications. IEEE Trans. Nucl. Sci. **53**(1), 270–278 (2006) https://doi.org/10.1109/TNS.2006.869826

[43] Allison, J., *et al.*: Recent developments in geant4. Nucl. Instrum. Methods Phys.

Res. A **835**, 186–225 (2016) https://doi.org/10.1016/j.nima.2016.06.125

[44] Gomez-Cadenas, J.J.: The NEXT experiment. Nuclear and Particle Physics Proceedings **273-275**, 1732–1739 (2016) https://doi.org/10.1016/j.nuclphysbps.2015.09.279

[45] Martín-Albo, J., Vidal, J.M., Ferrario, P., Nebot-Guinot, M., Gómez-Cadenas, J.J., et al.: Sensitivity of NEXT-100 to neutrinoless double beta decay. JHEP **2016**(5) (2016) https://doi.org/10.1007/jhep05(2016)159

[46] Renner, J., et al.: Background rejection in next using deep neural networks. JINST **12** (2017) https://doi.org/10.1088/1748-0221/12/01/T01004

[47] O'Shea, K., Nash, R.: An Introduction to Convolutional Neural Networks. arXiv (2015). https://doi.org/10.48550/arXiv.1511.08458

[48] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. arXiv (2017). https://doi.org/10.48550/arXiv.1706.03762

[49] Kekic, M., Adams, C., al.: Demonstration of background rejection using deep convolutional neural networks in the next experiment. J. High Energ. Phys. **189** (2021) https://doi.org/10.1007/JHEP01(2021)189

[50] Lin, T., Wang, Y., Liu, X., Qiu, X.: A survey of transformers. AI Open **3**, 111–132 (2022) https://doi.org/10.1016/j.aiopen.2022.10.001

[51] Workman, R.L., [Particle Data Group]: Review of Particle Physics. PTEP **2022**, 083–01 (2022) https://doi.org/10.1093/ptep/ptac097

[52] Silvela, J., Portillo, J.: Breadth-first search and its application to image processing problems. IEEE transactions on image processing : a publication of the IEEE Signal Processing Society **10 8**, 1194–9 (2001)

[53] Neyman, J., Pearson, E.S.: On the Problem of the Most Efficient Tests of Statistical Hypotheses. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character **231**, 289–337 (1933)

[54] Bilal, A., Jourabloo, A., Ye, M., Liu, X., Ren, L.: Do convolutional neural networks learn class hierarchy? IEEE Transactions on Visualization and Computer Graphics **24**(1), 152–162 (2018) https://doi.org/10.1109/tvcg.2017.2744683

[55] Ioffe, S., Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv (2015). https://doi.org/10.48550/arXiv.1502.03167

[56] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.:

Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**(56), 1929–1958 (2014)

[57] Xu, B., Wang, N., Chen, T., Li, M.: Empirical Evaluation of Rectified Activations in Convolutional Network. arXiv (2015). https://doi.org/10.48550/arXiv.1505.00853

[58] Kouretas, I., Paliouras, V.: Hardware Implementation of a Softmax-Like Function for Deep Learning. Technologies **8**(3), 46 (2020) https://doi.org/10.3390/technologies8030046

[59] Raffel, C., Ellis, D.P.W.: Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems (2016). https://doi.org/10.48550/arXiv.1512.08756

[60] Brauwers, G., Frasincar, F.: A general survey on attention mechanisms in deep learning. IEEE Trans. Knowl. Data. Eng. **35**(4), 3279–3298 (2023) https://doi.org/10.1109/TKDE.2021.3126456

[61] Li, Y., Tsang, T., Thorn, C., Qian, X., Diwan, M., Joshi, J., Kettell, S., al.: Measurement of longitudinal electron diffusion in liquid argon. Nucl. Instrum. Methods Phys. Res. A **816**, 160–170 (2016) https://doi.org/10.1016/j.nima.2016.01.094

[62] Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization (2017). https://doi.org/10.48550/arXiv.1412.6980

[63] Bradley, A.P.: The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognit **30**(7), 1145–1159 (1997) https://doi.org/10.1016/S0031-3203(96)00142-2