AP-SVM: Unsupervised Data Cleaning for the LEGEND Experiment

Esteban León*
University of North Carolina at Chapel Hill,
Triangle Universities Nuclear Laboratory
esleon97@unc.edu

Julieta Gruszko
University of North Carolina at Chapel Hill,
Triangle Universities Nuclear Laboratory
jgruszko@unc.edu

Aobo Li
University of California San Diego,
Halıcıoğlu Data Science Institute
aol002@ucsd.edu

Brady Bos
University of North Carolina at Chapel Hill,
Triangle Universities Nuclear Laboratory
bradybos@live.unc.edu

Miguel Angel Bahena Schott
University of North Carolina at Chapel Hill
bamsch@email.unc.edu

John F. Wilkerson
University of North Carolina at Chapel Hill,
Triangle Universities Nuclear Laboratory
jfw@unc.edu

Reyco Henning
University of North Carolina at Chapel Hill,
Triangle Universities Nuclear Laboratory
rhenning@unc.edu

Matthew Busch
Duke University,
Triangle Universities Nuclear Laboratory
matthew.busch@duke.edu

Eric L. Martin
Duke University,
Triangle Universities Nuclear Laboratory
eric.l.martin@duke.edu

Guadalupe Duran
University of North Carolina at Chapel Hill,
Triangle Universities Nuclear Laboratory
glduran@email.unc.edu

Jason R. Chapman
University of North Carolina at Chapel Hill,
Triangle Universities Nuclear Laboratory
jaschap@unc.edu

Abstract

The Large Enriched Germanium Experiment for Neutrinoless Double-Beta Decay (LEGEND) will deploy up to 200 kg of High-Purity Germanium (HPGe) detectors in its first stage to search for neutrinoless double-beta decay ($0\nu\beta\beta$). In this study, we present a data-driven approach to remove anomalous events captured by HPGe detectors powered by artificial intelligence (AI). We utilize Affinity Propagation (AP) to cluster signals based on their shape and a Support Vector Machine (SVM) to classify them into different categories. We train, optimize, and test the model on data taken from a HPGe detector installed in a liquid argon test stand. We demonstrate that our model gives maximum physical event sacrifice of $0.016^{+0.005}_{-0.004}\%$ when performing data cleaning cuts. The AP-SVM model can be

*Corresponding Author

Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

applied to classification of unlabeled time-series data from a variety of sources, and is being used to accelerate data cleaning development for LEGEND-200.

1 Introduction

LEGEND [1] is an international, large-scale, and phased experiment that attempts to discover $0\nu\beta\beta$ [2] utilizing HPGe detectors. LEGEND combines the best technologies from the previous Germanium-based experiments, namely, the GErmanium Detector Array (GERDA) [3] and the MAJORANA DEMONSTRATOR (MJD) [4].

Signals captured by HPGe detectors pass through an amplifying electronics chain before being digitized and saved to memory. The digitized time-series signals are also referred to as waveforms. Since LEGEND operates in a low-background environment, a considerable fraction of the recorded data corresponds to non-physical waveforms caused by electronic noise and transient anomalies in the data acquisition (DAQ) system. In order to analyze the data, these anomalous events must be tagged during digital signal processing. This process is referred to as data cleaning.

Traditional data cleaning methods rely on procedures where the scientist must browse through a comprehensive sample of the data to find all the existing types of anomalous events. The scientist must then develop parameters that capture solely non-physical events, and perform cuts based on these parameters to tag anomalous events. These parameters can vary over time and with detector type. LEGEND-200 will run for five years utilizing four detector types of different geometries. Additionally, different run conditions and hardware setups, including those used in detector characterization [5, 6] and other test stands [7], affect the performance and stability of these parameters as well. Thus, data cleaning with traditional procedures requires a significant amount of time and human effort.

Consequently, we have devised a data cleaning mechanism based on AI. AI has proven to be a successful tool for Ge-based experiments in the search for $0\nu\beta\beta$ [8–10]. Our model consists on a clustering with user intervention + classification scheme that distinguishes between physical and anomalous waveforms. With this method, we can tag different types of signals in a data-driven and semi-autonomous manner.

Our model has been applied to different run configurations and setups containing HPGe detectors that use varying acquisition electronics, proving its flexibility. For this study, we evaluate the performance of our model on data taken from from a HPGe detector installed in a local liquid argon test stand.

2 Model summary and training

The AI-powered data cleaning mechanism we have developed consists of three steps: (1) extract pulse shape information from waveforms, (2) group similar waveforms with an unsupervised learning model and re-label them based on user input, (3) extend clustering to larger datasets with a supervised learning model. For the first step, we utilize a Discrete Wavelet Transform (DWT). The DWT decomposes the waveform into mutually orthogonal down-sampled sets of time-series coefficients by convolving the input signal with a given type of wavelet. In this case, Haar wavelets are used. The DWT can be performed multiple times on the same input signal, resulting in a multilevel decomposition with a down-sampling factor of 2^l, where l is the level or number of decompositions to be performed. Thus, the DWT serves to de-noise and reduce the dimension of the waveform. Since our waveforms contain 8,192 features, we use a value of $l = 5$ to reduce the dimension to 256 features. We take the resulting approximate coefficients as a lower-dimensional representation of the input waveform and normalize them by the absolute value of their maximum or minimum amplitude, whichever one is greater, such that every signal lies between the $[-1, 1]$ range.

In the second step, we group and label waveforms using AP. AP is an unsupervised learning algorithm that clusters inputs based on a message-passing method between data points [11]. The number of clusters is autonomously computed by AP, which allows new clusters to appear automatically in the data over time with varying run conditions and experimental setups. We choose AP over DBSCAN since our data contains clusters of varying densities, as discussed in Appendix A. Each cluster center found by AP is referred to as an "exemplar." We utilize a representative dataset containing $N = 10,000$ waveforms for training. The hyperparameters we optimize for AP are the preference p and the damping factor λ. p is the overridden value of the diagonal of the similarities matrix

S, which stores pairwise distances between input waveforms. λ is an algorithmic convergence hyperparameter that reduces the influence of the previous iteration's values on the next iteration's. We perform a search over 100 grid points spanning $\lambda \in [0.85, 0.99]$ and $p \in [\min(S), -100]$ to find the hyperparameter combination that gives the closest to 100 exemplars. We found that obtaining ~ 100 exemplars yields an overarching representation of our data, and it allows the user to "hand-label" each exemplar according to data cleaning categories of Figure 1, as shown in Figure 2. Every grid point of the search uses $\sim$ 12 GB of RAM, requiring the code to run on multiple CPU cores in parallel. One of the shortcomings of AP is that, once trained, it cannot be applied to data outside the training dataset. Therefore, we must train a separate classifier to extend the clustering power of AP to larger datasets.

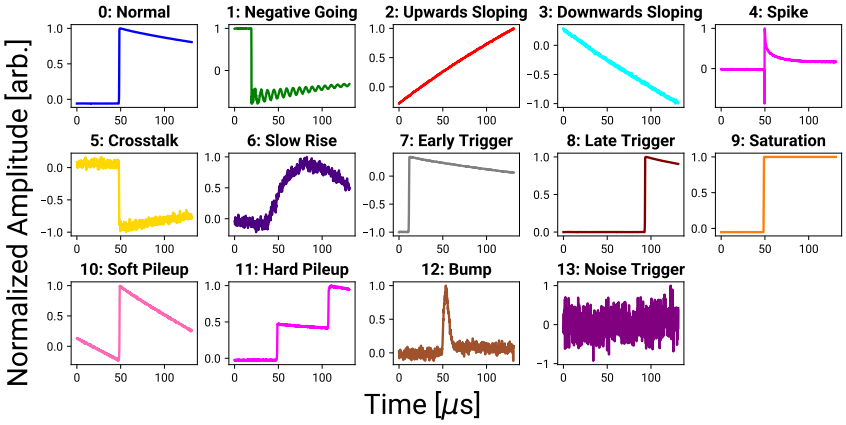


Figure 1: A comprehensive set of data cleaning categories corresponding to different types of physical and anomalous waveforms captured by HPGe detectors. Each category is assigned a unique tag ID running from 0 to 13. If AP finds a new population during training, it can be added to this set

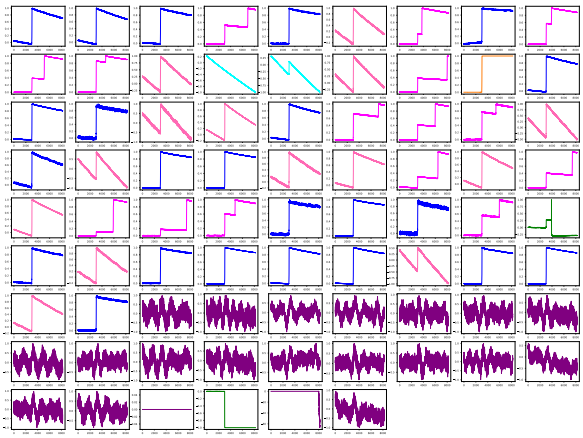


Figure 2: Exemplar waveforms found by AP in the training dataset re-labeled according to data cleaning categories.

Lastly, we proceed to the third stage where we classify waveforms based data cleaning tags using a SVM. The SVM is a supervised learning algorithm that classifies inputs by finding the hyperplane that best separates the training data into different classes [12]. We choose a SVM over other supervised learning classifiers as it excels at handling high-dimensional data (unlike k-Nearest Neighbors or Gaussian Processes) while providing a directly interpretable and intuitive explanation of how inputs are separated (unlike Random Forests or Neural Networks). Since our data contains multiple classes and is not linearly separable, we utilize a multiclass SVM [13] with a radial basis function (RBF) kernel [14]. The hyperparameters we optimize for the SVM are the inverse regularization parameter C and the RBF kernel scaling factor γ. We perform a random grid search over a broad range of values spanning several orders of magnitude for both hyperparameters. We use a 5-fold cross-validated

accuracy as the figure-of-merit for the optimization, splitting the training and validation datasets with a 80:20 ratio. The hyperparameter combination that gives the cross-validated accuracy closest to 1 is used . A visualization of the training dataset and the trained SVM in 3D is provided in Appendix A. The training and optimization for AP-SVM was conducted at a local computing cluster, requiring 264 GB of RAM and 18 CPU hours. The model utilizes code mainly from `scikit-learn` [15] and the `pygama` framework [16–18] for digital signal processing.

3 AP-SVM performance

To assess the performance of our model, we first define an AI data cleaning cut to apply to our data. Since we aim to keep only waveforms caused by physical interactions in HPGe detectors, we define the AI data cleaning cut according to Eq. 1.

$$\mathrm{AIDataCleaningCut} = \mathrm{SVMPrediction} \in \{0, 9\} \tag{1}$$

Waveforms from the Normal (0) category encompass events caused by energy depositions in HPGe detectors from α, β, and γ particles. Waveforms from the Saturation (9) category are caused by muons that deposit energies in the HPGe detector greater than the dynamic range of our DAQ system. All other waveform types correspond to electronic effects like cross-talk or saturation recovery, or to pile-up events in which multiple radiation interactions occur in close succession.

The main purpose of our model is to remove all anomalous waveforms while keeping all physical ones in our datasets. Consequently, we assess its performance in terms of physical signals that are incorrectly tagged as non-physical, defined as the model sacrifice (akin to the false negative rate), and anomalous signals that are accepted, defined as the model leakage (akin to the false positive rate). A high leakage of anomalous waveforms is preferred to a high sacrifice of physics waveforms. Anomalous signals that leak into our datasets are typically eliminated by pulse shape discrimination (PSD) cuts at later stages of the analysis chain [19, 20]. We maintain the labels defined by the data categories of Figure 1 as opposed to assigning binary physical vs anomalous labels to waveforms since the former option allows us to easily monitor the rate of each signal type in LEGEND.

To measure model sacrifice and leakage, we pre-apply traditional data cleaning cuts to all the data, creating pure datasets of the selected waveform type. We utilize combinations of different traditional data cleaning parameters to curate datasets for each waveform category. Thus, we do not provide a direct performance comparison of AP-SVM to traditional methods as the latter are used to define "true positive" and "true negative" waveforms for all physical and anomalous populations. The sacrifice, s, for a given dataset of physical waveforms is defined as the ratio of rejected events N_r to total events N. The leakage, l, for a given dataset of anomalous waveforms is defined as the ratio of accepted events N_a to total events N. The uncertainties on s and l, which are Binomial proportions themselves, are statistical and calculated using 90% Clopper-Pearson confidence intervals [21].

Table 1: Sacrifice and leakage of AI data cleaning cuts for the categories found by AP during training.

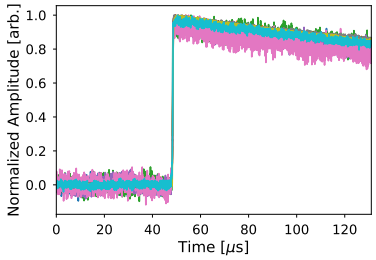
Category (Tag ID)	N	s (%)	l (%)
Normal (0)	200,000	$0.016^{+0.005}_{-0.004}$	-
Saturation (9)	23,659	$0.00^{+0.01}_{-0.00}$	-
Negative Going (1)	319	-	$0.0^{+0.9}_{-0.0}$
Downwards Sloping (3)	1,151	-	$0.0^{+0.3}_{-0.0}$
Soft Pileup (10)	24,531	-	13.4 ± 0.4
Hard Pileup (11)	4,570	-	15.5 ± 0.9
Noise Trigger (13)	203	-	$0.5^{+1.8}_{-0.5}$

Table 1 summarizes the estimates for the physics event sacrifice and anomalous event leakage of our model for all datasets. The model presents moderate anomalous event leakages in the Soft and Hard Pileup Categories of $13.4 \pm 0.4\%$ and $15.5 \pm 0.9\%$, respectively. The sacrifices and leakages of all other datasets lie below 1%. Figure 3 presents a confusion matrix of sample waveform plots of between the datasets with the highest sacrifice (Normal) and highest leakage (Hard Pileup).

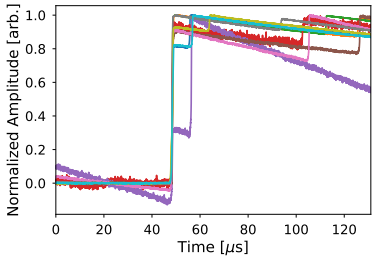
The model has a physical waveform sacrifice of $0.016^{+0.005}_{-0.004}\%$. The rejected waveforms of the Normal dataset (false negatives) have a low signal-to-noise ratio, which makes it difficult to disentangle

physics signals from anomalous populations. These waveforms are also characterized by a slow charge collection component at the top of the rising edge and on the tail, suggesting that the underlying events originated in the detector's surface [22, 23].

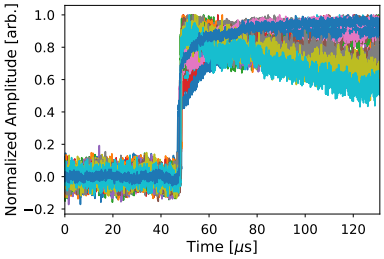
Waveforms in the pile-up categories occur when two particles deposit energy in the HPGe detector at closely spaced times, causing two overlapping pulses. Soft pileup occurs when the second pulse occurs during the falling tail of a previous pulse, and hard pileup occurs when multiple pulse rises are seen within a digitization window. The accepted Hard Pileup waveforms are characterized by a low-energy event occurring late in the digitization window, during the decaying tail of the first event. The timing and small magnitude of the second event compared to the first causes the SVM to classify these waveforms as Normal.



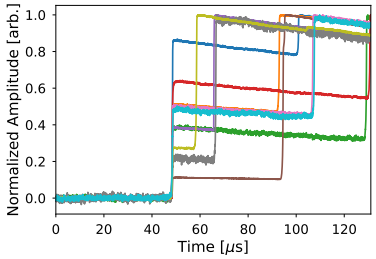
(a) Accepted Normal waveforms (true positives).



(b) Accepted Hard Pileup waveforms (false positives).



(c) Rejected Normal waveforms (false negatives).



(d) Rejected Hard Pileup (true negatives).

Figure 3: Confusion matrix of sample waveform plots between the datasets with the highest sacrifice (Normal) and highest leakage (Hard Pileup).

4 Conclusions

In this study, we have presented an AI-powered data cleaning mechanism for the LEGEND experiment. Utilizing data from an HPGe detector test-stand, we have demonstrated that our mechanism efficiently separates physical and anomalous signals, minimizing the sacrifice of physical waveforms when performing data cleaning cuts. In addition, we have successfully applied our model to data from several experimental setups with different configurations including the integration and commissioning stages of LEGEND-200, the current deployment with 142 kg of HPGe detectors, and in detector characterization test stands. We are also testing our model's performance on signals obtained from silicon photo multiplier (SiPM) detectors [24]. The versatility of our model allows it to be utilized in a variety of experiments yielding time-series waveform data.

Performing data cleaning procedures with traditional parameters requires significant time and human effort for even a single detector and can require frequent modification, particularly as run conditions change. AP-SVM has been deployed in the large-scale ($\approx \mathcal{O}(\mathrm{TB})$) data production chain of LEGEND. Our method accelerates traditional data cleaning development for LEGEND by identifying new anomalous populations and allowing the selection of events that the traditional method fails to identify. is also being used as the main data cleaning method in LEGEND's Julia-based secondary software stack: `Juleana` [25]. The AP-SVM model can thus be utilized for data cleaning on its own, to cross-validate traditional methods, or in conjunction with traditional procedures to provide a robust data cleaning method for LEGEND. The data and code used for obtaining the results presented in this study can be found in this Zenodo and GitHub repository, respectively.

5 Broader impacts

The AP-SVM model trains on time-series signals and serves the purpose of clustering with human supervision + classification. Thus, it can be applied to experiments that yield waveform time-series data. As described in Section 4, our model is currently being tested on waveforms from SiPM detector data. Additionally, our model can also be applied to experiments in fields outside physics with waveform signals, such as electroencephalograms (EEG) and electrocardiograms (ECG) in the medical field, as a tool for coarse separation of the data. Furthermore, as mentioned in Section 2, our model consumes 18 CPU hours for training and optimization. In developing the model for this particular study, we conducted the training and optimization approximately 5 times. Assuming a power consumption of a single CPU core of 65 W and a carbon intensity of the electricity of 626 lbs/MWh in North Carolina [26], we estimated the carbon footprint of training and optimizing AP-SVM to be 1.66 kg of CO_2.

Acknowledgments and Disclosure of Funding

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Nuclear Physics, under Award Numbers DE-SC0022339, DE-FG02-97ER41041, and DE-FG02-97ER41033. We acknowledge support from the Nuclear Physics Program of the National Science Foundation through grant number PHY-1812374. This work is done in support of the LEGEND-200 experiment and we thank our collaborators for their input. We would like to thank the University of North Carolina at Chapel Hill and the Research Computing group for providing computational resources and support that have contributed to the results presented in this study.

References

[1] N. Abgrall, I. Abt, M. Agostini, et al. LEGEND-1000 Preconceptual Design Report, 2021. URL https://arxiv.org/abs/2107.11462.

[2] Matteo Agostini, Giovanni Benato, Jason A. Detwiler, Javier Menéndez, and Francesco Vissani. Toward the discovery of matter creation with neutrinoless $\beta\beta$ decay. *Rev. Mod. Phys.*, 95:025002, May 2023. doi: 10.1103/RevModPhys.95.025002. URL https://link.aps.org/doi/10.1103/RevModPhys.95.025002.

[3] M. Agostini, G. R. Araujo, A. M. Bakalyarov, et al. Final Results of GERDA on the Search for Neutrinoless Double-β Decay. *Phys. Rev. Lett.*, 125:252502, Dec 2020. doi: 10.1103/PhysRevLett.125.252502. URL https://link.aps.org/doi/10.1103/PhysRevLett.125.252502.

[4] I. J. Arnquist, F. T. Avignone, A. S. Barabash, et al. Final Result of the MAJORANA DEMONSTRATOR'S Search for Neutrinoless Double-β Decay in ^{76}Ge. *Phys. Rev. Lett.*, 130:062501, Feb 2023. doi: 10.1103/PhysRevLett.130.062501. URL https://link.aps.org/doi/10.1103/PhysRevLett.130.062501.

[5] M. Agostini, G. Araujo, A. M. Bakalyarov, et al. Characterization of inverted coaxial ^{76}Ge detectors in GERDA for future double-β decay experiments. *The European Physical Journal C*, 81(6), June 2021. ISSN 1434-6052. doi: 10.1140/epjc/s10052-021-09184-8. URL http://dx.doi.org/10.1140/epjc/s10052-021-09184-8.

[6] Erica Andreotti. Characterization of BEGe detectors in the HADES underground laboratory. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 718:475–477, 2013. ISSN 0168-9002. doi: https://doi.org/10.1016/j.nima.2012.11.053. URL https://www.sciencedirect.com/science/article/pii/S0168900212013903.

[7] Iris Abt, Chris Gooch, Felix Hagemann, et al. A novel wide-angle Compton Scanner setup to study bulk events in germanium detectors. *The European Physical Journal C*, 82(10), October 2022. ISSN 1434-6052. doi: 10.1140/epjc/s10052-022-10884-y. URL http://dx.doi.org/10.1140/epjc/s10052-022-10884-y.

[8] I. J. Arnquist, F. T. Avignone, A. S. Barabash, et al. Interpretable Boosted-Decision-Tree Analysis for the MAJORANA DEMONSTRATOR. *Phys. Rev. C*, 107:014321, Jan 2023. doi: 10.1103/PhysRevC.107.014321. URL https://link.aps.org/doi/10.1103/PhysRevC.107.014321.

[9] P. Holl, L. Hauertmann, B. Majorovits, O. Schulz, M. Schuster, and A. J. Zsigmond. Deep learning based pulse shape discrimination for germanium detectors. *The European Physical Journal C*, 79(6):

450, May 2019. doi: 10.1140/epjc/s10052-019-6869-2. URL https://doi.org/10.1140/epjc/s10052-019-6869-2.

[10] Aobo Li, Julieta Gruszko, Brady Bos, Thomas Caldwell, Esteban León, and John Wilkerson. Ad-hoc Pulse Shape Simulation using Cyclic Positional U-Net, 2022. URL https://arxiv.org/abs/2212.04950.

[11] Brendan J. Frey and Delbert Dueck. Clustering by Passing Messages Between Data Points. *Science*, 315 (5814):972–976, February 2007. doi: 10.1126/science.1136800. URL https://www.science.org/doi/10.1126/science.1136800.

[12] Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, September 1995. ISSN 1573-0565. doi: 10.1007/BF00994018. URL https://doi.org/10.1007/BF00994018.

[13] Koby Crammer and Yoram Singer. On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines. *Journal of Machine Learning Research*, 2(Dec):265–292, 2001. ISSN ISSN 1533-7928. URL https://jmlr.csail.mit.edu/papers/v2/crammer01a.

[14] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel Methods in Machine Learning. *The Annals of Statistics*, 36(3), June 2008. ISSN 0090-5364. doi: 10.1214/009053607000000677. URL http://arxiv.org/abs/math/0701907. arXiv:math/0701907.

[15] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL http://jmlr.org/papers/v12/pedregosa11a.html.

[16] Matteo Agostini, Jason Detwiler, Luigi Pertoldi, et al. pygama, May 2024. URL https://doi.org/10.5281/zenodo.11403692. (v2.0.1).

[17] Ian Guinn, Luigi Pertoldi, Jason Detwiler, et al. dspeed, March 2024. URL https://doi.org/10.5281/zenodo.10731119. (v1.3.0).

[18] Jason Detwiler, Luigi Pertoldi, Ian Guinn, Grace Song, Sam Borden, Moritz Neuberger, and Patrick Krause. legend-pydataobj, May 2024. URL https://doi.org/10.5281/zenodo.11147394. (v1.7.0).

[19] M. Agostini, G. Araujo, A. M. Bakalyarov, et al. Pulse shape analysis in GERDA Phase II. *The European Physical Journal C*, 82(4):284, April 2022. ISSN 1434-6052. doi: 10.1140/epjc/s10052-022-10163-w. URL https://doi.org/10.1140/epjc/s10052-022-10163-w.

[20] S. I. Alvis, I. J. Arnquist, F. T. Avignone, et al. Multisite event discrimination for the MAJORANA DEMONSTRATOR. *Phys. Rev. C*, 99:065501, Jun 2019. doi: 10.1103/PhysRevC.99.065501. URL https://link.aps.org/doi/10.1103/PhysRevC.99.065501.

[21] Clopper C. J. and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 12 1934. doi: 10.1093/biomet/26.4.404. URL https://doi.org/10.1093/biomet/26.4.404.

[22] Frank Edzards, Lukas Hauertmann, Iris Abt, et al. Surface characterization of p-type point contact germanium detectors. *Particles*, 4(4):489–511, 2021. doi: 10.3390/particles4040036. URL https://www.mdpi.com/2571-712X/4/4/36.

[23] I. J. Arnquist, F. T. Avignone, A. S. Barabash, et al. α-event characterization and rejection in point-contact HPGe detectors. *The European Physical Journal C*, 82(3):226, March 2022. ISSN 1434-6052. doi: 10.1140/epjc/s10052-022-10161-y. URL https://doi.org/10.1140/epjc/s10052-022-10161-y.

[24] Rosanna Deckert, Igor Abritta, Gabriela Araujo, et al. The LEGEND-200 Liquid Argon Instrumentation: From a simple veto to a full-fledged detector. In *XVIII International Conference on Topics in Astroparticle and Underground Physics*, Proceedings of science. Sissa, January 2024. doi: 10.22323/1.441.0256. URL https://doi.org/10.5167/uzh-256263.

[25] Oliver Schulz. Juleana.jl, June 2024. URL https://github.com/legend-exp/Juleana.jl.

[26] North Carolina Electricity Profile 2022, 2022. URL https://www.eia.gov/electricity/state/NorthCarolina/index.php.

[27] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

A 3D SVM visualization

In this section we present a visualization of the SVM in 3D space. Since the inputs to our model contain 256 features as described in Section 2, we must apply dimensionality reduction to obtain a visual representation of the SVM decision regions. For this, we use a t-distributed Stochastic Neighbor Embedding (t-SNE) [27] algorithm to reduce both the input waveforms and the SVM decision boundaries into 3D space.

We first perform a hyperparameter optimization “by eye" on the t-SNE to obtain a representation that accurately clusters the data in 3D. The hyperparameters of the t-SNE that we consider are the perplexity and the learning rate. We conduct a search of 16 grid points spanning values of the perplexity in the [30,210] interval, and values of the learning rate in the [200, 800] interval. We plot the 3D representations of the grid, and the user chooses the hyperparameter combination that gives an accurate 3D clustering of the input waveforms. We then train a separate SVM on the 3D embedded data and optimize it in the same manner as discussed in Section 2. The trained 3D SVM is used to predict labels on a 3D mesh of voxels to simulate the separating hyperplanes. Figure 4 shows the 3D embedded waveforms of the training dataset as points and the 3D SVM decision boundaries.

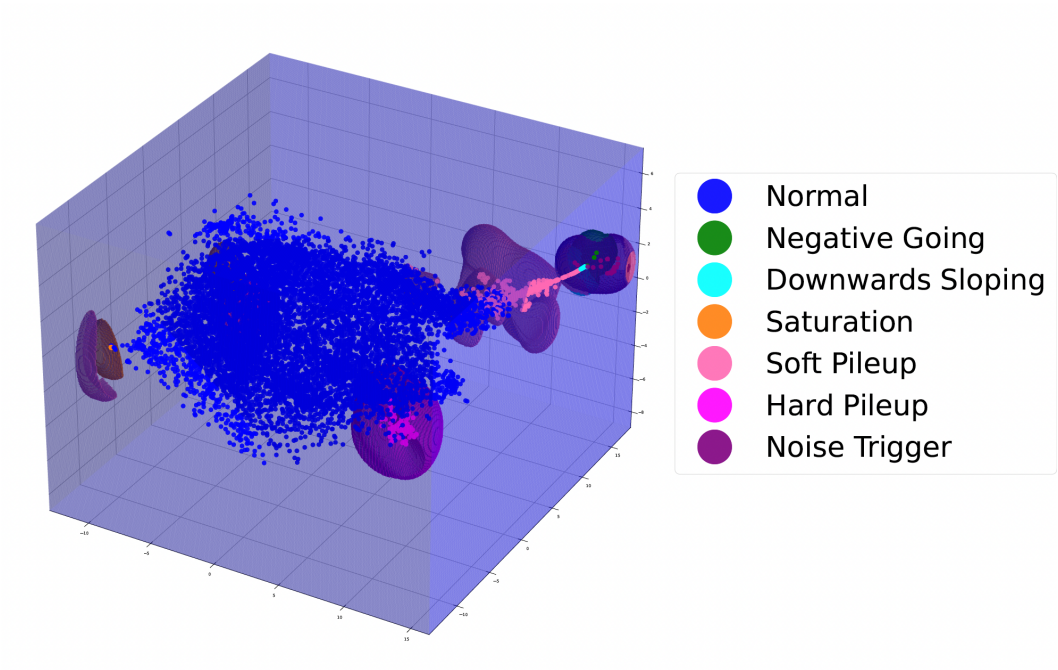


Figure 4: 3D representation of the training dataset with the SVM decision boundary hyperplanes.

B Sample accepted and rejected waveforms

In this section we include sample waveform plots of accepted and rejected events after applying the AI data cleaning cut of Eq. 1 for every dataset listed in Table 1. Since sample waveform plots of the Normal and Hard Pileup categories are presented in Figure 3, we do not include them in this section.

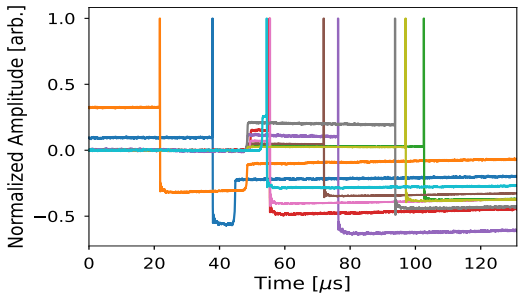


Figure 5: Sample rejected Negative Going waveforms. No Negative Going waveforms were accepted by the AI data cleaning cut.

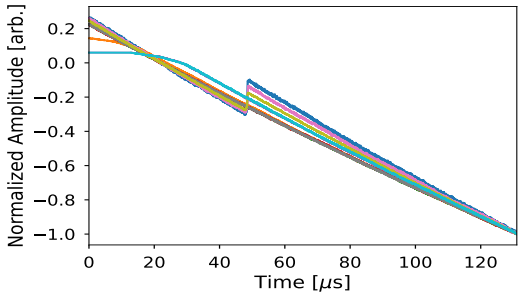


Figure 6: Sample rejected Downwards Sloping waveforms. No Downwards Sloping waveforms were accepted by the AI data cleaning cut.

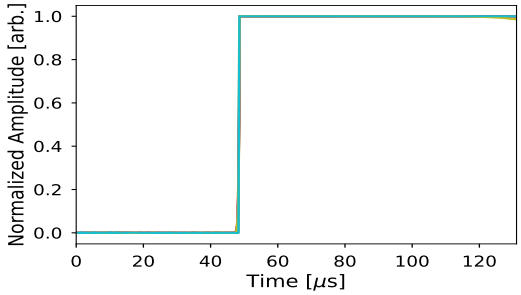
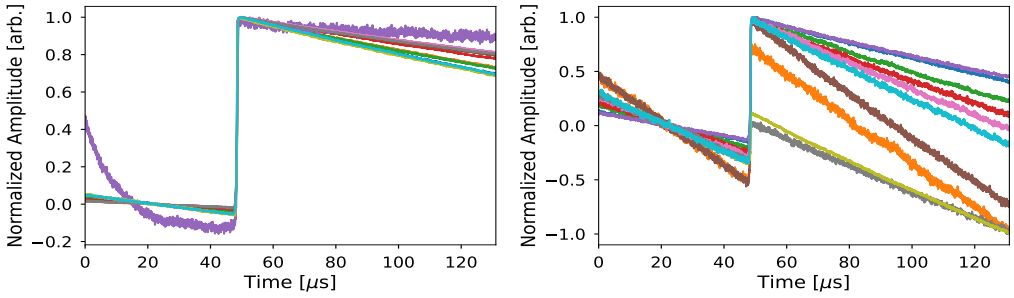


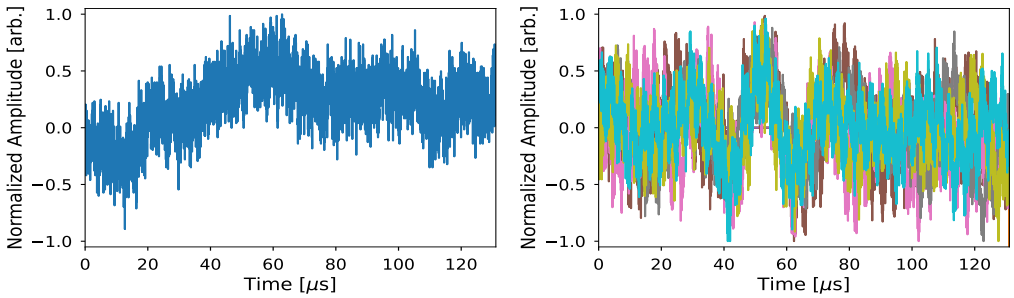
Figure 7: Sample rejected Saturation waveforms. No Saturation waveforms were rejected by the AI data cleaning cut.



(a) Accepted (false positives)

(b) Rejected (true negatives).

Figure 8: Sample accepted and rejected Soft Pileup waveforms.



(a) Accepted (false positives)

(b) Rejected (true negatives).

Figure 9: Sample accepted and rejected Noise Trigger waveforms.