Eur. Phys. J. C (2019) 79:450
https://doi.org/10.1140/epjc/s10052-019-6869-2

THE EUROPEAN PHYSICAL JOURNAL C

Regular Article - Experimental Physics

Deep learning based pulse shape discrimination for germanium detectors

P. Holl[1,2,a], **L. Hauertmann**[1], **B. Majorovits**[1], **O. Schulz**[1], **M. Schuster**[1], **A. J. Zsigmond**[1]

[1] Max Planck Institute for Physics, Föhringer Ring 6, 80805 Munich, Germany
[2] *Present Address:* Department of Informatics, Technical University of Munich, Boltzmannstr. 3, 85748 Garching, Germany

Received: 11 March 2019 / Accepted: 8 April 2019 / Published online: 28 May 2019

Abstract Experiments searching for rare processes like neutrinoless double beta decay heavily rely on the identification of background events to reduce their background level and increase their sensitivity. We present a novel machine learning based method to recognize one of the most abundant classes of background events in these experiments. By combining a neural network for feature extraction with a smaller classification network, our method can be trained with only a small number of labeled events. To validate our method, we use signals from a broad-energy germanium detector irradiated with a ^{228}Th gamma source. We find that it matches the performance of state-of-the-art algorithms commonly used for this detector type. However, it requires less tuning and calibration and shows potential to identify certain types of background events missed by other methods.

1 Introduction

Searches for rare processes, like neutrinoless double beta ($0\nu\beta\beta$) decay, critically depend on almost perfect background suppression. The leading semiconductor based experiments in this field, GERDA [1] and MAJORANA [2], search for $0\nu\beta\beta$ decay of ^{76}Ge using high-purity germanium (HPGe) detectors. Extremely low background rates of 10^{-3} cts/(keV kg year) are required in order to reach their target sensitivity on the $0\nu\beta\beta$ decay half-life. A significant contribution to their background budget are Compton scattered gamma rays originating from radioactive impurities in the experimental setup [3]. In many cases, these deposit energy in multiple locations in the germanium detector, producing so-called multi-site (MS) events. The electrons from $0\nu\beta\beta$ decays, on the other hand, deposit their energy within about 1 mm^3 of the detector volume, resulting in single-site (SS) events. The ability to discriminate between these two event types is therefore crucial. Experiments planned for the future, such as LEGEND [4], will depend even more on the efficiency of background reduction techniques.

Various pulse shape discrimination (PSD) algorithms have been developed in order to identify MS events by analyzing the digitized signal traces of semiconductor detectors in general, and HPGe detectors in particular [5–9]. Due to differences in geometry and electric field configurations, different germanium detector types exhibit different pulse shape characteristics. Broad-energy germanium (BEGe) detectors are particularly well suited for PSD since over a large fraction of the detector volume, the shape of the current-signal from SS events is almost independent of the location of the energy deposition. This results in a nearly fixed ratio between peak amplitude, A, and integral, E, of the current-signal for SS events, but not for MS events. By using the ratio A/E for PSD, both GERDA [6] and MAJORANA [8,9] achieved efficient background rejection for $0\nu\beta\beta$ decay search for BEGe type detectors. However, the A/E classifier is not able to detect all identifiable MS events since it does not take the full signal shape into account. It is also not well suited for coaxial HPGe detectors. Therefore, we look for alternative approaches.

Machine learning techniques based on deep neural networks have been used with increasing success for signal processing and analysis in recent years [10]. The ability of a neural network to fit a desired function scales primarily with the number of parameters that can be adjusted during training. However, if the number of trainable parameters is comparable to the amount of training data, a neural network can overfit, i.e. remember the training data without recognizing the underlying patterns. Large neural networks therefore require large amounts of training data. In our application, only a small set of labeled training data, i.e. data with SS/MS tagging, is available. This reflects the typical situation of larger-scale experiments that have to balance calibration and physics data collection time. It is also difficult to gener-

[a] e-mail: pholl@mpp.mpg.de

ate labeled data synthetically since the realistic behavior of a full germanium detector system, from charge drift to electronics response, is highly complex and simulations often do not fully describe measured signals. Therefore, the maximum size of a neural network used for event classification is limited by the amount of labeled data and, consequently, only a small number of input values are possible.

In previous work, this problem was solved by manually extracting a selected number of features from the signals of labeled events [6,7,11]. By using only these features as input, the artificial neural networks could be built with fewer parameters, at the cost of not being able to make use of all the information contained in the signal.

In this paper, we propose a classification scheme based on two neural networks that are trained independently. While the number of available labeled detector signals is limited, there are ample unlabeled data. The first network, an autoencoder, is trained on a large part of these unlabeled data in an unsupervised fashion. It learns to represent the shape of waveforms as low-dimensional feature vectors. The second network, the classification network, is then trained on labeled data transformed into this feature representation. This classification network can be built with a small number of trainable parameters due to its low-dimensional input. Using this two-stage process, all information present in the signal (except for its noise component) can be exploited for event classification, while preventing overfitting.

We first introduce the experimental setup and data taking in Sect. 2. Section 3 describes the applied PSD technique and Sect. 4 its verification. Section 5 provides a comparison with the state-of-the-art A/E PSD method. Conclusions are drawn in Sect. 6.

2 Experimental setup

We demonstrate our PSD technique using a prototype segmented BEGe detector [12] irradiated with a 1.1 kBq ^{228}Th source for a period of about 8 hours. The event count from this exposure is similar to what GERDA and MAJORANA record in a typical calibration cycle. The radial dimensions of the detector and the position of the source are shown in Fig. 1. The detector is made from a 40 mm high cylindrical n-type crystal with five readout electrodes: the n^+ electrode is called the core and the four p^+ surface electrodes the segments. Segments 1, 2 and 3 cover equally spaced slices of the surface while segment 4 covers the remaining area between them. A ring around the core contact is passivated. The detector is enclosed inside a cryostat and the ^{228}Th source is placed on the cryostat wall at the side of the detector, centered vertically. The source is centered on segment 1, about 20 mm from the detector surface (see Fig. 1). A more detailed description of the detector and the setup can be found in [12].

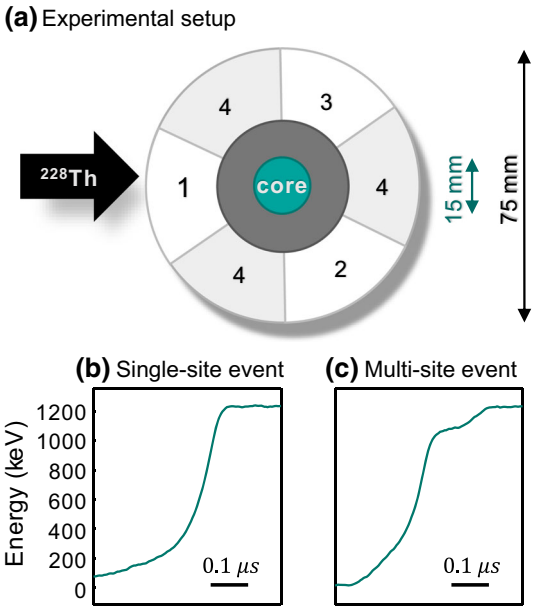

Fig. 1 **a** Schematic of the top of the segmented BEGe detector with the n^+ electrode (core) and the 4 segments. The ring around the core is passivated. The diameters of the detector and the core contact are 75 mm and 15 mm, respectively. Also shown is the azimuthal position of the ^{228}Th source. **b**, **c** The rising part of the charge waveforms recorded by the core electrode from a typical SS and a MS event, both with an energy of 1242 keV

The signals from the core and segment electrodes are amplified with charge-sensitive amplifiers and digitized with a sampling rate of 250 MHz. The recorded pulses have a length of 20 μs and are centred around the rising edge so that the recorded charge-pulses are long enough for a reliable energy reconstruction. The deposited energy, E, of each event is determined from the total increase in the charge-signal during charge-drift (see Fig. 1b, c). The part of the waveform before the rise is used to determine the baseline level for each event. To obtain the energy, the baseline level is subtracted from the mean value of the waveform after charge drift, correcting for the decay time of the signal and the cross-talk between segments and core. The energy is calibrated using the known gamma-line energies of the ^{228}Th source [12]. The detector resolution worsens with increasing energy and has a value of around 7 keV and 8 keV (FWHM) for the ^{208}Tl double-escape peak and ^{212}Bi full-energy peak, respectively. The amplitude, A, of the current waveform is determined using the procedure described in [6].

Our pulse shape analysis only takes the signals from the core electrode as input. This way, our method is compatible with the hardware deployed in existing $0\nu\beta\beta$ decay experiments, where typically only the core electrode of the germanium detectors is instrumented. We use the signals recorded from the segment electrodes only to validate the output of our PSD method.

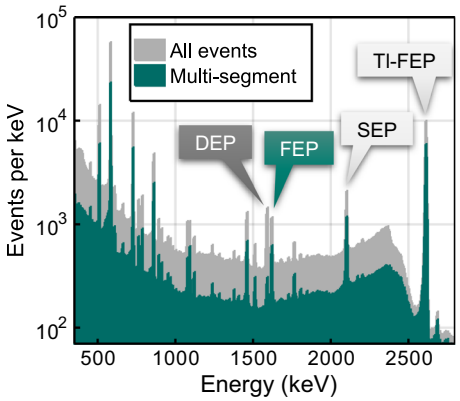

Fig. 2 Calibrated energy spectrum from the ^{228}Th source, reconstructed from the core electrode signals. The gray spectrum shows all events. The green spectrum shows only those events with energy depositions in multiple segments of the detector. The ^{208}Tl double-escape peak (DEP 1593 keV), the ^{212}Bi full-energy peak (FEP 1621 keV), the ^{208}Tl single-escape peak (SEP 2104 keV) and the ^{208}Tl full-energy peak (Tl-FEP 2615 keV) are marked. Events from the DEP and FEP are used as examples of signal and background events during training of our classifier, the labels are colored accordingly

The observed core energy spectrum (Fig. 2) shows a Compton continuum, background lines and multiple gamma lines from the ^{228}Th decay chain, the most prominent of which is the ^{208}Tl line at 2615 keV. In the following, we analyze one million events with an energy higher than 1000 keV.

We use the following gamma lines to label events:

- the ^{212}Bi full-energy peak (FEP) at 1621 keV, mostly MS
- the ^{208}Tl single-escape peak (SEP) at 2104 keV, mostly MS
- the ^{208}Tl full-energy peak (Tl-FEP) at 2615 keV, mostly MS
- the ^{208}Tl double-escape peak (DEP) at 1593 keV, predominantly single-site (SS).

Events within ±4 keV of one of these peaks are assigned the respective SS/MS label. Due to the intrinsic SS / MS mixture of the gamma lines, not all of the assigned labels are correct. Compton events cause an additional impurity in the labelling because they occur everywhere in the spectrum. The window size of 8 keV is chosen as a compromise between purity and size of the resulting labeled datasets. It is comparable to the FWHM of the DEP and FEP while resulting in datasets of similar size. Table 1 shows the event counts of the labeled datasets as well as the estimated SS fraction in each dataset. The estimated SS fractions are based on Monte Carlo (MC) simulations performed with GEANT4. In the simulations, SS events are defined by $R_{90} < 1$ mm, as in [13] and detector resolution as well as Compton events are taken into account.

In addition to the SS / MS labels, the segmentation of the detector allows us to label events as either single-segment – events which deposit energy only in segment 1 – or multi-segment – those which additionally deposit energy in another segment. These labels can be used as alternative approximations for SS and MS event labels. They are neither used in network training nor for filtering the training datasets. Instead, they serve to independently verify the classification outputs of our method (see Sect. 5).

Table 1 Overview of the energy-based datasets and associated SS and MS classification including event count and SS fraction with statistical uncertainty obtained from MC simulation

Dataset (keV)	Event count (10^3)	SS (%)	Label
DEP ± 4	10.8	85.5 ± 0.3	SS
FEP ± 4	10.0	20.1 ± 0.4	MS
SEP ± 4	15.0	12.8 ± 0.2	MS
Tl-FEP ± 4	73.4	6.3 ± 0.1	MS

3 Method

We perform a number of signal processing steps on the raw waveforms before passing them to the neural networks for further analysis. The neural network analysis consists of two stages: In the first stage, the autoencoder extracts features from all preprocessed waveforms of unlabeled events and stores them in a low-dimensional feature vector. The feature vectors of labeled events are then passed to the classifier network in the second stage.

To train both networks, the total dataset is split into 60% for training and 40% for testing. Both the autoencoder and classifier networks are then trained and evaluated on subsets of these two datasets.

3.1 Preprocessing

The raw charge-waveforms, digitized from the core electrode, span a time of 20 µs (Fig. 3a). In the first preprocessing step, the baseline is subtracted and the waveforms are normalized to their total charge (Fig. 3b). The normalized waveforms are then aligned in time so that all of them reach a value of 0.5 at the same sample number. We then trim the waveforms to a symmetric 1 µs window around the alignment point (Fig. 3c). The aligned and trimmed waveforms consist of 256 samples that cover the entire rise of the signal. Finally, a differentiation step is performed to obtain the current-waveform from the charge-waveform (Fig 3d) where individual peaks correspond to spatially separated energy depositions in the detector. Thus, events with one distinguishable peak are assumed to be SS events while events with multiple peaks are MS events.

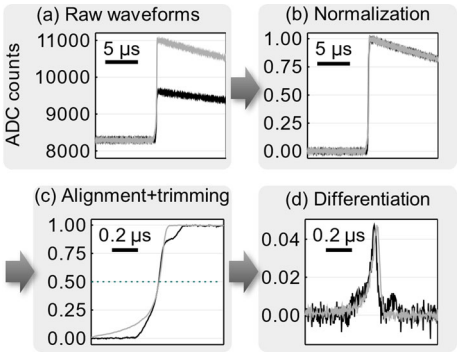

Fig. 3 Preprocessing steps shown for the MS event from Fig. 1b (black curve) and a presumed SS event with 2578 keV energy (gray curve). The raw waveforms (**a**) are baseline-subtracted and normalized to a charge of one (**b**), aligned at the intersection with a threshold of 0.5, then trimmed to a 1 µs long window (**c**), and finally differentiated yielding the current-signal (**d**)

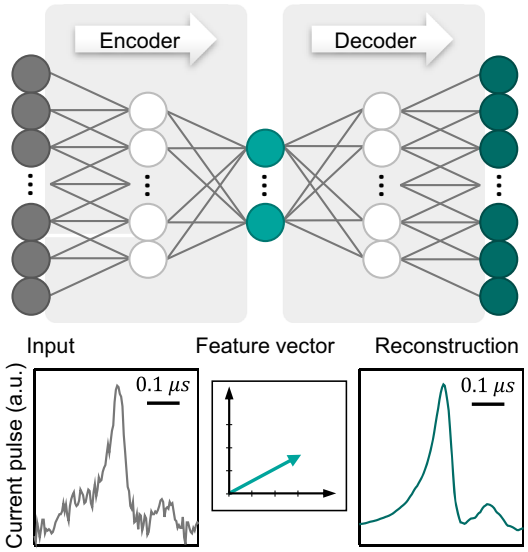

Fig. 4 Working principle of the autoencoder network. The symmetric network first encodes an input current-signal (left) into a low-dimensional feature vector (center) before decoding, i.e. reconstructing the input from the feature vector (right)

3.2 Autoencoder

After preprocessing, the resulting current-waveforms are used as input to the autoencoder. The autoencoder is a convolutional neural network. Its layout consists of two parts, shown in Fig. 4: the encoder extracts the important features from the waveform, storing them in a low-dimensional feature vector, and the decoder attempts to reconstruct the waveform from the feature vector.

In the encoder, a convolutional layer first performs two convolutions with trainable filters on the 256-dimensional input vector, producing two vectors of the same length. Both filters have a length of nine samples plus a constant bias. The convolutional layer is followed by an activation, applying a rectifying linear unit, $\mathrm{ReLU}(x) \equiv \max(0, x)$, to each value x of its input. Next, a pooling operation quarters the time-resolution from 256 to 64 by picking the maximum value of each 4 neighbouring samples across the vectors. A fully connected layer then transforms the reduced vectors into a low-dimensional vector. This operation is implemented as a matrix multiplication where all entries of the $(2 \cdot 64 \times$ feature vector dimension$)$ matrix are trainable. Another ReLU is applied to produce the feature vector. The use of convolution, activation and pooling has become common in computer vision research, where 2D convolutions on images are employed instead of 1D temporal convolutions [14]. Since key operations of the encoder depend on trainable parameters, the encoding step is flexible and can map a wide variety of possible functions. All trainable parameters are randomly initialized before training. It is not possible to predict what information each individual entry of the feature vector will represent after training, and there is no obvious interpretation of the feature representation.

Trials established that seven parameters in the feature vector are sufficient as input to a lightweight classifier network. Seven parameters are also enough to ensure that all waveforms are reconstructed with sufficient accuracy and that the training converges reliably. The encoder, with only two hidden layers, proves to be powerful enough to capture the underlying structure of the waveforms, as will be discussed in Sect. 4.

The layout of the decoder mirrors the encoder: it consists of a fully connected layer followed by a ReLU activation, a four times upsampling operation and a deconvolution. The goal of the decoder during training is to reconstruct the original waveform from the feature vector. The mean squared error (MSE) is used to measure the accuracy of the reconstruction and as the loss function to be minimized during training:

$$L_{\mathrm{MSE}} = \frac{1}{2NM} \sum_{n \in \mathcal{D}} \sum_{i \in \mathcal{S}} \left(x_{n,i} - x^*_{n,i} \right)^2 , \tag{1}$$

where, $\mathcal{D}$ denotes the training data containing N events, $\mathcal{S}$ the set of the $M = 256$ sample indices of each waveform and x and x^* represent a value of the reconstructed and the original waveform, respectively. Both encoder and decoder are trained together as a single network that tries to reproduce the input waveform. This way, the encoder learns to extract the information from the waveform that yields the most faithful reconstruction, focusing on the underlying structure rather than the noise of the signal.

In principle, all recorded events could be used to train the autoencoder since no labels are required. However, we discard events with energies lower than 1000 keV as they are

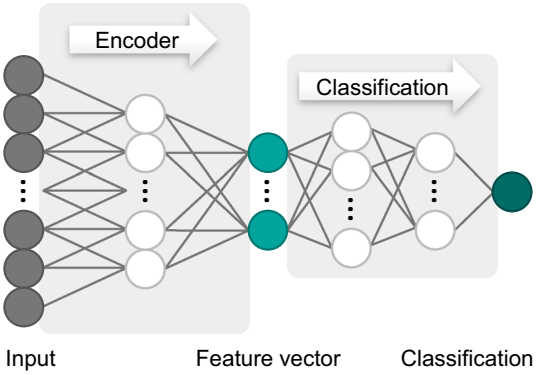

Fig. 5 Architecture of the combined network consisting of the encoder to extract feature vectors followed by the smaller classification network to identify events as SS or MS

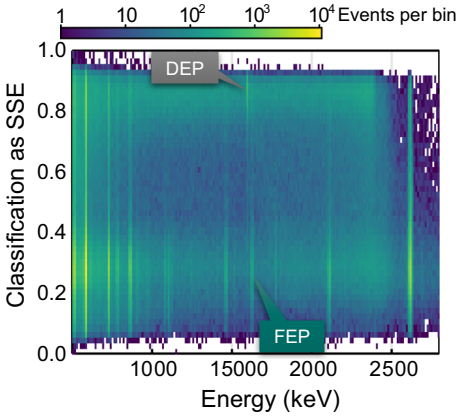

Fig. 6 Distribution of the output, c, of the combined encoder+classifier network for the test dataset (bin size X: 10 keV, Y: 0.02). The autoencoder is trained on events above 1000 keV and the classifier on events from the DEP (mostly SS) and FEP (mostly MS). The DEP events cluster around 0.9 while the peaks containing mostly MS events are centered at about 0.3

less relevant for $0\nu\beta\beta$ decay searches and are more affected by noise.

Because of their more complex pulse shapes, MS events require more information to model than SS events. However, our dataset contains more similar-looking SS events than MS events. To counteract this, we drop a fraction of SS events to balance the datasets, leaving 725k events. This filter is based on the A/E value and drops a large fraction of SS events which look almost identical except for noise. Discarded events are chosen randomly to prevent introduction of a bias.

Examples of current waveforms and their reconstructions are shown in Figs. 4 and 10. The reconstructions exhibit the same shape as the original waveforms but lack the high-frequency noise due to its high entropy. A quantitative analysis of the reconstruction quality is presented in Sect. 4.

3.3 Classifier

The training of the autoencoder is followed by the training of the classifier network. The DEP and FEP events serve as the SS and MS training datasets, respectively (see Sect. 2). Their small difference in energy ensures that the noise level is very similar for the two peaks, an additional safeguard that prevents the training to be influenced by varying signal-to-noise ratios.

The classification network takes the low-dimensional feature vector as input. Two fully connected layers consisting of 10 and 5 neurons with ReLU activation functions process the feature vector and an output layer produces a single value, $c \in (0, 1)$, which is correlated to the probability of a given event to be SS (see Fig. 5).

With the described network architecture, the classifier has a total of 141 trainable parameters. The training set contains around 30 times as many events per class to ensure that the classifier cannot overfit and remember individual events from the training dataset. This lightweight network architecture is only possible because the underlying structure of the raw waveforms has already been extracted by the autoencoder. Again, we use MSE loss (see Eq. 1) for training but adjust only the parameters of the classifier network, leaving the previously trained autoencoder unchanged.

The output values of the classifier are shown in Fig. 6 for the complete test dataset. They demonstrate that the peaks are classified as expected: DEP events are clustered around 0.9, while events from MS peaks cluster below 0.5.

4 Verification

We verify our method using the test dataset, which has 405k events above 1000 keV. Out of these, 232k events deposited energy in segment 1 and are therefore labeled as single-segment or multi-segment (see Sect. 2). First, the reconstruction accuracy of the autoencoder is examined on the whole test dataset before the discrimination performance of the combination of encoder and classifier (E + C) is evaluated on events with single-segment/multi-segment labeling.

The autoencoder is trained to keep as much relevant information of the waveform as possible when encoding. To assess its performance, we define the reconstruction error, ε_n, of an event n as the normalized RMS difference between the original and reconstructed waveform

$$\varepsilon_n \equiv \frac{1}{\sigma_n}\sqrt{\sum_{i\in\mathcal{S}}(x_{n,i} - x^*_{n,i})^2}. \tag{2}$$

Here, σ_n denotes the noise level of the normalized waveform from event n, so ε is constructed to be independent of this noise. A small ε thus indicates that much of the information

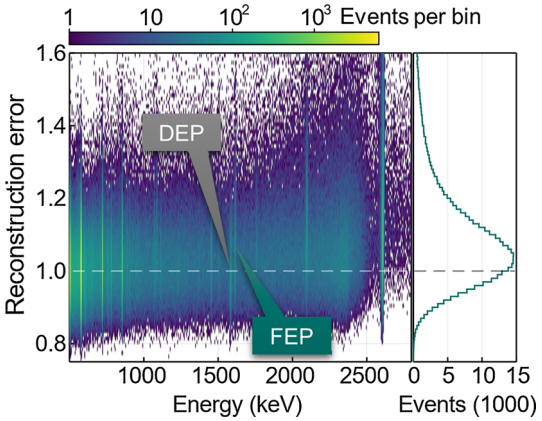

Fig. 7 Distribution of the normalized reconstruction error, ε, of the autoencoder as a function of energy (left) and its marginalized distribution for events with an energy between 1000 and 3000 keV (right)

from the original pulse is contained in the reconstruction. For $\varepsilon = 1$, the reconstruction error is equivalent to the deviation expected from noise only.

Figure 7 shows the distribution of ε as a function of energy. Single-site events are reconstructed with high accuracy since their consistent shape makes them easy to learn. As a result, SS events from the DEP form a Gaussian-like distribution around $\varepsilon \approx 1$. Multi-site events, on the other hand, are more difficult to reconstruct since they comprise events with multiple peaks in the current waveform. Therefore, the gamma lines dominated by MS events, like the FEP, contain more events with less accurate reconstructions.

Generally though, ε is small and even complex signal structures of MS or rare outlier events are usually reconstructed well, indicating that the feature vector encodes all important waveform characteristics. In addition, studies of similar network architectures have shown that the information loss of autoencoder reconstructions can primarily be attributed to the decoder part of the network [15], which is not used for our classification. For these reasons, the extracted feature vectors constitute an ideal basis for the second classification stage and classification performance is not affected by the information loss.

To evaluate the performance of the classifier, we introduce a variable discrimination threshold. Events above the threshold are classified as SS events and accepted, while the ones below are classified as MS events (background) and rejected. Varying this threshold results in different survival fractions for each gamma peak. The SS fractions of the test datasets resulting from a specific threshold are detailed in Table 2.

Figure 8 shows how the survival fraction of DEP events and the rejection of FEP events vary depending on the discrimination threshold. All survival fractions are obtained from a binned fit of a linear background plus a Gaussian to the peaks in the energy spectrum. From our MC simulations,

Table 2 Classification of events in the test datasets. The class assignment is based on the output, c, of the combined encoder+classifier network. Events are counted as SS if $c > 0.6$ (maximum class separation threshold). The given uncertainties are statistical only

Dataset (keV)	Event count (10^3)	Classified SS (%)
DEP $\pm$ 4	4.3	82.8 $\pm$ 1.4
FEP $\pm$ 4	4.1	28.4 $\pm$ 0.8
SEP $\pm$ 4	6.0	21.7 $\pm$ 0.6
Tl-FEP $\pm$ 4	29.4	14.1 $\pm$ 0.2

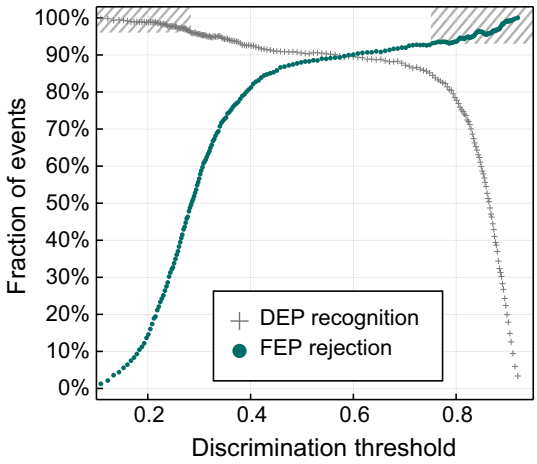

Fig. 8 SS event recognition and MS event rejection efficiencies as a function of the discrimination threshold on the classifier output for the DEP (mostly SS) and the FEP (mostly MS). Inside the highlighted areas, more events are accepted or rejected than the peak contains of the associated class

Table 3 Classifications by the neural network (discrimination threshold of 0.6) of single-segment and multi-segment events

	Single-site (10^3)	Multi-site (10^3)
Single-segment	65	35
Multi-segment	26	106

we know that 93% of true DEP events (excluding Compton background) are SS and 96% of true FEP events are MS. Higher recognition or rejection rates necessarily accept or reject too many events. These areas are highlighted in Fig 8.

For our classifier, a discrimination threshold of 0.6 yields maximum class separation efficiency with 90% FEP rejection at 90% DEP recognition. At this threshold, 90k of the 232k events in the test dataset which have energy depositions in segment 1 are classified as SS and 141k events as MS. In the same dataset, 100k events are labeled as single-segment and 132k as multi-segment. Table 3 shows the number of events classified as SS and MS for both single-segment and multi-segment events.

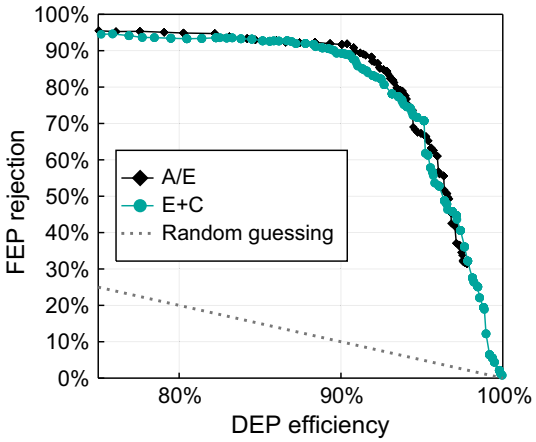

Fig. 9 Rejection power of the combined encoder+classifier network (E + C) compared to the A/E discrimination algorithm

Of all single-segment events, 65% are classified as SS and 80% of the multi-segment events are classified as MS. The relatively high fraction of single-segment events that are classified as multi-site is not surprising because gammas can cause multiple energy depositions within one segment, resulting in single-segment MS events.

5 Discussion

In order to assess the overall discrimination performance of our method, we compare it to the performance of the A/E technique currently employed for BEGe type detectors by $0\nu\beta\beta$ experiments (see Sect. 1). The A/E survival fractions are calculated with the same fitting method, described in Sect. 4. Figure 9 compares the rejection power of the combined E + C network with the power of the A/E method for equal DEP survival fractions.

Despite the similarity of the two performance curves, the classifiers are fundamentally different and often do not agree in their classifications. While A/E is based on a single parameter (the maximum current divided by the total deposited energy), the E + C network takes the whole waveform into consideration. Using a threshold chosen to result in a 90% DEP survival fraction, the two methods assign different classes to about 10% of all events. The two methods can therefore be regarded as complementary. The events classified as SS by the E + C network and as MS by A/E account for about three quarters of these events and 52% of them are labeled as single-segment. On the other hand, 79% of the events classified as SS by A/E but classified as MS by E + C are multi-segment, and therefore almost certain to be true MS events. This demonstrates that the E + C network can identify certain types of MS events that are incorrectly classified as SS by the A/E method.

These differences in classification become clear when examining specific waveforms. Figure 10a shows a 2618 keV event that deposited all energy in segment 1. The rise in the charge-pulse starts more abruptly than with most SS events and the current-waveform shows larger-than-normal noise fluctuation just before its peak. While A/E classifies this as SS, the E + C network classifies it as MS.

The waveform of the multi-segment 1334 keV event in Fig. 10b, classified in the same manner, looks similar except for multiple small substructures in the signal. However, as only about half of the energy is deposited in segment 1, it is almost certainly multi-site. Part of the energy may be deposited close to the contact, causing a high peak in the current waveform which in turn causes A/E to misclassify it as SS. Our test dataset contains about 4500 events of this type.

Figure 10c shows an event with all energy deposited in segment 1. Events at such high energies are almost exclusively coincidence events, i.e. events caused by two or more gamma rays depositing energy at the same time. It can also

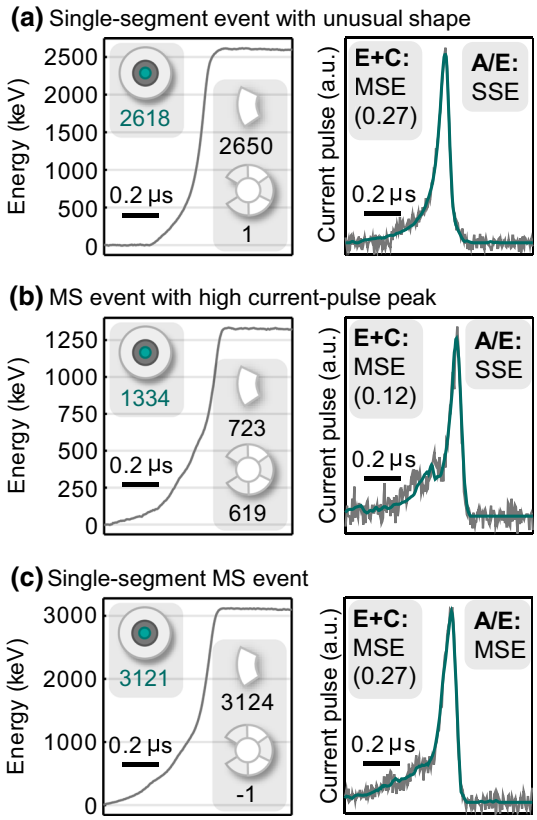

Fig. 10 Example pulses illustrating different classifications of the A/E and the encoder+classifier (E + C) methods. Left: The charge-signal waveforms, also indicating the core energy (in keV), the energy in segment 1 and the energy in other segments. Right: The preprocessed current-signal waveforms (gray curves) with their autoencoder reconstructions (green curves), indicating the event classifications and output value of the E + C network and the class predicted by the A/E method

clearly be identified as MS from the multiple-peak structure in the current-pulse and both A/E and E + C classify it as such. This demonstrates that segment information alone is not sufficient to detect all MS events.

All machine learning techniques heavily rely on the quality of their training data since no prior knowledge of physical processes is assumed. It is therefore remarkable that the E + C network matches the performance of the physics-based A/E method despite our small and impure training dataset.

The electronics of our detector system has a relatively low noise level compared to larger scale experiments. It has been demonstrated, using GERDA data, that our classification is robust in the presence of different types of noise or variation in waveform shape due to changes of detector or amplifier characteristics over time [16]. This robustness stems from the fact that the classifier only depends on the extracted features equivalent to the denoised waveform. This is an advantage over the A/E method, of which the classification performance directly depends on the noise level: The low noise level of the data used in this work can be seen as a best-case scenario for A/E.

6 Conclusions

We have demonstrated the use of two different deep-learning based neural networks in combination to achieve state-of-the-art discrimination performance for single-site/multi-site recognition for germanium detectors. By splitting the discrimination method into two independent stages, a feature extraction and a classification stage (E + C), only a small subset of all training data needs to be labeled. The first stage is a feature-extraction performed by an autoencoder. It drastically reduces the dimensionality of the data while retaining the essential characteristics of the waveform. This network is trained in an unsupervised fashion, so no class labels are required. Using only seven feature parameters, the waveforms are approximated with sufficient accuracy.

The classification network that operates on the extracted feature vectors has been shown to be competitive in discrimination performance with the widely used A/E algorithm, despite being trained on a small and impure dataset.

Our method is currently limited by the volume and purity of the training data. We are working on an accurate pulse shape simulation that could provide arbitrary amounts of high-purity synthetic training data. Assuming the simulation reaches a sufficient level of accuracy, one might also be consider to train the autoencoder on measured data and the classifier on simulated data. This would combine the properties of a real detector system, especially its noise characteristics, with a classifier trained on synthetic data with pure labels.

Due to their fundamentally different approach it may also be profitable to combine our E + C method with the A/E method, this may result in an even more powerful background rejection scheme. One way to achieve this would be to adjust the classification thresholds and only accept an event if both methods classify it as SS. As has been shown, the A/E technique fails to reject certain types of multi-site events, e.g. events with energy depositions close to the detector contact. Since the E + C network has access to the whole waveform shape, it can identify them like any other MS event.

The E + C method presented here is powerful enough to encode and classify events in an automated fashion for a large number and different types of detectors without manual corrections [16]. It has the potential to be a valuable background rejection technique for the next generation of $0\nu\beta\beta$ decay experiments, e.g. LEGEND [4].

Acknowledgements We would like to thank Iris Abt for her valuable comments to this manuscript. The data used in this paper can be requested from the authors.

Data Availability Statement This manuscript has no associated data or the data will not be deposited. [Authors' comment: The data used in this paper can be requested from the authors.]

Funded by SCOAP[3].

References

1. K.H. Ackermann et al., Eur. Phys. J. C **73**, 2330 (2013). https://doi.org/10.1140/epjc/s10052-013-2330-0
2. C.E. Aalseth et al., Phys. Rev. Lett. **120**, 132502 (2018). https://doi.org/10.1103/PhysRevLett.120.132502
3. M. Agostini et al., Eur. Phys. J. C **74**, 2764 (2014). https://doi.org/10.1140/epjc/s10052-014-2764-z
4. N. Abgrall et al., AIP Conf. Proc. **1894**, 020027 (2017). https://doi.org/10.1063/1.5007652
5. F. Petry et al., Nucl. Instrum. Meth. A **332**, 107 (1993). https://doi.org/10.1016/0168-9002(93)90746-5
6. M. Agostini et al., Eur. Phys. J. C **73**, 2583 (2013). https://doi.org/10.1140/epjc/s10052-013-2583-7
7. A. Caldwell, F. Cossavella, B. Majorovits, D. Palioselitis, O. Volynets, Eur. Phys. J. C **75**, 350 (2015). https://doi.org/10.1140/epjc/s10052-015-3573-8
8. C. Cuesta et al., J. Phys. Conf. Ser. **888**, 012240 (2017). https://doi.org/10.1088/1742-6596/888/1/012240
9. S.I. Alvis, et al., (2019). arXiv:1901.05388
10. S. Delaquis et al., JINST **13**, P08023 (2018). https://doi.org/10.1088/1748-0221/13/08/P08023
11. B. Majorovits, H.V. Klapdor-Kleingrothaus, Eur. Phys. J. A **6**, 463 (1999). https://doi.org/10.1007/s100500050370
12. I. Abt et al., Nucl. Instrum. Meth. A **925**, 172 (2019). https://doi.org/10.1016/j.nima.2019.02.005
13. I. Abt, M.F. Altmann, A. Caldwell, K. Kroninger, X. Liu, B. Majorovits, L. Pandola, C. Tomei, Nucl. Instrum. Meth. A **570**, 479 (2007). https://doi.org/10.1016/j.nima.2006.10.188

14. Z.Q. Zhao, P. Zheng, S.t. Xu, X. Wu, (2018). arXiv:1807.05511
15. S. Palacio, J. Folz, J. Hees, F. Raue, D. Borth, A. Dengel, in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018)
16. P. Holl. Deep Learning Based Pulse Shape Analysis for GERDA. (2017). https://publications.mppmu.mpg.de/2017/MPP-2017-247/FullText.pdf