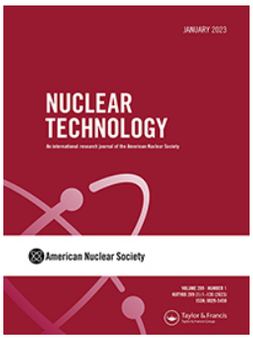

Nuclear Technology

ISSN: (Print) (Online) Journal homepage: https://ans.tandfonline.com/loi/unct20

# Explanation of Deep Learning–Based Radioisotope Identifier for Plastic Scintillation Detector

Byoungil Jeon, Jinhwan Kim & Myungkook Moon

**To cite this article:** Byoungil Jeon, Jinhwan Kim & Myungkook Moon (2023) Explanation of Deep Learning–Based Radioisotope Identifier for Plastic Scintillation Detector, Nuclear Technology, 209:1, 1-14, DOI: 10.1080/00295450.2022.2096389

**To link to this article:** https://doi.org/10.1080/00295450.2022.2096389

Published online: 23 Aug 2022.

NUCLEAR TECHNOLOGY · VOLUME 209 · 1–14 · JANUARY 2023

DOI: https://doi.org/10.1080/00295450.2022.2096389



# Explanation of Deep Learning–Based Radioisotope Identifier for Plastic Scintillation Detector

Byoungil Jeon,[a] Jinhwan Kim,[b] and Myungkook Moon[c]*

[a]*Korea Atomic Energy Research Institute, Artificial Intelligence Application and Strategy Team, Daejeon, 34507, Korea*
[b]*Korea Atomic Energy Research Institute, HANARO Utilization Division, Daejeon, 34507, Korea*
[c]*Korea Atomic Energy Research Institute, Neutron Science Division, Daejeon, 34507, Korea*

Received February 15, 2022
Accepted for Publication June 21, 2022

**Abstract** — *Radioisotope identification (RIID) is a representative application of deep learning for radiation measurements. Deep learning-based RIID models have been implemented in various types of radiation detectors; however, very few of these models have been interpreted using explainable artificial intelligence (XAI) methods. This paper presents an explanation of a deep learning–based RIID model for a plastic scintillation detector. The RIID task is defined as a multilabel binary classification problem, and the dataset is generated using a random sampling procedure. The identification performance is verified using experimental data. The experimental results demonstrate that the performance of the RIID models increased with the increase in the total counts of the dataset. Additionally, XAI methods are implemented, and their explanatory performance is verified for the spectral input. The domain knowledge of RIID for the plastic scintillation detector is that patterns near the Compton edge can be used as evidence for the existence of radioisotopes. Among the implemented XAI methods, integrated gradient and layerwise relevance propagation exhibited concurrence with the domain knowledge, with the Shapley value explanation method presenting the most reliable results.*

**Keywords** — *Deep learning, model interpretation, explainable artificial intelligence, radioisotope identifier, plastic scintillation detector.*

**Note** — *Some figures may be in color only in the electronic version.*

## I. INTRODUCTION

Deep learning technologies have been applied in various fields, including radiation measurements, in the emerging deep learning era. Radioisotope identification (RIID) is one of the most important applications of deep learning research on radiation measurements. Various studies have been conducted on deep learning–based RIID for silicon and inorganic scintillation detectors.[1–7] These studies have been focused on automatic and rapid analysis rather than conventional full-energy-peak analysis, despite the spectroscopic capabilities of these detectors. Furthermore, deep learning techniques have been applied to RIIDs for plastic scintillation detectors.[8–13] The explanation performance of the deep learning techniques on RIID demonstrated in these studies has been promising even for plastic scintillation detectors, which have poor spectroscopic capability owing to the absence of full-energy peaks and poor energy resolution.

Explainable artificial intelligence (XAI) has been developed to improve the reliability of deep learning techniques; it has also been implemented in related research fields to interpret deep learning applications. Several studies conducted on deep learning RIID models[14,15] for inorganic scintillation detectors have demonstrated that deep learning models identify radioisotopes (RIs) by concentrating on the full-energy-peak regions in a spectrum, which is

*E-mail: moonmk@kaeri.re.kr

analogous to conventional spectroscopic analysis. However, there has been no XAI research conducted on the application of deep learning–based RIID models for plastic scintillation detectors.

This study presents an explanation of a deep learning–based RIID model for a plastic scintillation detector. An RIID task is defined as a multilabel binary classification problem in this study, and deep learning models are implemented for the task. Counting statistics can be considered as a metric to assess the quality of the data in spectral data. Two datasets, whose total counts differed by almost 10 times, were generated to determine the performance variations based on the counting statistics, i.e., the quality of the spectral data. The RIID models were trained, and their performance was verified using experimental datasets. Subsequently, several XAI methods were implemented, and their performance for the RIID model with a spectral input was verified.

## II. MATERIALS AND METHODS

### II.A. Dataset Generation

#### *II.A.1. Experimental Setup*

The gamma spectra were measured using a plastic scintillator (EJ-200, Eljen Technology) of 30-mm diameter and 50-mm height. The scintillator was coupled with a photomultiplier tube (PMT, R2228, Hamamatsu). A digital pulse processor (DP5G, Amptek) was used for data acquisition, and the operating voltage was supplied by a high-voltage supply unit (NHQ 224M, Iseg). Eight RIs, i.e.,$^{22}$Na, $^{54}$Mn, $^{57}$Co, $^{60}$Co, $^{109}$Cd, $^{133}$Ba, $^{137}$Cs, and $^{152}$Eu (standard gamma sources, Eckert & Ziegler), were used as isotopes whose spectra are to be measured; all the experiments were conducted inside a dark box to enhance the optical shielding.

Table I presents detailed information on the RIs used in the study. The box was composed of 10-mm-thick

TABLE I

Detailed Information on RIs Used in the Study

| Radioisotopes | Gamma Energy (MeV) | Compton Edge Energy (MeV) | Intensity (%) |
|---|---|---|---|
| $^{22}$Na | 0.511 | 0.341 | 180.76 |
| | 1.275 | 1.062 | 99.94 |
| $^{54}$Mn | 0.835 | 0.639 | 99.98 |
| $^{57}$Co | 0.014 | 0.001 | 9.16 |
| | 0.122 | 0.039 | 85.6 |
| | 0.137 | 0.048 | 10.68 |
| $^{60}$Co | 1.173 | 0.963 | 99.9 |
| | 1.333 | 1.119 | 99.98 |
| $^{109}$Cd | 0.088 | 0.023 | 3.64 |
| $^{133}$Ba | 0.081 | 0.019 | 32.9 |
| | 0.276 | 0.143 | 7.16 |
| | 0.303 | 0.164 | 18.34 |
| | 0.356 | 0.207 | 62.05 |
| | 0.384 | 0.231 | 8.94 |
| $^{137}$Cs | 0.662 | 0.478 | 85.1 |
| $^{152}$Eu | 0.344 | 0.197 | 1.09 |
| | 0.411 | 0.253 | 1.09 |
| | 0.444 | 0.282 | 1.11 |
| | 0.678 | 0.492 | 1.29 |
| | 0.779 | 0.587 | 1.41 |
| | 0.867 | 0.670 | 4.25 |
| | 0.964 | 0.762 | 14.6 |
| | 1.086 | 0.879 | 10.21 |
| | 1.090 | 0.883 | 1.73 |
| | 1.112 | 0.904 | 13.64 |
| | 1.299 | 1.085 | 1.62 |
| | 1.408 | 1.192 | 21 |

aluminum plates. Its internal dimensions were as follows: 440 mm (width) × 400 mm (height) × 900 mm (depth). The distance from the source to the detector window was set to 1.25 cm, and the measurement period was set to 1 h to minimize the statistical fluctuations in the spectra, which were measured for 512 channel bins. However, the first 12 channels were deactivated by a low-level discrimination (LLD) threshold.

### II.A.2. Simulation Setup

Many data are required to train a deep learning model, but obtaining the required data experimentally is a laborious task. Therefore, Monte Carlo simulation was used to generate large portions of our dataset. The Monte Carlo N-Particle (MCNP) version 6.2 code[16] was employed to simulate the spectra of the plastic scintillation detector. The geometry of the simulation was implemented in a manner analogous to the experimental environment. The material information, composition ratios, and densities were defined according to a material data report.[17] A pulse-height tally with a Gaussian energy broadening (GEB) card was employed to simulate the gamma spectra. A parametric optimization method[18] with measured spectra was used for channel-to-energy calibration and to determinate the GEB coefficient. Three cases of GEB coefficients found in parametric optimization were used to generate a dataset that is robust against changes in the energy broadening levels. Table II lists the GEB parameters used in this study. The use of three GEB cases was effective in developing the RIID model whereas the model could not be trained with the dataset of one GEB case despite similarities between the two approaches.

### II.A.3. Dataset Generation

The spectra for each RI were simulated for three GEB cases with a history of one billion, to obtain good counting statistics, and were normalized by their integral values to be represented as probabilistic density functions (PDFs). The dataset was then generated by iterating the random sampling procedure with the spectral PDF of each RI as follows. First, the ratios and GEB cases of the RIs to be mixed were determined randomly. The sum of the RI ratios was normalized to retain the characteristics of the PDF even when the RI PDFs were mixed. Second, the PDFs were mixed based on the determined RI ratios. Third, random numbers were sampled using a mixed PDF for a given number of samples. Therefore, the number of samplings represents the number of counts, the sampled random number represents the event of energy reading, and the histogram of the numbers represents the energy spectrum of a detector. Figure 1 depicts a flowchart of this procedure.

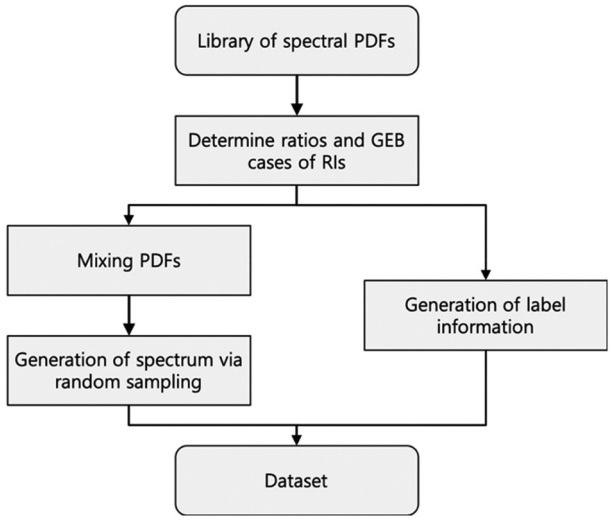


Fig. 1. Flowchart of dataset generation procedure.

Two types of datasets, whose number of samplings differed by 10 times, were prepared to compare the performances and explanations of RIID models based on the counting statistics of the input spectra. Table III

TABLE II

Three Cases of GEB Parameters Used for Dataset Generation

| GEB Parameters | *a* | *b* | *c* |
|---|---|---|---|
| Case A | −0.0007 | 0.3944 | −0.4999 |
| Case B | 0.0068 | 0.3549 | −0.4999 |
| Case C | 0.0004 | 0.3706 | −0.4999 |

TABLE III

Conditions of Generated Datasets

| Dataset Type | Number of Samplings | Number of Gamma Rays Emitted at Source Position[a] |
|---|---|---|
| A | 1000 to 3000 | $2.43 \times 10^3$ to $7.29 \times 10^3$ |
| B | 10 000 to 30 000 | $2.43 \times 10^4$ to $7.29 \times 10^4$ |

[a]The number of gamma rays was converted using the average detection efficiency calculated by MCNP simulations for the eight RIs. It can be converted to radioactivity when the composition ratio of the RIs and their gamma emission intensities are known.

summarizes the conditions of the datasets. For each condition, 70 000 spectra were generated and used as a dataset to develop the deep learning models.

## II.B. Deep Learning Model RIID

In this study, the RIID task was considered as a multilabel binary classification problem. The input data in the multilabel classification problem can have multiple labels. Because eight RIs were used in this study, our datasets had eight labels with the binary classes of true or false. For example, the gamma spectra of $^{22}$Na and $^{60}$Co were labeled as two true classes for the $^{22}$Na and $^{60}$Co labels and six false classes for the other labels. For this problem definition, the the RIID models were developed to output RI labels, which indicate the existence of a single RI or multiple RIs, when the gamma spectra are provided as input.

A convolutional neural network (CNN) was implemented as an RIID model in a Python environment using the Tensorflow 2.4 library.[19] The model comprises two convolution units: one flattened layer and one hidden layer. Each convolution unit consists of two convolution layers that are one-dimensional (1-D) and one max pooling layer that is 1-D. The number of convolution filters was set to 16, and the length of the filters was set to three for all the convolution layers. The flattened layer is deployed after the convolution units to reduce the dimension of the output from two-dimensional to 1-D. The sigmoid function is used as the activation function of the final hidden layer, and the rectified linear unit (ReLU) functions are used for the convolution layers. Figure 2 presents a schematic diagram of the proposed RIID model structure.

The RIID models, A and B, were trained using the datasets, A and B, respectively. A binary cross-entropy function was used as the loss function to train the RIID model. The 70 000 spectra in the dataset were divided as follows: 50 000 were used as the training set, 15 000 as the validation set, and 5000 as the test set. The RIID models were trained using the Adam optimizer[20] for 1000 epochs, and the cyclical learning rate was applied to obtain stable model parameters. The learning rate ranged from $1 \times 10^{-5}$ to $1 \times 10^{-3}$. The validation loss was monitored during the epochs to obtain a model with minimum loss; this model was saved and then used as the final model.

## II.C. Explanation of Deep Learning Model

Deep learning models are typically considered to be black boxes; this means that the model produces an output from the input data, but the detailed information on the procedure used to obtain the output is inaccessible. With the rapid advancement of deep learning, novel XAI technology has been developed to explain deep learning models. Feature attribution methods, which determine the amount of contribution of each feature to the output, are particularly useful in explaining deep learning models. These methods can be used to determine the quantitative attributes of the input features used to obtain the output in a model.

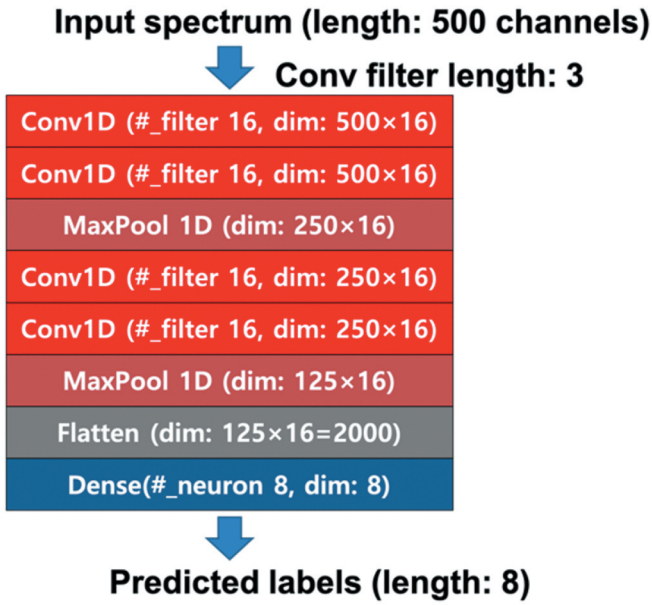


Fig. 2. Structure of RIID model.

The explanation performance of the XAI methods varies for different types of data.[21–23] A method suitable for spectral data is required as spectral data were used in this study. Therefore, several feature attribution-based XAI methods were implemented, and their performances in explaining the deep learning–based RIID model that analyzed spectral data were compared. These methods include the following: Gradient-weighted Class Activation Mapping (Grad-CAM) (Ref. 24), Integrated Gradients (IG) (Ref. 25), Local Interpretable Model-agnostic Explanations (LIME) (Ref. 26), Layer-wise Relevance Propagation (LRP) (Ref. 27), and SHapley Additive exPlanations (SHAP) (Ref. 28). Open-source libraries were utilized for the LIME (Ref. 26) and SHAP (Ref. 28) methods, and the Grad-CAM, IG, and LRP methods were implemented using the Tensorflow 2.4 library.[19]

### *II.C.1. Grad-CAM Method*

Grad-CAM is a method used to calculate feature attributions using the gradient backpropagation of the

model. After a deep learning model is trained, the model parameters are no longer variable. Therefore, the input vectors are the only variables for the model in the use phase. A feature time gradient can be used to calculate feature attribution in this phase. This implies that feature attributions can be calculated using layerwise gradient backpropagation. This calculation is expressed as follows:

$$L^{c}_{\text{Grad-CAM}}(i) = \text{ReLU}\left(\sum_{k} a^{c}_{k} A^{k}\right), \tag{1}$$

where

$c$ = class

$k$ = depth of the hidden layers

$A^k$ = activation of the $k$'th hidden layer

$a_k{}^c$ = weighting value of the node in the $k$'th hidden layer for class $c$.

This weighting was calculated using Eq. (2):

$$a^{c}_{k} = \sum_{i} \frac{\partial y^{c}}{\partial A^{k}_{i}}, \tag{2}$$

where $y^c$ denotes the gradient of class $c$.

### II.C.2. Integrated Gradients Method

The IG method is typically analogous to gradient-based methods; however, it is used to calculate the integration of the backpropagating gradients from the input to the baseline. In the IG calculation, the model can be represented as a function, $F$: $\mathbf{R}^n \rightarrow [0,1]$. A straight-line path from the baseline to the input can be considered when $x$ is the input and $x$' is the baseline input, and the gradients at all points along the path can be computed. Here, the baseline input represents a meaningless input; it can be considered a vector whose elements are zero. The IG was calculated by integrating these gradients. This calculation is expressed as follows:

$$\text{IG}_i(x) = (x - x') \int_{\alpha=0}^{1} \frac{\partial F(x' + n(x - x'))}{\partial x_i} d\alpha, \tag{3}$$

where $i$ represents the dimension along the path and $\partial F(x)/\partial x_i$ represents the gradient of $F(x)$ on the $i$'th path.

### II.C.3. LIME Method

LIME is a method used to calculate feature attributions using surrogated analysis of a deep learning model. In the LIME algorithm, the input features are permutated, and the distances between the permutated features and the original input are measured. Subsequently, the model is used to select several permutated features with the maximum likelihood output. Simplified models, such as linear models, are trained using the selected features, and the weighting values of the features are calculated using a deep learning model, simplified model, and feature distances. This calculation is expressed as follows:

$$L(x) = \underset{g \in G}{\text{argmin}}\, L(f, g, \pi_x), \tag{4}$$

where

$f$ = deep learning model

$g$ = simplified model

$\pi_x$ = distance between the permutated and original features

$L$ = loss function.

### II.C.4. LRP Method

The LRP method is used to calculate the relevance of the input features by decomposing a deep learning model. If a model is completely decomposed to achieve an output from the activation functions of all the hidden layers, relevance propagation can be used to obtain element-wise attributions from the input to the output. This calculation is expressed as follows:

$$R^{(l,l+1)}_{i \leftarrow j} = R^{(l,l+1)}_{j} \frac{a_i w_{ij}}{\sum_h a_h w_{hj}}, \tag{5}$$

where

$l$ = depth of the hidden layers

$i$ = node in the $l$'th hidden layer

$j$ = node in the $(l + 1)$'th hidden layer

$h$ = all the nodes in the $l$'th hidden layer

$a_i$ = output from node $i$

$w_{ij}$ = weighting value corresponding to nodes $i$ and $j$.

Here, the final output from the model is used as the initial values for $a$ and $R$ of the last layer, and the relevance is calculated sequentially from the output layer to the input layer.

### II.C.5. SHAP Method

The SHAP method is used to calculate the feature attributions in cooperative tasks. The SHAP value is calculated using the averaged changes for various combinations of features with and without individual features. SHAP involves a heavier computational load than the other XAI methods as it calculates feature attributions for all possible subsets. However, it can be used to calculate the appropriate feature attributions even when the features are interdependent. The SHAP calculation can be expressed as follows:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \{F_{S \cup \{i\}} - F_S\}, \qquad (6)$$

where

$\phi_i$ = Shapley value of feature $i$

$N$ = set of all features

$S$ = subset of the features consisting of all the features in $N$ apart from $i$

$F_{S \cup \{i\}}$, $F_S$ = attributes of subset $S$, including and excluding feature $i$, respectively.

## III. RESULTS AND DISCUSSION

### III.A. Identification Results

First, the trained models were evaluated using test sets in the datasets. Figures 3 and 4 present the evaluation results on the test sets. Figure 3 depicts the accuracies of the averaged and individual labels of the RIID models for test sets A and B. Figure 4 presents an example of the identification results in the form of confusion matrices. The accuracy of the RIID models was observed to have increased with the increase in the number of samplings, as shown in Fig. 3. This implies that counting statistics correspond to the quality of the spectrum, and a deep learning model presents good performance when a high-quality dataset is used. The RIID accuracies for the RIs that emit relatively lower gamma-ray energies were poorer than others for individual labels, as shown in Figs. 3 and 4. This can be attributed to the truncation of the low-energy region. Among the 512 channels, or energy bins, in the spectra, the first 12 channels were truncated owing to large noise counts. This truncation may result in a loss of informative features in their spectra as the spectra for low-energy RIs could have effective counts in these low-energy regions.

Second, the performance of the RIID models was verified using the experimental data. Datasets were generated using the spectra measured in the configured experimental setup described in Sec. II.A.1; each RI was placed 1.25 cm way from the detector window, and the measurement period was fixed at 1 h to obtain spectra with good counting statistics. The background spectra were also measured 10 times during

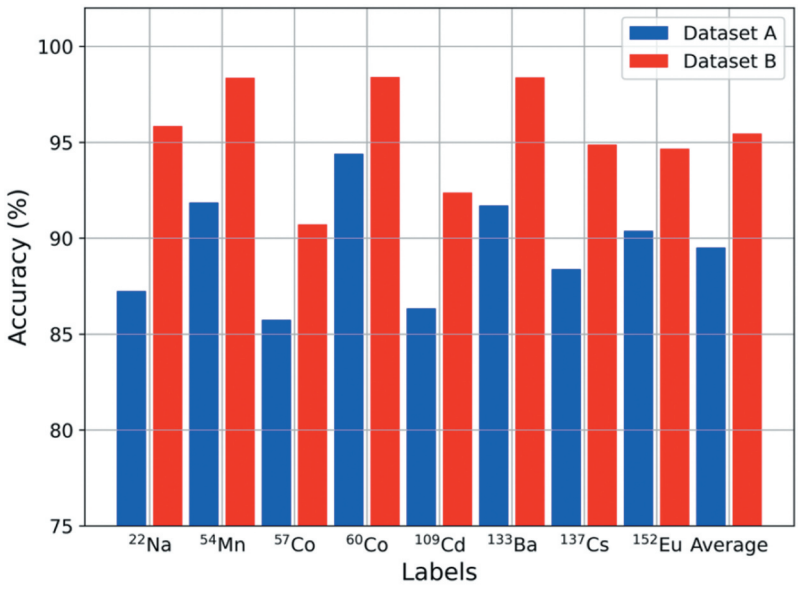


Fig. 3. Individual and averaged accuracies of two RIID models for each test set.

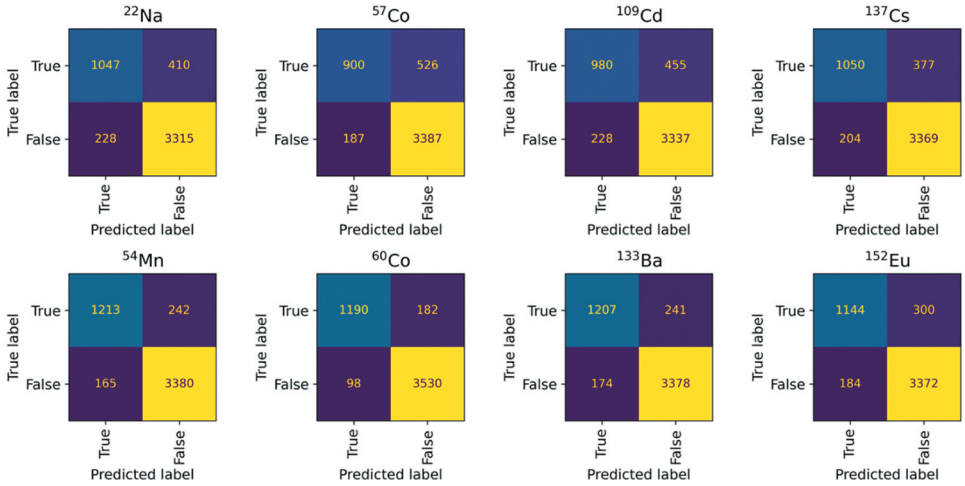


Fig. 4. Confusion matrices of an RIID model for test set A.

identical measurement periods, and the average was used to subtract the background contributions in the spectrum of each RI. The background-subtracted spectra were then divided by their total counts and used as PDFs for dataset generation. A dataset was generated from the experimental PDFs using the procedure depicted in Fig. 1. Subsequently, 800 spectra were generated under the conditions listed in Table II. Figure 5 depicts the individual and averaged accuracies of the RIID model for the experimental data.

The obtained results (the increment of accuracy with the increment of the number of counts and poor accuracies for labels of low-energy RIs) were similar to the evaluation results of the test set, but the overall accuracies were decreased. This decrease can be attributed to the difference between the simulated and the experimental spectra. We attempted to simulate the spectrum of a plastic scintillation detector as closely as possible; however, the simulated spectrum cannot be identical to the measured spectrum as the simulation results vary based on the simulation setup (e.g., material definition and cross-section library). Furthermore, the channel energy calibration cannot be conducted perfectly. Nevertheless, it was confirmed that the average accuracies of the RIID models were over 85% for all types of experimental data.

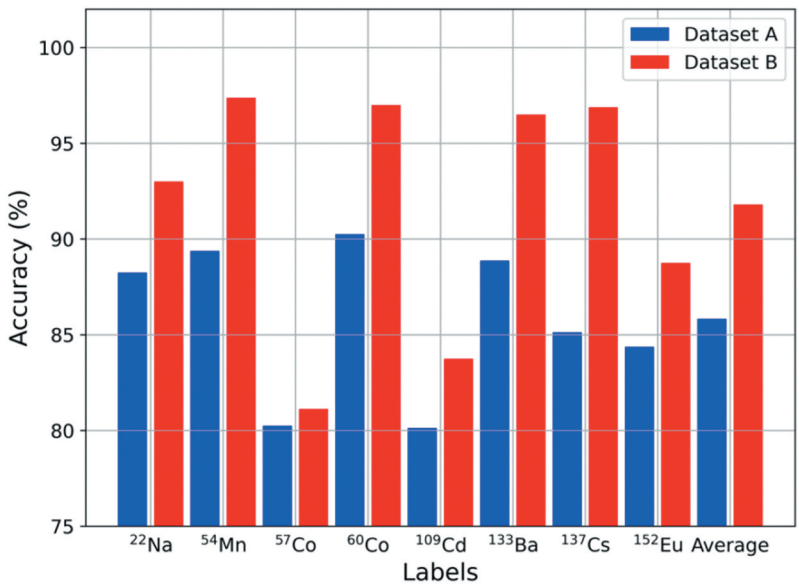


Fig. 5. Individual and averaged accuracies of two RIID models for different types of experimental data.

## III.B. RIID Model Explanation

The feature attributions of the RIID models for the two types of experimental data were calculated using XAI methods.

### *III.B.1. Comparison of XAI Methods*

The trained RIID models were analyzed by implementing XAI methods for the measured spectra of a single RI: $^{137}$Cs. The Compton maximum (CM), which is the broadened peak near the Compton edge (CE) caused by the energy broadening effect, is the only factor that can determine the existence of an RI as full-energy peaks do not appear in the spectrum of a plastic scintillation detector.[29–31] Therefore, the spectrum of $^{137}$Cs presented a broadened peak shape near its CE energy of 0.478 MeV, which proves the existence of $^{137}$Cs. This information is called domain knowledge in applied deep learning research.

Figure 6 depicts the spectra of $^{137}$Cs in the experimental data with good counting statistics and model explanation results obtained by the XAI methods. The results of the IG, LRP, and SHAP methods indicate that the energy bins near the CM have higher attributions than other regions; however, Grad-CAM and LIME failed to output reasonable feature attributions, as shown in Figs. 6a through 6e. Among the successful XAI methods, the SHAP results presented the most concentrated feature attribution near the CM

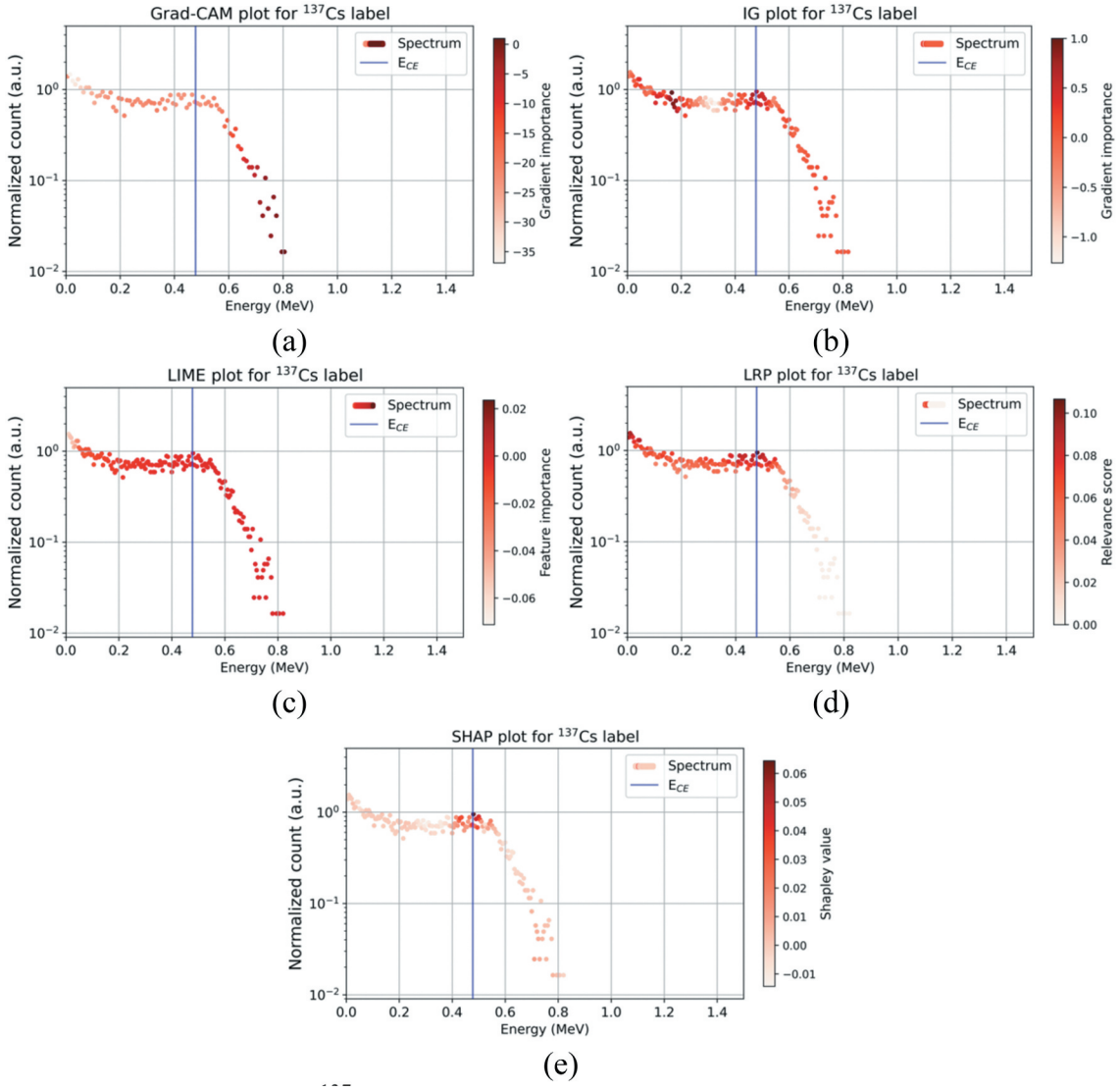


Fig. 6. Spectra of $^{137}$Cs and their explanation results of XAI methods for RIID model and experimental data type B. Each subplot shows results of (a) Grad-CAM, (b) IG, (c) LIME, (d) LRP, and (e) SHAP.

energy. In the case of the IG and LRP methods, attribution maps were activated in energy bins near the CM, along with energy bins in other regions, which could be uninformative. These results can be explained by the fact that the RIID model was able to determine the existence of $^{137}$Cs with attention to the input date near the CM and the domain knowledge of the RIID task.

### *III.B.2. Explanation of RIID Model Using SHAP*

Detailed explanations of the RIID models were achieved using the SHAP method as it presented the most reliable results when compared to other XAI methods. The RIID results from the models were analyzed for the measured spectra of the individual RIs. Figures 7 and 8 depict the spectra of eight RIs with SHAP values for RIID models and the experimental data types A and B, respectively. The figures show that most of the feature attributions of SHAP were accurately concentrated near the CE energy of each RI label regardless of the quality of the counting statistics as follows: low-energy regions for $^{57}$Co, $^{109}$Cd, and $^{133}$Ba; middle-energy regions for$^{22}$Na, $^{54}$Mn,$^{137}$Cs, and $^{152}$Eu; and high-energy regions for $^{22}$Na, $^{60}$Co, and $^{152}$Eu.

The RIID results from the models were also analyzed for the measured spectra of the multiple RIs. Figure 9 depicts three cases of spectra for two RIs that emit analogous energies of gamma rays with SHAP values for the RIID model and experimental data type A. In the low-energy case, the CE energies of $^{57}$Co are 0.001, 0.039, and 0.048 MeV, and that of $^{109}$Cd is 0.023 MeV. The feature attributions of the $^{109}$Cd label were activated in a lower-energy region than those of the $^{57}$Co label, as shown in Fig. 9a. The lowest CE energy of $^{57}$Co was not

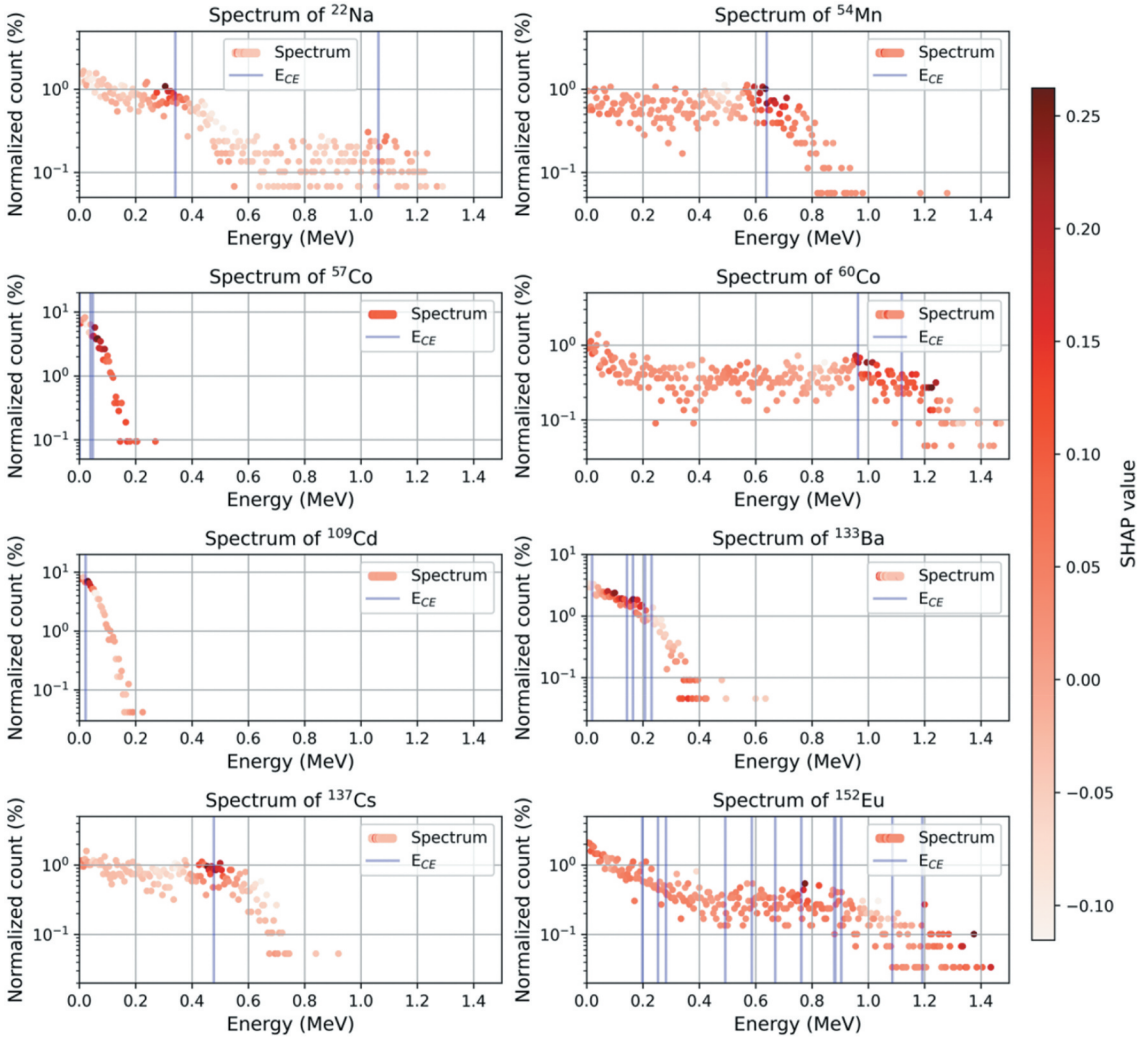


Fig. 7. Spectra of eight RIs with SHAP values for RIID model and experimental data type A.

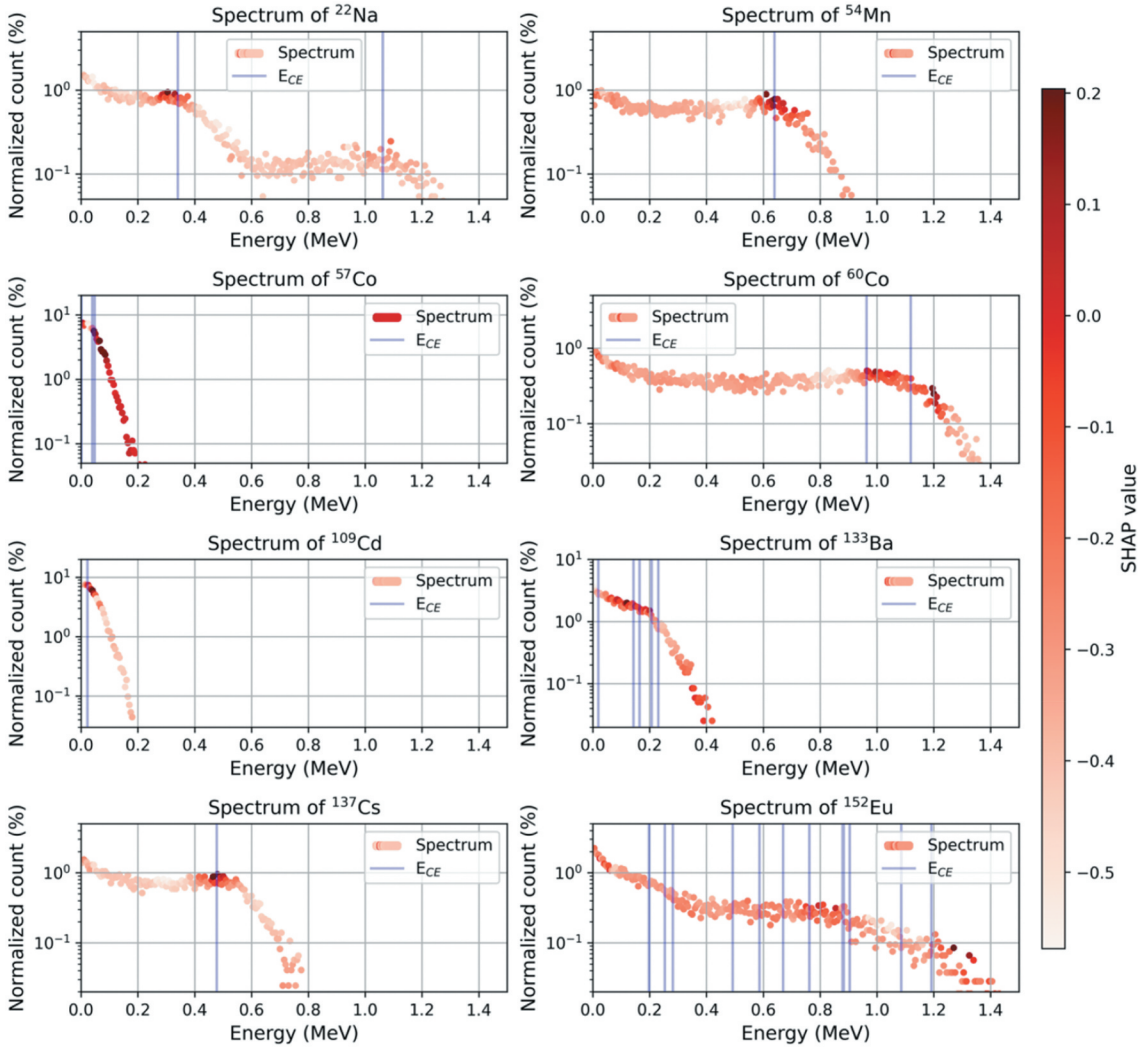


Fig. 8. Spectra of eight RIs with SHAP values for RIID model and experimental data type B.

activated because it was truncated by the LLD threshold. In the middle-energy case, the CE energies of $^{54}$Mn and $^{137}$Cs were 0.639 and 0.478 MeV, respectively. The activated feature attributes of the $^{54}$Mn label were higher than those of the $^{137}$Cs label, as shown in Fig. 9b. In the high-energy case, the CE energies of $^{60}$Co are 0.963 and 1.117 MeV, respectively. The two CEs appear as one CM in the spectrum of the plastic scintillation detector since the gap between the energies does not vary significantly. Despite the dozens of gamma rays emitted by $^{152}$Eu, the representative CE energies are 0.039, 0.119, 0.197, 0.587, 0.762, 0.904, and 1.192 MeV. Therefore, the $^{152}$Eu spectrum has several CMs in the low- to high-energy regions. Consequently, it can be stated that the feature attributions of the $^{60}$Co label must be high near a CM and that the feature attributions of the $^{152}$Eu label must be high near three CMs. The feature attributions of the $^{60}$Co label were high around 1 MeV, and those of the $^{152}$Eu label were high in three regions, around 0.05, 0.8, and 1.3 MeV, as shown in Fig. 9c.

### *III.B.3. Sanity Check of SHAP*

The adequacy of SHAP was evaluated using the sanity check.[32] The check consists of model parameter and data randomization tests and allows one to confirm whether explanation results are dependent on both model parameters and data or not. If results of the check are independent of both model parameters and data, this implies that the explanation method works just like a pattern extractor (e.g., edge detector or peak finder). For the model randomization test, the parameters of the trained RIID model were initialized using a random number generator. For the data randomization test, the RIID model was

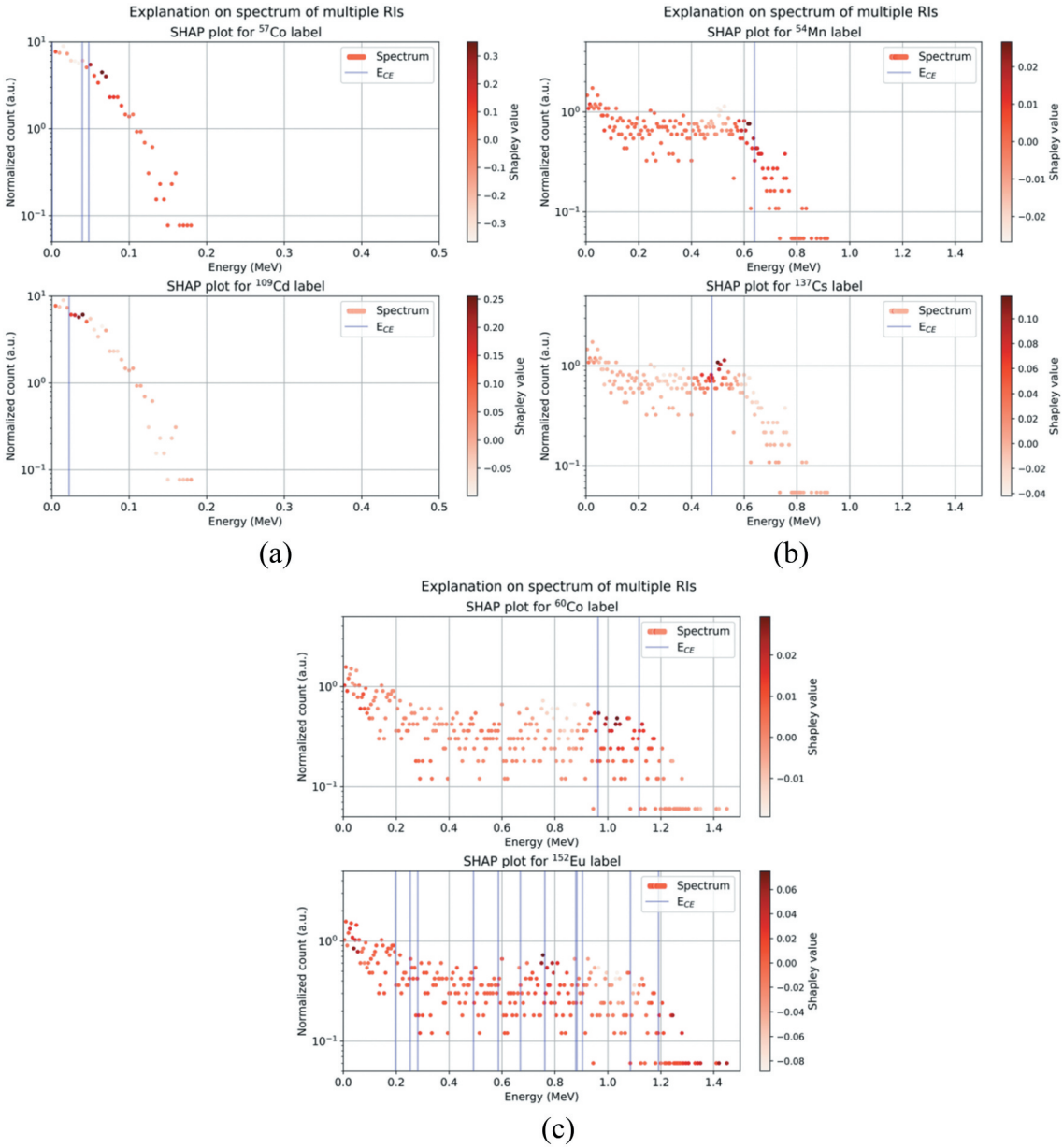


Fig. 9. Three RIID cases for spectra of two RIs with SHAP values for RIID model and experimental data type A. Each subplot shows results for (a) low-energy pair ($^{57}$Co and $^{109}$Cd), (b) middle-energy pair ($^{54}$Mn and $^{137}$Cs), and (c) high-energy pair ($^{60}$Co and $^{152}$Eu).

trained with permutated labels one after another, and it was confirmed that the trained RIID model with label permutation has a similar level of averaged accuracy with the original RIID model. Figure 10 depicts sanity checks for SHAP on trained model, model parameter, and data randomizations. As shown in Fig. 10, SHAP plots were varied for the randomization tests. This implies that the SHAP results are dependent on the data and model parameters and that SHAP found adequate feature attributions of the RIID model.

### III.C. Discussion

The XAI methods explain deep learning models in different ways. Their explanation results varied for identical models and inputs; the SHAP method presented the most reliable results as shown in Figs. 6 and 7. The major cause for this difference can be attributed to the capability of the SHAP method to calculate cooperative attributions. In the SHAP method, the feature attributions are obtained by measuring changes with and without individual features.

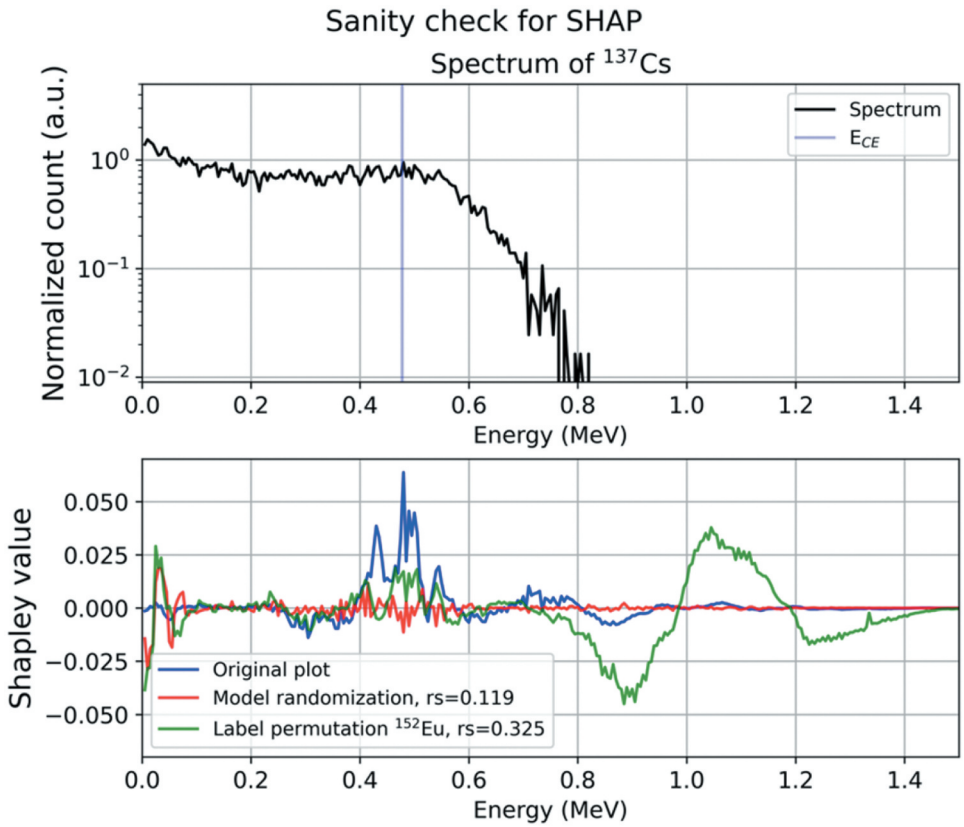


Fig. 10. Sanity check for SHAP on trained model, model parameter randomization, and data label randomization with Spearman rank correlation *rs*.

Therefore, the attributions that consider the dependency among features can be obtained; however, this approach involves large computational loads. Other methods calculate feature attributions by data backpropagation or surrogate modeling without considering the dependency among the features.

Another reason may be related to the types of the explainable data for each method. The SHAP method presented good explanation performance for deep learning models for various types of data, including radiation measurements,[15] whereas the major applications of gradient-based methods (Grad-CAM and IG) are models for visual data. LRP and LIME are also known to provide reliable explanation performance for visual data. Visual data contain various informative features, such as edges, contrasts in pixel values, and clear shapes of features. The performance of gradient-based methods was verified for such data, and they have been considered as one of the representative XAI methods to explain deep learning models for visual tasks. However, the quality of the input data used in this study was poorer than that of the visual data. When considering the spectrum from the perspective of visual data, it has no clear features, edges, or contrasts in the counting values.

Although the SHAP method showed the most reliable performance compared to the others to explain the RIID model for the spectrum of a plastic scintillation detector, this does not mean that SHAP always has better explanation performance than the others. The performance can be varied according to the type of input data as mentioned in Sec. II.C. For example, if the input spectrum is from inorganic or silicon detectors, which have a full-energy peak and good energy resolution, other XAI methods could show good performance even for the identical RIID model. Our result implies that when the result of an XAI method is inappropriate, it is necessary to review the deep learning model or prepared dataset, but cross-validation using other methods is also necessary.

We addressed the development, and explanation of the RIID models has been presented, but the development of a practical RIID application for plastic scintillation detectors still faces several limitations. First, the performance limitations must be analyzed. Specific patterns, regardless of clarity, must appear on a

spectrum for trained models to perform the RIID task; that is, the minimum radioactivity or the number of counts is required for RIID. This is associated with the minimum radioactivity for a single RI or the minimum relative radioactivity for multiple RIs. Second, robustness against spectral variation must be evaluated. A gain shift in the electronics of the detector system may be observed based on the experimental environment, causing a peak shift in the spectrum. This effect can be resolved through correction methods,[33–35] but minor shifts remain. Further studies must be conducted on these tasks for the development of practical RIID applications.

## IV. CONCLUSION

This study presented the development and explanation of a deep learning model for the RIID task of the spectrum of a plastic scintillation detector. The dataset generates an iterative random sampling procedure using the simulated spectra of the MCNP code. Various types of datasets with different numbers of counts were established to confirm the change in performance based on the total counts. We defined an RIID task as a multilabel binary classification problem in this study; CNN models were implemented and trained using established datasets. Experimental datasets were also generated using a random sampling procedure to determine the RIID performance for the experimental data. The experimental results demonstrate that the performance of the RIID models increased with increase in the total counts of the dataset. The RIID models were analyzed using XAI methods. Several methods exhibited good concurrence with domain knowledge, with the SHAP method presenting the most reliable results. The main limitations of this study include performance limitations and limitations of robustness against spectral variation. In the future, correction methods must be developed to resolve the gain shift effect in the electronics of the detector system.

## Disclosure Statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported by the Ministry of Oceans and Fisheries (KIMST project number 20200611).

## ORCID

Myungkook Moon http://orcid.org/0000-0003-4513-8217

## References

1. P. E. KELLER and R. T. KOUZES, "Gamma Spectral Analysis via Neural Networks," *Proc. 1994 IEEE Nuclear Science Symposium (NSS'94)*, Norfolk, Virginia, October 30–November 5, 1994, Vol. 1, p. 341, IEEE (1994); https://doi.org/10.1109/NSSMIC.1994.474365.

2. E. YOSHIDA et al., "Application of Neural Networks for the Analysis of Gamma-Ray Spectra Measured with a Ge Spectrometer," *Nucl. Instrum. Meth. Phys. Res. A*, **484**, *1–3*, 557 (2002); https://doi.org/10.1016/S0168-9002(01)01962-3.

3. J. KIM et al., "Quantitative Analysis of NaI (Tl) Gamma-Ray Spectrometry Using an Artificial Neural Network," *Nucl. Instrum. Meth. Phys. Res. A*, **944**, 162549 (2019); https://doi.org/10.1016/j.nima.2019.162549.

4. M. KAMUDA, J. STINNETT, and C. J. SULLIVAN, "Automated Isotope Identification Algorithm Using Artificial Neural Networks," *IEEE Trans. Nucl. Sci.*, **64**, *7*, 1858 (2017); https://doi.org/10.1109/TNS.2017.2693152.

5. G. DANIEL et al., "Automatic and Real-Time Identification of Radionuclides in Gamma-Ray Spectra: A New Method Based on Convolutional Neural Network Trained with Synthetic Data Set," *IEEE Trans. Nucl. Sci.*, **6**, *4*, 644 (2020); https://doi.org/10.1109/TNS.2020.2969703.

6. S. M. GALIB et al., "A Comparative Study of Machine Learning Methods for Automated Identification of Radioisotopes Using NaI Gamma-Ray Spectra," *Nucl. Eng. Technol.*, **53**, *12*, 4072 (2021); https://doi.org/10.1016/j.net.2021.06.020.

7. A. N. TURNER et al., "Convolutional Neural Networks for Challenges in Automated Nuclide Identification," *Sensors*, **21**, *15*, 5238 (2021); https://doi.org/10.3390/s21155238.

8. L. J. KANGAS et al., "The Use of Artificial Neural Networks in PVT-Based Radiation Portal Monitors," *Nucl. Instrum. Meth. Phys. Res. A*, **587**, *2–3*, 398 (2008); https://doi.org/10.1016/j.nima.2008.01.065.

9. J. FOMBELLIDA et al., "Neural Network Based Radioisotope Discrimination on Polyvinyl Toluene Radiation Portal Monitors," *Proc. 22nd Mediterranean Conf Control and Automation*, Palermo, Italy, June 16–19, 2014, p. 1099, IEEE (2014); https://doi.org/10.1109/MED.2014.6961521.

10. J. KIM, K. PARK, and G. CHO, "Multi-Radioisotope Identification Algorithm Using an Artificial Neural Network for Plastic Gamma Spectra," *Appl. Radiat. Isot.*, **147**, 83 (2019); https://doi.org/10.1016/j.apradiso.2019.01.005.

11. B. JEON et al., "Reconstruction of Compton Edges in Plastic Gamma Spectra Using Deep Autoencoder," *Sensors*, **20**, *10*, 2895 (2020); https://doi.org/10.3390/s20102895.

12. B. JEON et al., "Pseudo-Gamma Spectroscopy Based on Plastic Scintillation Detectors Using Multitask Learning," *Sensors*, **21**, *3*, 684 (2021); https://doi.org/10.3390/s21030684.

13. B. T. KOO et al., "Development of a Radionuclide Identification Algorithm Based on a Convolutional Neural Network for Radiation Portal Monitoring System," *Radiat. Phys. Chem.*, **180**, 109300 (2021); https://doi.org/10.1016/j.radphyschem.2020.109300.

14. M. GOMEZ-FERNANDEZ et al., "Isotope Identification Using Deep Learning: An Explanation," *Nucl. Instrum. Meth. Phys. Res. A*, **988**, 164925 (2021); https://doi.org/10.1016/j.nima.2020.164925.

15. J. RYU et al., "Development of Neural Network Model with Explainable AI for Measuring Uranium Enrichment," *IEEE Trans. Nucl. Sci.*, **68**, *11*, 2670 (2021); https://doi.org/10.1109/TNS.2021.3116090.

16. C. J. WERNER et al., "MCNP Version 6.2 Release Notes," LA-UR-18-20808, Los Alamos National Laboratory (2018).

17. R. J. MCCONN et al., "Compendium of Material Composition Data for Radiation Transport Modeling," Pacific Northwest National Laboratory (2011).

18. B. JEON et al., "Parametric Optimization for Energy Calibration and Gamma Response Function of Plastic Scintillation Detectors Using a Genetic Algorithm," *Nucl. Instrum. Meth. Phys. Res. A*, **930**, 8 (2019); https://doi.org/10.1016/j.nima.2019.03.003.

19. M. ABADI et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," TensorFlow (2015); arxiv.org/abs/1603.04467 (current as of Feb. 15, 2022).

20. D. P. KINGMA and J. BA, "Adam: A Method for Stochastic Optimization" (2014); arXiv Prepr. arXiv1412.6980 (current as of Feb. 15, 2022).

21. A. B. ARRIETA et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI," *Inf. Fusion*, **58**, *82*, 153 (2020); https://doi.org/10.1016/j.inffus.2019.12.012.

22. M. SAHAKYAN, Z. AUNG, and T. RAHWAN, "Explainable Artificial Intelligence for Tabular Data: A Survey," *IEEE Access*, **9**, 135392 (2021); https://doi.org/10.1109/ACCESS.2021.3116481.

23. P. LINARDATOS, V. PAPASTEFANOPOULOS, and S. KOTSIANTIS, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy*, **23**, *1*, 18 (2021); https://doi.org/10.3390/e23010018.

24. R. R. SELVARAJU et al., "Grad-Cam: Visual Explanations from Deep Networks via Gradient-Based Localization," *Proc. IEEE Int. Conf. Computer Vision*, Venice, Italy, October 22–29, 2017, p. 618, IEEE (2017); https://doi.org/10.1109/ICCV.2017.74.

25. M. SUNDARARAJAN, A. TALY, and Q. YAN, "Axiomatic Attribution for Deep Networks," *Proc. 34th Int. Conf. Machine Learning (PMLR 70)*, Sydney, Australia, August 6–11, 2017, p. 3319 (2017).

26. M. T. RIBEIRO, S. SINGH, and C. GUESTRIN, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, California, August 13–17, 2016, p. 1135, Association for Computing Machinery (2016); https://doi.org/10.1145/2939672.2939778.

27. G. MONTAVON et al., "Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition," *Pattern Recognit.*, **65**, 211 (2017); https://doi.org/10.1016/j.patcog.2016.11.008.

28. S. M. LUNDBERG and S.-I. LEE, "A Unified Approach to Interpreting Model Predictions," *Proc. 31st Int. Conf. Advances in Neural Information Processing Systems*, Long Beach, California, December 4–9, 2017, p. 4768 (2017).

29. E. R. SICILIANO et al., "Energy Calibration of Gamma Spectra in Plastic Scintillators Using Compton Kinematics," *Nucl. Instrum. Meth. Phys. Res. A*, **594**, *2*, 232 (2008); https://doi.org/10.1016/j.nima.2008.06.031.

30. L. SWIDERSKI et al., "Measurement of Compton Edge Position in Low-Z Scintillators," *Radiat. Meas.*, **45**, *3–6*, 605 (2010); https://doi.org/10.1016/j.radmeas.2009.10.015.

31. S. ASHRAFI and M. G. GOL, "Energy Calibration of Thin Plastic Scintillators Using Compton Scattered γ-Rays," *Nucl. Instrum. Meth. Phys. Res. A*, **642**, *1*, 70 (2011); https://doi.org/10.1016/j.nima.2011.04.003.

32. J. ADEBAYO et al., "Sanity Checks for Saliency Maps," *Proc. 32nd Conf. Neural Information Processing Systems*, Montreal, Canada, December 2–8, 2018, p. 9525 (2018).

33. M. PALACZ et al., "A New Method of Shift and Gain Correction for Arbitrary Spectra," *Nucl. Instrum. Meth. Phys. Res. A*, **383**, *2–3*, 473 (1996); https://doi.org/10.1016/S0168-9002(96)00756-5.

34. L. STAVSETRA, E. A. HULT, and J. P. OMTVEDT, "Real-Time Gain Shift Correction for On-Line Alpha-Liquid Scintillation Spectroscopy," *Nucl. Instrum. Meth. Phys. Res. A*, **551**, *2–3*, 323 (2005); https://doi.org/10.1016/j.nima.2005.06.045.

35. R. CASANOVAS, J. J. MORANT, and M. SALVADÓ, "Temperature Peak-Shift Correction Methods for NaI (Tl) and LaBr3 (Ce) Gamma-Ray Spectrum Stabilisation," *Radiat. Meas.*, **47**, *8*, 588 (2012); https://doi.org/10.1016/j.radmeas.2012.06.001.