Highlights

Comparative study of machine learning and statistical methods for automatic identification and quantification in γ-ray spectrometry

Dinh Triem Phan, Jérôme Bobin, Cheick Thiam, Christophe Bobin

- Open-source benchmark with dataset, code and metrics for γ-ray spectrometry
- Radionuclide identification and quantification by CNN and spectral unmixing
- Hybrid machine learning spectral unmixing for spectral variability
- Statistical approach outperforms CNN for identification and quantification
- Statistical approach can control the false alarm rate

Comparative study of machine learning and statistical methods for automatic identification and quantification in γ-ray spectrometry

Dinh Triem Phan[a,*], Jérôme Bobin[b], Cheick Thiam[a], Christophe Bobin[a]

[a] *Université Paris-Saclay, CEA, List, Laboratoire national Henri Becquerel (LNE-LNHB), Palaiseau, 91120, France*
[b] *IRFU, CEA, Universite Paris-Saclay, Gif-sur-Yvette, 91191, France*

Abstract

During the last decade, a large number of different numerical methods have been proposed to tackle the automatic identification and quantification in γ-ray spectrometry. However, the lack of common benchmarks, including datasets, code and comparison metrics, makes their evaluation and comparison hard. In that context, we propose an open-source benchmark that comprises simulated datasets of various γ-spectrometry settings, codes of different analysis approaches and evaluation metrics. This allows us to compare the state-of-the-art end-to-end machine learning with a statistical unmixing approach using the full spectrum. Three scenarios have been investigated: (1) spectral signatures are assumed to be known; (2) spectral signatures are deformed due to physical phenomena such as Compton scattering and attenuation; and (3) spectral signatures are shifted (*e.g.*, due to temperature variation). A large dataset of 2.10^5 simulated spectra containing nine radionuclides with an experimental natural background is used for each scenario with multiple radionuclides present in the spectrum. Regarding identification performance, the statistical approach consistently outperforms the machine learning approaches across all three scenarios for all comparison metrics. However, the performance of the statistical approach can be significantly impacted when spectral signatures are not modeled correctly. Consequently, the full-spectrum statistical approach is most effective with known or well-modeled spectral signatures, while end-to-end machine learning is a good alternative when measurement conditions are uncertain for radionuclide identification. Concerning the quantification task, the statistical approach provides accurate estimates of radionuclide counting, while the machine learning methods deliver less satisfactory results.

Keywords: Gamma-ray spectrometry, Spectral variability, Hybrid algorithm, Machine Learning, Spectral unmixing, Convolutional neural networks

*Corresponding author
Email address: dinh-triem.phan@cea.fr (Dinh Triem Phan)

1. Introduction

Gamma-ray spectrometry is a widely used technique for identifying and quantifying γ-emitting radionuclides, with various applications in nuclear physics including nuclear security, environmental monitoring and radiological characterization for decommissioning of nuclear facilities. Specific applications, such as in situ environmental analysis following a nuclear accident or the detection of illicit nuclear material trafficking, require rapid and reliable radionuclide identification under challenging conditions, including short measurement durations (*i.e.,* low statistics). Meeting these demands necessitates methods that are accessible to non-experts and capable of making robust decisions. Traditional methods [1], such as peak-based regression assuming Gaussian noise statistics, rely heavily on expert intervention and struggle to address these challenges, particularly in scenarios involving complex mixtures of radionuclides or low count rates.

Recent advances in machine learning (ML) have been applied to γ-ray spectrometry, offering automated, end-to-end solutions to overcome existing challenges. Techniques such as convolutional neural networks (CNN) and multi-layer perceptrons (MLP) [2–10] utilize large training datasets of different radionuclide mixtures to automate the identification and quantification process. These methods take γ-spectra as input, the presence of radionuclides as output for identification and their mixing weights as output for quantification. ML approaches have been effectively applied across various detectors and applications, yielding robust results in radionuclide identification.

Besides this approach, the statistical method based on full-spectrum analysis with Poisson likelihood has also shown solid performance. A notable example is the full-spectrum unmixing technique, which models an observed γ-spectrum $y \in \mathbb{R}^M$ where M is the number of channels as a Poisson distribution of Xa:

$$y \sim \mathcal{P}(Xa) \tag{1}$$

Here, $X \in \mathbb{R}^{M \times N}$ is a matrix of the normalized spectral signature of all N radionuclides, including the natural background (Bkg). Each column of X represents the detector's characteristic response to γ-photon emissions from a specific radionuclide. The vector $a \in \mathbb{R}^N$ contains the corresponding counting for each radionuclide. A regularized Maximum Likelihood Estimation (MLE) is then employed to estimate the counting of all radionuclides. This method has been successfully applied to NaI(Tl) and HPGe detectors and demonstrated to give better performance in activity estimation and decision thresholds than traditional methods [11–14]. For instance, experimental results obtained on real aerosol filter samples revealed that this approach enables rapid detection of ^{137}Cs at extremely low activity levels (mBq) around 1.5 days after sampling, while traditional methods typically require around eight days [12, 13].

Most existing studies lack publicly available code and datasets, making it difficult to reproduce results and perform fair comparisons between methods. To address this issue, we introduce an open-source benchmark that includes the full implementation of ML and spectral unmixing methods, as well as a simulated dataset and standardized evaluation metrics. This benchmark establishes a

framework for fair and consistent comparison while allowing easy integration of new methods and flexible adaptation to various datasets and use cases.

The benchmark compares the ML and statistical unmixing approaches in three scenarios: (1) spectral signatures are assumed to be known, corresponding to well-defined measurement conditions; (2) spectral signatures are deformed due to physical phenomena such as Compton scattering and attenuation; and (3) gain shift (*e.g.*, due to temperature variations). The evaluation uses a large dataset of 200000 simulated spectra containing nine radionuclides with an experimental natural background (Bkg). These simulated spectra are generated by varying the number of radionuclides present and the counting associated with each radionuclide. Both methods are evaluated based on their performance in identifying and quantifying γ-emitting radionuclides. To ensure fairness in comparison, the CNN architecture and hyperparameters are optimized, and classification thresholds are calibrated to achieve a false positive rate close to the predefined value.

All materials for this work are available on the GammaBench GitHub repository. This includes the Geant4-simulated spectral signatures for all radionuclides across the three scenarios. The repository also provides the full codebase developed in Python using the open-source PyTorch library [15], covering data generation, ML architectures, hyperparameter optimization, pre-trained models and spectral unmixing algorithms.

The article is organized as follows:

- Section 2 provides a description of the dataset and the evaluation metrics for identification and quantification.
- Section 3 presents the end-to-end ML and statistical unmixing full-spectrum approaches
- Section 4 shows the results of these two approaches for three scenarios for identification and quantification.

2. Datasets and evaluation metrics

2.1. Datasets

Automatic identification and quantification in γ-ray spectrometry can be applied using different types of detectors, including CdTe, LaBr3, NaI(Tl) and HPGe detectors. In that context, NaI(Tl) detectors are widely used due to their low cost and high detection efficiency allowing short measurement time. As a result, this study utilizes a 3"×3" NaI(Tl) detector to compare ML and statistical unmixing methods. While the analysis focuses on this specific detector, the methodologies presented are readily adaptable to other detector types with different energy resolutions.

In general, training an ML model requires a large and representative dataset that captures all key parameters characterizing the data. In the context of radionuclide identification and quantification, this means including a wide range

of γ-ray spectra that account for both the diversity of radionuclides and their varying contributions. Obtaining such a dataset experimentally can be difficult and time-consuming due to the complexity of possible spectral combinations. To address this, a common approach in the literature is to simulate spectral signatures for various radionuclides using Monte Carlo codes such as Geant4 [16] or MCNP [17]. Synthetic spectra are then generated by following the Poisson distribution of a linear combination of normalized spectral signatures. The synthetic dataset is constructed by randomly selecting the number of radionuclides, the total counting (counting of all radionuclides) and the mixing weights representing each radionuclide's contribution.

Three datasets were created to address different scenarios: (1) ideal conditions with known spectral signatures, (2) deformed spectral signatures due to physical phenomena and (3) gain shift.

Scenario 1: Known spectral signatures. This scenario represents idealized conditions where measurement conditions are well-defined and spectral signatures are assumed to be known. The library of radionuclides has nine radionuclides covering a large range of energies between 20 keV and 2 MeV: ^{57}Co, ^{60}Co, ^{99m}Tc, ^{123}I, ^{131}I, ^{133}Ba, ^{137}Cs, ^{152}Eu and ^{241}Am. Most of these radionuclides are classical γ-emitters for handheld instruments or spectroscopy-based portal monitors (e.g., ANSI N42.38-2015 [18]) and widely used in the literature. Spectral signatures were generated from Geant4 [16], a Monte Carlo toolkit for simulating particle interactions with matter. A point source was located 20 cm from the 3"×3" NaI(Tl) detector. The energy resolution of the NaI(Tl) detector was considered (6.5% at 662 keV). Each spectral signature consists of 1024 channels, with a binning of 2 keV per channel and a low-energy cut-off of 20 keV.

Scenario 2: Deformed spectral signatures. This scenario illustrates the deformation of spectral signatures when a source is enclosed within a steel sphere of varying thickness. Although conceptually simple, it effectively captures the impact of physical phenomena-such as attenuation and Compton scattering-resulting from the interaction of γ-photon with the surrounding steel. The dataset is based on simulations presented in [19], where a point source was placed inside a steel sphere with thickness varying from 0.001 mm to 30 mm. In this case, 96 spectral signatures of each radionuclide were simulated by Geant4 for the NaI(Tl) detector. Due to the complexity of spectral variability, the library was reduced from 9 radionuclides in the previous scenario to 8 by excluding ^{131}I from the originally defined library. The evolution of the spectral signatures of ^{137}Cs as a function of the sphere's thickness is shown in Fig. 1.

Scenario 3: Gain shift. Another factor contributing to spectral variability is gain shift, which can be caused, for example, by temperature variation in scintillation detectors. This gain shift phenomenon has also been studied in recent literature for its impact on automatic radionuclide identification and quantification [2, 6, 7, 10]. For this scenario, the same radionuclides and spectral signatures as those used in the known-signatures case were applied. For

each radionuclide, Geant4 provides a list of deposited energies, and the spectral signature is represented as a histogram of these energies across a specific energy range. The shifted spectral signatures are described by the relation: $e_s = e_r \times (1 - \alpha)$ where e_s is the shifted energy, e_r is the energy of the reference spectral signature, and α is the shift factor. The shifted spectral signatures are then constructed based on these shifted energies. An example of spectral signatures of ^{137}Cs with different values of the shift factor α is shown in Fig. 1.

All three scenarios discussed above were used to create synthetic datasets. For each scenario, a dataset consisting of 200000 simulated spectra was generated through the linear combination of spectral signatures (including Bkg) and Poisson noise. In this work, Bkg was supposed to be known. The total counting was sampled from a log-uniform distribution ranging from 200 to 100000: $log(\sum a) \sim U(2 + log(2), 5)$. The number of present radionuclides in the simulated spectra was randomly chosen between 0 and 4. Bkg is always present in the spectrum and contributes at least 10%. The mixing weights $z = a/\sum(a)$ of the radionuclides were selected randomly, with the constraint that their sum equals one. The minimum counting for ^{60}Co, ^{137}Cs and ^{152}Eu is 100, while for other radionuclides, the minimum is 50. The dataset was then divided into training, validation, and test sets in a 64:16:20 ratio. For the spectral deformation scenario, the thicknesses of the steel spheres were randomly selected from 96 values ranging from 0.001 mm to 30 mm. In the case of spectral shifts, the shift factor was uniformly varied between -10% and 10%, accounting for a wide range of potential shifts.

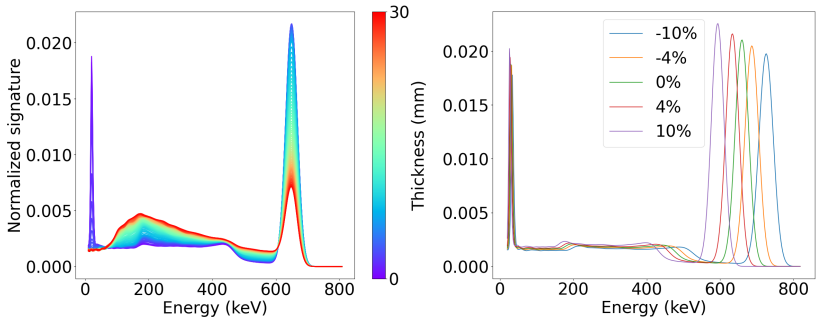

Fig. 1. The left panel (Scenario 2) shows spectral signatures of ^{137}Cs as a function of steel thickness, with deformation caused by γ-photon interactions with the surrounding sphere. The right panel (Scenario 3) displays spectral signatures of ^{137}Cs as a function of gain shift.

2.2. Evaluation metrics

2.2.1. Identification

For the identification task, let $(y_1, l_1), .., (y_K, l_K)$ represent the test set where y_i is the i-th γ-spectrum in test set and l_i is its associated label. The label $l_i = (l_{i,2}, .., l_{i,N})$ is defined such that $l_{i,j} = 1$ if the radionuclide j is present in the spectrum y_i and $l_{i,j} = 0$ otherwise. The index j starts from 2 since Bkg is always present in the simulated spectrum. Concerning the regression task,

(y_i, z_i) is the i-the γ-spectrum and its corresponding mixing weight. The mixing weight $z_i = (z_{i,1}, .., z_{i,N})$ is defined as the radionuclide counting divided by the total counting $z_{i,j} = a_{i,j} / \sum_{j=1}^{N}(a_{i,j})$.

Since radionuclide identification is a multi-label classification problem (multiple radionuclides (labels) can be present in a single γ-spectrum), the following metrics are used for performance evaluation:

- False positive rate (FPR) for all radionuclides: the proportion of cases where at least one radionuclide is incorrectly predicted to be present while it is absent.
- False negative rate (FNR) for all radionuclides: the proportion of cases where at least one radionuclide present in the radioactive source is not identified.
- Perfect prediction rate (PPR): the proportion of spectra in the dataset with no false positives or negatives. It is defined as: $PPR = (\sum_{i=1}^{K} \mathbb{1}_{\hat{l}_i = l_i})/K$ where $\hat{l}$ is the label predicted by the model.
- False prediction rate (FPrR) for all radionuclides: the proportion of spectra with any false predictions, including false positives, false negatives, or both. This is related to PPR as: $FPrR = 1 - PPR$
- Accuracy: for each radionuclide, the accuracy is defined as the ratio of correct predictions to the total number of samples. The accuracy for the multi-label problem is calculated as the average accuracy across all radionuclides: $acc = 1/(N-1) \times \sum_{j=2}^{N}(\sum_{i=1}^{K} \mathbb{1}_{\hat{l}_{i,j} = l_{i,j}})/K$
- Recall (probability of detection) for each radionuclide: proportion of present radionuclides that are correctly identified: $Re_j = \sum_{i=1}^{K} 1/K \times \mathbb{1}_{\hat{l}_{i,j}=1 | l_{i,j}=1}$

While additional metrics like precision and F1-score can also be used, they can be derived from the above-mentioned metrics and are not explicitly included in this comparison.

Let us notice that for multi-label classification, the interpretation of accuracy differs from that of binary or multi-class classification. Each radionuclide is associated with a confusion matrix, and accuracy is calculated as the average of the binary classification accuracy for all radionuclides. In this work, the number of radionuclides present in each mixture is randomly chosen between 0 and 4, with radionuclides selected at random from a dictionary of 9. As a result, each radionuclide is present in only around 20% of the cases, creating an imbalance among the labels. As a result, a high accuracy score can be misleading - for example, a naive model predicting the absence of all radionuclides would still achieve 80% accuracy.

2.2.2. Quantification

For quantification tasks, standard regression metrics such as relative error and mean squared error (MSE) are used to measure the difference between the estimated and true values of mixing weights or counting.

- MSE on mixing weights: $\sum_{i=1}^{K} 1/K \times \sum_{j=1}^{N} 1/N \times (\hat{z}_{i,j} - z_{i,j})^2$ where $\hat{z}$ is the estimated mixing weight.
- Relative absolute error on counting for present radionuclides: $\sum_{i=1}^{K} 1/K \times \sum_{j \in S(i)} |\hat{a}_{i,j} - a_{i,j}|/a_{i,j}/dim(S(i))$ where $S(i)$ is the set of present radionuclide for the ith spectrum.
- Relative error for individual radionuclide: The relative error is further analyzed for each radionuclide as a function of radionuclide counting.

3. Automatic identification and quantification methods in γ-ray spectrometry

3.1. Machine learning method

3.1.1. State of the art

Radionuclide identification can be considered as a multi-label classification problem in ML where multiple radionuclides (labels) can be present in a single γ-spectrum. This differs from multi-class classification, which assumes that each spectrum contains only one radionuclide. In the state of the art of ML, several approaches have been proposed to address this multi-label problem [20, 21].

One widely used approach is problem transformation, which converts the multi-label classification problem into simpler tasks. For instance, the binary relevance method treats each label independently, creating a binary classification problem for each radionuclide (in our case, each model predicts whether a specific radionuclide is present or absent). While straightforward, this method requires training multiple models and does not capture label dependencies (*e.g.,* correlation between radionuclides). Some works in γ-ray spectrometry used this technique are [2, 3]. Another problem transformation technique is Label Powerset, which converts the task into a single-label classification problem in which each unique combination of labels is treated as a single class. Although effective for small numbers of radionuclides, it becomes impractical with a large label set (*e.g.,* 2^N possible combinations for N labels (radionuclides)). This method has been employed in γ-ray spectrometry with four radionuclides in [4]. Less common transformation methods include Classifier Chains and Pairwise Transformation, which have yet to see significant application in γ-ray spectrometry.

The second major approach is algorithm adaptation, which modifies existing ML algorithms to handle multi-label classification directly. This approach has been applied to traditional ML methods such as K-nearest neighbors, logistic regression, decision trees and support vector machines. For deep learning models, a common method is to use an average of binary cross-entropy (BCE) over all radionuclides as the loss function with sigmoid activation for the output layer. This configuration allows the network to simultaneously account for all radionuclides and produce outputs between 0 and 1 for each radionuclide, which can then be used to identify radionuclides. This method has been widely employed in recent γ-ray spectrometry studies [5–7]. Additionally, in multi-label classification, recent studies have explored adaptations of loss functions in Deep

Learning such as Focal Loss [22] and Asymmetric Loss [23]. However, these new techniques have not yet been applied to γ-ray spectrometry.

While radionuclide identification benefits from the detection of peaks (similar to regions of interest), radionuclide quantification is challenging. Few studies in the literature have focused on γ-ray spectrometry quantification [7, 8, 10], which involves estimating mixing weights. This task is treated as a regression problem in ML, where common loss functions are mean square error (MSE) or BCE with softmax activation, ensuring that the sum of the mixing weights equals one. Quantification results can also be used for identification by applying appropriate thresholds [9, 10].

Concerning ML algorithms used in the γ-ray spectrometry literature, traditional ML methods have been employed previously. Recently, deep learning approaches such as multilayer perception (MLP) [4, 9, 10] and convolutional neural network (CNN) [2, 3, 5–8] have become increasingly popular. Among these, the CNN architecture has been demonstrated to provide better performance with high accuracy in radionuclide identification in recent studies and is then used in this work.

3.1.2. Machine learning methods in the benchmark

Two commonly used approaches for automatic identification in γ-ray spectrometry, binary relevance and multi-label classification with the BCE loss function, were evaluated in this work. For simplicity, multi-label classification applying the BCE loss function is referred to as "multi-label", and the binary relevance method is referred to as "mono-label" in the remainder of this study.

The architecture employed is CNN, which is composed of several convolutional layers followed by fully connected layers (FCL). This well-established architecture is commonly applied in ML for classification and has been widely adopted in γ-ray spectrometry research [3, 5–8]. The CNN structure is illustrated in Fig. 2. Each convolutional layer is designed with the following components:

- A 1D convolutional layer to extract useful features such as peaks and attenuation patterns.
- A max-pooling layer to reduce spatial dimensions.
- Batch normalization to help train faster and in a more stable manner.
- ReLU as the activation function.

The output of the convolutional layers is flattened and passed through several FCL (dense) layers. Dropout is applied after each FCL layer to regularize the model and avoid overfitting.

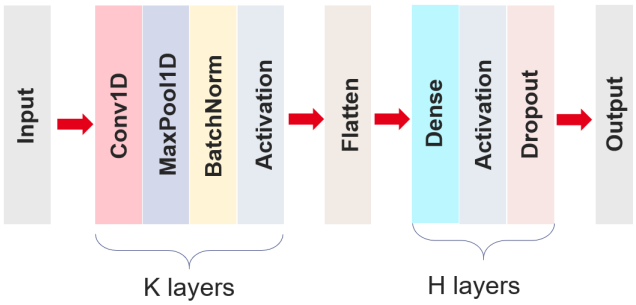

Fig. 2. The CNN architecture consists of several convolutional layers followed by fully connected layers.

The performance of CNN models is highly influenced by their hyperparameters, such as the number of convolutional and fully connected layers, filter sizes, learning rate, etc. Careful selection of these parameters is essential to achieve optimal results. In ML, this process-known as hyperparameter fine-tuning-typically involves testing various combinations to identify the most effective configuration. To achieve optimal ML models and ensure a fair comparison with the statistical approach, hyperparameters are fine-tuned using Ray Tune [24] on the validation dataset described in Section 2.1. For the mono-label approach, since N independent models are trained for N radionuclides in each scenario, fine-tuning all models would be computationally intensive. Therefore, the same hyperparameters optimized for the multi-label model were used, and only the models with significantly worse performance compared to the multi-label model were further fine-tuned. Additional details on training the ML models are provided on the GammaBench GitHub repository.

The performance of the ML models for the identification task highly depends on the threshold value that applies to the output of the neural network. Typically, this threshold is set to 0.5 [2–4, 6, 7], and when the output is greater than this value, the radionuclides are predicted to be present in the radioactive source. For a fair comparison with the statistical approach discussed in the next section, which depends on an expected false positive rate, this has been adjusted to achieve a false positive rate close to the desired value by applying the model to the validation set. In the following experiments, the expected FPR used in this work is set to 1%.

For the quantification task, the same CNN architecture is used, with the output representing the mixing weights of all radionuclides. The loss function is the BCE with the softmax activation. As in the identification task, hyperparameters are optimized using Ray Tune.

3.2. Spectral unmixing method

3.2.1. Quantification

In γ-ray analysis, the spectral unmixing method aims to decompose an observed spectrum y into individual spectral signatures of various radionuclides

and Bkg. Mathematically, an observed spectrum y is modeled by a Poisson distribution $y \sim \mathcal{P}(Xa)$, where, $X = [X_{Bkg}, X_2, .., X_N] \in \mathbb{R}^{M \times N}$ stands for the matrix of spectral signatures (including Bkg), and $a \in \mathbb{R}^N$ is the vector of counting. In this work, Bkg is supposed to be known.

The cost function can be defined as the negative log-likelihood of the Poisson mixture model described in Eq.1:

$$L(y, X, a) = \sum_{m=1}^{M} ((Xa)_m - y_m log((Xa)_m)) \tag{2}$$

When the spectral signatures X are known, the spectral unmixing problem involves estimating the non-negative vector a that minimizes the cost function L:

$$\hat{a} = \underset{a \geq 0}{\operatorname{argmin}}\, L(y, X, a) \tag{3}$$

To efficiently solve this problem, the non-negative Poisson unmixing (NNPU) algorithm [25] was developed, utilizing the multiplicative update rule.

However, in many practical scenarios, the spectral signatures X are unknown and can be modified depending on the measurement conditions. For instance, in the reference [19], spectral signatures are mostly deformed due to attenuation and Compton scattering when the radioactive source is placed in a steel sphere of varying thickness. Addressing such spectral variability requires building a surrogate model of the spectral signatures that captures their variability. To this end, a particular ML model called Interpolating AutoEncoder (IAE) [19, 26]) can be employed to model the spectral deformation, which can be derived from Geant4 simulations. The advantage of this IAE model is that it can be used in a generative way; the spectral signatures $X_{j \geq 2}$ can be modeled by a non-linear function of a latent variable λ learned by IAE: $X_{j \geq 2} \approx f(\lambda)\,; \lambda \in [0, 1]$. Mathematically, the spectral unmixing problem is reformulated to jointly estimate the latent variable λ (capturing spectral deformation) and the counting vector a that minimizes the cost function L :

$$\hat{\lambda}, \hat{a} = \underset{\lambda \in [0,1], a \geq 0}{\operatorname{argmin}}\; L(y, f(\lambda), a) \quad ; \quad \hat{X}_{j \geq 2} = f(\hat{\lambda}) \tag{4}$$

The semi-blind spectral unmixing (SEMSUN) algorithm [19] (Github) was developed to solve this above optimization problem.

For the spectral shift, X can be modeled by a non-linear function of a shift factor α: $X \approx g(\alpha)$. Therefore, the spectral unmixing problem is as follows:

$$\hat{\alpha}, \hat{a} = \underset{\alpha, a \geq 0}{\operatorname{argmin}}\, L(y, g(\alpha), a) \quad ; \quad \hat{X} = g(\hat{\alpha}) \tag{5}$$

This optimization problem is similar to the spectral variability case and can be solved using the same algorithm. Additional details are provided in Appendix A.

3.2.2. Identification

The quantification result can be directly used to identify the radionuclide by calculating the decision threshold based on radionuclide counting. However, minimizing the cost function in Eq.2 for the quantification problem often leads to noisy solutions. Indeed, additional radionuclides in the library, which are actually not present in the radioactive source, will tend to make the algorithm overfit the noise of the observed spectrum. This overfitting can lead to an under-estimation of the counting for all radionuclides and then lead to some false positive (alarm) detections.

In order to overcome these issues, extra regularisation is necessary. A common approach involves enforcing sparsity on the counting vector a. Penalization of the number of identified radionuclides is equivalently reformulated as penalizing the non-zero entries of the counting vector a. To that end, the optimization problem can be reformulated as:

$$\hat{\theta} = \underset{\theta \in C}{\operatorname{argmin}} L(y, \theta) \text{ subject to } ||a||_0 = K \tag{6}$$

where θ is (a) for known spectral signatures, θ is (λ, a) for spectral deformation or θ is (α, a) for spectral shift. The constraint space C includes all feasible parameter values θ, satisfying the conditions outlined in the previous sections.

To solve this problem, algorithms such as Poisson Orthogonal Matching Pursuit (P-OMP) [25] and Model selection with spectral variability based on manifold learning (MoSeVa)[27] have been developed for cases involving known or deformed spectral signatures. These algorithms implicitly enforce sparsity thanks to a greedy model selection procedure, where models with increasing dimensionality are tested, and a statistics-based criterion is employed to stop the selection procedure. By using the likelihood ratio test (LRT) [28], this approach can correctly identify the radionuclide with the ability to control the false positive rate (see [25] for more details). For the case of spectral shift, the problem formulation is similar to that of spectral deformation, allowing the same type of algorithm to be employed.

Note that in Scenarios 2 and 3, errors in modeling spectral signatures can lead to false positives at high statistics, although the corresponding radionuclide counting is very low. To keep the false positive rate close to the expected value, an additional constraint is introduced: a radionuclide is considered present only if it contributes at least 1% of the total counts $\sum_{i=2}^{N} a_i$ from all radionuclides, excluding background. This threshold is based on prior knowledge used during dataset generation.

4. Results

4.1. Identification performances

4.1.1. Scenario 1: Known spectral signatures

Table 1 shows the identification performance of the multi-label and mono-label CNN techniques, compared with the spectral unmixing method. The FPR

obtained by all methods is close to the expected value of 1%. Concerning the other metrics, the spectral unmixing method gives a lower FNR, as well as higher accuracy and PPR than the ML techniques. As a result, the unmixing method outperforms the ML methods in terms of identification performance.

Fig. 3 illustrates the FPR of all methods as a function of total counting and the number of present radionuclides. While the average FPR of the ML methods remains close to the expected level across the entire dataset, differences from the expected value appear when the total counting or the number of present radionuclides varies. In contrast, the spectral unmixing method, based on a statistical hypothesis test, maintains the FPR within the error bars of the expected value. Note that, to ensure a fair comparison, this work applies the same threshold to all radionuclides, with its value calibrated to keep the FPR on the validation set close to a expected value. However, the FPR still varies with the radionuclides present and their counting-an expected outcome, as false positives are influenced by the spectral composition. Achieving robust FPR control would require more adaptive thresholding methods that should take into account radionuclides and counting.

To provide a clearer understanding of the FNR, Fig. 4 shows the recall (1-FNR) of each radionuclide as a function of its counting. When the statistics are sufficient (counting greater than 1000 per radionuclide), both methods can accurately identify radionuclides. On the other hand, at lower statistical levels, the spectral unmixing method has a better detection capability, as it is based on physical properties such as the Poisson distribution and the linear nature of the mixing model.

	Accuracy	PPR	FPrR	FPR	FNR
CNN multi	99.49	95.53	4.47	0.94	3.60
CNN mono	99.44	95.11	4.89	0.99	3.96
Unmixing	99.68	97.19	2.81	0.98	1.88

Table 1: Identification performance for the test dataset for Scenario 1.

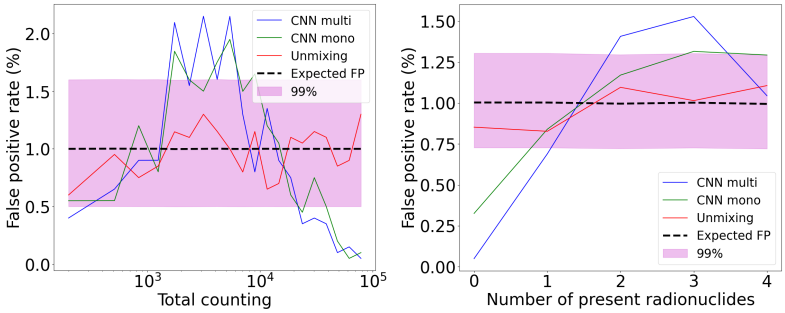

Fig. 3. Scenario 1: FPR as a function of total counting (left) and number of present radionuclides (right).

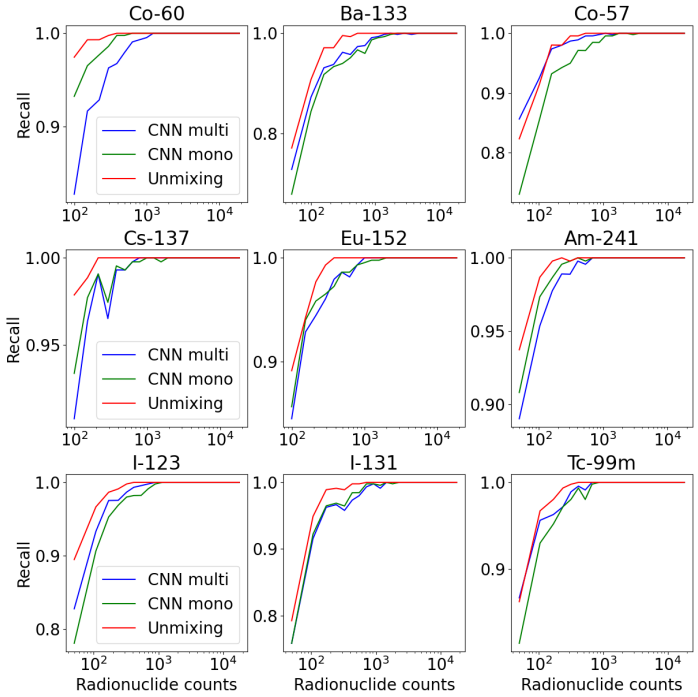

Fig. 4. Scenario 1: Recall (1-FNR) for each radionuclide as a function of radionuclide counts.

4.1.2. Scenario 2: Spectral deformation

The first numerical experiment examines the robustness of these methods in the presence of spectral variability. Specifically, the spectral signatures employed are the known (incorrect) signatures from Scenario 1 tested on a dataset including spectral deformation. Pre-trained CNN models based on the known spectral signatures from Scenario 1 and the P-OMP algorithm are employed for evaluation. The test dataset for this case introduces spectral variability by varying the steel thickness in a small range, from 0.001 mm to 1 mm. Even in this range, spectral signatures are considerably deformed, particularly for low-energy peaks that are absorbed by the attenuation phenomenon. Table 2 shows the result of both methods in this case. Since the spectral signatures used are incorrect and do not take into account the spectral deformation, the quality of the results for all methods decreases significantly. The spectral unmixing method is more sensitive than ML methods since it tries to fit the observed γ-spectrum using the individual spectral signatures. When the spectral signatures are incorrect, the spectral unmixing algorithm tends to fit the spectrum by the spectral signatures of non-present radionuclides, leading to a significantly high FPR. In contrast, ML techniques are similar to the region of interest, highly depending on the peak's position, making them less sensitive to incorrect spectral signatures.

To ensure more robust identification, spectral variabilities must be taken

into account. For this purpose, new CNN models are trained using the dataset described in Section 2.1, which includes spectral variability. To that end, the MoSeVa algorithm is employed instead of P-OMP to tackle the spectral unmixing task. Table 3 features the identification evaluation of these methods. Similar to Scenario 1, all methods achieve the FPR close to the expected values. Additionally, the spectral unmixing method achieves a lower FNR and higher accuracy and PPR than ML methods.

Fig. 5 illustrates the FPR of all methods as a function of total counting and the number of present radionuclides. When these parameters vary, the FPR obtained by ML methods occasionally exceeds the expected values. Conversely, the spectral unmixing method keeps the FPR below the expected value. Fig. 6 displays the recall of each radionuclide as a function of its counting. With sufficient statistics, both methods accurately identify radionuclides. At lower statistical levels, the spectral unmixing method demonstrates superior detection capability.

Identification performance for all thickness values and different counting levels are shown in Fig. 7. As the thickness increases, attenuation effects become more important, leading to an increase in FNR for all methods. With regard to the total counting, the FPR of the ML methods is higher at low statistical levels but decreases significantly when counts increase. Notably, high steel thickness values (greater than 20 mm) tend to be the most difficult scenarios for radionuclide identification for all methods.

	Accuracy	PPR	FPrR	FPR	FNR
CNN multi	97.94	85.04	14.96	4.53	11.02
CNN mono	98.31	87.55	12.45	3.62	9.27
Unmixing	97.20	82.82	17.18	12.22	5.86

Table 2: Scenario 2: Identification performance in the case of spectral variability with known incorrect spectral signatures. Steel thicknesses range from 0.001 mm to 1 mm.

	Accuracy	PPR	FPrR	FPR	FNR
CNN multi	98.74	90.68	9.32	0.92	8.62
CNN mono	98.66	90.15	9.85	1.15	8.86
Unmixing	99.12	93.55	6.45	0.89	5.70

Table 3: Scenario 2: Identification performance in the case of spectral variability with unknown spectral signatures. Steel thickness values range from 0.001 mm to 30 mm.

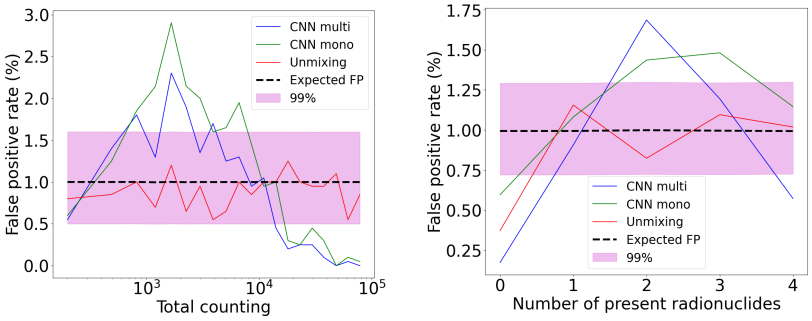

Fig. 5. Scenario 2: FPR as a function of total counting (left) and number of present radionuclides (right).

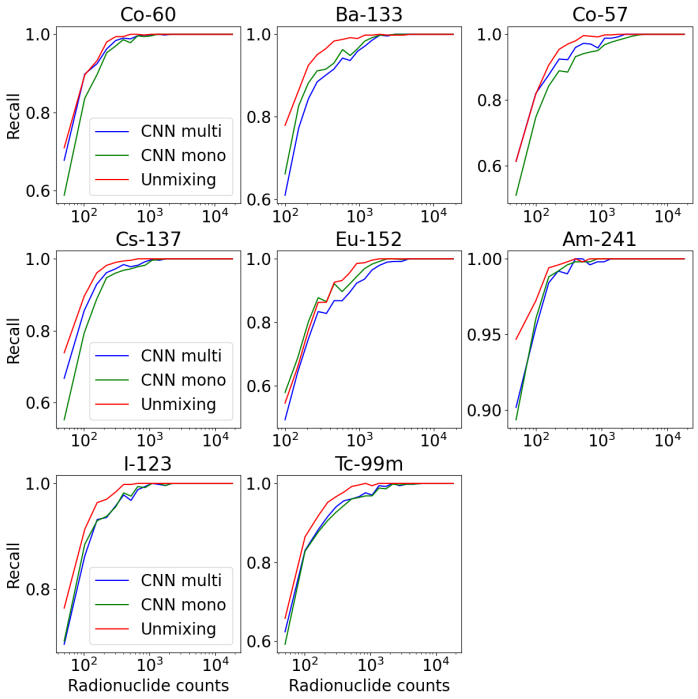

Fig. 6. Scenario 2: Recall (1-FNR) for each radionuclide as a function of radionuclide counts.

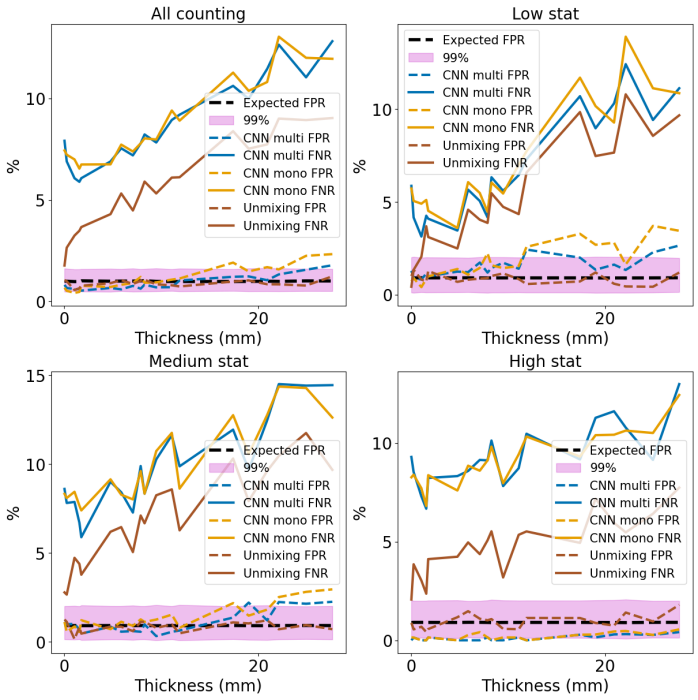

Fig. 7. Scenario 2: FPR and FNR for different thickness values at various counting levels. The top-left panel shows results for all counting levels combined. The other panels show results for low, medium, and high statistics corresponding to counting ranges of 200-3604, 3604-19596, and 19596-99973, respectively. Each range represents one-third of the dataset.

4.1.3. Scenario 3: Gain shift

The robustness of ML and spectral unmixing methods is evaluated with respect to gain shift. Specifically, the spectral signatures employed are the known (incorrect) signatures from Scenario 1 tested on a dataset that includes gain shifts. Pre-trained CNN models based on the known spectral signatures from Scenario 1 and the P-OMP algorithm are used for evaluation. The test dataset includes the spectral shift with the shift factor varying from -2 to 2%. Table 4 shows the results of both methods in this case. As the spectral signatures used are incorrect and do not take into account the spectral shift, all methods show a significant decrease in performance. The spectral unmixing method is more sensitive than the ML methods, similar to the behavior observed in Scenario 2.

To ensure robust identification performance, the spectral shift must be considered. To this end, new CNN models are trained using the dataset described in Section 2.1, which includes gain shifts. For the spectral unmixing method, the P-OMP algorithm with gain shift consideration described in Appendix A is employed, replacing the original P-OMP. Table 5 summarizes the identification results of these methods. Similarly to Scenario 1, all methods achieve an FPR close to the expected value, and the spectral unmixing method achieves a lower

FNR and higher accuracy and PPR than ML methods. Fig. 8 demonstrates that the spectral unmixing method allows the FPR to be lower than the expected value while the FPR obtained by ML methods occasionally exceeds the expected values. Fig. 9 shows the recall of each radionuclide as a function of its counting. At lower statistical levels, the spectral unmixing method demonstrates superior detection capability.

Identification performance for different shift factors and counting levels are shown in Fig. 10. High gain shift factors (close to 10%) present the most challenging cases for radionuclide identification for ML methods, as the spectral signatures are significantly modified.

	Accuracy	PPR	FPrR	FPR	FNR
CNN multi	98.21	85.52	14.48	10.64	4.46
CNN mono	97.82	82.49	17.51	13.72	4.59
Unmixing	95.15	67.40	32.60	31.10	2.22

Table 4: Scenario 3: Identification performance in the case of gain shift with known incorrect spectral signatures. The shift factor is between -2% and 2%.

	Accuracy	PPR	FPrR	FPR	FNR
CNN multi	99.27	93.75	6.25	1.11	5.35
CNN mono	99.19	93.13	6.87	1.04	6.00
Unmixing	99.60	96.59	3.41	0.90	2.60

Table 5: Scenario 3: Identification performance in the case of gain shift with unknown spectral signatures. The shift factor is between -10% and 10%.

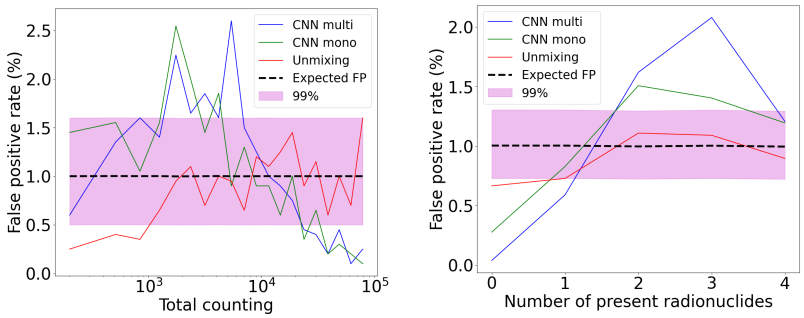

Fig. 8. Scenario 3: FPR as a function of total counting (left) and number of present radionuclides (right).

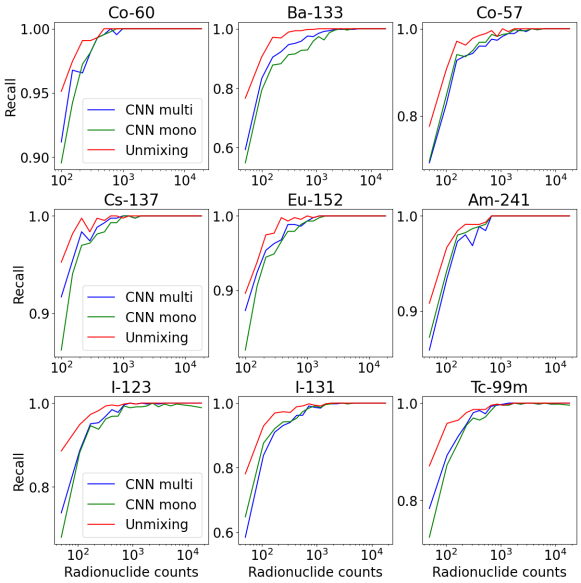

Fig. 9. Scenario 3: Recall (1-FNR) as a function of radionuclide counts.

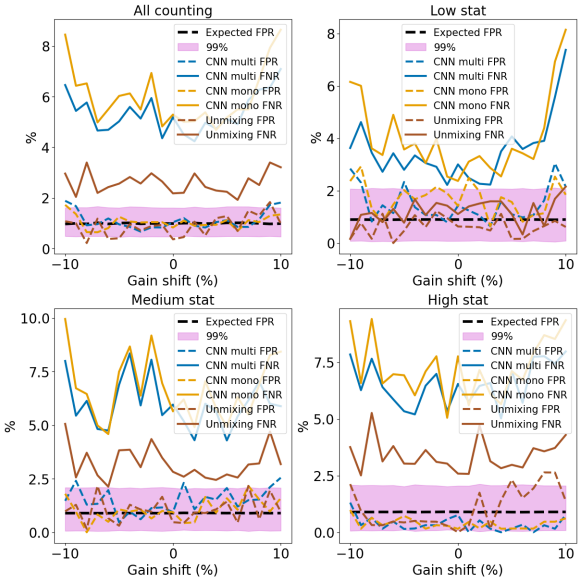

Fig. 10. Scenario 3: FPR and FNR for different gain shift factors with different counting levels. The top-left panel shows results for all counting levels combined. The other panels show results for low, medium, and high statistics, respectively.

4.2. Quantification performance

Table 6 shows the quantification metrics MSE and RAE for both the ML and spectral unmixing methods across the three scenarios. The spectral unmixing method outperforms the ML method, achieving the lowest error for all metrics. Fig.11, 12 and 13 display the relative error of estimated counting for each radionuclide in the three scenarios. The spectral unmixing method consistently provides estimated radionuclides counting with means close to the expected values and very low variability. In contrast, the ML method exhibits significant bias and high variability in its estimates across all scenarios. For instance, for Scenario 1 illustrated in Fig. 11, the mean of relative error of spectral unmixing method is close to 0 with 90% of relative error falling below 1.5% when the radionuclide counting is high (10^5) and below 10% when the radionuclide counting is greater than 1000. For the ML method, the estimated counting is less accurate, with the mean of relative error being 6% and 90% of relative error falling below 11% even at radionuclide counting of 10^5.

Similar to the robustness evaluation in Section 4.1, the quantification performance is also calculated using the known incorrect spectral signatures, as shown in Table 7. As demonstrated earlier, the spectral unmixing method is more affected than the ML method for the identification task when the spectral signatures are incorrectly modeled. However, for the quantification task, the spectral unmixing method outperforms the ML method, yielding lower errors for all metrics and scenarios.

These results highlight a key limitation of CNNs for quantification. While radionuclide identification benefits from the detection of peaks (similar to regions of interest), radionuclide quantification presents a more complex challenge. Unlike identification, which involves classification (presence or absence), quantification requires accurate estimation of continuous values. For identification, CNNs are trained with binary targets (0 or 1), and a threshold is applied to their outputs to determine radionuclide presence. For quantification, the raw CNN outputs are used directly to estimate the mixing weights. However, due to prediction errors, CNN outputs can be biased, even at high statistics, leading to inaccurate estimates of radionuclide contributions.

	Scenario 1		Scenario 2		Scenario 3	
	MSE(1e-5)	RAE(%)	MSE(1e-5)	RAE(%)	MSE(1e-5)	RAE(%)
CNN	14.1	10.24	29.5	13.21	20.1	12.27
Unmixing	3.3	4.42	11.8	7.46	4.9	5.03

Table 6: Quantification performances.

	Scenario 2		Scenario 3	
	MSE(1e-5)	RAE(%)	MSE(1e-5)	RAE(%)
CNN	163.2	21.11	28.4	13.65
Unmixing	82.7	14.61	7.8	6.89

Table 7: Quantification performance used known incorrect spectral signatures.

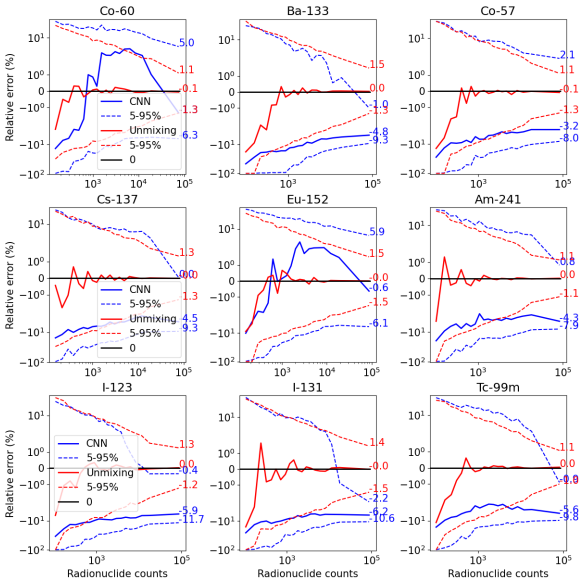

Fig. 11. Scenario 1: Relative error of the estimated counting for each radionuclide. For each radionuclide, the 5-95% range represents the interval between the 5th and 95th percentiles of the relative errors, with the continuous line indicating the average value.

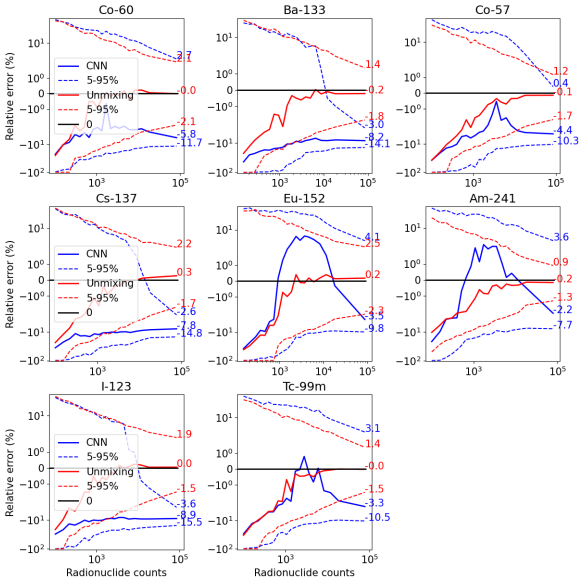

Fig. 12. Scenario 2: Relative error of estimated counting for each radionuclide.

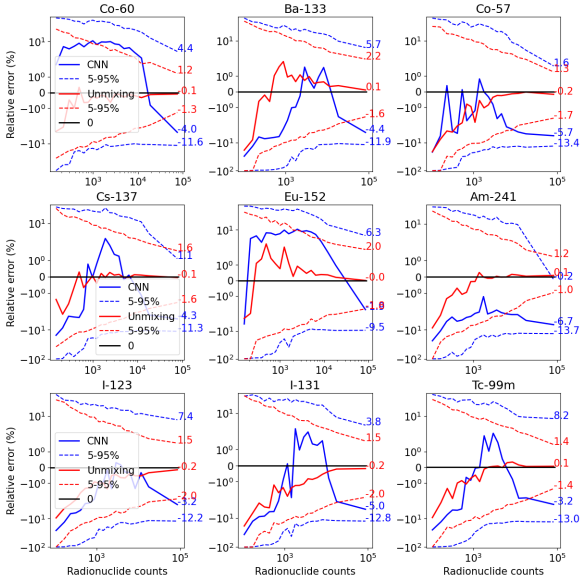

Fig. 13. Scenario 3: Relative error of estimated counting for each radionuclide.

5. Conclusion

This work introduces an open-source benchmark with data, code and metrics to compare the automatic identification and quantification performance of the end-to-end ML and statistical unmixing methods in γ-ray spectrometry. These methods are evaluated across three scenarios: (1) spectral signatures are assumed to be known, (2) spectral signatures are deformed due to physical phenomena like Compton scattering and attenuation, and (3) gain shift. A library of Geant4 simulated spectral signatures for nine radionuclides, alongside an experimental natural background, is utilized in this study.

Regarding identification performance, the statistical unmixing approach consistently outperforms the ML approaches across all three scenarios in terms of accuracy and false prediction rate, especially at low statistics. Additionally, the statistical method effectively controls the false alarm rate, keeping it close to the predefined value, which is less robust for ML methods. However, the performance of the statistical approach can be impacted when spectral signatures are not modeled correctly.

For quantification, the statistical approach provides accurate estimates of radionuclide counting or mixing weights, while the ML methods deliver less satisfactory results. When measurement conditions are not well-defined or difficult to model, the statistical approach still yields better outcomes than ML.

In conclusion, the statistical unmixing approach is most effective for the identification task when spectral signatures are known or spectral variability is correctly modeled. In cases where measurement conditions are not well-defined or challenging to model, end-to-end ML provides a suitable alternative. For quantification tasks, the statistical method is the preferred choice.

Appendix A. Algorithm for gain shift

Let $s^{(s)}$ denote the shifted spectral signature of a radionuclide corresponding to a gain shift factor α, and $s^{(r)}$ represents the reference spectral signatures (without shift). In this work, the form of relation function that induces the spectral shift is assumed to be known $e^{(s)} = h(e^{(r)}, \alpha) = e^{(r)} \times (1 - \alpha)$; only the parameter α needs to be estimated. Since this relation applies to energy rather than channel, we can construct an 'artificial' list with associated energies from the reference spectral signature $s^{(r)}$.

- $s_i^{(r)}$ is the value for channel i of reference spectral signature, corresponding to the energy range $[2i + 20, 2i + 2 + 20]$ (binning of 2 keV per channel and a low-energy cut-off of 20 keV)
- The normalized spectrum $s^{(r)}$ is multiplied by a large number (*e.g,* 10^6) to bring the spectrum down to high statistics: $s^{(*)} = s^{(r)} \times 10^6$. In this case, $s_i^{(*)}$ can be interpreted as the number of times the energy falls within the interval $[2i + 20, 2i + 20 + 2]$.

- For each channel i, an energy list is defined as $\{2i + 20, 2i + 20 + d, 2i + 20 + 2d, .., 2i + 20 + 2\}$, where the number of points corresponds to $s_i^{(*)}$.
- An energy list $\{\hat{e}_j^{(r)}\}$ can be obtained by cumulating all channels.

Using this energy list derived from the reference spectral signature, the shift function can be applied to compute the shifted energy list $\{e_j^{(s)}\}$: $e_j^{(s)} = h(\hat{e}_j^{(r)}, \alpha)$. The spectral signature is then obtained by constructing a histogram from this list. Consequently, the shifted spectral signatures of all radionuclides $X^{(s)}$ can be modeled as a function of α: $X^{(s)} = g(X^{(r)}, \alpha)$.

This optimization problem is similar to the case of spectral deformation. To address this, the BCD minimization scheme implemented in SEMSUN is used. It is important to note that this problem is non-differentiable because it involves converting the energy list into a histogram (bin $= \sum 1_{2i<e_j<2i+2}$). The Nelder-Mead algorithm [29], which relies on direct search methods, is well-suited to solving this without requiring derivatives.

References

[1] G. Gilmore, Practical gamma-ray spectroscopy, John Wiley & Sons, 2008.

[2] D. Liang, P. Gong, X. Tang, P. Wang, L. Gao, Z. Wang, R. Zhang, Rapid nuclide identification algorithm based on convolutional neural network, Annals of Nuclear Energy 133 (2019) 483–490.

[3] G. Daniel, F. Ceraudo, O. Limousin, D. Maier, A. Meuris, 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 Transactions on Nuclear Science 67 (4) (2020) 644–653.

[4] J. Kim, K. Park, G. Cho, Multi-radioisotope identification algorithm using an artificial neural network for plastic gamma spectra, Applied Radiation and Isotopes 147 (2019) 83–90.

[5] B. T. Koo, H. C. Lee, K. Bae, Y. Kim, J. Jung, C. S. Park, H.-S. Kim, C. H. Min, Development of a radionuclide identification algorithm based on a convolutional neural network for radiation portal monitoring system, Radiation Physics and Chemistry 180 (2021) 109300.

[6] A. N. Turner, C. Wheldon, T. K. Wheldon, M. R. Gilbert, L. W. Packer, J. Burns, M. Freer, Convolutional neural networks for challenges in automated nuclide identification, Sensors 21 (15) (2021) 5238.

[7] W. Kim, K. Ko, J. Park, S. Lee, H. Yun, G. Cho, Deep learning-based gamma spectroscopic analysis considering multiple variables for in situ applications, Radiation Physics and Chemistry 226 (2025) 112261.

[8] M. Kamuda, J. Zhao, K. Huff, A comparison of machine learning methods for automated gamma-ray spectroscopy, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 954 (2020) 161385.

[9] M. Kamuda, J. Stinnett, C. Sullivan, Automated isotope identification algorithm using artificial neural networks, IEEE Transactions on Nuclear Science 64 (7) (2017) 1858–1864.

[10] M. Kamuda, C. J. Sullivan, An automated isotope identification and quantification algorithm for isotope mixtures in low-resolution gamma-ray spectra, Radiation Physics and Chemistry 155 (2019) 281–286.

[11] J. Xu, J. Bobin, A. de Vismes Ott, C. Bobin, Sparse spectral unmixing for activity estimation in γ-ray spectrometry applied to environmental measurements, Applied Radiation and Isotopes 156 (2020) 108903.

[12] P. Malfrait, J. Bobin, A. de Vismes Ott, Spectral unmixing of multitemporal data in gamma-ray spectrometry, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 1045 (2023) 167547.

[13] P. Malfrait, J. Bobin, A. de Vismes Ott, Online spectral unmixing in gamma-ray spectrometry, Applied Radiation and Isotopes 201 (2023) 111011.

[14] C. Mano, C. Chapelle, A. D. M. Kabakian, P. Gross, H. Paradis, O. Delaune, L. Patryl, Algorithm development for low level radioxenon 2D spectra analysis: A first case of study using spectral unmixing for a β-γ detector, Applied Radiation and Isotopes 203 (2024) 111079.

[15] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035.

[16] S. Agostinelli, J. Allison, K. a. Amako, J. Apostolakis, H. Araujo, P. Arce, M. Asai, D. Axen, S. Banerjee, G. Barrand, et al., GEANT4—a simulation toolkit, Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment 506 (3) (2003) 250–303.

[17] R. Forster, T. Godfrey, MCNP-a general Monte Carlo code for neutron and photon transport, in: Monte-Carlo Methods and Applications in Neutronics, Photonics and Statistical Physics: Proceedings of the Joint Los Alamos National Laboratory-Commissariat à l'Energie Atomique Meeting Held at Cadarache Castle, Provence, France April 22–26, 1985, Springer, 2006, pp. 33–55.

[18] American national standard for performance criteria for spectroscopy-based portal monitors used in homeland security, ANSI N42.38-2015 (Revision of ANSI N42.38-2006) (2016) 1–59.

[19] D. T. Phan, J. Bobin, C. Thiam, C. Bobin, A hybrid machine learning unmixing method for automatic analysis of γ-spectra with spectral variability, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 1060 (2024) 169028.

[20] G. Tsoumakas, I. Katakis, Multi-label classification: An overview, Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications (2008) 64–74.

[21] J. Read, F. Perez-Cruz, Deep learning for multi-label classification, arXiv preprint arXiv:1502.05988 (2014).

[22] T. Lin, Focal loss for dense object detection, arXiv preprint arXiv:1708.02002 (2017).

[23] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, L. Zelnik-Manor, Asymmetric loss for multi-label classification, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 82–91.

[24] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, I. Stoica, Tune: A research platform for distributed model selection and training, arXiv preprint arXiv:1807.05118 (2018).

[25] R. André, C. Bobin, J. Bobin, J. Xu, A. de Vismes Ott, Metrological approach of γ-emitting radionuclides identification at low statistics: application of sparse spectral unmixing to scintillation detectors, Metrologia 58 (1) (2021) 015011.

[26] J. Bobin, R. C. Gertosio, C. Bobin, C. Thiam, An autoencoder-based model for learning regularizations in unmixing problems, Digital Signal Processing (2023) 104058.

[27] D. T. Phan, J. Bobin, C. Thiam, C. Bobin, Automatic identification and quantification of γ-emitting radionuclides with spectral variability using a hybrid machine learning unmixing method, Radiation Physics and Chemistry 232 (2025) 112654.

[28] S. S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses, The annals of mathematical statistics 9 (1) (1938) 60–62.

[29] J. C. Lagarias, J. A. Reeds, M. H. Wright, P. E. Wright, Convergence properties of the Nelder–Mead simplex method in low dimensions, SIAM Journal on optimization 9 (1) (1998) 112–147.