Nuclear Science and Techniques (2023) 34:154
https://doi.org/10.1007/s41365-023-01304-1

Novel algorithm for detection and identification of radioactive materials in an urban environment

Hao-Lin Liu[1,2] · Hai-Bo Ji[1] · Jiang-Mei Zhang[3] · Jing Lu[4] · Cao-Lin Zhang[2] · Xing-Hua Feng[2]

Received: 15 May 2023 / Revised: 20 June 2023 / Accepted: 2 July 2023 / Published online: 27 October 2023

Abstract

This study introduces a novel algorithm to detect and identify radioactive materials in urban settings using time-series detector response data. To address the challenges posed by varying backgrounds and to enhance the quality and reliability of the energy spectrum data, we devised a temporal energy window. This partitioned the time-series detector response data, resulting in energy spectra that emphasize the vital information pertaining to radioactive materials. We then extracted characteristic features of these energy spectra, relying on the formation mechanism and measurement principles of the gamma-ray instrument spectrum. These features encompassed aggregated counts, peak-to-flat ratios, and peak-to-peak ratios. This methodology not only simplified the interpretation of the energy spectra's physical significance but also eliminated the necessity for peak searching and individual peak analyses. Given the requirements of imbalanced multi-classification, we created a detection and identification model using a weighted k-nearest neighbors (KNN) framework. This model recognized that energy spectra of identical radioactive materials exhibit minimal inter-class similarity. Consequently, it considerably boosted the classification accuracy of minority classes, enhancing the classifier's overall efficacy. We also executed a series of comparative experiments. Established methods for radionuclide identification classification, such as standard KNN, support vector machine, Bayesian network, and random tree, were used for comparison purposes. Our proposed algorithm realized an F1 measure of 0.9868 on the time-series detector response data, reflecting a minimum enhancement of 0.3% in comparison with other techniques. The results conclusively show that our algorithm outperforms others when applied to time-series detector response data in urban contexts.

Keywords Gamma-ray spectral analysis · Nuclide identification · Urban environment · Temporal energy window · Peak-ratio spectrum analysis · Weighted KNN

1 Introduction

Nuclear technology and science have enriched the lives of millions globally, with advancements in areas such as clean energy, cancer treatment, food security, and pest control. However, it is imperative that nuclear and radioactive materials employed in these beneficial applications remain secure to prevent potential misuse [1]. Data from the incident and trafficking database (ITDB) of the international atomic energy agency (IAEA) reveal that between 1993 and 2020, there were 3686 reported incidents worldwide. Of these, 290 were confirmed or suspected cases of trafficking or malicious use. Notably, 12 incidents involved highly enriched uranium (HEU), and 2 featured plutonium [2]. The detection and identification of illegal radioactive materials in an urban environment is crucial to ensure the safe and legal use

This work was supported by the National Defense Fundamental Research Projects (Nos. JCKY2020404C004 and JCKY2022404C005) and Sichuan Science and Technology Program (No. 22NSFSC0044).

✉ Hao-Lin Liu
lhl0502@mail.ustc.edu.cn

[1] Department of Automation, University of Science and Technology of China, Hefei 230026, China

[2] School of Information Engineering, Southwest University of Science and Technology, Mianyang 621010, China

[3] Fundamental Science on Nuclear Wastes and Environment Safety Laboratory, Southwest University of Science and Technology, Mianyang 621010, China

[4] School of Automation and Information Engineering, Sichuan University of Science and Engineering, Zigong 643000, China

of radioactive materials, prevent their illegal transfer, and protect the safety of the state and its citizens [3, 4].

Numerous researchers delved into the detection and identification of radioactive materials. Most studies focus on conditions, where the detector and nuclear material maintain a static position relative to each other. In these cases, the radioactive source is often scaled proportionally and linearly superimposed onto a measured background. However, real measurement environments rarely exhibit a consistent background. Thus, simulations using a constant background intensity do not adequately represent the complexities encountered in actual measurement contexts [5–7].

During routine monitoring of radioactive events, or when responding to specific incidents involving uncontrolled radioactive material, imagine a detection scenario within an urban block. Experimenters traverse this block, seeking subtle indications of radioactive materials to ascertain their presence. Notably, in the backdrop of this urban environment, the most dominant element is the naturally occurring radioactive materials (NORM) found in various construction materials such as brick, granite, and concrete [8, 9]. The concentration of NORM varies near different buildings due to the unique composition of each structure and the environmental conditions surrounding it. Clearly, the background radiation within an urban environment fluctuates based on neighboring structures and prevailing environmental factors [10]. Additionally, radioactive materials can sometimes exhibit low intensity, with their gamma rays being attenuated by any shielding or dense materials surrounding the source. The energy spectra derived from these scenarios may be further complicated by cumulative and peak effects [11]. Given these complexities, traditional methods often struggle to effectively detect illicit radioactive materials concealed within buildings or accurately determine their types. Regrettably, false positives in radioactive material detection can lead to grave repercussions, wasting valuable time and posing potential health risks to researchers and the local populace. Consequently, algorithms designed for detecting and identifying radioactive materials should be resilient against diverse background conditions and shielding setups [12].

The task of detecting and identifying illicit radioactive materials presents significant challenges, and various studies have pursued techniques to address them. From a hardware equipment standpoint, Flanagan et al. [13] recommended the use of mobile, distributed sensors to detect nuclear materials in transit. Their research evaluated the efficacy of a mobile sensor network in detecting radioactive materials by melding radiation transport with geographic information systems.

Tran-Quang et al. [14] introduced an internet of radiation sensor system (IoRSS) designed for the detection of unregulated radioactive materials in scrap metal recycling and production facilities. This system enhances the detection, localization, and identification of radioactive materials by assimilating data from an array of portable radiation detectors. Meanwhile, Li et al. [15] pioneered the nuclide identification and quantitative analysis system (NIQAS) aimed at identifying hazardous substances via MCNP simulations. Central to this system are a D-T neutron generator and an HPGe detector. Various modules within the system were fine-tuned utilizing a signal-to-noise ratio (SNR) assessment method.

Conversely, when faced with hardware constraints, the onus shifts to the development of effective algorithms for energy spectrum analysis. A myriad of machine learning techniques, designed to emulate human cognition, have made significant strides in various domains. These include medical diagnosis [16], signal processing [17–19], and text classification [20–23]. Within the realm of radioisotope identification and radiation detectors, Pfund et al. [7] delved into defining energy region boundaries and decision metrics for gamma-ray spectra. Their research illuminated that selecting specific energy regions can augment the probability of detection in scenarios with low-count or obscured sources. Concurrently, Li et al. [24] proposed a groundbreaking approach for radionuclide identification in urban settings, harnessing a feature enhancer coupled with a one-dimensional neural network. Their methodology adeptly preprocesses the input energy spectrum data via the feature enhancer and seizes nonlinear information using the neural network.

Wu et al. [25] devised a peak searching technique using a generative adversarial network (GAN) tailored for urban environments characterized by low-count rates and brief measurements of single nuclide spectra. This GAN-centric approach outperforms the symmetric zero-area (SZA) method in accurately pinpointing characteristic peaks. By significantly reducing both the likelihood and number of false peaks, it bolsters the overall efficacy of peak recognition. Nonetheless, the quest to detect and identify illicit radioactive materials faces enduring challenges, including diminished detection sensitivity and the sway of environmental factors. As such, ongoing research is imperative to refine the precision and dependability of these techniques.

This study introduces a novel algorithm for the detection and identification of radioactive materials within urban environments. Our approach aims to offer a fresh solution to detect and identify radioactivity against the backdrop of complex urban settings, both during routine monitoring and in scenarios involving the uncontrolled dispersal of radioactive substances. Initially, the time-series detector response data, collected from an urban setting, were segmented using a temporal energy window. We then extracted distinct features from the energy spectra, drawing on the formation mechanism and measurement principle inherent to gamma-ray instrument spectra. These key features encompass aggregated counts, peak-to-flat ratios, and peak-to-peak

ratios. Given the need for imbalanced multi-classification, we crafted a detection and identification model grounded in the weighted KNN architecture.

2 Method

The proposed method unfolds in three pivotal steps: (a) To contend with the variability of backgrounds and accentuate the primary information from the radiation source, the time-series detector response data were segmented using a temporal energy window. (b) For a comprehensive analysis and to elucidate the physical implications of an energy spectrum, distinct features were drawn from the energy spectra. This extraction leaned on the formation mechanism and measurement principle of the gamma-ray instrument spectrum, incorporating features such as aggregated counts, peak-to-flat ratios, and peak-to-peak ratios. (c) With the aim of enhancing the resilience and precision of the model for detection and identification tasks within urban settings, we fashioned a model rooted in the weighted KNN architecture. The sequence of our proposed algorithm is illustrated in Fig. 1.

2.1 Temporal energy window

In this subsection, the temporal energy window is proposed for sample processing of time-series detector response data. Samples were partitioned into multiple segments with consideration of the sample type.

Urban landscapes teem with roads and structures composed of natural and man-made substances. Naturally occurring radioactive materials (NORMs) are inherent in these substances, with concentrations differing across materials. Predominantly, NORM comprises isotopes, such as ^{40}K, ^{238}U, and ^{232}Th, along with the radioactive daughter products of the latter two, commonly denoted as KUT [26]. As detectors navigate the search zone, particularly when radioactive substances are unmonitored, the makeup of the neighboring structures and their ensuing radioactive signatures shift with each locale [27]. Consequently, the cumulative gamma photon count rate and spectra recorded by detectors might demonstrate notable fluctuations [28]. Adding to the complexity, illicit radioactive substances may be concealed, leading to attenuated detection signals that are challenging to identify. The interplay between gamma photons and diverse substances, mediated by various physical processes, amplifies the dynamism of the observed radiation background signal. Hence, the time-series detector response data acquired in urban settings are profoundly shaped by ambient conditions, often overshadowing the distinctive peaks that mark the presence of radioactive materials in the energy spectra.

To counteract the effects of variable backgrounds, enhance the integrity and dependability of the energy spectrum data, and streamline subsequent data processing, we segmented the time-series detector response data using a temporal energy window. This strategy primarily underscores the features of faint radioactive materials.

In an urban setting, the detection system operates under two potential conditions: with or without the presence of an auxiliary radiation source, which is contextualized against the background radiation. Consequently, the time-series detector response dataset encompasses active and

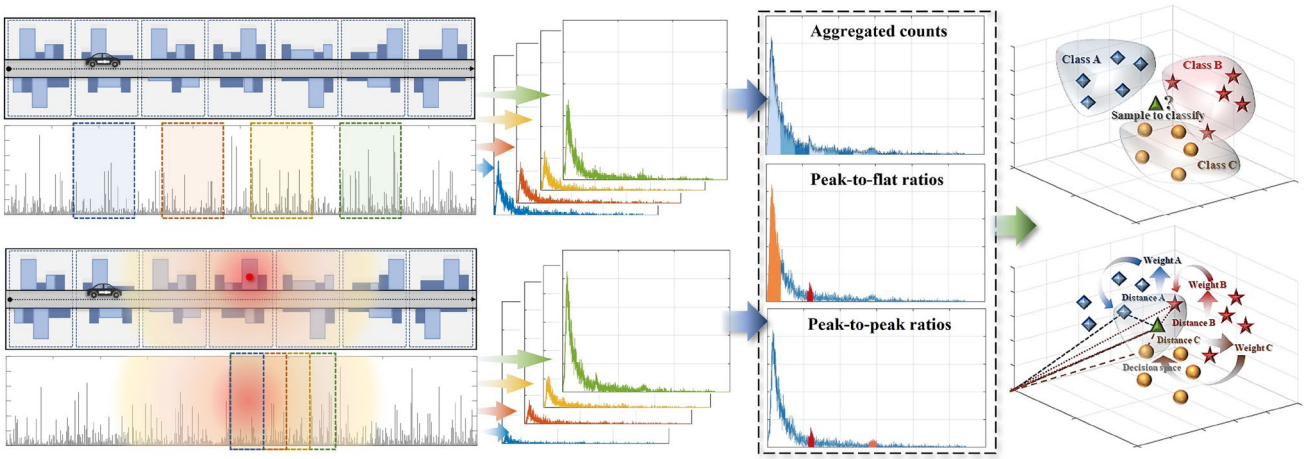

Fig. 1 (Color online) Block diagram of the proposed method. Partitioning the time-series detector response data using temporal energy windows, and converting the resulting corresponding segments into spectral form. Extracting features, such as aggregated counts, peak-to-flat ratios, and peak-to-peak ratios, based on the formation mechanism and measurement principle of the gamma-ray instrument spectrum. Constructing a weighted KNN-based detection and identification model for the imbalanced multi-classification problem in urban environment radiation detection

passive samples. An active sample pertains to the detector response data captured in the presence of a radioactive source, while a passive sample relates to data collected in an environment devoid of any radioactive source.

The time-series detector response dataset is defined as $\boldsymbol{T}$.

$$\boldsymbol{T} = \{(\boldsymbol{S}_1, y_1, loc_1), (\boldsymbol{S}_2, y_2, loc_2), \ldots, (\boldsymbol{S}_M, y_M, loc_M)\}$$

$\boldsymbol{S}_i$ is the matrix of time-series detector response data and defined by Eq. 1. Furthermore, M denotes the number of samples in the dataset. $i = 1, 2, \ldots, M$.

$$\boldsymbol{S}_i = [\boldsymbol{t}^i, \boldsymbol{e}^i] = \begin{bmatrix} t_1^i & e_1^i \\ t_2^i & e_2^i \\ \vdots & \vdots \\ t_m^i & e_m^i \end{bmatrix} \tag{1}$$

$\boldsymbol{t}^i$ denotes the time series, $t_j^i \in \mathbb{Q}^+$, and $\boldsymbol{e}^i$ denotes the energy value recorded by the detector over time, and $e_j^i \in \mathbb{Q}^+$. m denotes the length of $\boldsymbol{S}_i$, with $j = 1, 2, \ldots, m$.

When $\boldsymbol{S}_i$ is an active sample, $y_i \in \mathcal{Y}$ is the class label of $\boldsymbol{S}_i$. Furthermore, $loc_i \in \mathbb{Q}^+$ denotes the time point when the detector is closest to the radioactive source position during the movement, $t_1^i < loc_i < t_m^i$. When $\boldsymbol{S}_i$ is a passive sample, $y_i = 0$ and $loc_i = 0$.

The temporal energy window is proposed for sample processing of time-series detector response data. The quantity and length of a temporal energy window were defined as $w_\mathrm{q} \in \mathbb{Z}^+$ and $w_\mathrm{l} \in \mathbb{Z}^+$. The key of utilizing a temporal energy window for processing time-series detector response data lies in determining the temporal origin of the window, which refers to the initial point from which the temporal energy window conducts the partition task, thereby determining the position of the window within the time series. The temporal origin of an energy window is defined as: $t_{j'}$ and j' is calculated by Eq. 2.

$$\begin{aligned} j' = &\left\{ [0, w_\mathrm{q} - 1] \times \left\lfloor \frac{m}{w_\mathrm{q}} \right\rfloor + \left\lfloor \frac{m}{2 \times w_\mathrm{q}} \right\rfloor \right\} \\ &\times \bar{b} + \left\{ \underset{j' \in [1,m]}{\operatorname{argmin}} \left| loc_i - t_{j'} \right| - \lfloor \frac{1}{2} \times w_l \rfloor \right\} \times b \end{aligned} \tag{2}$$

In Eq. 2, b denotes a Boolean variable. Assuming that the present sample is active, b is true, whereas if the present sample is passive, then b is false. Obviously, $t_{j'}$ of a passive sample is distributed evenly at several different locations in the time axis, while $t_{j'}$ of an active sample is fixed due to the demand for obtaining energy fragments as close to the source as possible.

The segmented time-series detector response dataset processed by a temporal energy window is denoted as $\boldsymbol{T}_\mathrm{seg}$ and represented by Eq. 3. Samples in $\boldsymbol{T}_\mathrm{seg}$ may have been expanded in comparison with $\boldsymbol{T}$, which is dependent on w_q.

$$\begin{aligned} \boldsymbol{T}_\mathrm{seg} &= \{(\boldsymbol{S}_{\boldsymbol{seg}}^1, y_1, loc_1), (\boldsymbol{S}_{\boldsymbol{seg}}^2, y_2, loc_2), \ldots, \\ &\quad (\boldsymbol{S}_{\boldsymbol{seg}}^M, y_M, loc_M)\} \\ &= \{(\boldsymbol{S}_{\mathrm{seg}_1}^1, y_1, loc_1), (\boldsymbol{S}_{\mathrm{seg}_2}^1, y_1, loc_1), \\ &\quad \ldots, (\boldsymbol{S}_{\mathrm{seg}_{w_\mathrm{q}}}^1, y_1, loc_1), \ldots, (\boldsymbol{S}_{\mathrm{seg}_{w_\mathrm{q}}}^M, y_M, loc_M)\} \end{aligned} \tag{3}$$

In Eq. 3, $\boldsymbol{S}_{\boldsymbol{seg}}^i$ denotes the group of segmented time-series detector response data of $\boldsymbol{S}_i$ and represented by Eq. 4, $\boldsymbol{seg} = [\mathrm{seg}_1, \mathrm{seg}_2, \ldots, \mathrm{seg}_{w_\mathrm{q}}]$.

$$\boldsymbol{S}_{\boldsymbol{seg}}^i = \{[\boldsymbol{t}_{\boldsymbol{seg}}^i, \boldsymbol{e}_{\boldsymbol{seg}}^i]\} \tag{4}$$

Employing $\boldsymbol{S}_{\mathrm{seg}_k}^i$ to symbolize each individual segmented sample of $\boldsymbol{S}_{\boldsymbol{seg}}^i$. $\boldsymbol{S}_{\mathrm{seg}_k}^i$ is denoted by Eq. 5. Obviously, the length of each $\boldsymbol{S}_{\mathrm{seg}_k}^i$ is correlated with the length of the temporal energy window w_l.

$$\boldsymbol{S}_{\mathrm{seg}_k}^i = \left[\boldsymbol{t}_{\mathrm{seg}_k}^i, \boldsymbol{e}_{\mathrm{seg}_k}^i\right] = \begin{bmatrix} t_{j'}^{(i,k)} & , & e_{j'}^{(i,k)} \\ t_{j'+1}^{(i,k)} & , & e_{j'+1}^{(i,k)} \\ \vdots & & \vdots \\ t_{j'+w_l}^{(i,k)} & , & e_{j'+w_l}^{(i,k)} \end{bmatrix} \tag{5}$$

2.2 Peak-ratio spectrum analysis

In this subsection, we delve into the formation mechanism and measurement principle of the gamma-ray instrument spectrum. These are leveraged as the foundation for extracting spectral features. Key features include aggregated counts, peak-to-flat ratios, and peak-to-peak ratios. This type of an approach aids in the analysis and interpretation of the intrinsic significance of energy spectra.

After processing through the temporal energy window, the segmented time-series detector response data are converted into an energy spectrum format, easing the subsequent feature extraction. The energy spectrum provides a distribution curve mapping the count rate against particle energy, a pivotal tool in detecting and identifying radioactive nuclear materials.

For the context of this study, the relative distance between the detector and radiation source is in constant flux due to the detector's movement. It is essential to underline that this study primarily focuses on scenarios with static radiation sources. Dynamics, such as the continuous movement of the source or its dissolution in water, have not been contemplated. Owing to the finite number of photon counts within the full-energy peak, statistical fluctuations become pronounced. Consequently, the channel with the peak counts might not align with the expected value of a Gaussian distribution [29, 30]. To mitigate the effects of these statistical fluctuations, spectral data are reorganized into multiple bins along the energy axis. Each bin encompasses an energy

range, and counts within this range are consolidated to create a novel feature vector.

The transformed spectrum dataset is denoted as $\boldsymbol{T}_{\text{spe}}$.

$$\boldsymbol{T}_{\text{spe}} = \{(\boldsymbol{x}^1_{\boldsymbol{seg}}, y_1), (\boldsymbol{x}^2_{\boldsymbol{seg}}, y_2), \ldots, (\boldsymbol{x}^M_{\boldsymbol{seg}}, y_M)\}$$

$\boldsymbol{x}^i_{\boldsymbol{seg}}$ represents the transformed spectrum data from $\boldsymbol{S}^i_{\boldsymbol{seg}}$, $i = 1, 2, \ldots, M$ and $\boldsymbol{seg} = \{\text{seg}_1, \text{seg}_2, \ldots, \text{seg}_{w_q}\}$. $y_i \in \mathcal{Y}$ is the class label of $\boldsymbol{x}^i_{\boldsymbol{seg}}$. During the process of transformation, information of loc and t are discarded. The transformed spectrum data $\boldsymbol{x}^i_{\boldsymbol{seg}}$ are in the form of vector and represented by Eq. 6.

$$\boldsymbol{x}^i_{\boldsymbol{seg}} = \{(\boldsymbol{x}^i_{\text{seg}_1}, y_1), (\boldsymbol{x}^i_{\text{seg}_2}, y_1), \ldots, (\boldsymbol{x}^i_{\text{seg}_{w_q}}, y_1)\} \tag{6}$$

$\boldsymbol{x}^i_{\text{seg}_k}$ represents the transformed energy spectrum of the kth segment of the ith sample of the time-series detector response dataset.

$$\boldsymbol{x}^i_{\text{seg}_k} = \{x^{(i,k)}_1, x^{(i,k)}_2, \ldots, x^{(i,k)}_n\} \tag{7}$$

$i = 1, 2, \ldots, M$ and $k = 1, 2, \ldots, w_q$. $x^{(i,k)}_n$ is the aggregated count of the nth bin of $\boldsymbol{x}^i_{\text{seg}_k}$.

The length of each $\boldsymbol{x}^i_{\text{seg}_k}$, i.e., n, is the same because the same energy range is precomputed before the transformation process and a fixed number of bins is selected uniformly.

The value of n is determined based on the expected value of the maximum and minimum energy values across all samples in $\boldsymbol{T}$. The expected value of the maximum and minimum energy values is denoted as $\lceil \mathbb{E}[\max(\boldsymbol{e}^i)] \rceil$ and $\lfloor \mathbb{E}[\min(\boldsymbol{e}^i)] \rfloor$, $i = 1, 2, \ldots, M$. $n \in \mathbb{Z}^+$ and the value of n is calculated by Eq. 8.

$$n = \lceil \mathbb{E}[\max(\boldsymbol{e}^i)] \rceil - \lfloor \mathbb{E}[\min(\boldsymbol{e}^i)] \rfloor \tag{8}$$

Furthermore, $x^{(i,k)}_n$ indicates the photon count in $\boldsymbol{e}^i_{\text{seg}_k}$ in the corresponding energy interval, and $x^{(i,k)}_n$ is calculated by Eq. 9. Specifically, countif(A, B) is a function that searches the range A for items that match condition B and counts them. Additionally, α is applied to represent the energy range to fine-tune the transform accuracy of the energy spectrum.

$$\begin{aligned} x^{(i,k)}_n = \text{countif}(\boldsymbol{e}^i_{\text{seg}_k}, \lfloor \mathbb{E}[\min(\boldsymbol{e}^i)] \rfloor + (n-1) \\ \times \alpha \leqslant e^i_{\text{seg}_k} < \lfloor \mathbb{E}[\min(\boldsymbol{e}^i)] \rfloor + n \times \alpha) \end{aligned} \tag{9}$$

Detection and identification of radioactive materials primarily hinge on nuclear radiation detectors, which capture gamma rays emitted during the decay process. The measurement of gamma-ray energy is determined by registering the energy dispersed within the detector. The main mechanisms driving gamma energy spectrum measurements encompass three interactions between gamma rays and the detector medium: the photoelectric effect, the Compton effect, and pair production.

Low-energy gamma rays (0 – hundreds of keV) predominantly undergo the photoelectric effect, resulting in at least one distinct photoelectric peak. Medium-energy gamma rays (hundreds of keV – 3 MeV) primarily interact through the Compton effect. Conversely, high-energy gamma rays (5–10 MeV and beyond) are primarily subject to pair production. The photoelectric peak, when the energy of the incident gamma radiation is below 1.02 MeV, is often termed as the full-energy peak. This peak is traditionally considered as the primary hallmark for identifying specific radioactive nuclides. The full-energy peak arises from the sum of the photoelectric peak's energy combined with energy from Compton electrons and photoelectrons stemming from Compton scattering interactions. In the spectrum of low-to-medium energy gamma rays, pair production is negligible. Instead, the energy spectrum is characterized by a Compton continuum and photoelectric peaks. When gamma rays possess intermediate energy, incident gamma photons experience multiple successive Compton scatterings. The energy from the recoil electrons, produced from these scatterings, is deposited in the detector. Notably, the cumulative energy of these recoil electrons can surpass the energy transfer's upper limit in a single scattering event, filling regions between the Compton edge and photoelectric peaks [9].

From the prior discussion on the formation mechanism and measurement principle of the gamma-ray instrument spectrum, it is clear that gamma energy spectra contain both photoelectric peaks and the Compton continuum. Conventionally, the photoelectric peaks serve as the primary identifiers for radionuclides. Conversely, the Compton continuum, which often exhibits similar shapes across different contexts, is usually overlooked. However, relying solely on characteristic peaks for radioactive nuclide identification may fall short in complex background situations [31–33]. Drawing inspiration from Ref. [7], this subsection introduces peak-to-flat ratios and peak-to-peak ratios as descriptors for the spectral features.

Equation 7 defines the form of energy spectrum after binning. Here, $x^{(i,k)}_n$ indicates the photon counts in the corresponding energy bins and can be calculated by Eq. 9. Based on this, specific bins are selected according to the decay properties of the radionuclide material, which correspond to the area of theoretical Compton continuum, characteristic peaks, and auxiliary peaks, respectively. For $\boldsymbol{x}^i_{\text{seg}_k}$, the area of the theoretical Compton continuum, characteristic peaks, and auxiliary peaks are represented as a_c, a_f, and a_u, respectively, and defined as Eq. 10– Eq. 12, respectively. $i = 1, 2, \ldots, M$ and $k = 1, 2, \ldots, w_q$.

$$a_{\mathrm{c}} = \sum_{n' \in [c_{\mathrm{l}}, c_{\mathrm{r}}]} (x_{n'}^{(i,k)}) \tag{10}$$

$$a_{\mathrm{f}} = \sum_{n' \in [f_{\mathrm{l}}, f_{\mathrm{r}}]} (x_{n'}^{(i,k)}) \tag{11}$$

$$a_{\mathrm{u}} = \sum_{n' \in [u_{\mathrm{l}}, u_{\mathrm{r}}]} (x_{n'}^{(i,k)}) \tag{12}$$

The boundaries of the Compton continuum are denoted by c_{l} and c_{r}, while the characteristic peaks are bounded by f_{l} and f_{r}, and the auxiliary peaks are bounded by a_{l} and a_{r}. These boundaries are selected based on the decay properties of the radionuclide material. The peak-to-flat ratio r_1 and peak-to-peak ratio r_2 are defined as Eq. 13.

$$r_1 = a_{\mathrm{f}}/a_{\mathrm{c}}, \quad r_2 = a_{\mathrm{u}}/a_{\mathrm{f}} \tag{13}$$

Based on a macroscopic perspective, r_1 characterizes the capability to discern low-energy weak peaks amidst a complex background, while r_2 measures the likelihood of gamma rays experiencing multiple interactions within the detector, culminating in their contribution to the full-energy peaks.

2.3 Classification

In this subsection, considering the requirements for imbalanced multi-classification, we developed a detection and identification model using the weighted KNN architecture. By capitalizing on the inherent trait that energy spectra from identical radioactive materials exhibit minimal inter-class variability, the model significantly boosts the classification accuracy for underrepresented classes and improves the overall efficacy of the classifier.

Following the peak-ratio spectrum analysis, we derive a set of feature vectors comprised of aggregated counts, peak-to-flat ratios, and peak-to-peak ratios, denoted as $\boldsymbol{T}_{\mathrm{app}}$.

$$\boldsymbol{T}_{\mathrm{app}} = \{(\boldsymbol{c}_{\boldsymbol{seg}}^{1}, y_1), (\boldsymbol{c}_{\boldsymbol{seg}}^{2}, y_2), \ldots, (\boldsymbol{c}_{\boldsymbol{seg}}^{M}, y_M)\}$$

$\boldsymbol{c}_{\boldsymbol{seg}}^{i}$ represents the set of feature vectors from $\boldsymbol{x}_{\boldsymbol{seg}}^{i}$, which is denoted by Eq. 14. $i = 1, 2, \ldots, M$ and $\boldsymbol{seg} = \{\mathrm{seg}_1, \mathrm{seg}_2, \ldots, \mathrm{seg}_{w_{\mathrm{q}}}\}$. $y_i \in \mathcal{Y}$ is the class label.

$$\boldsymbol{c}_{\boldsymbol{seg}}^{i} = \{(\boldsymbol{c}_{\mathrm{seg}_1}^{i}, y_i), (\boldsymbol{c}_{\mathrm{seg}_2}^{i}, y_i), \ldots, (\boldsymbol{c}_{\mathrm{seg}_{w_{\mathrm{q}}}}^{i}, y_i)\} \tag{14}$$

Here, $\boldsymbol{c}_{\mathrm{seg}_k}^{i}$ represents the feature vector of the kth segment of the ith sample in the original time-series detector response dataset.

$$\boldsymbol{c}_{\mathrm{seg}_k}^{i} = [\boldsymbol{x}_{\mathrm{seg}_k}^{i}, r_1^{(i,k)}, r_2^{(i,k)}] \tag{15}$$

$i = 1, 2, \ldots, M$ and $k = 1, 2, \ldots, w_{\mathrm{q}}$. $\boldsymbol{x}_{\mathrm{seg}_k}^{i}$, $r_1^{(i,k)}$, and $r_2^{(i,k)}$ denote the aggregated counts, peak-to-flat ratio, and peak-to-peak ratio of the kth segment of the ith sample in the original time-series detector response dataset, respectively.

The sample to be classified is represented as $\boldsymbol{S}_0$, with its corresponding feature vector symbolized as $\boldsymbol{c}_0$. $\boldsymbol{c}_0$ and $\boldsymbol{c}_{\mathrm{seg}_k}^{i}$ are n-dimensional vectors, i.e., $\boldsymbol{c}_0 \in \mathbb{R}^n$ and $\boldsymbol{c}_{\mathrm{seg}_k}^{i} \in \mathbb{R}^n$.

The function f evaluated at the sample point $\boldsymbol{c}_{\mathrm{seg}_k}^{i}$ is y_i, i.e., $y_i = f\left(\boldsymbol{c}_{\mathrm{seg}_k}^{i}\right)$. The vector of observations is defined as:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix}$$

To construct a surrogate model $\hat{f}$ of a function f, sample points $\boldsymbol{c}_{\mathrm{seg}_k}^{i}$ are acquired into the matrix. The dimension of the matrix $\boldsymbol{C}$ is $M \times w_{\mathrm{q}} \times n$.

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{c}^{(1,1)^{\mathrm{T}}} \\ \vdots \\ \boldsymbol{c}^{(M,w_{\mathrm{q}})^{\mathrm{T}}} \end{bmatrix}$$

Notation $\boldsymbol{c}^{(i,k)}$ signifies $\boldsymbol{c}_{\mathrm{seg}_k}^{i}$. Here, K denotes the number of nearest neighboring sample points. $\boldsymbol{Z}$ denotes the set of K sample points that are closest to $\boldsymbol{c}_0$ in terms of distance, which is denoted by Eq. 16. $\boldsymbol{Z} \subseteq \boldsymbol{C}$ and $|\boldsymbol{Z}| = K$.

$$\boldsymbol{Z} = \left\{\boldsymbol{c}^{(i,k)} \in \boldsymbol{C} : \mathrm{rank}\left(d\left(\boldsymbol{c}_0, \boldsymbol{c}^{(i,k)}\right)\right) \leq K\right\} \tag{16}$$

Function $\mathrm{rank}(d(\boldsymbol{c}_0, \boldsymbol{c}^{(i,k)}))$ represents the ranking of the distance $d(\boldsymbol{c}_0, \boldsymbol{c}^{(i,k)})$ in ascending order.

The surrogate model $\hat{f}$ of function f is defined by Eq. 17.

$$\hat{f}(\boldsymbol{c}_0) = \frac{\sum_{\boldsymbol{c}^{(i,k)} \in \boldsymbol{C}} w\left(\boldsymbol{c}_0, \boldsymbol{c}^{(i,k)}\right) y_i}{\sum_{\boldsymbol{c}^{(i,k)} \in \boldsymbol{C}} w\left(\boldsymbol{c}_0, \boldsymbol{c}^{(i,k)}\right)} \tag{17}$$

$w(\cdot)$ is the inverse distance weight function, which is defined by Eq. 18.

$$w\left(\boldsymbol{c}_0, \boldsymbol{c}^{(i,k)}\right)_{\boldsymbol{c}^{(i,k)} \in \boldsymbol{C}} = \frac{1}{\left[d\left(\boldsymbol{c}_0, \boldsymbol{c}^{(i,k)}\right)\right]^q} \tag{18}$$

where q denotes a normalization power and $d(\boldsymbol{c}_0, \boldsymbol{c}^{(i,k)})$ denotes the distance between the target point $\boldsymbol{c}_0$ and sample point $\boldsymbol{c}_{\mathrm{seg}_k}^{i}$. The metric used to measure this distance is L_p-norm, which is defined by Eq. 19.

$$d\left(\boldsymbol{c}_0, \boldsymbol{c}^{(i,k)}\right) = \left\|\boldsymbol{c}_0 - \boldsymbol{c}^{(i,k)}\right\|_p = \left(\sum_{\beta=1}^{m} \left|\boldsymbol{c}_{0,\beta} - \boldsymbol{c}_{\beta}^{(i,k)}\right|^p\right)^{1/p} \tag{19}$$

Here, β denotes the index of the dimension of the vector. $\boldsymbol{c}_{\beta}^{(i,k)}$ denotes the βth dimension of $\boldsymbol{c}_{\mathrm{seg}_k}^{i}$, and $\boldsymbol{c}_{0,\beta}$ denotes the βth dimension of $\boldsymbol{c}_0$. Specifically, L_1-norm (where $p = 1$) represents the rectangular distance, while L_2-norm (where $p = 2$)

represents the Euclidean norm. Furthermore, L_∞-norm (where $p \to \infty$) represents the maximum norm.

To limit the effect of farther samples points and also avoid divisions by zero, the distance function is implemented as follows. If $\|\boldsymbol{c}_0 - \boldsymbol{c}^{(i,k)}\|_p = 0$, then distance $d\left(\boldsymbol{c}_0, \boldsymbol{c}^{(i,k)}\right)$ is set to ε. If $0 < \|\boldsymbol{c}_0 - \boldsymbol{c}^{(i,k)}\|_p \le R$, then distance $d(\boldsymbol{c}_0, \boldsymbol{c}^{(i,k)})$ is calculated by Eq. 19. If $\|\boldsymbol{c}_0 - \boldsymbol{c}^{(i,k)}\|_p > R$, then distance $d(\boldsymbol{c}_0, \boldsymbol{c}^{(i,k)}) = 0$, where ε is a small number and R is the radius of the distance function $d(\cdot)$. Table 1 summarizes the overall flow of the algorithm.

3 Experiments and analysis

This section detailed the processes of data acquisition and preprocessing and established a series of comparative experiments to validate the efficacy of the proposed algorithm. All experiments conducted in this section utilized tenfold cross-validation to guarantee the reliability of the results.

3.1 Introduction of data source

The experimental data utilized in this study originated from a time-series detector response dataset, representing a NaI(Tl) detector's movement within a simulated city block using the Monte Carlo method. This dataset was curated by J. M. Ghawaly Jr and his team at Oak Ridge National Laboratory (ORNL) [34]. Figure 2 offers a visual representation capturing the core features of the dataset. The model simulated seven interconnected city blocks in three dimensions, encompassing various buildings, sidewalks, roads, parking areas, and other urban elements. The naturally occurring radioactive materials (NORMs) incorporated included ^{40}K, ^{232}Th and its progeny, as well as ^{238}U/^{235}U and their respective offspring. The concentration of each component within the KUT (potassium, uranium, and thorium) might vary depending on the specific material. Each block's background radiation was individually computed. The radioactive materials were potentially concealed in 15 distinct spots. Each radioactive source could exist in one of two states: either unshielded or shielded by 1 cm of lead. A NaI(Tl) detector navigated through these city blocks without the interference of cars or other forms of clutter.

The dataset comprises radioactive materials from two categories: special nuclear materials (SNMs) and common sources. The SNMs are represented by highly enriched uranium (HEU) and weapons-grade plutonium (WGPu), while the common sources are technetium-99m (^{99m}Tc), iodine-131 (^{131}I), and cobalt-60 (^{60}Co). Both HEU and WGPu are characterized by energy spectra dominated by prompt fission neutrons and prompt gamma rays, which are emitted during fission. These gamma rays possess a broad energy range, spanning from several hundred keV up to multiple MeV. Conversely, ^{99m}Tc releases gamma rays that

Table 1 Overall flow of the proposed method

Algorithm	Algorithm for detection and identification of radioactive materials in an urban environment
Input	$\boldsymbol{S}_i$: Time-series detector response data y_i: Class label loc_i: Time point (detector closest to the source) w_q: Quantity of the temporal energy window w_l: Length of the temporal energy window α: Energy range of bins b: Sample state, true for active, false for passive $\boldsymbol{S}_0$: Sample to be classified K: Number of nearest neighbors
Begin	1. Compute the length of $\boldsymbol{S}_i$ M=length($\boldsymbol{S}_i$) 2. Search the temporal origin of the energy window $j' = f_p \times \bar{p} + f_a \times p$ Here, $f_p = \left\{ [0, w_q - 1] \times \left\lfloor \frac{m}{w_q} \right\rfloor + \left\lfloor \frac{m}{2 \times w_q} \right\rfloor \right\}$ $f_a = \left\{ \underset{j' \in [1,m]}{\operatorname{argmin}} \left| loc_i - t_{j'} \right| - \lfloor \frac{1}{2} \times w_l \rfloor \right\}$ 3. Preprocess the time-series detector response data to obtain segmented pulses $\boldsymbol{S}^i_{seg} = \left[\boldsymbol{t}^i_{seg}, \boldsymbol{e}^i_{seg} \right]$ 4. Calculate the count value in bins $x_n^{(i,k)} = countif(\boldsymbol{e}^i_{\mathrm{seg}_k}, Cond)$ $Cond = E + (n-1) \cdot \alpha \leqslant e^i_{\mathrm{seg}_k} < E + n \cdot \alpha$, in which, $n = \lceil \mathbb{E}[\max(\boldsymbol{e}^i)] \rceil - \lfloor \mathbb{E}[\min(\boldsymbol{e}^i)] \rfloor$ and $E = \lfloor \mathbb{E}[\min(\boldsymbol{e}^i)] \rfloor$ 5. Segmented energy spectrum is obtained $\boldsymbol{x}^i_{\mathrm{seg}_k} = \left\{ x_1^{(i,k)}, x_2^{(i,k)}, \ldots, x_n^{(i,k)} \right\}$ 6. Calculate boundary values of Compton continuum characteristic peaks and auxiliary peaks $a_c = \sum_{n' \in [c_l, c_r]} (x_{n'}^{(i,k)})$ $a_f = \sum_{n' \in [f_l, f_r]} (x_{n'}^{(i,k)})$ $a_u = \sum_{n' \in [u_l, u_r]} (x_{n'}^{(i,k)})$ 7. Obtain peak-to-flat ratio and peak-to-peak ratio $r_1 = a_f / a_c, \quad r_2 = a_u / a_f$ 8. Obtain feature vector $\boldsymbol{c}^i_{\mathrm{seg}_k} = [\boldsymbol{x}^i_{\mathrm{seg}_k}, r_1^{(i,k)}, r_2^{(i,k)}]$ 9. Calculate the distance between c_0 and sample points $d\left(\boldsymbol{c}_0, \boldsymbol{c}^{(i,k)}\right) = \left\| \boldsymbol{c}_0 - \boldsymbol{c}^{(i,k)} \right\|_p$ 10. Search K sample points that are closest to $\boldsymbol{c}_0$ in terms of distance $\boldsymbol{Z} = \left\{ \boldsymbol{c}^{(i,k)} \in \boldsymbol{C} : \operatorname{rank}\left(d\left(\boldsymbol{c}_0, \boldsymbol{c}^{(i,k)}\right) \right) \le K \right\}$ 11. Calculate the weight of K sample points $w\left(\boldsymbol{c}_0, \boldsymbol{c}^{(i,k)}\right)_{\boldsymbol{c}^{(i,k)} \in \boldsymbol{C}} = \frac{1}{\left[d\left(\boldsymbol{c}_0, \boldsymbol{c}^{(i,k)}\right) \right]^q}$

Table 1 (continued)

Algorithm	Algorithm for detection and identification of radioactive materials in an urban environment
End	
Output	
	Predict the class label of $\boldsymbol{S}_0$
	$\hat{f}(\boldsymbol{c}_0) = \frac{\sum_{c^{(i,k)} \in \boldsymbol{C}} w(\boldsymbol{c}_0, \boldsymbol{c}^{(i,k)}) y_i}{\sum_{c^{(i,k)} \in \boldsymbol{C}} w(\boldsymbol{c}_0, \boldsymbol{c}^{(i,k)})}$

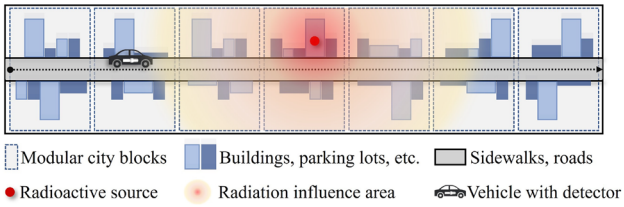

Fig. 2 (Color online) Schematic diagram of the fundamental characteristics of the dataset. This model consisted of seven modular city blocks, and the order of the blocks can be adjusted. Size of the model was 989–1047 m × 201 m × 158 m. For each component of the blocks, every NORM isotope in each material (asphalt, brick, granite, concrete, and soil) in its composition was modeled. These data form the background of the urban environment. A 2″ × 4″ × 16″ NaI(Tl) detector traversed the city block in the absence of cars or other forms of clutter. The velocity of the detector was a value in the range of 1–13.4 m/s and remains constant. The walls of the buildings in the model were 6 in (15.24 cm) thick [34]

Table 2 Radionuclide library

Label	Radioactive materials	Capacity
0	Background	4900
1	HEU	800
2	WGPu	800
3	^{131}I	800
4	^{60}Co	800
5	^{99m}Tc	800
6	HEU+^{99m}Tc	800

predominantly linger around the 140-keV energy mark. ^{131}I emits mainly beta particles accompanied by gamma rays; the beta particles peak at energies near 606 keV. The emitted gamma rays have varying energy levels, with the most notable peaks observed at 364 keV and 637 keV. Finally, ^{60}Co radiates gamma rays that prominently feature two energy peaks, one at 1.17 MeV and the other at 1.33 MeV [9].

Specifically, 9700 samples were labeled and are listed in Table 2, of which 4900 were background samples without any radioactive materials, while the remaining 4800 samples contained radioactive materials.

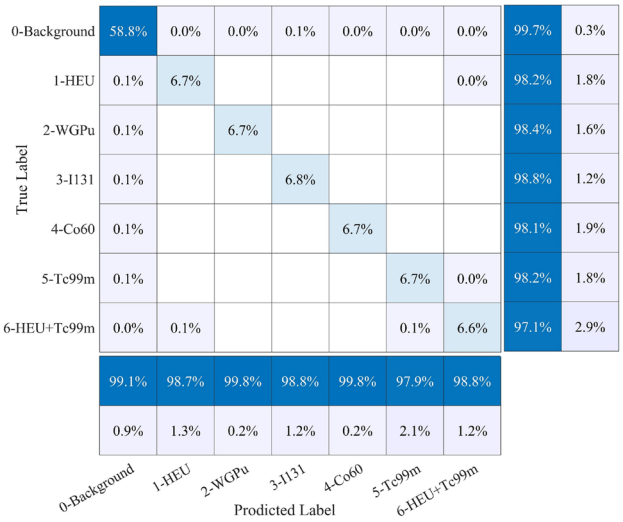

Fig. 3 Confusion matrix for the proposed algorithm's test results. Each cell within the matrix's core is normalized according to the total observations of the respective class, illustrating the proportion of correctly identified samples within the whole dataset. The column summary indicates the percentage of correct and incorrect classifications for each predicted class, scaled by the overall observations of that predicted class. Similarly, the row summary portrays the percentage of correct and incorrect classifications for each actual class, adjusted by the total observations of that specific class

3.2 Comparative experiments

In this subsection, to optimize and assess the model's performance while ensuring its practical applicability, the time-series detector response dataset was partitioned into three distinct subsets: training (60%), validation (20%), and testing (20%). This division was subjected to tenfold cross-validation. A stratified random split was adopted, guaranteeing a balanced representation of radioactive materials across all subsets. The model was implemented in Python, leveraging the capabilities of the PyTorch framework. For comparative analysis, the Weka machine learning toolkit was employed. All experiments were executed on a system furnished with an Intel Core i7 processor, 16 GB of RAM, and an NVIDIA GeForce RTX 3070 graphics card.

To streamline the discussion, the proposed algorithm in the experimental results will be referred to as TPW. In the experiments detailed in this subsection, the values of K, p, and q were set to 5, 2, and 2, respectively. A comprehensive examination and discussion regarding the selection of these parameter values is given in Sect. 3.5. The testing accuracy achieved was 99.1%, with an F1 score of 0.9868. The confusion matrix derived from the test data is viewed in Fig. 3.

In summary, the TPW algorithm has shown promising results in both passive backgrounds and active scenarios. Analysis of the column and row summaries indicates that the most notable misclassifications occur between class 1(HEU),

5(^{99m}Tc), and 6(HEU+^{99m}Tc). The primary reason for this is the presence of radioactive material in the detection scene of class 6, which is also present in classes 1 and 5, causing ambiguity in the identification process.

Furthermore, to provide a comprehensive comparison, the standard KNN (KNN) [16, 35], support vector machine (SVM) [17, 36], Bayesian network (BayesNet) [18, 37], random tree (RandomTree) [19, 38], and the proposed algorithm (TPW) were applied for evaluation. The aforementioned methods were commonly utilized for radionuclide identification classification in recent years. Comparative experiments were conducted using the Weka machine learning toolkit [39] with a batch size of 100 and tenfold cross-validation. The main parameters used for each method were as follows: For standard KNN, the number of neighbors was set to 5 with no distance weighting. For SVM, the Poly Kernel was used as the kernel function. The complexity and tolerance parameters were set to 1.0 and 0.001, respectively. For Bayesian network, the initial network structure used for learning was the naive Bayes network and the maximum number of parents that a node in the Bayes net can have was limited to 1. For Random Tree, the random number seed used for selecting attributes was 1, and the minimum total weight of instances in a leaf was set to 1.0. The maximum depth of the tree was unlimited.

Given the imbalanced nature of the dataset, relying solely on traditional classification accuracy can be misleading. For instance, a model might achieve high accuracy simply by categorizing all samples as the majority class, in this case, "Background." Hence, a variety of evaluation metrics were utilized in this subsection to provide a holistic view of model performance. Figure 4 presents these metrics for different models. Comparing the TPW algorithm with four other methods (standard KNN, support vector machine, Bayesian network, and random tree) across five distinct evaluation metrics, it became evident that TPW excels. Specifically, TPW consistently showcased superior accuracy, F1 score, MCC, ROC area, and PRC area when compared to its counterparts. This underlines TPW's enhanced efficacy and reliability in tasks related to radionuclide identification.

We further performed individual tests for each class of samples, contrasting the TPW algorithm's performance with the four other methods using the F1 measure. Figure 5 showcases the classification results across different models for every sample class. Overall, the TPW algorithm emerged as the top performer among all the tested methods. In particular, it exhibited a commendable capacity to accurately classify samples from every class, underscoring its robustness and adaptability to various sample types.

Examining the results in depth, we note that the efficacy of different methods varies considerably across classes. In particular, for some classes, the TPW algorithm notably surpasses the F1 measures of its competitors, while for others, the performance differences are more nuanced. This indicates that the TPW algorithm is especially adept at processing certain sample types, although its relative advantage might be less distinct for other sample types. Notably, the F1 measure for the samples of class 0 (background), 3 (^{131}I), and 4 (^{60}Co) is higher, whereas it is somewhat subdued for class 1 (HEU), 2 (WGPu), 5 (^{99m}Tc), and 6 (^{99m}Tc). By analyzing the peak energies of these radioactive materials, we discerned that their characteristic peaks are all below 200 keV. This suggests that their accurate detection and identification might be compromised by the Compton continuum. Nevertheless, the TPW algorithm excels over other models, surpassing them by at least 0.18% in multi-isotope scenarios like HEU+^{99m}Tc, and showcases lower variability than other models when detecting radioactive materials.

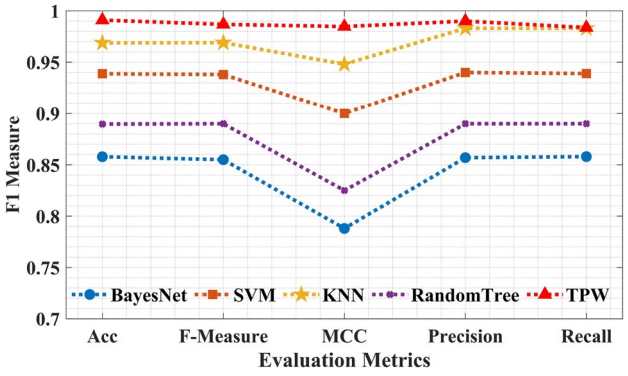

Fig. 4 Multiple evaluation metrics across various models. The *x*-axis represents the evaluation metrics including accuracy (Acc), F1 measure (F-measure), Matthews correlation coefficient (MCC), receiver operating characteristic area (ROC area), and precision–recall curve area (PRC area). The y-axis shows the performance values for each method under the corresponding metrics for every class of samples

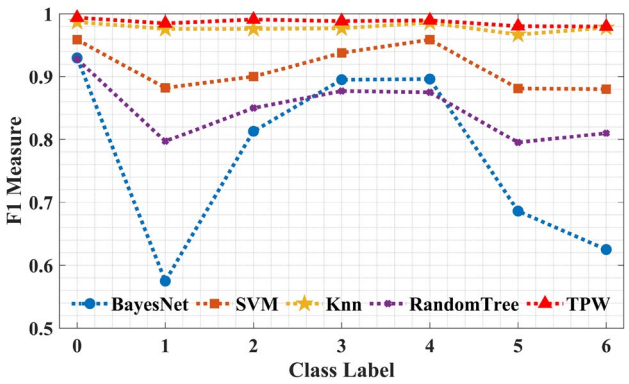

Fig. 5 F1 measure for each class of samples across various models. The F1 measure values of different methods for each class of samples are plotted on the y-axis of the plot, while the x-axis of the plot indicates the corresponding class of samples as listed in Table 2

3.3 Discussion on temporal energy window

In this subsection, we explored the effects of varying parameters associated with the temporal energy window, specifically focusing on the number and duration of these windows. The movement of the detector poses challenges in determining the ideal length for an energy window. A brief window, such as 1 s, might not capture sufficient relevant data, while an extended window, such as 20 s, could introduce substantial noise, potentially overshadowing crucial signals. Figure 6 depicts the energy spectra for a sample from class 4 (^{60}Co) at varying energy window durations. As the window length increased, there was a noticeable rise in the count of the energy spectrum. However, the count values across different energy points did not grow linearly with the expansion of the window length, possibly due to statistical fluctuations and other influencing factors [29, 30]. For a deeper insight into the nuances of the spectral lines, Fig. 7 highlights the variations in the morphology and attributes of the energy spectra, as the window length transitions across five distinct durations.

Figure 8 illustrates the model's performance changes in relation to varying window lengths and quantities. The data indicate that as the number of temporal energy windows increases, the model's prediction accuracy also rises, particularly when the window quantity equals 5. With the increase in the number of temporal energy windows, the proposed algorithm captures a richer representation of the signal, enhancing the differentiation between signal and noise, and thus providing a more precise target prediction. Additionally, the model's prediction accuracy trends upward with longer window durations. However, there is a significant decline in accuracy with excessively long windows. These experiments shed light on the influence of window lengths and quantities on classification efficacy, emphasizing the need for optimal temporal scale selection and feature extraction methods for accurate classification in the given task.

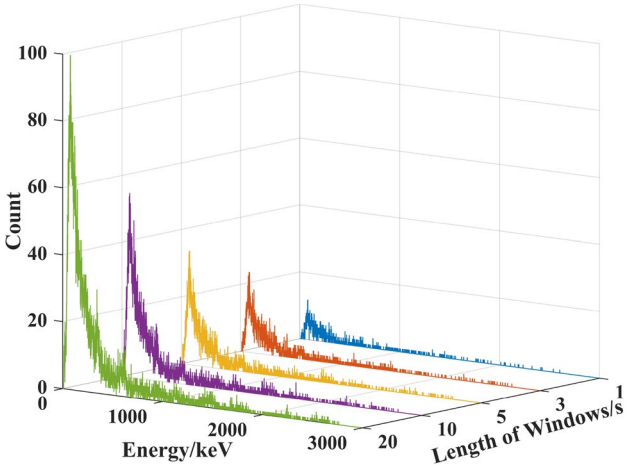

Fig. 6 (Color online) Energy spectra with respect to different temporal energy window lengths. This figure illustrates the energy spectra of ^{60}Co for different temporal energy window lengths of 1, 3, 5, 10, and 20 s. The energy range shown is between 0 and 3000 keV. The count value displayed on the *Z*-axis shows an increasing trend with the varying window lengths

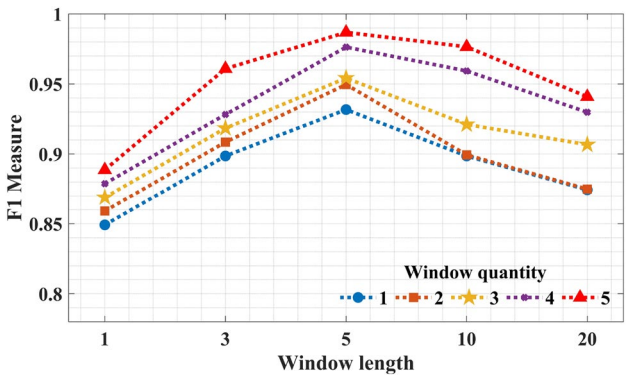

Fig. 8 Performance variations of models with different window lengths and quantities. The plot displays the F1 measure values for various parameter combinations, with x-axis representing the energy window lengths

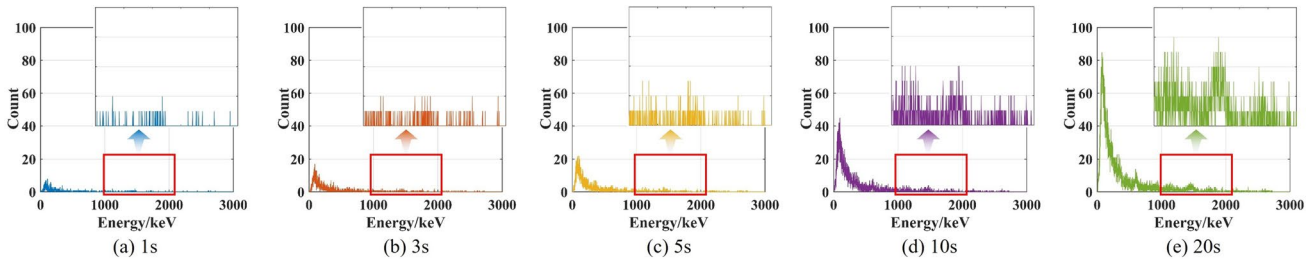

Fig. 7 (Color online) Detailed energy spectra across varied window lengths. These five diagrams offer an in-depth examination of the energy spectra of ^{60}Co, captured at different temporal energy window durations, complementing Fig. 6. The inset images in the upper right magnify the spectrum within the highlighted red boxes. The irregular growth in count values at various energy addresses, as the window length expands, can be linked to statistical fluctuations and other potential influences

Table 3 Results of ablation experiments

No	Feature	Length	Test F1 measure
1	Aggregated counts	99	0.9843
2	Peak-flat ratio	22	0.9245
3	Peak-peak ratio	30	0.8435
4	Joint feature	151	0.9868

3.4 Discussion on peak-ratio spectrum analysis

In this subsection, we explored the influence of individual and combined features on the classification performance using ablation experiments. Table 3 contrasts the classification outcomes of the aggregated energy spectrum counts, peak-flat ratio, and peak-peak ratio features against those using the joint features, shedding light on their individual and combined impacts.

The F1 measure comparisons reveal that combined features outperform their individual counterparts. This superior performance of joint features arises from their capacity to seamlessly assimilate energy spectrum data from diverse viewpoints. By harmoniously harnessing the unique attributes of each feature, joint features amplify classification precision, overshadowing the results achieved by singular features. Additionally, joint features adeptly counteract challenges intrinsic to individual features, such as noise interference, data sparsity, or lack of comprehensive representation. Conversely, singular features often struggle to offer a holistic and resilient information foundation for classification [23]. Therefore, combined features furnish a more holistic and richer data representation, bolstering classification efficiency. In essence, the empirical findings underscore the merit of deploying joint features for more accurate radionuclide identification.

3.5 Discussion on classification model

In this subsection, we examined the proposed algorithm by assessing the impact of three factors: the distance metric, value of K, and distance weight. We experimented with various distance metrics, including Euclidean, City block, Chebyshev, Correlation, Spearman, Hamming, Jaccard, and Cosine. The value of K was varied between 1 and 20. For distance weighting, we considered three approaches: "Equal distance" (ED), which did not incorporate any weight; "Inverse distance" (ID), where the weight was based on the inverse of the distance to the data point; and "Inverse distance squared" (IDS), where the weight was determined by the inverse of the squared distance to the data point. The experimental results are listed in Table 4. During the experiments, K was set to 10 and all the samples were subjected to standardization, bold numbers in Table represent the best results under this experimental setting.

Table 4 Results of discussion on classification model

No	Distance metric	Distance weight	Veri F1	Test F1
1	Euclidean	IDS	0.9672	0.9763
2	City block	IDS	0.9727	**0.9814**
3	Chebyshev	IDS	0.8761	0.8872
4	Correlation	IDS	**0.9749**	0.9762
5	Spearman	IDS	0.9639	0.9644
6	Hamming	IDS	0.3575	0.3811
7	Jaccard	IDS	0.3575	0.3811
8	Cosine	IDS	0.9707	0.9711
9	Euclidean	ID	0.9505	0.9595
10	Euclidean	ED	0.9238	0.9333

Based on the observed F1 measure during validation and testing, the distance metric of Correlation was superior in the validation experiments, while the distance metric of City block outperformed the other distance metrics in terms of classification accuracy in the testing experiments. In terms of distance weighting, the IDS emerged as the superior performer. By significantly reducing the influence of far-off points on the classification decision, IDS led to more precise and dependable results.

Selecting an appropriate value of K was crucial for optimal model performance. A very small K makes the model vulnerable to noise in the feature points, which can greatly influence classification outcomes. Conversely, an overly large K dilutes the specificity of the model as the neighborhood around the training instance becomes too expansive, increasing the likelihood of misclassifications [16, 35]. Thus, striking a balance between noise resistance and model precision by carefully adjusting the K value is imperative. We undertook a series of tests to discern the effect of different K values, ranging from 1 to 20, on the efficacy of the proposed algorithm. The outcomes of these tests are depicted in Fig. 9.

From the results, it can be observed that the F1 score initially increases as the value of K increases from 1 to 5, reaching a peak value of 0.9868 at K=5. Then, F1 score slightly fluctuates and then starts to decline as K further increases. Additionally, the results suggest that the model has a high overall performance, with F1 scores consistently above 0.95 for all values of K. This indicates that the model is effective in accurately predicting the class labels of the input data. This pattern of results suggests that increasing the value of K can lead to better classification performance up to a certain point, beyond which overfitting may occur, resulting in a decline in performance.

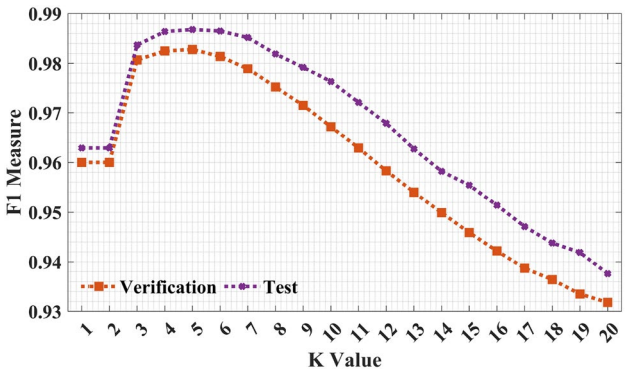

Fig. 9 Performance variations based on different K values. The figure showcases the F1 measure for the proposed algorithm for K values spanning from 1 to 20, plotted on the X-axis. The Y-axis indicates the F1 measure. Two distinct curves denote the outcomes from verification and test experiments, respectively

4 Conclusion

This study introduces a novel approach for detecting and identifying radioactive materials within urban settings.

From the conducted comparative experiments, we derive the following key conclusions: (1) Detector response data, when viewed as a time series, are effectively segmented using temporal energy windows. Segmenting the data in this manner mitigates the impact of shifting backgrounds, enhances the reliability and quality of energy spectrum data, and streamlines the downstream data processing. This segmentation yields an energy spectrum that emphasizes pivotal information pertaining to the radioactive materials. (2) Our feature extraction strategy taps into the formation mechanisms and the measurement principles of gamma-ray instruments, yielding deep insights into the physical nature of energy spectra. The features we extract, including aggregated counts, peak-to-flat ratios, and peak-to-peak ratios, offer a comprehensive view of the sample's multidimensional attributes. This approach negates the need for individual peak analysis and peak searches, thereby enhancing the efficiency and precision of data processing. (3) The custom weighted KNN model crafted for detection and identification capitalizes on subtle variations in energy spectrum classes for identical nuclides. This shifts the classification challenge into a task of partitioning the feature space. By incorporating weights, our model counters the issues posed by imbalanced datasets, meets the requirements of real-time and multi-classification detection, and fortifies the robustness of the detection model, especially when operating against intricate urban backgrounds.

However, this technology does come with certain limitations. It struggles when tasked with measuring systems affected by improvised nuclear devices (INDs) or tactical nuclear artifacts. The detectors are vulnerable to electromagnetic pulses (EMPs) and prompt gamma rays, which span a broader energy spectrum. It is our hope that subsequent research will overcome these hurdles. Additionally, this study's limitation lies in its exclusive reliance on simulated spectra for training and testing the algorithm. Future endeavors need to validate the findings against spectra measured from actual detectors. Moving forward, our research will prioritize the practical application of this method in real-life scenarios. This includes refining detection accuracy, minimizing false positives, and bolstering the algorithm's computational prowess. Moreover, this technique holds promise for broader applications, potentially benefiting areas such as nuclear safety, environmental conservation, and public health.

Acknowledgements Thanks for the https://doi.org/10.13139/ORNLNCCS/1597414 dataset which is provided by the Oak Ridge National Laboratory.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Hai-Bo Ji, Jiang-Mei Zhang, Cao-Lin Zhang, Jing Lu, and Xing-Hua Feng. The first draft of the manuscript was written by Hao-Lin Liu, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data availability The data that support the findings of this study are openly available in Science Data Bank at https://doi.org/10.57760/sciencedb.10892 and http://cstr.cn/31253.11.sciencedb.10892.

Declarations

Conflict of interest The authors declare that they have no competing interests.

References

1. X. Li, C. Dong, Q. Zhang et al., Research and design of a rapid nuclide recognition system. J. Instrum. **17**(06), T06008 (2022). https://doi.org/10.1088/1748-0221/17/06/T06008
2. IAEA Incident and Trafficking Database (ITDB), Incidents of nuclear and other radioactive material out of regulatory control 2020 Fact Sheet. *Paper Presented at the Nuclear Security Plan 2022–2025* (USA 15 September 2021). https://www.iaea.org/sites/default/files/gc/gc65-24.pdf
3. X. Li, Q. Zhang, H. Tan et al., Research of nuclide identification method based on background comparison method. Appl. Radiat. Isot. **192**, 110596 (2023). https://doi.org/10.1016/j.apradiso.2022.110596
4. L. Li, G. Huang, S. Xi et al., Application of fuzzy probability factor superposition algorithm in nuclide identification. J. Radioanal. Nucl. Chem. **331**(5), 2261–2271 (2022). https://doi.org/10.1007/s10967-022-08318-w
5. D. Liang, P. Gong, X. Tang et al., Rapid nuclide identification algorithm based on convolutional neural network. Ann. Nucl. Energy **133**, 483–490 (2019). https://doi.org/10.1016/j.anucene.2019.05.051
6. W. Zhao, R. Shi, X.G. Tuo et al., Novel radionuclides identification method based on Hilbert–Huang transform and convolutional neural network with gamma-ray pulse signal. Nucl.

Instrum. Methods Phys. Res. A. **1051**, 168232 (2023). https://doi.org/10.1016/j.nima.2023.168232
7. D.M. Pfund, R.C. Runkle, K.K. Anderson et al., Examination of count-starved gamma spectra using the method of spectral comparison ratios. IEEE Trans. Nucl. Sci. **54**(4), 1232–1238 (2007). https://doi.org/10.1109/TNS.2007.901202
8. Z. Szabó, P. Völgyesi, H.É. Nagy et al., Radioactivity of natural and artificial building materials -a comparative study. J. Environ. Radioact. **118**, 64–74 (2013). https://doi.org/10.1016/j.jenvrad.2012.11.008
9. D.M. Abrams, in *Radiation Detection and Measurement*. ed. by J. Welter, D. Matteson (Wiley, New York, 2010), p.625
10. M.W. Swinney, D.E. Peplow, B.W. Patton et al., A methodology for determining the concentration of naturally occurring radioactive materials in an urban environment. Nucl. Technol. **203**(3), 325–335 (2018). https://doi.org/10.1080/00295450.2018.1458558
11. D. E. Archer, D. E. Hornback, J. O. Johnson et al., Systematic assessment of neutron and gamma backgrounds relevant to operational modeling and detection technology implementation. (Oak Ridge National Lab. (ORNL), Oak Ridge, TN (United States) 1 Jan 2010) https://doi.org/10.2172/1185844
12. L.A.O. Giraldo, *Special Nuclear Material and Radiological Sources Detection in Urban Settings* (The Pennsylvania State University, Degree of Master of Science, 2015)
13. R.R. Flanagan, L.J. Brandt, A.G. Osborne et al., Detecting nuclear materials in urban environments using mobile sensor networks. Sensors. **21**(6), 2196 (2021). https://doi.org/10.3390/s21062196
14. V. Tran-Quang, H. Dao-Viet, An internet of radiation sensor system (IoRSS) to detect radioactive sources out of regulatory control. Sci. Rep. **12**(1), 7195 (2022). https://doi.org/10.1038/s41598-022-11264-y
15. J. Li, W. Jia, D. Hei et al., Research on the NIQAS device for hazardous goods identification based on PGNAA technology. Appl. Radiat. Isot. **169**, 109445 (2021). https://doi.org/10.1016/j.apradiso.2020.109445
16. M.A. Calin, F.G. Elfarra, S.V. Parasca, Object-oriented classification approach for bone metastasis mapping from whole-body bone scintigraphy. Phys. Med. **84**, 141–148 (2021). https://doi.org/10.1016/j.ejmp.2021.03.040
17. G. Kusuma, R. M. Saryadi, S. K. Wijayaet al., Radionuclide identification analysis using machine learning and GEANT4 simulation. *Paper Presented at the Proceedings of International Conference on Nuclear Science, Technology, and Application 2020* (Jakarta, Indonesia 23–24 November 2020) https://doi.org/10.1063/5.0067593
18. J.W. Wang, W.G. Gu, H. Yang et al., Analytical method for γ energy spectrum of radioactive waste drum based on deep neural network. Nucl. Tech. **45**, 040501 (2022). https://doi.org/10.11889/j.0253-3219.2022.hjs.45.040501 (**in Chinese**)
19. D. Pérez-Loureiro, J. Alexander, Radioisotope identification using CLYC detectors. *Paper Presented at the 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE* (Nassau, Bahamas 12–14 December 2022) https://doi.org/10.1109/ICMLA55696.2022.00214
20. A. Onan, Bidirectional convolutional recurrent neural network architecture with group-wise enhancement mechanism for text sentiment classification. J. King. Saud. Univ.-Com. **34**(5), 2098–2117 (2022). https://doi.org/10.1016/j.jksuci.2022.02.025
21. A. Onan, S. Korukoğlu, H. Bulut et al., A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. Inf. Process. Manag. **53**(4), 814–833 (2017). https://doi.org/10.1016/j.ipm.2017.02.008
22. A. Onan, Mining opinions from instructor evaluation reviews: a deep learning approach. Comput. Appl. Eng. Educ. **28**(1), 117–138 (2020). https://doi.org/10.1002/cae.22179
23. A. Onan, An ensemble scheme based on language function analysis and feature engineering for text genre classification. J. Inf. Sci. **44**(1), 28–47 (2018). https://doi.org/10.1002/cae.22179
24. C. Li, S. Liu, C. Wang et al., A new radionuclide identification method for low-count energy spectra with multiple radionuclides. Appl. Radiat. Isot. **175**, 110219 (2022). https://doi.org/10.1016/j.apradiso.2022.110219
25. S. Wu, X. Tang, P. Gong et al., Peak-searching method for low count rate spectrum under short-time measurement based on a generative adversarial network. Nucl. Instrum. Methods Phys. Res. A. **1002**, 165262 (2021). https://doi.org/10.1016/j.nima.2021.165262
26. S. Croft, I. Hutchinson, The measurement of U, Th and K concentrations in building materials. Appl. Radiat. Isot. **51**(5), 483–492 (1999). https://doi.org/10.1016/S0969-8043(99)00064-0
27. W. Yao, Z.M. Liu, Y.P. Wan et al., Energy spectrum nuclide recognition method based on long short-term memory neural network. Nucl. Eng. Technol. **54**, 4684–4692 (2022). https://doi.org/10.1016/j.net.2022.08.011
28. R. Trevisi, S. Risica, M. D'Alessandro et al., Natural radioactivity in building materials in the European Union: a database and an estimate of radiological significance. J. Environ. Radioact. **105**, 11–20 (2012). https://doi.org/10.1016/j.jenvrad.2011.10.001
29. Y.L. Song, F.Q. Zhou, Y. Li et al., Methods for obtaining characteristic c-ray net peak count from interlaced overlap peak in HPGe c-ray spectrometer system. Nucl. Sci. Tech. **30**, 11 (2019). https://doi.org/10.1007/s41365-018-0525-7
30. Z.D. Li, H.Q. Zhang, J.Y. Liu et al., Implementation and analysis of Gaussian shaping method for digital nuclear pulse signal. Nucl. Tech. **42**, 060403 (2019). https://doi.org/10.11889/j.0253-3219.2019.hjs.42.060403 (**in Chinese**)
31. T. Wang, X. He, L. Ge et al., Neural Network Radionuclide Identification Algorithm Based on Exponential Smoothing. *Paper Presented at the 2022 International Conference on Computation, Big-Data and Engineering (ICCBE)* (Yunlin, Taiwan, China 27-29 May 2022) https://doi.org/10.1109/ICCBE56101.2022.9888230
32. R. Shi, X.G. Tuo, H.L. Li et al., Unfolding analysis of LaBr 3: Ce gamma spectrum with a detector response matrix constructing algorithm based on energy resolution calibration. Nucl. Sci. Tech. **29**, 1 (2018). https://doi.org/10.1007/s41365-017-0340-6
33. Y. Yuan, L.Q. Zhang, X.L. Luo et al., A real-time peak detection method for nuclear pulse signal and energy spectrum analysis. Nucl. Tech. **42**, 020404 (2019). https://doi.org/10.11889/j.0253-3219.2019.hjs.42.020404 (**in Chinese**)
34. J.M. Ghawaly Jr., A.D. Nicholson, D.E. Peplow et al., Data for training and testing radiation detection algorithms in an urban environment. Sci. Data. **7**(1), 328 (2020). https://doi.org/10.1038/s41597-020-00672-2
35. S. Qi, W. Zhao, Y. Chen et al., Comparison of machine learning approaches for radioisotope identification using NaI(TI) gamma-ray spectrum. Appl. Radiat. Isot. **186**, 110212 (2022). https://doi.org/10.1016/j.apradiso.2022.110212
36. S. Qi, W. Zhao, Y. Chen et al., Comparison of machine learning approaches for radioisotope identification using NaI (TI) gamma-ray spectrum. Appl. Radiat. Isot. **186**, 110212 (2022). https://doi.org/10.1016/j.apradiso.2022.110212
37. Z. Wu, B. Wang , J. Sun, Design of radionuclides identification algorithm based on sequence bayesian method. *Paper Presented at the 2nd International Conference on Advanced Materials, Intelligent Manufacturing and Automation* (China 17–19 May 2019) https://doi.org/10.1088/1757-899X/569/5/052047
38. J.R. Romo, K.T. Nelson, M. Monterial et al., Classifier Comparison for Radionuclide Identification from Gamma-ray Spectra.

Paper Presented at the Proceedings of the INMM & ESDARSA Joint Virtual Annual Meeting (Vienna, Austria, 23–26 August & 30 August–1 September, 2021). https://www.osti.gov/servlets/purl/1818402

39. I.H. Witten, E. Frank, M.A. Hall, et al., *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd edn. (New Zealand, 2014), pp. 403–406