ORNL/TM-2023/2906

Metrics and Methods for Radiation Detection Algorithm Characterization for Nuclear/Radiological Source Search

May 2024

OAK RIDGE National Laboratory

ORNL IS MANAGED BY UT-BATTELLE LLC FOR THE US DEPARTMENT OF ENERGY

ORNL/TM-2023/2906

Physics Division

METRICS AND METHODS FOR RADIATION DETECTION ALGORITHM CHARACTERIZATION FOR NUCLEAR/RADIOLOGICAL SOURCE SEARCH

Mark Bandstra[6]
Carl Britt[2]
James Ghawaly[1]
Thomas Grimes[2]
Tom Haard[3]
Peter Heimberg[3]
Tenzing Joshi[6]
Heidi Komkov[8]
Simon Labov[4]
Noah McFerran[4]
Tyler Morrow[8]
Andrew Nicholson[1]
Marc Paff[5]
Brian Quiter[6]
Michael Reed[7]
Gregory Thoreson[8]

[1]Oak Ridge National Laboratory
[2]Pacific Northwest National Laboratory
[3]John Hopkins University Applied Physics Laboratory
[4]Lawrence Livermore National Laboratory
[5]Los Alamos National Laboratory
[6]Lawrence Berkeley National Laboratory
[7]Nevada National Security Site
[8]Sandia National Laboratory

January 2024

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831
managed by
UT-BATTELLE LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AIPT	Algorithm Improvement Program Team
ANSI	American National Standards Institute
BARNI	Benchmark Algorithm for RadioNuclide Identification
BKG	background
CWMD	Countering Weapons of Mass Destruction Office
DHS	US Department of Homeland Security
DOCA	distance of closest approach
DoD	US Department of Defense
DOE	US Department of Energy
DNN	Defense Nuclear Nonproliferation
DRAG	Detection Radiation Algorithms Group
DTRA	Defense Threat Reduction Agency
FPR	false positive rate
FSA	full-spectrum analysis
GADRAS-DRF	Gamma Detector Response and Analysis Software–Detector Response Function
GCR	gross count rate
HEU	highly enriched uranium
HPGe	high-purity germanium
ID	identification
IRQ	interquartile range
JH-APL	Johns Hopkins University Applied Physics Laboratory
KUT	naturally occurring potassium, uranium, and thorium (^{40}K, ^{238}U and daughters, and ^{232}Th and daughters)
ML	machine learning
NMF	non-negative matrix factorization
NNSA	National Nuclear Security Administration
NORM	naturally occurring radioactive material
PD	probability of detection
PFA	probability of false alarm
PFID	probability of false identification
PID	probability of identification
prompt eq.	prompt equilibrium
PVT	polyvinyl toluene

R&D	research and development
RESTful API	representational state transfer application programming interface
RIID	radioisotope identification device
ROC	receiver operating characteristic
RPM	radiation portal monitor
sec. eq.	secular equilibrium
SME	subject matter expert
SNM	special nuclear material
SNR	signal-to-noise ratio
SPRT	sequential probability ratio test

ABSTRACT

This report presents a series of recommendations for data to train and evaluate radiation-detection algorithms and performance metrics to evaluate these algorithms. These recommendations were formed via a community consensus approach through the Detection Radiation Algorithms Group (DRAG), a multi-institution collaboration spanning eight US Department of Energy (DOE) laboratories and Johns Hopkins University Applied Physics Laboratory (JH-APL). This report includes recommendations on background data variability and metrics to quantify that variability, sources and shielding configurations to include in data collection campaigns, and detector response variability. This report also describes several anomaly detection and identification algorithms and recommends metrics to report their performance. Finally, this report ends with a discussion of machine learning (ML) algorithms.

1. OVERVIEW

1.1 INTRODUCTION

Several national and international programs rely on radiation-detection technologies to achieve their mission goals. The objectives can span from fundamental, such as confirming the presence of any radiation above background, to more complex, such as quantifying the mass of bulk nuclear materials. In most scenarios, signals from radiation detectors must be analyzed and interpreted before they can be useful. The science of radiation-detection algorithms is a broad field that seeks to automate this interpretation to produce useful output for the mission. For example, the output may be an intermediate step that filters output to a human analyst, or it may be the final step from a completely automated analysis used to drive strategic and tactical decisions.

Investment in radiation-detection algorithms has been an ongoing effort at the international level for many decades. In the United States, a long-term, multi-agency effort to further radiation detection and associated algorithms has been spurred by major world events, including the Manhattan Project, nuclear testing and nuclear forensics, the fall of the Union of Soviet Socialist Republics and the associated nonproliferation concerns, and modern countering of weapons of mass destruction. As a consequence of nearly a century of evolving missions and priorities, the field of radiation detection is incredibly diverse, and the algorithm requirements are highly mission specific.

Advancement in radiation-detection algorithms is hindered by the lack of clear requirements, evaluation metrics, evaluation datasets, and baseline algorithms for comparison. The American National Standards Institute (ANSI) standards address some of these needs [1, 2], but these standards were intended as a minimum set of requirements and do not represent the most current challenging scenarios.

To develop and promote the use of a commonly accepted and appropriately challenging evaluation framework, metrics along with benchmark datasets with which to evaluate these algorithms, a consensus approach was developed by an established group of subject matter experts (SMEs) from multiple organizations.

1.2 HISTORY OF THE DETECTION RADIATION ALGORITHMS GROUP

DRAG was formed in 2021 by the Office of Defense Nuclear Nonproliferation (DNN) Research and Development (R&D) Near-Field Detection Portfolio to identify and prioritize radiation-detection algorithm needs across the United States government. DRAG is a working group of SMEs that perform R&D for detection algorithms or have programmatic responsibility for missions that rely on radiation detection. The core DRAG group, currently composed of the authors of this report, is responsible for

organizing the larger DRAG community, which currently comprises approximately 200 scientists and leaders from a dozen US national laboratories and multiple federal agencies, including DOE/National Nuclear Security Administration (NNSA), the US Department of Defense (DoD)/Defense Threat Reduction Agency (DTRA), and the US Department of Homeland Security (DHS)/Countering Weapons of Mass Destruction Office (CWMD).

To organize algorithm needs for radiation detection, a set of 11 Mission Area Focus Groups that have a radiation-detection component were identified. The mission areas are not exhaustive, but DRAG attempted to cover as many mission spaces as possible. Furthermore, the missions are not orthogonal—in some cases, significant overlap exists between missions. In most cases, the missions are not specific to any single agency or office, although specific tasks in each mission area may vary by agency. A partial list of stakeholders for each mission area is provided in Table 1.

Table 1. Stakeholders for each of the DRAG-defined mission areas

Mission Area Focus Group	US government stakeholder
Checkpoint Monitoring	DHS/CWMD, DOE/NNSA/NA-213
Consequence Management	DOE/NNSA/NA-84
Facility Protection	DOE/NNSA/NA-10, DOE/NNSA/NA-243, NNSA Production Office
Forensics	DOE/NNSA/NA-80, DoD/DTRA
Nuclear Battlefield	DoD
Diagnostics	DOE/NNSA/NA-80, DoD/DTRA
Nuclear Safeguards	DOE/NNSA/NA-241
Search	DOE/NNSA/NA-84, DHS/CWMD
Site Characterization	DoD
Arms Control (Treat Verification)	DOE/NNSA/NA-243
Waste Management	DOE Office of Environmental Management

Each Mission Area Focus Group was tasked to identify objectives, sensors, challenges, and needs/gaps for algorithm-related mission spaces in a mission summary. From these mission summaries, several algorithm topics emerged that are needed in one or more mission areas that are a high priority for DNN R&D. The following mission topic areas were identified:

- Detection, classification, and identification
- Nuclear materials analysis
- Localization and mapping
- Imaging
- Plume and dispersion modeling
- Data fusion (cross cutting)
- Testing and metrics (cross cutting)
- Data processing (cross cutting)
- Uncertainty quantification (cross cutting)

Using these algorithm topic areas, DRAG organized a workshop held on June 21–22, 2021, to gather experts from the national laboratories, industry, academia, and government to discuss mission summaries

to create consensus and prioritize R&D. The following prioritized list of R&D needs was identified by DRAG and workshop attendees and delivered to DNN R&D:

1. Consensus algorithm evaluation framework and data
2. Data fusion algorithms
3. Enabling capabilities for inverse modeling

From this list, the DNN R&D Near-Field Detection Portfolio included #2 data fusion algorithms in the DNN R&D FY 2023 Call for Proposals and tasked DRAG with pursuing #1 consensus algorithm evaluation framework and data for the radiation search via a multi-laboratory consensus approach in 2022 and beyond.

Standard algorithm performance evaluation metrics are needed to enable informative algorithm performance comparisons to quantify return on investment, operational readiness, and areas for improvement. DRAG identified four algorithm types most relevant to the source-search mission. For each type, a set of performance evaluation metrics and example baseline algorithms were identified. Furthermore, datasets must contain enough background variations and real-world detector response variability and must include sources that which span operational ranges. These needs are shown schematically in Figure 1.

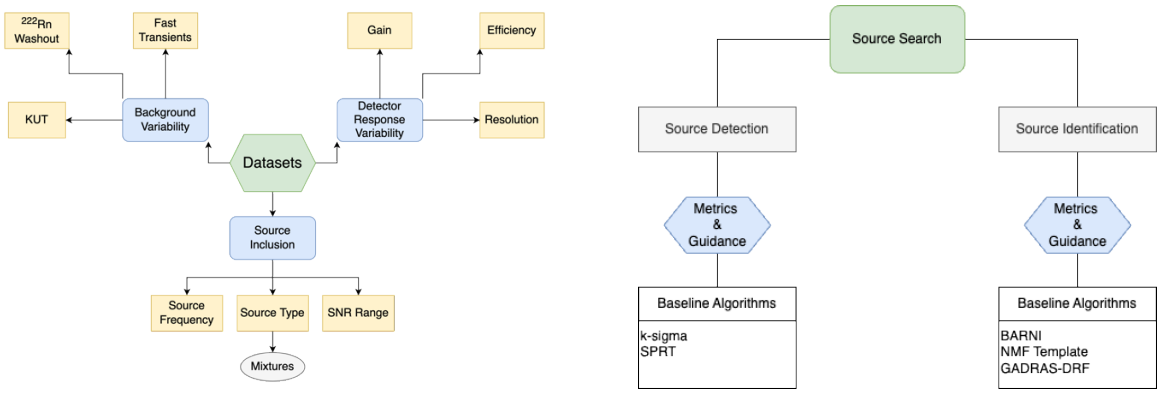


Figure 1. Dataset and source-search schematic. (left) Primary attributes that data collectors and curators should consider when collecting or generating datasets for developing and evaluating algorithms for the source-search mission. (right) Two algorithm types identified as most relevant to the source-search mission, with performance metrics, guidelines, and baseline algorithms.

On June 28–30, 2022, DRAG held a workshop on algorithm metrics and dataset benchmarks for the source-search mission space to garner input from the greater radiation-detection community. This document reports the findings of this workshop and details community consensus recommendations for datasets and metrics for radiation algorithm evaluation for the search problem.

1.3 DOCUMENT SCOPE

This document addresses the evaluation of algorithms and datasets for the nuclear/radiological source-search mission space. Specifically, this document provides recommendations on metrics for quantifying radiation-detection algorithm performance and methods for interpreting the results. Likewise, this document provides recommendations regarding the characteristics that datasets should contain and methods for evaluating this content.

1.4 DOCUMENT PURPOSE

This document provides recommendations for quantifying algorithm performance and dataset characteristics for the nuclear/radiological source-search mission space, thereby enabling reliable quantitative algorithm comparison for evaluation of operational readiness, return of R&D investment, and identification of areas for future work. From the perspective of the algorithm developer, this document can be used to produce results for publications and reports on algorithm development and testing.

1.5 DEFINITIONS

- *Detection*: Determining a source is present that is not part of the background radiation and its normal fluctuations.
- *Classification*: Determining the type of source found (e.g., naturally occurring radioactive material [NORM], medical, industrial, fissile) or broader classification such as benign vs. threat or allowable vs. source of interest.
- *Identification (ID)*: Determining the radionuclide or mixtures of radionuclides within a source.
- *Quantification*: Determining the amount (mass or activity) of the source(s) identified.
- *Localization*: Determining the position of the source.
- *Directionality*: Determining the direction from the detector to the source.
- *False detection*: Declaring a source is present when only background fluctuations are present.
- *False classification or ID*: Declaring a source to be one type when it is actually another.
- *Nuisance alarm*: An alarm caused by false detection or false classification of background, benign, or allowable sources as threats or sources of interest.
- *False alarm*: A false detection, sometimes combined with false classification of source type requiring alarm (usually false classification of a benign source as a threat source).
- *Benign source*: A source that does not pose a significant hazard (e.g., NORM, medical, industrial). A subset of these sources may be defined as nuisance sources.
- *Threat source*: A source type that could pose a significant hazard. This type includes fissile materials (nuclear threats) and high-activity radionuclides (radiological threats).
- *Shielding*: The atomic number (Z) and areal density (AD) of intervening materials between the source and the detector.
- *Multiplication*: The average number of neutrons produced via fission per source neutron in the system.

2. DATASETS

This section outlines the primary attributes that data collectors and curators should consider when collecting or generating datasets intended for the development and evaluation of radiation-detection algorithms for the source-search mission. Three main categories of attributes are considered: background variations, sources, and detector response variations. Metrics for quantifying each of these three attributes are included and can be used to generate quantitative dataset descriptors.

2.1 BACKGROUND VARIATIONS

Source-search missions are often required to operate in environments with dynamic background radiation count rates and spectral variations, resulting from the combined effects of variations in the potassium, uranium with daughters, and thorium with daughters (KUT) concentrations in different manmade and natural materials, environmental effects such as precipitation-induced ^{222}Rn washout, and cosmic rays. In mobile detectors, KUT variations manifest as variations in the solid angle of nearby structures and materials such as bridges, tunnels, buildings, raised berms, and road material transitions. The following sections describe these dynamics, recommendations for dataset generation, and metrics for quantifying each.

2.1.1 Recommendations for Background Variability

Potassium, Uranium, and Thorium Variability

The absolute and relative concentrations of ^{40}K, ^{238}U and daughters, and ^{232}Th and daughters (the three primary sources of naturally occurring radioactivity) can vary between different material types and sources. Likewise, some materials such as granite can have significant absolute quantities of KUT that present as fast transients in the detection data. In a mobile detector source search, this variability can lead to dynamics in both the detected gross count rate (GCR) and gamma-ray energy spectral shape. Consequently, datasets should include a variety of KUT conditions to test algorithms against. Typical choices of materials to consider are concrete, asphalt, soil, bricks, natural stone (e.g., granite, sandstone, marble), and variations in the type and source of each. Typical KUT concentrations for different materials have been tabulated in several studies [3, 4, 5, 6]. Table 2 lists baseline typical activity concentrations for a variety of common materials, determined by combining data from these studies. To induce variability within a given dataset for algorithm evaluation, the recommendation is to scale the baseline KUT component activities within ±80% of the baseline values of Nicholson et al. [7], except for seawater, which should only be varied ±20%.

Table 2. Typical mean activity concentrations for common materials for ^{40}K, ^{238}U (secular equilibrium), and ^{232}Th-232 (secular equilibrium)

Material	^{40}K (Bq/kg)	^{238}U (Bq/kg)	^{232}Th (Bq/kg)
Concrete	383	142	30
Brick	547	53	54
Asphalt	130	24	19
Granite	1045	93	82
Soil	428	40	29
Seawater	0	0	12

The GCR variability (as a result of KUT variations) largely depends on the environment of deployment (urban vs. rural). Recommendations for the magnitude and frequency of this variability are listed in Table 3. As an example for interpretation, for a detector moving throughout a dense urban environment it is recommended that the mean GCR vary by a factor of 0.5 to 2 times the mean count rate at an *average* rate of 3 times per 100 m of travel. This rule is not absolute, and exceptions will always occur. These rate variations should occur randomly rather than at a fixed rate.

Table 3. Recommended GCR variability (from KUT variations) for a mobile detector in rural, suburban, and dense urban environments. These values do not consider other forms of variability such as rain, fast transients, or maritime considerations, which are described later in the report.

Measurement environment	Variation magnitude
Rural	0.8–1.1× mean count rate
Dense urban	0.5–2.0× mean count rate
Urban	0.5–2.0× mean count rate
Suburban	0.9–1.2× mean count rate

Some examples of real-world measured background variations are presented below. In the first example, a single 2 × 4 × 16 in. Na(Tl) detector was driven through the Knoxville, Tennessee area inside of a 16-passenger van. The resulting total background count rate is shown in Figure 2 along with the detector path and histograms of total count rates for different regions of the path. The downtown Knoxville area was labeled as "urban," strip malls were labeled as "suburban," and highways and rural roads were labeled as "rural." The average rural area count rate was found to be 592 counts/s, the average suburban rate was 647 counts/s, and the average urban rate was 816 counts/s. The average from all three environments was 689.3 counts/s. The distributions are not normal, and the urban histogram is bimodal.

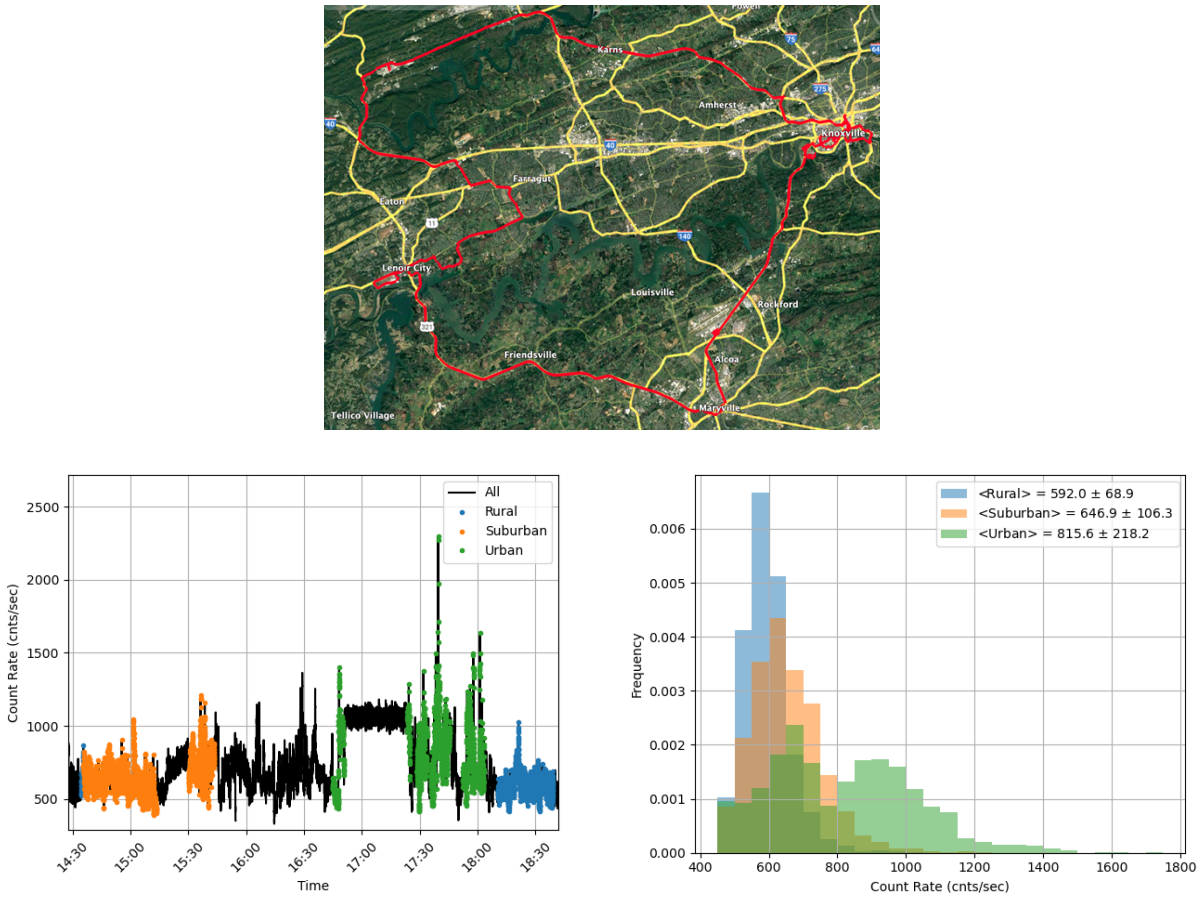


Figure 2. Gamma-ray backgrounds taken with a 2 × 4 × 16 in. Na(Tl) detector around the Knoxville, Tennessee area. (top) Map of data collected. (bottom left) Total count rates as a function of time. Data around downtown Knoxville is labeled as "urban," data taken around Maryville, Tennessee, and Farragut, Tennessee areas are labeled as "suburban," and data taken along highways and back roads are labeled as "rural." (bottom right) Normalized histograms of detector count rates for each environment. The total average count rate over all three environments is 689.3 counts/s.

Another example is background from the Washington, D.C., and Baltimore, Maryland, urban area in the Algorithm Improvement Program Team (AIPT) M.4 dataset provided by JH-APL. These data were collected using four 2 × 4 × 16 in. Na(Tl) detectors (all summed together) and are summarized in Figure 3. These data were separated into "dense urban," "urban," "suburban," and "rural" areas. This dataset differs from the Knoxville dataset in that the variation in the rural histogram is much larger and exhibits much more variation than was seen in Knoxville. This result reinforces the fact that dataset recommendations are really a rule of thumb, and variations from location to location are largely unique to the local environment.

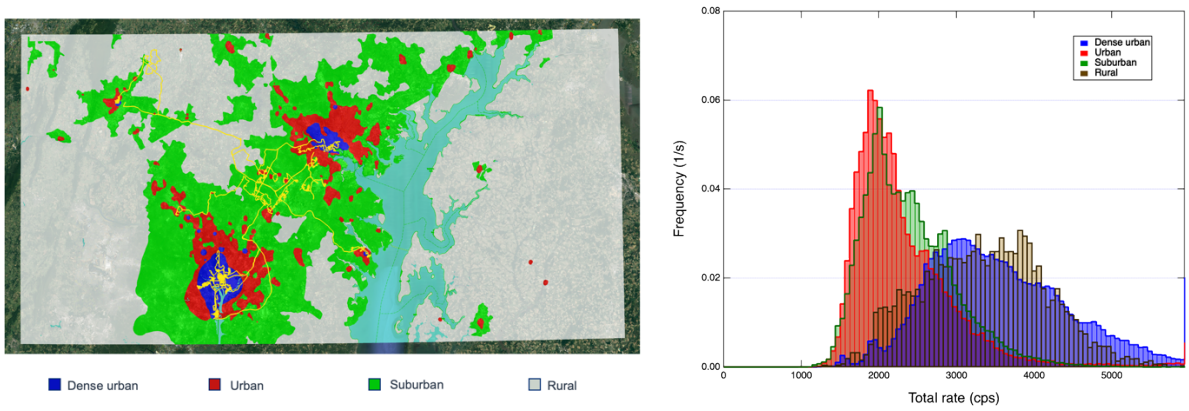


Figure 3. Gamma-ray background for four detectors taken around the Washington, D.C., and Baltimore, Maryland, area. (left) Map of the measurement area with the four environments shown: dense urban, urban, suburban, and rural. (right) Count rate histograms, all four 2 × 4 × 16 in. Na(Tl) detectors summed, for each environment.

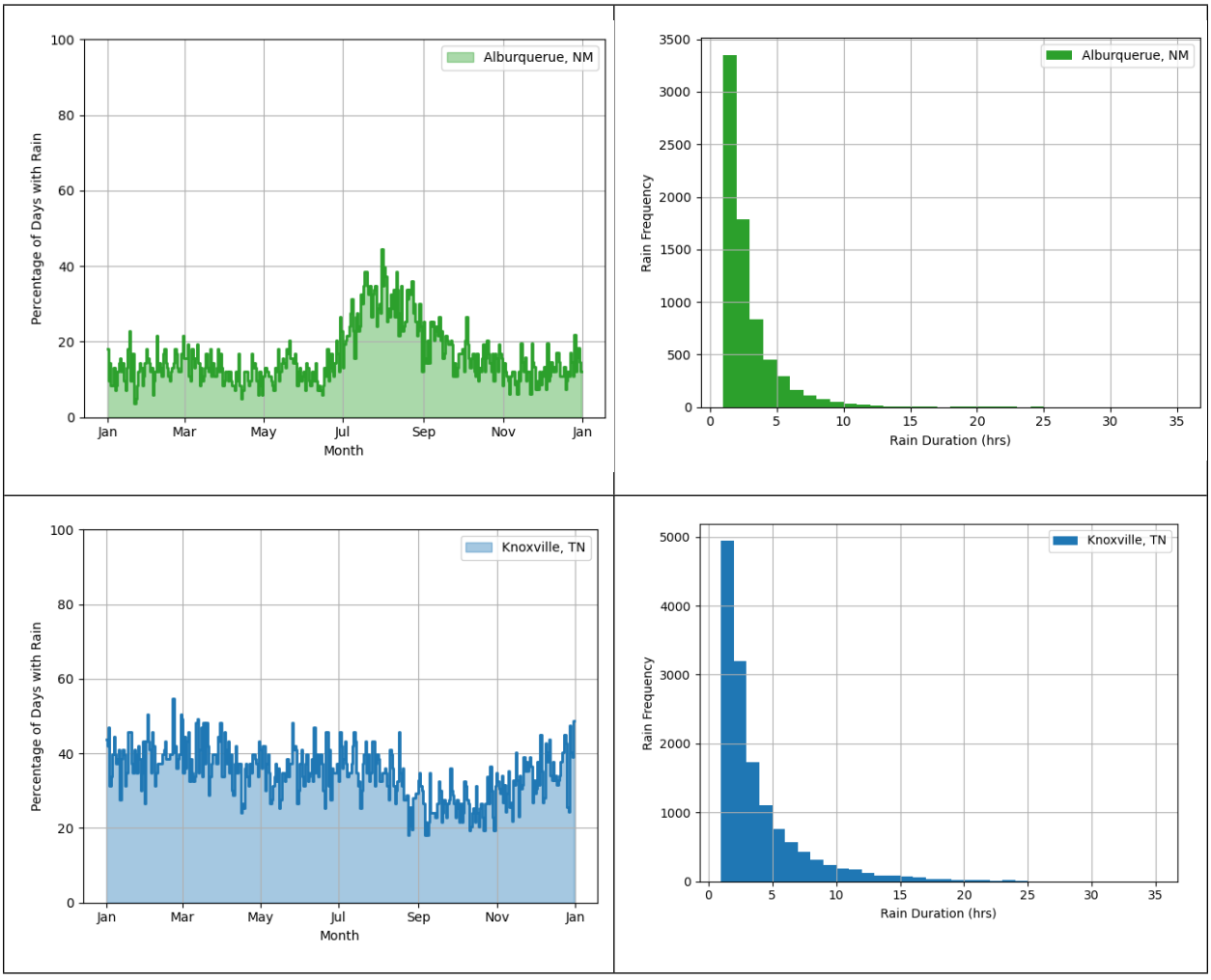
Radon-222 Rainout and Washout

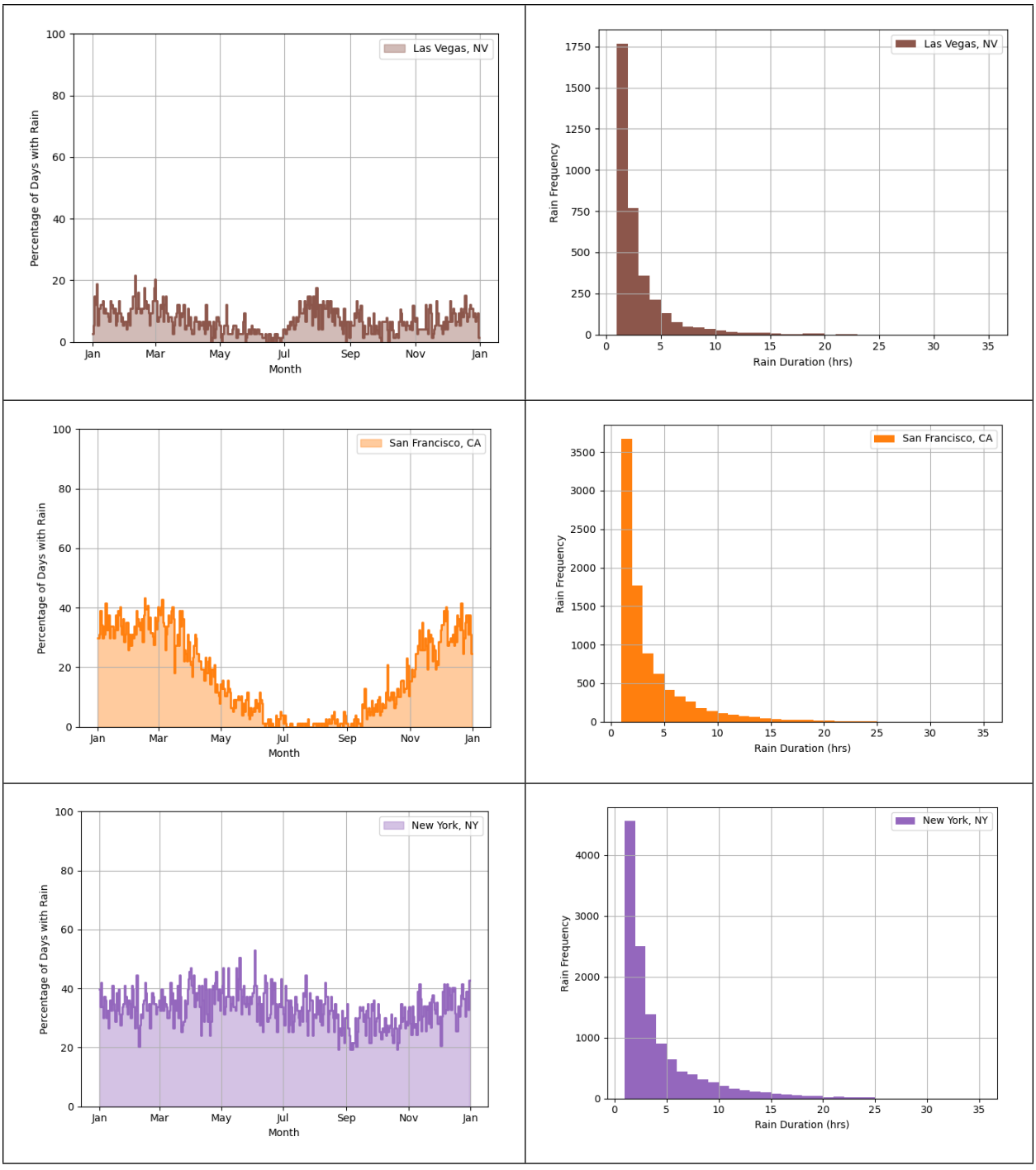
Radon-222 (half-life of 3.8 days) is a naturally occurring radioactive gas that is present in the atmosphere as part of the ^{238}U decay chain. Radon-222 and its daughters collect in clouds and fall to the ground surface via a process called rainout. Likewise, they can be captured by rainfall below the cloud level and deposited on the surface via a process called washout. In both cases, a gamma detector at the ground level can experience increases in the GCR up to a factor of 3 from nominal levels. The primary contributors to this increase are gamma-rays emitted by the ^{222}Rn daughters, ^{214}Pb and ^{214}Bi. Radon-222 decays to ^{214}Pb (half-life of 26.8 min) through a series of alpha decays. The ^{214}Pb in turn beta decays to ^{214}Bi (half-life of 20 min) [8]. Prominent photopeaks from ^{214}Pb include those at 295 and 352 keV. Likewise, prominent photopeaks from ^{214}Bi include 609, 1,120, 1,764, and 2,204 keV. Although these photopeaks are present in typical gamma-ray background spectra, they are amplified during precipitation. In natural uranium, ^{214}Pb and ^{214}Bi are typically in secular equilibrium, but ^{222}Rn progeny in the atmosphere are not typically in equilibrium, and the ratio of ^{214}Pb to ^{214}Bi can even fluctuate during a rain event [8,9,10,11,12]. In one study [10], the authors measured the concentrations of ^{214}Pb and ^{214}Bi in the Kumatori village, Japan, throughout 24 rain events and found that the activity concentration of ^{214}Pb varied between 0.12 and 2.7 Bq/cm^3, and ^{214}Bi varied between 0.11 and 0.27 Bq/cm^3. Major gamma-ray contributors to the gamma-ray GCR increase are listed in Table 4.

Table 4. Predominant gamma-rays produced from ^{222}Rn washout [13]

^{222}Rn daughter	Half-life (min)	Predominant gamma-rays	
		Energy (keV)	Intensity (%)
^{214}Pb	27.06	295.2	18.5
		351.93	35.7
^{214}Bi	19.71	609.3	45.4
		1120.3	14.9
		1764.5	15.3
		2204.1	4.9

The rainfall frequency largely depends on local weather patterns and can be estimated using historical weather data. The National Oceanic and Atmospheric Administration hosts the Global Historical Climatology Network daily database, which contains historical records from more than 100,000 weather stations in 180 countries and territories [14, 15]. From this database, precipitation data from weather stations based in Albuquerque, New Mexico, Knoxville, Tennessee, Las Vegas, Nevada, San Francisco, California, New York, New York, Seattle, Washington, and Washington, D.C., were queried from 1940 to 2022. The left column in Figure 4 shows the percentage of rainy days in each year calculated from ~80 years of daily records contained in the Global Historical Climatology Network database. The National Oceanic and Atmospheric Administration hosts an additional database with hourly precipitation data (Hourly Precipitation Database [16]), which is a bit more limited: data are available for 1948–2014, although this range varies by location. The same weather station was used in each database, usually based at an airport, except for New York, New York, where the data were collected in Central Park. These data were used to determine the average duration of rain events, as illustrated by the histograms in the right column of Figure 4. Summaries of the statistics in Figure 4 are shown in Table 5, including the average amount of time each location experienced rainfall. For example, Knoxville, Tennessee, has an average of 127 days of precipitation per year, with average rain durations of about 3.2 h, for an average of about 48 in. per year. The average percentage of time it is raining in Knoxville is 4.8%.




Las Vegas, NV
Percentage of Days with Rain
Month
Jan
Mar
May
Jul
Sep
Nov
Jan
Las Vegas, NV
Rain Frequency
Rain Duration (hrs)
San Francisco, CA
Percentage of Days with Rain
Month
San Francisco, CA
Rain Frequency
Rain Duration (hrs)
New York, NY
Percentage of Days with Rain
Month
New York, NY
Rain Frequency
Rain Duration (hrs)

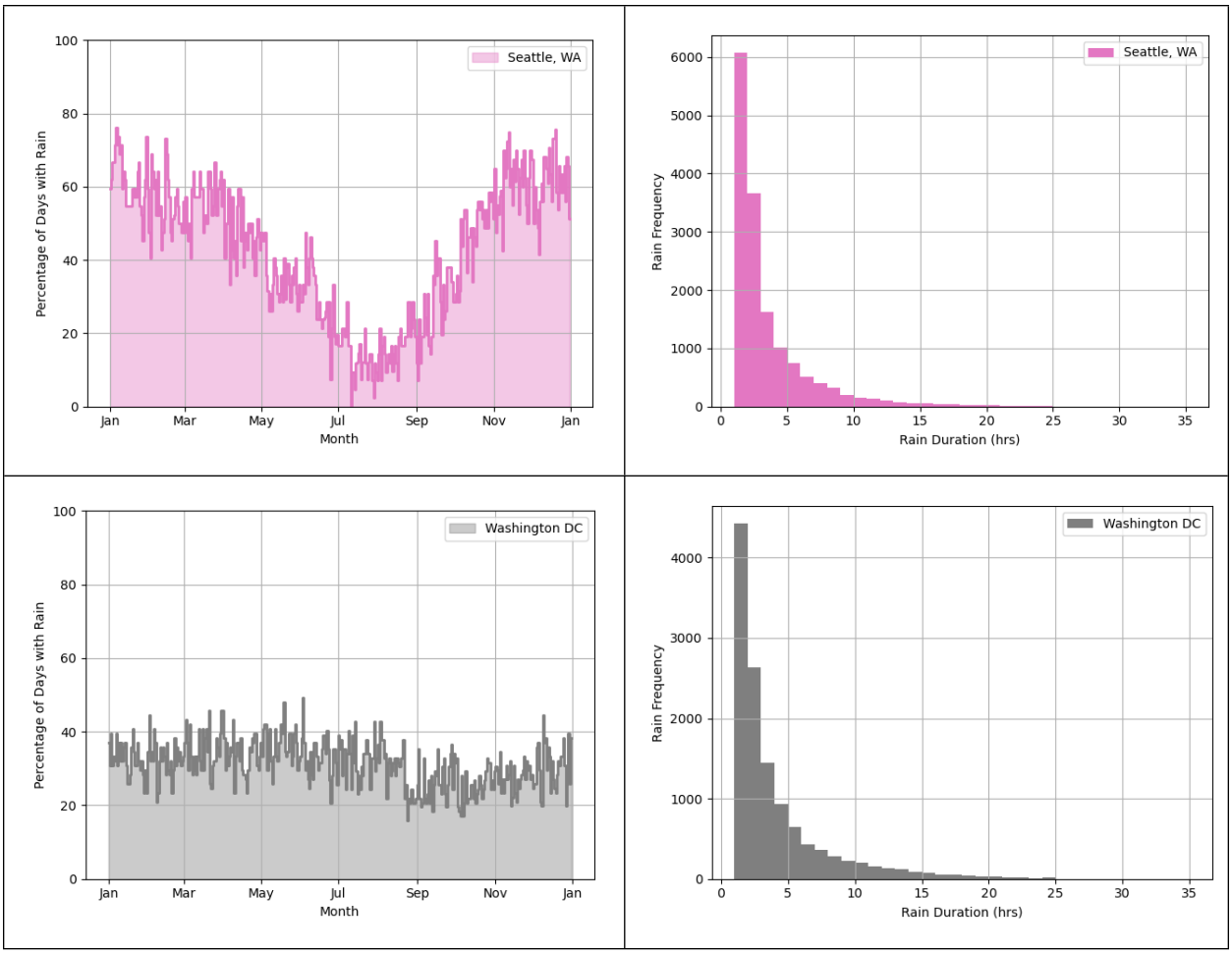


Figure 4. Precipitation data collected from weather stations located in Albuquerque, New Mexico, Knoxville, Tennessee, Las Vegas, Nevada, San Francisco, California, New York, New York, Seattle, Washington, and Washington, D.C. (left column) Percentage of days with rain based on ~80 years of daily data gathered from the Global Historical Climatology Network daily database [15]. (right column) Rain frequency, in hours, gathered from ~60 years of data gathered from the Hourly Precipitation Database for each location [16].

Table 5. Summary precipitation data for Albuquerque, New Mexico, Knoxville, Tennessee, Las Vegas, Nevada, San Francisco, California, New York, New York, Seattle, Washington, and Washington, D.C.

Location	Start	End	Percentage of days with rain (%) [15]			Annual rainfall (in.) [15]			Average rain duration (h) [16]	Recommended rain frequency (%) [16]
			Avg.	Min.	Max.	Avg.	Min.	Max.		
Albuquerque, New Mexico	01/1940	11/2022	16.1	3.6	44.6	8.6	4.1	15.9	2.2	1.1
Knoxville, Tennessee	01/1940	11/2022	34.7	18.1	54.9	48.3	32.5	69.3	3.2	4.8
Las Vegas, Nevada	09/1948	11/2022	7.0	0.0	21.6	4.0	0.3	9.9	2.6	0.6
San Francisco, California	01/1940	11/2022	16.1	0.0	44.4	14.0	2.8	32.4	3.0	2.7

Table 5. Summary precipitation data for Albuquerque, New Mexico, Knoxville, Tennessee, Las Vegas, Nevada, San Francisco, California, New York, New York, Seattle, Washington, and Washington, D.C. (continued)

Location	Start	End	Percentage of days with rain (%) [15]			Annual rainfall (in.) [15]			Average rain duration (h) [16]	Recommended rain frequency (%) [16]
			Avg.	Min.	Max.	Avg.	Min.	Max.		
New York, New York	011940	11/2022	33.2	19.3	53.0	46.8	26.1	80.6	3.7	5.3
Seattle, Washington	01/1948	11/2022	42.5	0.0	76.2	34.4	7.2	50.4	3.1	5.5
Washington, D.C.	01/1940	11/2022	31.3	15.9	49.4	39.5	0.0	66.3	3.4	4.5

The detector response to a short-duration ("delta"), simple rain event is typically characterized by a sharp spike followed by an exponential decay tail, as shown in the plot on the left of Figure 5. Analytical expressions for the detector response to a delta rain event have been derived in previous literature for some portal monitors [8]. For rain events in which the duration is not significantly shorter than the half-lives of ^{214}Pb and ^{214}Bi, the observed temporal variation can become much more complicated. Additionally, multiple consecutive rain events can stack in time to create complex count-rate profiles, as shown in the right-hand panel of Figure 5.

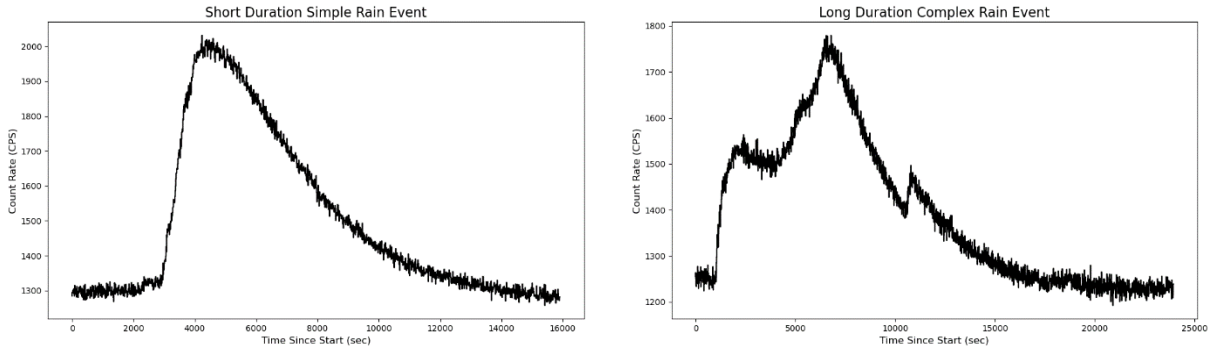


Figure 5. Detector response as a function of time. (left) Example of 2 × 4 × 16 in. NaI(Tl) detector response as a function of time for a short-duration rain event. (right) Example of 2 × 4 × 16 in. NaI(Tl) detector response for a long-duration rain event.

Fast Transients:

In a mobile detector scenario, a variety of geographic structures can cause rapid changes in GCR and spectral shape that should be considered for inclusion (and quantification) in datasets for algorithm evaluation.

1. Bridges: Crossing a bridge over water often leads to a rapid decrease in GCR owing to the low KUT content in both fresh- and saltwater environments. Regarding the KUT contents of water, the concentration of radioisotopes in the U and Th decay series is essentially negligible. However, the ^{40}K concentration, while low compared with that of most terrestrial materials, is still on the order of 13.6 Bq/kg [17]. Consequently, the ratio between the 1,460 keV photopeak and other background photopeaks in the gamma-ray spectrum will appear amplified. Because of both the rapid change in

GCR and the unique change in spectral shape, this scenario should be included in algorithm evaluation datasets aimed at mobile source search. In contrast to crossing bridges over water, crossing bridges over land can lead to other effects and can even increase backgrounds.

2. Tunnels: Underground tunnels can lead to rapid increases in background GCR, depending on the KUT environment that the detector system was in before entering the tunnel and on the KUT concentrations of the material used to construct the tunnel. This situation causes a rapid step change in GCR and a possible change in spectral shape. This effect can generally be considered temporally unique compared with the GCR and spectral deviations arising from typical KUT dynamics in urban search because of the step change followed by a period of uniformity in the count rate and spectral shape as the detector travels through the tunnel. Upon leaving the tunnel, the detector will again rapidly enter a new KUT environment.

3. Overpasses: Traveling underneath an overpass can have a similar response to that encountered in a tunnel, although it is generally shorter in duration, and the variations are generally smaller in amplitude.

Real-world examples of these fast transients—bridges, tunnels, and overpasses—on the total gamma-ray count rate are shown in Figure 6. Data were collected using four summed 2 × 4 × 16 in. NaI(Tl) detectors.

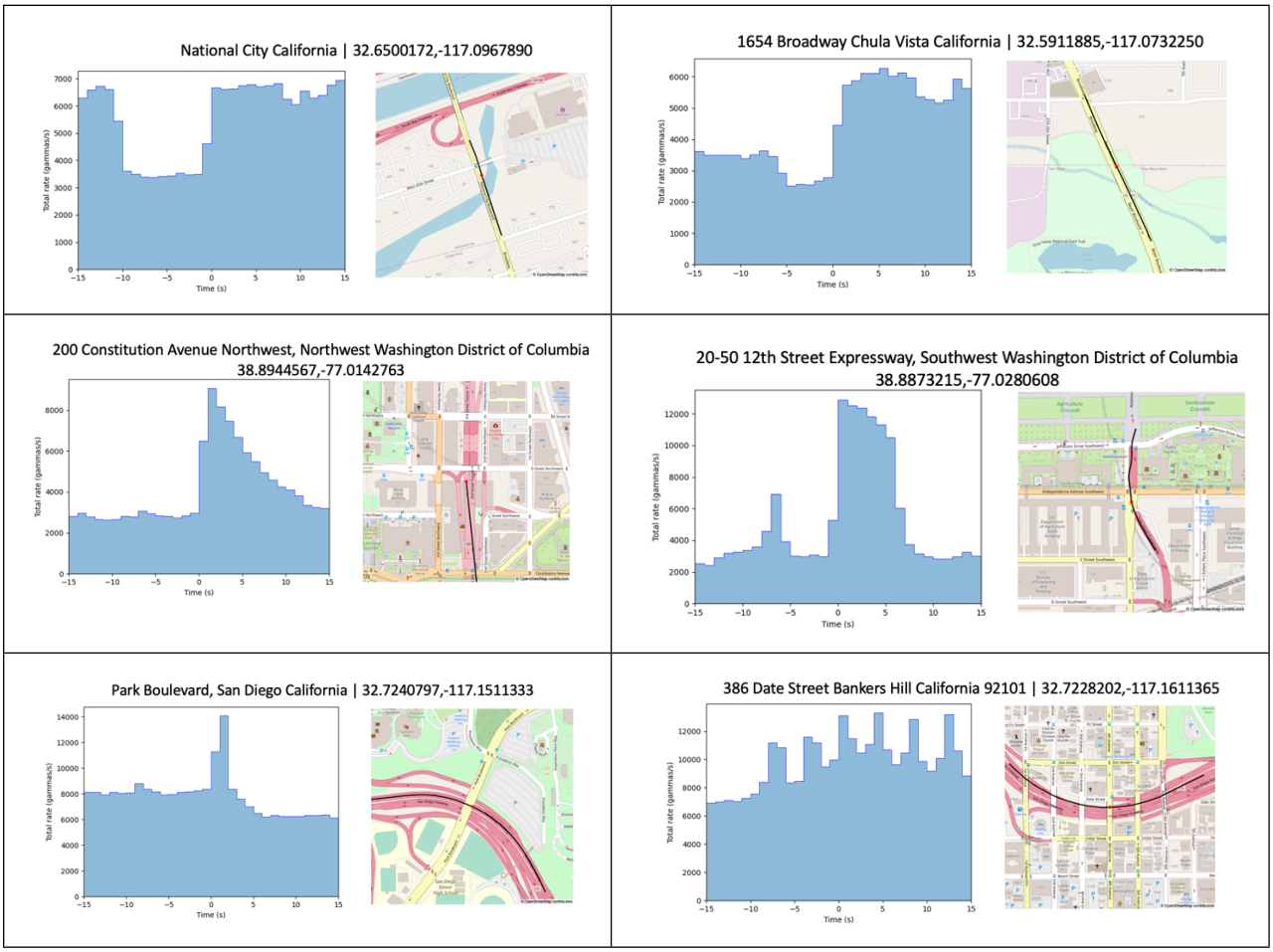


Figure 6. Fast transient examples. (top row) bridges, (middle row) tunnels, and (bottom row) overpasses. Total count rate data were collected using four summed 2 × 4 × 16 in. NaI(Tl) detector systems. The red dot in each map

shows the position of the detector system at time $t = 0$, and the black line shows where the detector was located ±15 s from $t = 0$.

Seawater Characteristics

Maritime

Maritime environments are often characterized by a unique KUT profile that should be considered when developing datasets for algorithm characterization. Both U and Th products generally precipitate in seawater conditions and thus occur at very low concentrations in seawater compared with terrestrial conditions. By contrast, K is present in the form of soluble salts, and its concentration increases with the salinity of the water. The global average activity concentration of ^{40}K in seawater has been reported to be 13.6 Bq/kg [17]. This concentration tends to be relatively uniform across open seas, and variations are generally within ±20% of the average value (Walker, 1990). Extremes can occur in hypersaline bodies of water such as the Dead Sea (178 Bq/kg) and in bodies of water with high levels of freshwater incursion such as the Baltic Sea (4 Bq/kg) [17]. As a general recommendation, datasets should consider including backgrounds with very little U and Th series isotopes and ^{40}K present at an activity of 10.88–16.32 Bq/kg (i.e., average value ± 20%).

Shipping Ports and Other Nearshore Environments

Very nearshore environments such as shipping ports or maritime border crossings can have background radiation characteristics that are further complicated by the effects of the tide. Tides affect the amount of seafloor that is exposed: high tide will decrease and low tide will increase the ratio of terrestrial to seawater surface exposure. This cyclical variation can lead to changes in both the background GCR and spectral shape that will be measured by a nearshore detector. The magnitude of these dynamics will be especially significant for detectors placed at nearshore locations that experience large tide changes, depending on a combination of factors involving the declination of the moon, local seafloor topography, and the size of the body of water in consideration. For algorithms intended to operate on detectors in nearshore environments, training and evaluation datasets should include data with varying terrestrial/seawater exposure ratios. In general, this effect can be achieved by oscillating the count rate in a sinusoidal pattern ±5% from the average count rate. This oscillation will have two peaks (low tides) and two troughs (high tides) during a 24 h period. Figure 7 shows an example of the count rate measured by a seaside radiation portal monitor during a 4 day period. The tidal influences are clear, along with a precipitation-induced radon daughter peak on day 2.

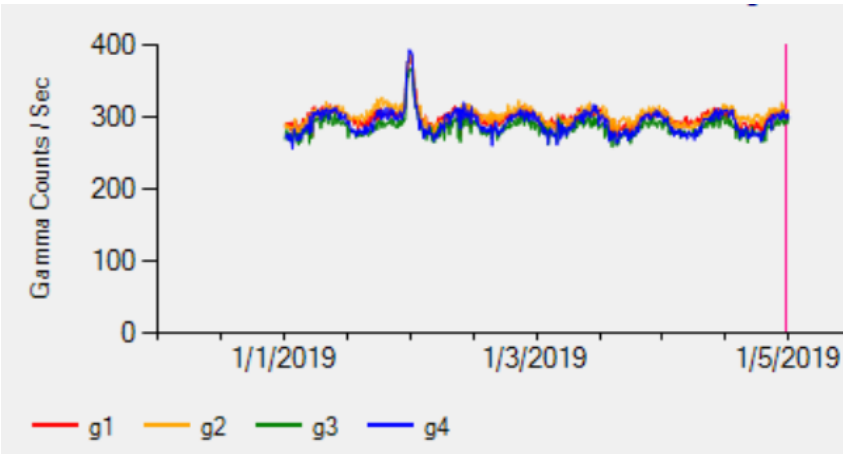


Figure 7. Gamma-ray GCR in a seaside radiation portal monitor measured during a 4 day period, demonstrating tidal-induced sinusoidal count rate fluctuations and a rain event on day 2.

Cosmic Rays

Cosmic-ray primaries (the vast majority of which are solar protons) continually impinge on Earth's atmosphere and create showers of secondary particles. The most well-known cosmic-ray secondaries at the surface are muons, which have a flux of ~1 particle/(cm^2/s). However, cosmic-ray showers also produce photons at the surface, largely in a diffuse power-law continuum with an observed index of ~1.3 in 4 × 4 × 16 in. NaI(Tl) detectors [18] and an additional line feature at 511 keV caused by pair production by the high-energy continuum photons. Although the cosmic component is usually the smallest background contribution to the overall count rate, it is typically the dominant source of background above the 2.614 MeV line of ^{208}Tl. Besides photons and their associated electrons and positrons, other cosmic-ray particles, predominantly muons, may result in detector signatures as well.

Recent findings indicate that the cosmic background component had some correlation with a detector's view of the sky as it traveled through an urban area [19]. This result is expected because buildings and other material can attenuate the photons. Other sources of variability that may occur are temporal variability resulting from changes in the flux of the cosmic-ray primaries, such as during the day–night cycle or on longer scales such as the 11 year solar cycle, although such changes are typically small.

2.1.2 Metrics for Background Variability

Two sets of metrics can be defined for quantifying (1) general count rate and spectral variability across a given dataset that captures the effects of all background contributors and (2) how well a dataset covers specific events in terms of number and type.

Spectral Variability

This section defines a single quantitative measure of spectral variability within gamma-ray spectral datasets. Gamma-ray spectra can be treated as discrete Poisson distributions. Therefore, approaches from information theory can be leveraged to describe the information contained within each spectrum and, more importantly for this application, the relative information among multiple spectra. A symmetrized and smoothed version of the Kullback–Leibler divergence (KLD or D_{KL}), called the Jensen–Shannon divergence (JSD), can be used as a metric for measuring the (dis)similarity between two probability distributions, P and Q, both defined on the same probability space χ.

Let $D_{KL}(P||Q)$ be the divergence from gamma-ray spectrum Q to a reference gamma-ray spectrum P as defined in Eq. (1).

$$D_{KL}(P \parallel Q) = \sum_{(x \in \chi)} P(x) \log\left(\frac{P(x)}{Q(x)}\right), \#(1)$$

Because the KLD is asymmetric ($D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$) and does not satisfy the triangle inequality, it cannot be treated as a true metric. The JSD, however, is a true metric and can be used to measure the similarity between two probability distributions, in this case, two gamma-ray energy spectra, S_0 and S_1. For the JSD to be valid, the integral of each spectrum should sum to 1, which can be achieved by dividing the count spectrum by the total number of counts in the spectrum, N, such that

$$\hat{S}_0 = \frac{S_0}{N_0}, \text{ and } \hat{S}_1 = \frac{S_1}{N_1}. \#(2)$$

The JSD between the two spectra is then defined by Eq. (3). The JSD is between 0 and 1, with 0 indicating exact similarity and 1 indicating no similarity. Although JSD is fairly robust to noise from low

counting statistics, the spectrum integration time, T_s, should be set such that the gross count uncertainty in each spectrum is less than or equal to 1% for consistency to enable dataset comparison.

$$\text{JSD} = \sqrt{\frac{D_{KL}(\hat{S}_0 \parallel \hat{S}_1) + D_{KL}(\hat{S}_1 \parallel \hat{S}_0)}{2}}. \#(3)$$

The JSD can be used to measure spectral variability within a dataset. First, the mean spectrum for the entire dataset, $\overline{S}$, is calculated. Then, the JSD between each spectrum in the dataset and $\overline{S}$ is calculated and recorded. All spectra should be integrated using the same integration time. The JSD of each spectrum can then be plotted on a box and whisker plot to view the distribution of spectral variability within the dataset. An example is shown in Figure 8, with interpretation provided below.

Interpretation of Spectral Variability and Guidance

- The JSD is always a positive value between 0 and 1, with 0 indicating no variability between the spectrum and the mean spectrum, and 1 indicating maximum variability.

- The dataset developer should consider the following metrics from the box plot:
  - Median variability
  - Interquartile range (IQR): majority of data fall within this block
  - Q1 − 1.5IQR: data below this value are considered outliers
  - Q3 + 1.5IQR: data above this value are considered outliers
  - Fliers above Q3 + 1.5IQR: these values indicate high-variability outliers, which can be an important challenge for algorithms

- The JSD metric is not completely independent of detector response and counting uncertainties (although it is fairly robust to these issues compared with other approaches). Thus, this metric should be computed on data from the same detector on spectra with the same integration time.

Figure 8 shows spectral variability among three datasets, each containing gamma-ray spectra (1,000 bins, 30–3000 keV) collected from the same 2 × 4 × 16 in. NaI(Tl) detector deployed over the course of several years in a static location in East Tennessee, USA. Two datasets were obtained, each containing 452,075 spectra sampled from times spanning 3 years. The first dataset contains data for times when it was raining, and the other contains data for times when it was not raining. The third dataset is a simple combination of the other two datasets. All spectra had an integration time of 10 s. The following items are an example of the types of observations that can be made from Figure 8:

- Although the rain-only dataset spanned a wide range of variability, most of the data are between 0.09 and 0.1. These values are lower than for the typical background (no-rain) dataset, which had a higher median but lower range in variability than the rain dataset.

- The rain dataset contained a significant number of high-variability fliers, whereas the typical background dataset did not.

- Combining both datasets yielded a dataset that had the high-median variability of the typical background dataset and the wider variability range, high-variability fliers of the rain dataset.

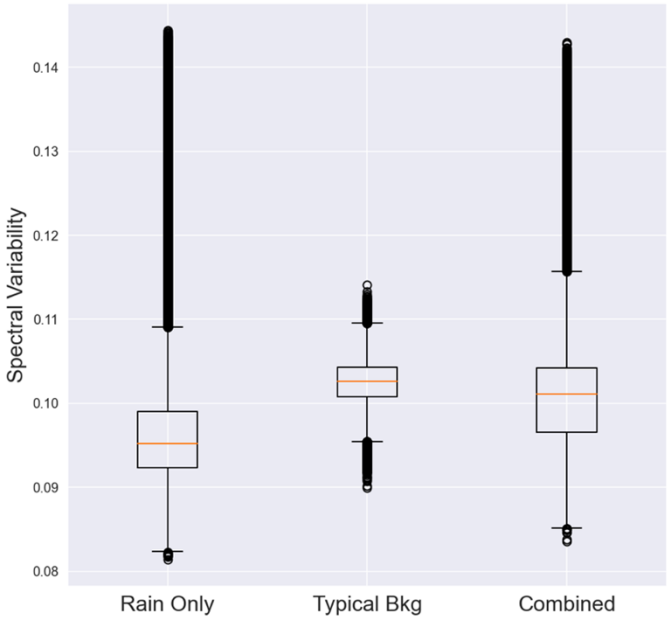


Figure 8. Spectral variability within three datasets, one containing gamma-ray spectra during rain events only, one containing typical non-rain background spectra spanning several years and seasons, and one that is a combination of both. Interpretation of this data is outlined in Section 2.1.2.

Metrics for Specific Conditions:

Beyond the spectral and gross count variability metrics, the dataset coverage should be quantified for the different specific conditions (e.g., KUT variations, rain, transients) that were outlined in Section 2.1.1.

- Mobile detector vs. static detector:
  - Every dataset should be labeled as static, mobile, or both.

- KUT variations:
  - For synthetic datasets, list the types of materials included, KUT distributions for each.
  - For real-world datasets, if detector is mobile, then list the types of environments spanned by the dataset: urban, suburban, rural, coastal maritime, near-shore maritime, offshore maritime.
    - If possible, outline dataset percent coverage for each environment type.

- Rainout and washout (^{222}Rn):
  - Number of precipitation events and their duration.
  - Percent dataset coverage.

- Fast transients
  - Number of bridges, tunnels, and overpasses.
  - Percent dataset coverage for each.

2.2 SOURCES

In addition to varying background, search instruments commonly encounter NORM, medical radionuclides (both in shipment and in vivo), and industrial sources. Bulk quantities of nuclear material or actinides are rarely encountered in the real world, but they are often the primary objective for detection

and identification and therefore important to include. Aside from a specific search mission in which a source set may be tailored, a general-purpose set to test algorithms should span a wide range of radionuclides and combinations, activities, ages, and shielding configurations. Although standards such as ANSI-N42.34-2021 [2] are useful for improving early phase-development algorithms, the current state of the art demands a more comprehensive and challenging suite to reflect realities of search missions and make meaningful comparisons of high-performing algorithms. For example, the variety of medical sources likely to be encountered greatly exceeds the ^{131}I, ^{67}Ga, ^{99m}Tc, and ^{201}Tl specified in the standard. Figure 9 illustrates the number of times different medical radionuclides were identified by a radioisotope identification device (RIID) on-board algorithm from a set of data collected from 2008 to 2019. This chart is not complete, nor does it necessarily represent the correct answer, but it demonstrates the variety in medical radionuclides.

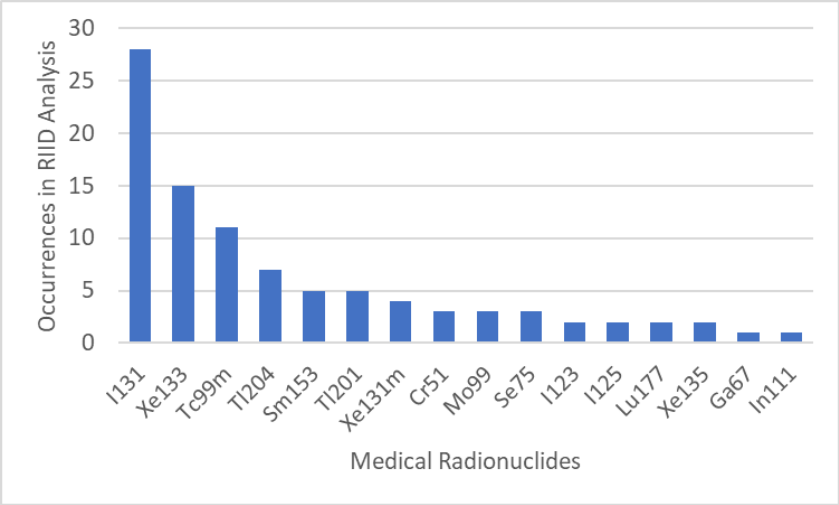


Figure 9. Medical radionuclides identified by RIID.

2.2.1 Recommendations

Signal

When considering sources, a signal-to-noise ratio (SNR) metric that describes the source encounter must be defined. A variety of ways to calculate SNR can be envisioned, however, in this section we describe the "optimal SNR", which is the maximum SNR experienced within the entire duration of the source *encounter* for a range of integration times. The source counts, S, and the background counts, B, are calculated sequentially over the duration of the source encounter with a given integration time, T, and SNR is calculated using Eq. (4). The "optimal SNR" is then the maximum SNR calculated during the source encounter across all integration times evaluated. Integration times spanning from 0.1 to 10 s are generally applicable for most moving detector/static source type encounters at typical urban driving speeds.

$$\mathrm{SNR} = \frac{S}{\sqrt{S+B}} . \#(4)$$

Nontrivial search missions are signal starved; only in that regime can high-performing algorithms be distinguished from one another. Therefore, this space is the most important and should comprise the bulk of the dataset. However, medium SNR should not be excluded, and the importance of high and very high SNR scenarios should not be discounted. In the very high SNR regime, systematic uncertainties dominate and nonlinear effects such as random pile-up and peak-shape distortion occur in gamma detectors. These

effects can be challenging for identification and classification algorithms. Thus, a useful search dataset should include mostly low SNR sources but should span all SNR ranges. Dataset creators, curators, and testers should be wary of datasets that contain mostly high SNR sources, because algorithm results on these datasets can instill false confidence. Narrow SNR range sets can also make comparisons between algorithms difficult if not meaningless.

Inorganic scintillators are the most common gamma detectors employed in this mission space because of their reasonable resolution, cost, and ability to grow large crystals. The most common neutron detector is a generic thermal detector based on thermal neutron capture by ^{3}He or ^{6}Li. The discussion is limited to these two types. Table 6 and Table 7 provide recommended SNR ranges for gamma and neutron detection, respectively. Table 8 lists recommended SNR ranges for identification (or classification). These recommendations should be good starting points, but the SNRs chosen for a comparison of algorithms should be low enough to provide a challenge to the algorithms (not perfect performance) but high enough that the algorithms provide some capability. These values will vary considerably depending on the complexity of the spectrum.

The difficulty in detecting or identifying radiation sources from a spectroscopic gamma detector is driven by how easily the radiation sources are confused with other sources, including background. In addition to the signal level, the uniqueness of the spectral shape significantly influences spectroscopic detection and identification algorithm performance. Because shielding drastically affects this shape, enumerating a subjective assessment on the uniqueness for all sources is likely too onerous a task. However, a source (and its shielding) can reasonably be categorized into three broad categories of spectral uniqueness: low, medium, and high. A prototypical low-uniqueness source is ^{226}Ra, especially when shielded and in the signal-starved regime. On the other end, monoenergetic sources with good penetration through shielding are easy to distinguish. Future work should include algorithms and metrics to quantify this uniqueness in a more systematic manner.

Generally, datasets should target 70% of the data in the low signal range, 20% in the medium range, and 10% in the high range.

Table 6. Recommended SNR ranges for gamma detection

		Source spectral uniqueness		
		Low (amorphous) (e.g., heavily shielded ^{226}Ra)	**Medium** (confusable) (e.g., lightly shielded ^{131}I)	**High** (distinct) (e.g., ^{137}Cs)
Signal level	**Low** (difficult)	>0–5	>0–3	>0–2
	Medium	5–9	3–6	2–5
	High (easy)	>9	>6	>5+

Table 7. Recommended SNR ranges for neutron detection

Signal level	**Low** (difficult)	>0–2
	Medium	2–5
	High (easy)	>5

Table 8. Recommended SNR ranges for gamma detection identification and classification

		Source spectral uniqueness		
		Low (amorphous) (e.g., shielded ^{226}Ra)	**Medium** (confusable) (e.g., shielded ^{133}Ba)	**High** (distinct) (e.g., ^{137}Cs)
Signal level	**Low** (difficult)	>0–6	>0–5	>0–3
	Medium	6–15	5–10	3–5
	High (easy)	>15	>10	>5

Selection and Shielding

Datasets should include a variety of radionuclides and combinations. A short list of radionuclides, such as the 14 sources (and 7 combinations) described in ANSI-N42.34-2021 [2], are intended to allow reasonable measurement campaigns to be conducted for testing. In practice, sensors and algorithms must be capable of performing with a much larger set of nuclides, combinations, and shielding conditions to reflect the full set of spectra that may be encountered. Success with the simpler set does not guarantee success with the full set. Developers could train or tune their algorithm to the simpler set, achieving what appears to be a very high level of performance while disregarding or deprioritizing most other radionuclides. The real world can and does present a much more diverse set, and datasets to test algorithms should be more comprehensive. It may not be possible to construct a single dataset that contains all recommended source configurations. Multiple datasets can be combined and augmented with simulated data to allow for a comprehensive assessment.

Shielding (and scattering which can vary with distance and other measurement configuration variations) can drastically alter the spectral shape measured by gamma detectors. In extreme cases, only shoulders are present and no photopeaks. Even common medical sources that emit gammas and x-rays over a wide range of energies can present completely different signatures when the low-energy region is readily shielded by a human body. Ranges of appropriate shielding for each radionuclide are presented in Tables 10 through 14, but a good dataset will not use these discrete points. Rather, it must contain a healthy range of intermediate shielding configurations and heterogeneous combinations of each.

Table 9 provides recommended sources, and each references a shielding configuration table, including example shield thicknesses. For simplification, the shielding thicknesses can be interpreted as spherical shells, but they can be almost any geometry. The spectral uniqueness for each source is assessed to give guidance on how unique each gamma-ray spectral signatures is. Isotopes with one or two prominent photopeaks which do not overlap with other sources on the list, are generally rated with higher spectral uniqueness and those with more photopeaks that overlap with other source photopeaks are rated with

lower spectral uniqueness. The spectral uniqueness gives guidance on how to combine sources and which sources are needed to span the relevant range of threat isotopes. For experimental or simulated datasets, simulation codes such as Gamma Detector Response and Analysis Software–Detector Response Function (GADRAS-DRF) and Gamma Designer can be very helpful in scoping the appropriate ranges of activities, distances, and times needed to provide spectra in the desired SNR ranges.

Several sources in Table 9 have an age or equilibrium specified in parenthesis after the source name. If no age is given, then the age can be assumed to be zero, except for the nuclear material.

Table 16 gives specific age recommendations for nuclear material. Secular equilibrium (sec. eq.) is a commonly understood term. This condition can be achieved when the half-life of the parent is much greater than that of all the daughters. After roughly seven half-lives of the longest-lived daughter, all radionuclides in the decay chain have equal activity. Prompt equilibrium (prompt eq.) is a less common term that describes a special condition in which the decay chain contains monotonically decreasing half-lives until a longer-lived daughter is met, breaking the monotonicity. If the maximum half-life of the monotonically decreasing daughters is less than the radium age, then a chemical separation could have occurred, and equilibrium could be achieved with only the shorter-lived daughters. The most familiar example is the ^{238}U decay chain, which can achieve prompt eq. with daughters ^{234}Th, ^{234m}Pa, and ^{234}Pa in roughly 7 months. The half-life of the ^{234}U daughter (245,000 years) breaks the monotonicity and requires millions of years to achieve equilibrium.

Although specific shielding recommendations are given in Table 10 through

Table 14, developers are strongly encouraged not to use those exact values. Instead, they should interpolate between the shielding types with at least a few different areal densities for each atomic number, especially if they are developing a simulation-based data set. Generally, a logarithmic distribution of shielding thicknesses is best to provide a more even distribution of spectral shapes—the spectral shape changes more quickly for thin shielding variations than for heavy shielding variations.

This source list, combined with the shielding configurations and signal ranges, can create an enormous dataset. Although this level of complexity is required to distinguish high-performing and mature algorithms, it may be too burdensome for newer algorithm developers. Thus, for each source, an inclusion level is defined:

- Inclusion Level 1: initial exploration of novel approaches.
- Inclusion Level 2: initial testing only for more complex algorithms.
- Inclusion Level 3: required for any performance evaluation intended to make claims of algorithm performance in nuclear security applications.

These inclusion levels are cumulative, so the full list of nuclides for Inclusion Level 2 also includes those listed as Level 1, and Level 3 includes all those listed as Level 1 and Level 2. As algorithms advance, they should increase the complexity of the dataset used to test. These levels may be usable as requirements to advance the technology readiness level of algorithms. However, **algorithms using different inclusion levels should never be compared**. A limited-scope Level 1 algorithm can easily outperform a general-purpose Level 3 algorithm when tested only against Level 1 sources. Likewise, a Level 3 algorithm can outperform a Level 1 algorithm when tested against all Level 3 sources. The comparisons are meaningless. Care should be taken to understand each algorithm's assumptions and the sources it was designed to identify.

Table 9. Recommended radionuclides and shielding.

Source	Source class	Shielding table	Spectral uniqueness	Inclusion level
^{40}K	NORM	Table 10	Medium	1
^{226}Ra (sec. eq.)	NORM	Table 10	Low	1
^{232}Th (sec. eq.)	NORM	Table 10	Low	1
^{227}Th (10 days)	NORM	Table 10	Medium	3
^{228}Th (sec. eq.)	NORM	Table 10	Low	2
^{138}La	NORM	Table 10	High	3
^{140}La	NORM	Table 10	Medium	3
^{176}Lu	NORM	Table 10	Medium	3
^{220}Rn (sec. eq.)	NORM	Table 10	High	2
^{222}Rn (prompt eq.)	NORM	Table 10	Low	2
^{18}F	Medical	Table 11	High	1
^{67}Ga	Medical	Table 11	Medium	1
^{90}Sr (sec. eq.)	Medical	Table 11	Low	1
^{99}Mo (sec. eq.)	Medical	Table 11	High	1
^{99m}Tc	Medical	Table 11	Medium	1
^{111}In	Medical	Table 11	High	1
^{114m}In	Medical	Table 11	Medium	1
^{123}I	Medical	Table 11	Medium	1
^{125}I	Medical	Table 11	Low	1
^{131}I	Medical	Table 11	Medium	1
^{177}Lu	Medical	Table 11	Medium	2
^{117m}Lu (sec. eq.)	Medical	Table 11	Medium	2
^{201}Tl	Medical	Table 11	Medium	1
^{202}Tl	Medical	Table 11	High	1
^{204}Tl	Medical	Table 11	Low	1
^{51}Cr	Medical	Table 11	High	2
^{82}Rb	Medical	Table 11	High	2
^{103}Pd (sec. eq.)	Medical	Table 11	Medium	2
^{131m}Xe	Medical	Table 11	Medium	3
^{133}Xe	Medical	Table 11	Medium	2
^{135}Xe (prompt eq.)	Medical	Table 11	Medium	3
^{153}Sm	Medical	Table 11	Medium	2
^{167}Ho (sec. eq.)	Medical	Table 11	Medium	3
^{223}Ra (sec. eq.)	Medical	Table 11	Medium	3
50 kVP x-ray	Medical	Table 11	Low	3
120 kVP x-ray	Medical	Table 11	Low	1
160 kVP x-ray	Medical	Table 11	Low	2
^{57}Co	Industrial	Table 12	Medium	1
^{60}Co	Industrial	Table 12	High	1
^{88}Y	Industrial	Table 12	High	1

Table 9. Recommended radionuclides and shielding (continued).

Source	Source class	Shielding table	Spectral uniqueness	Inclusion level
^{133}Ba	Industrial	Table 12	Medium	1
^{137}Cs (sec. eq.)	Industrial	Table 12	High	1
^{152}Eu	Industrial	Table 12	Low	1
^{154}Eu	Industrial	Table 12	Low	1
^{166m}Ho	Industrial	Table 12	Medium	1
^{192}Ir	Industrial	Table 12	Low	1
^{207}Bi (sec. eq.)	Industrial	Table 12	High	2
^{22}Na	Industrial	Table 12	High	2
^{46}Sc	Industrial	Table 12	High	2
^{54}Mn	Industrial	Table 12	High	3
^{56}Co	Industrial	Table 12	Medium	2
^{58}Co	Industrial	Table 12	High	3
^{65}Zn	Industrial	Table 12	High	2
^{75}Se (sec. eq.)	Industrial	Table 12	Medium	2
^{85}Kr	Industrial	Table 12	Medium	2
^{89}Zr (sec. eq.)	Industrial	Table 12	High	3
^{109}Cd (sec. eq.)	Industrial	Table 12	Low	3
^{110m}Ag (sec. eq.)	Industrial	Table 12	Medium	3
^{113}Sn (sec. eq.)	Industrial	Table 12	Medium	2
^{124}Sb	Industrial	Table 12	Medium	2
^{139}Ce	Industrial	Table 12	Medium	3
^{166}Ho	Industrial	Table 12	High	3
^{187}W	Industrial	Table 12	Medium	3
^{198}Au	Industrial	Table 12	High	3
^{252}Cf	Industrial (neutron)	Table 13	Low	1
AmBe	Industrial (neutron)	Table 13	Medium	1
^{232}U	Nuclear material	Table 14	Low	1
^{233}U	Nuclear material	Table 14	Low	2
^{235}U	Nuclear material	Table 14	Medium	1
^{238}U	Nuclear material	Table 14	Low	1
^{237}Np	Nuclear material	Table 14	Medium	2
^{238}Pu	Nuclear material	Table 14	High	1
^{239}Pu	Nuclear material	Table 14	Medium	1
^{241}Am	Nuclear material	Table 14	Medium	1

The inclusion levels can also be applied to the recommended shielding configurations. The general guidelines are as follows:

- Level 1: specified discrete shielding configurations used; can ignore heavy and extreme shielding.
- Level 2: all specified discrete shielding levels are covered to fully span the shielding range.

- Level 3: the shielding configurations are interpolated to include intermediate levels.

Table 10. Recommended NORM shielding configurations

Shielding type	Shielding level	Example shielding
Self-shielding matrix	none	N/A
	Light (20 g/cm^2 of low-*Z*)	20 cm wood
	Moderate (50 g/cm^2 of low-*Z*)	50 cm wood
	Heavy (100 g/cm^2 of mid-*Z*	13 cm iron
External shield	Bare	N/A
	Light (1 g/cm^2 of mid-*Z*)	1.9 mm iron
	Moderate (15 g/cm^2 of mid-*Z*)	2 cm iron

Table 11. Recommended medical shielding configurations

Shielding type	Shielding level	Example shielding
Self-shielding matrix	None	N/A
	Light (10 g/cm^2 of low-*Z*)	10 cm polyethylene
	Moderate (30 g/cm^2 of low-*Z*)	30 cm polyethylene
External shield	Bare	N/A
	Light (10 g/cm^2 of low-*Z*)	4 cm aluminum
	Moderate (50 g/cm^2 of low-*Z*)	50 cm polyethylene
	Moderate (30 g/cm^2 of mid-*Z*)	4 cm iron
	Heavy (20 g/cm^2 of high-*Z*)	2 cm lead

Table 12. Recommended industrial shielding configurations

Shielding type	Shielding level	Example shielding
Self-shielding matrix	None	N/A
External shielding	Bare	N/A
	Light (10 g/cm^2 of low-*Z*)	4 cm aluminum
	Moderate (50 g/cm^2 of low-*Z*)	50 cm polyethylene
	Moderate (10 g/cm^2 of mid-*Z*)	2 cm iron
	Moderate (30 g/cm^2 of mid-*Z*)	4 cm iron
	Heavy (30 g/cm^2 of high-*Z*)	3 cm lead
	Heavy (50 g/cm^2 of high-*Z*)	5 cm lead
	Extreme (160 g/cm^2 of high-*Z*)	9 cm depleted uranium

Table 13. Recommended neutron source shielding configurations

Shielding type	Shielding level	Example shielding
External shielding	Bare	N/A
	Light	1 cm polyethylene
	Heavy	8 cm polyethylene

	Heavy	8 cm borated polyethylene

Table 14. Recommended nuclear material shielding configurations

Shielding type	Shielding level	Example shielding
Self-shielding	Light	<1 g sphere
	Moderate	100 g–1 kg sphere
	Heavy	>1 kg sphere
External shielding	Bare	N/A
	Light (10 g/cm^2 of low-*Z*)	10 cm polyethylene
	Light (10 g/cm^2 of mid-*Z*	2 cm iron
	Moderate (30 g/cm^2 of low-*Z*)	30 cm polyethylene
	Heavy (30 g/cm^2 of mid-*Z*)	6 cm iron
	Heavy (20 g/cm^2 of high-*Z*)	2 cm lead

Combinations

Common medical sources contain contaminants in varying amounts, and as the contaminants decay with different radiological (and sometimes biological) half-lives; the spectral signature is dynamic. Medical and industrial sources are also shipped and stored together in bundles. Nuclear material isotopics can also vary based on the initial grade and age. Finally, concerns often arise about "masking," where a nuclide of concern (such as special nuclear material [SNM]) is placed near a more commonly found radioactive "nuisance" source such as medical nuclides, NORM, or industrial sources such as ^{60}Co or ^{137}Cs.

Thus, source combinations are commonly encountered in the real world, and identification and classification algorithms should be able to, at a minimum, identify the nuclide responsible for most of the measured gamma-rays, detect that a combination is present, and not provide incorrect assessments. For a masking configuration, the hidden nuclide should be found. However, all possible combinations of sources and shielding configurations create an unmanageably large dataset. Therefore, source bundling recommendations shown in Table 15 use the same inclusion levels as the source selection recommendations. Nuclear material combinations, which are specified as isotopic compositions, grades, enrichments, and ages, are shown in Table 16.

Furthermore, if a developer is creating a simulated dataset at Inclusion Level 3, then the creator should select sample pairs of individual sources and specified shielding configurations to test the algorithms' ability to distinguish masking configurations of concern. In addition to selecting a challenging SNR for the masked source, the ratio of counts measured from the masking source to the masked source should be varied to provide a reasonable challenge to the algorithms. Typically, this ratio will range from 3 to 100, depending on the SNR of the masked source. As such, tens of test combinations should be used for each masking pair, with masking pairs covering several of the most common nuisance sources of each type (medical, NORM, industrial), masking at least several configurations of sources of concern.

Combinations of Radionuclides in Dataset

Table 15 and Table 16 list the combinations of radionuclides that should be included for mixtures normally encountered, and the text provides guidelines for inclusion of mixtures that might be used to mask a radionuclide of concern with a strong signal from a radionuclide commonly found in legitimate activities. The metric for expected combinations is thus the fraction of the combinations suggested in

Table 15 and Table 16, and the metric for masking combinations is the number of masking nuclides provided with the appropriate range of masked materials and masking intensity ratios.

Table 15. Recommended NORM radionuclide combinations.

Source mixture	Source class	Justification	Inclusion level
^{238}U (prompt eq.) + ^{226}Ra (sec. eq.)	NORM	Uranium ore	1
^{226}Ra (age 0) + ^{222}Rn (age 0)	NORM	Radium with variable radon emanation	1
^{232}Th (age 1–20 years) + (variable daughters)	NORM	Freshly separated thorium	1
^{232}Th (sec. eq.) + ^{238}U (sec. eq.)	NORM	Monazite	1
^{226}Ra + ^{222}Rn + ^{228}Ra + ^{220}Rn	NORM	Propane tanks	2
^{227}Th + ^{138}La + ^{176}Lu	NORM	Mischmetal	3
^{138}La + ^{137}La + ^{140}La	NORM	Rare earth element	3
^{201}Tl + ^{200}Tl + ^{202}Tl	Medical	Common contaminant(s)	1
^{177}Lu + ^{177m}Lu	Medical	Common contaminant(s)	1
^{111}In + ^{114m}In	Medical	Common contaminant(s)	1
^{133}Xe + ^{133m}Xe + ^{131m}Xe	Medical	Common contaminant(s)	1
^{124}I + ^{123}I + ^{125}I	Medical	Common contaminant(s)	1
^{125}I + ^{126}I	Medical	Common contaminant(s)	1
^{153}Sm + ^{152}Eu + ^{154}Eu + ^{156}Eu	Medical	Common contaminant(s)	1
^{88}Y + ^{90}Y	Medical	Common contaminant(s)	2
^{131}Cs + ^{132}Cs + ^{136}Cs	Medical	Common contaminant(s)	2
^{131}Cs + ^{75}Zr	Medical	Real-world observed	3
^{123}I + ^{121}Te	Medical	Real-world observed	3
^{211}At + ^{211}Po	Medical	Real-world observed	3
^{56}Co + ^{57}Co + ^{58}Co	Medical	Common contaminant(s)	2
^{58}Co + ^{57}Co + ^{60}Co	Medical	Common contaminant(s)	2
^{64}Cu + ^{62}Cu	Medical	Real-world observed	3
^{223}Ra + ^{219}Ra + ^{211}Bi + ^{211}Pb	Medical	Real-world observed	3
^{117m}Sn + ^{113}Sn	Medical	Real-world observed	3
^{82}Br + ^{42}K + ^{24}Na	Medical	Real-world observed	3
^{56}Co + ^{52}Mn + ^{54}Mn + ^{48}V + ^{183}Re + ^{184}Re	Medical	Real-world observed	3
^{166m}Ho + ^{154}Eu	Medical	Real-world observed	3
^{86}Rb + ^{134}Cs	Medical	Real-world observed	3
^{90}Y + ^{152}Eu + ^{154}Eu + ^{57}Co + ^{60}Co	Medical	Real-world observed	3
^{252}Cf + ^{249}Cf	Industrial (Neutron)	Common contaminant(s)	1
^{252}Cf + ^{137}Cs	Industrial	Common bundling	1
AmBe + ^{137}Cs	Industrial	Common bundling	1
^{192}Ir + ^{46}Sc + ^{124}Sb	Industrial	Common bundling	1

Table 15. Recommended NORM radionuclide combinations (continued)

^{124}Sb + ^{90}Y	Industrial	Common bundling	1
^{241}Am + ^{243}Am	Industrial	Common bundling	1
^{60}Co + ^{57}Co + ^{58}Co	Industrial	Common contaminant(s)	1
^{137}Cs + ^{134}Cs	Industrial	Common contaminant(s)	1
Bi-213 + Tl-209	Industrial	Real-world observed	3
^{152}Eu + ^{154}Eu + ^{155}Eu	Industrial	Real-world observed	3
^{56}Mn + ^{241}Am	Industrial	Real-world observed	3
^{95}Zr + ^{181}Hf	Industrial	Real-world observed	3

Table 16. Recommended nuclear material radionuclide combinations and characteristics

Source	Characteristics	Recommended value(s)	Inclusion level
Plutonium	Grades	6% ^{240}Pu	1
		14% ^{240}Pu	1
		19% ^{240}Pu	1
		27% ^{240}Pu	2
		81% ^{238}Pu	2
	Ages	1 year	2
		10 years	1
		40 yefars	1
Uranium	Enrichments	0.2% ^{235}U	1
		0.72% ^{235}U	1
		5% ^{235}U	2
		20% ^{235}U	2
		90% ^{235}U	1
		90% ^{235}U + 100 ppt ^{232}U	2
	Ages	30 days	2
		20 years	1
^{233}U+^{232}U	U-232 Content	1 ppm	2
		10 ppm	2
		100 ppm	2
	Ages	20 years	2
^{237}Np	Ages	20 years	2

2.2.2 Metrics

Signal strengths in dataset

The most important metric for any dataset is the signal range. Without the appropriate range, testing high-performing algorithms will yield no useful information, regardless of the source selection.

Any single dataset is unlikely to contain all recommended sources. Thus, combinations of measured data, perhaps even from different detectors, along with substantial augmentation from simulations, must be

evaluated together for source and shielding completeness. Special considerations for augmenting with simulations for ML algorithms is discussed in Section 6.

A simple measure of coverage can be obtained by comparing the data coverage in each signal level category (low, medium, high) defined by SNR with each source uniqueness category (Section 2.2.1 and Table 9 contain definitions on both). The low-signal regime should comprise at least 70% of the dataset because this information is the most valuable. The high-signal regime should comprise less than 10% of the dataset. Each dataset should report this coverage as a metric for each source uniqueness category. A simple means to compare a dataset is to compute the ratio of the actual coverage to the ideal coverage. An example is shown in Table 17. Dataset users can quickly see where certain sources are underrepresented or sampled.

Table 17. Comparison of ideal signal coverage with actual

		Ideal			Example actual (dataset)			Score (ratio)		
		Source uniqueness								
		Low	Medium	High	Low	Medium	High	Low	Medium	High
Signal level	Low	23%	7%	3%	13%	13%	7%	57%	200%	200%
	Medium	23%	7%	3%	10%	13%	10%	43%	200%	300%
	High	23%	7%	3%	17%	13%	3%	71%	200%	100%

Sources Coverage in Dataset

For each dataset, the percent coverage of the source classes in each category should be reported within each inclusion level, thereby allowing data users to quickly ascertain the applicability. Furthermore, similar metrics should be generated and reported for all combined datasets used in an evaluation. An example coverage chart is given in Figure 10.

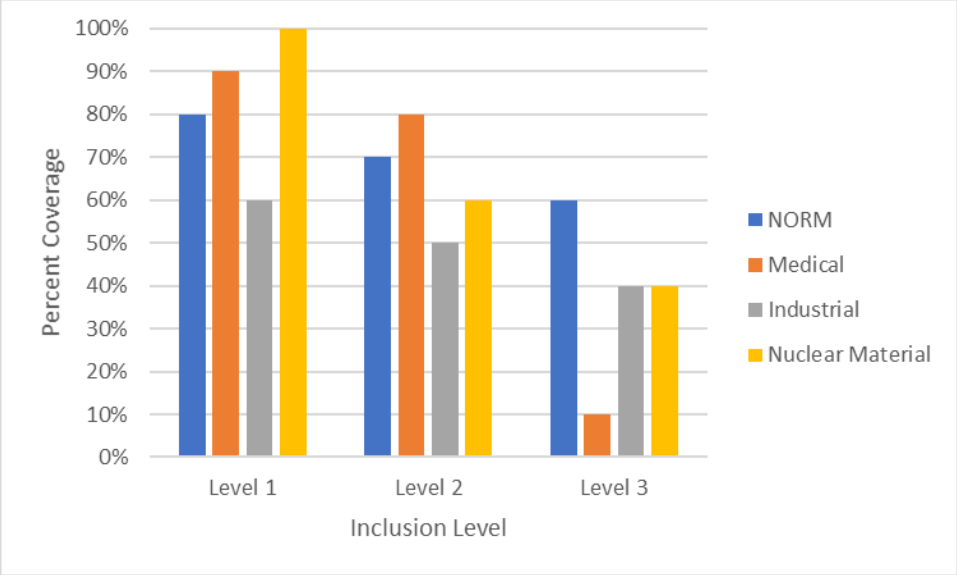


Figure 10. Example visualization of dataset percent source coverage for each inclusion level.

Shielding Coverage in Dataset

A considerable range of shielding atomic numbers and areal densities is recommended. The mean atomic numbers of materials (atomic number weighted by weight fraction) should span the very light elements (e.g., hydrogen and oxygen) to the very heavy (e.g., uranium). Furthermore, for each material or category, the thickness or areal density of the shielding should span from very lightly shielded to the extremely heavily shielded.

The attenuation and scatter from a shield are smoothly varying as a function of atomic number and areal density. Therefore, deviations in an actual dataset from an ideal distribution of shields are acceptable. Generally, recommended datasets comprise data points that are uniformly distributed across atomic numbers and gradually decreasing frequency as a function of areal density. This recommendation is not a requirement; however, if the coverage of the dataset has gaps in the atomic number or areal density map compared with Table 18, then this fact should be made clear in any use of the data, and the consequences on the algorithms should be carefully considered. The shielding recommendations in Table 9 through

Table 14 provide guidelines on the ranges of areal density needed for each nuclide of interest.

Table 18. Recommended shielding coverage

Shielding	Atomic numbers	Areal density (g/cm^2)			
		Light	Moderate	Heavy	Extreme
		0–20	20–50	50–100	100–200
Light	**0–20**	40%	30%	20%	10%
Moderate	**20–60**	40%	30%	20%	10%
Heavy	**60–100**	40%	30%	20%	10%

2.3 DETECTOR RESPONSE VARIATIONS

Temperature changes, aging, handling, unit-to-unit variation, and other environmental effects can change detector signals relative to a characterized exemplar. Limited-scope measured datasets and ideal simulated datasets both contain insufficient variation in the detector response compared with that experienced by real detectors. Many algorithms are based on an assumed detector response or characteristics. Such algorithms must account for these variations, and datasets should include them for testing and training. If an algorithm claims to be independent of any assumed detector characteristics, then datasets should be constructed to test this claim.

2.3.1 Recommendations

If a dataset is designed for a specific mission, then requirements such as temperature range and deployment time may be known and tested against, either in simulation or in a measurement campaign. However, the quantity of measured data required to capture the full range of these effects, especially when coupled with the recommended source variations, is typically prohibitively costly and time-consuming. As a result, measured data are often augmented with simulated data.

Absent specific requirements, a combination of measured and simulated data should produce variation consistent with the following requirements:

- Distribution of energy resolution and efficiency from unit-to-unit (alternatively, an algorithm can be designed for a specific unit)

- Varying operating temperatures
- Various detector ages
- Hydration from aging in humid environments
- Temperature cycling effects
- Varying radio-frequency interference from sources such as radios, cellular towers
- Expected source-to-detector distances
- Different scattering environments

Some of these variations' effects on a detector signal are difficult to predict based on first principles. Thus, to inform the simulations, measured data from a real detector are critical to understanding how the integrated system will respond to these changes. Three different approaches to use measured data to inform simulated detector variation are provided:

- Baseline
  - A single detector unit is used to measure calibration sources in nominal conditions.
  - A high-fidelity model (detector response function) is created based on calibration source measurements.
  - Simulated datasets use typical variations for detector type used to perturb the model (guidelines are given below)

- Better
  - A single detector unit is used to measure calibration sources in nominal conditions.
  - A high-fidelity model is created based on calibration source measurements.
  - The detector unit is used in a limited measurement campaign with known sources to spot-check the range of conditions.
  - Observed variations are used to create perturbation models to vary the detector response function.
  - Simulated datasets use perturbation models to extrapolate to the full range of conditions.
  - Measured data are used to validate perturbation models.

- Best
  - Multiple (or all) detector units are used to measure calibration sources in the full range of conditions.
  - High-fidelity models of detector response functions and variations are created based on interpolating measured conditions.
  - Simulated datasets use models to interpolate continuously across the full range of guidelines.

Without measurement data across the range of expected operating conditions, the following general guidelines may be used for a typical inorganic scintillator to introduce variation in the detector response for simulated datasets:

- Energy resolution
  - Energy resolution offset and 661 keV full width at half maximum varied from nominal to ±15% range (both by same amount)
- Gain drift (detector has built-in but imperfect calibration)
  - ±5%–10% energy drift at 661 keV
- Scattering environment
  - Very low: 20% room return (example: ^{137}Cs source 2 m from detector elevated 10 m)
  - Low (outdoors): 40% room-return from ^{137}Cs at ~2 m
  - Medium (indoors): 50% room return from ^{137}Cs at ~2 m
  - High (indoors, close quarters): 80% room return from ^{137}Cs at ~2 m

- Source-to-detector distance 1–50 m

The four scattering environments' descriptions by themselves may be too vague to describe the range of variation recommended. To give a more concrete example of the effect of the scattering environment, for each environment the ratio of room-scattered counts to total counts detected from a ^{137}Cs source at 2 m is also provided. Figure 11 illustrates the effect of these nominal environments on a simulated ^{137}Cs source when measured by a 3 × 3 in. NaI detector. The source-to-detector distance will also have a substantial effect on the scatter profile.

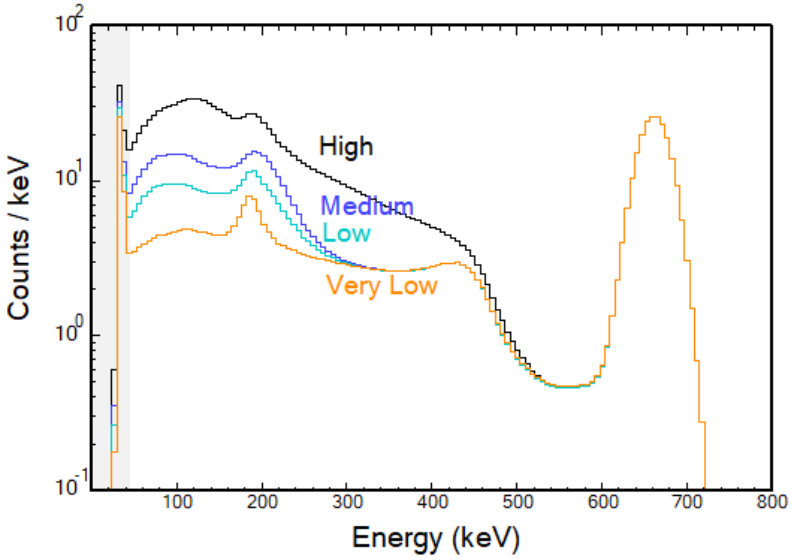


Figure 11. Illustration of effect of scattering environments on ^{137}Cs NaI spectrum (simulated).

3. ALGORITHM METRICS

This section describes quantitative metrics for measuring the performance of four algorithm types: source detection, radioisotope identification/classification, directionality, and localization. These metrics are intended to be the baseline minimum for reporting algorithm performance. In practice, these metrics should be supplemented with additional metrics based on the intended application and operational needs. Along with each metric's description, guidance on its interpretation and best practices are outlined.

3.1 SOURCE DETECTION

Datasets used to evaluate search algorithms typically comprise some number of synthetic, real, or semisynthetic "runs" that each represent a detector moving through some environment, with or without one or more sources present in the run. When quantifying the performance of an algorithm on such a dataset, "success" must be carefully defined, given the underlying statistics of the data. For example, when training spectral anomaly detection algorithms, sometimes a given source encounter is split into multiple spectra for inclusion in the training set. This configuration may be appropriate for some algorithm training, each spectrum taken from a given single source encounter should not be treated as independent samples. During testing and evaluation, this assumption could lead to counting success or failure multiple times for the same source. Analysis algorithms must aggregate data to create the most advantageous source measurement statistics.

3.1.1 Alarm Aggregation and Filtering

Algorithms typically analyze data at some predefined temporal rate to produce an alarm metric whose value is compared with a user-defined threshold to make a binary decision. Values above the threshold result in an alarm condition. A single-source encounter may produce an alarm condition for more than one time step. Consequently, the algorithm should produce the time series of alarm metrics continuously so the evaluation can be conducted at any threshold. Some algorithms will also output a time or time window for the alarm for a given alarm metric threshold. Generally, this capability is not necessary for performance evaluation because the maximum alarm metric during an alarm interval specified by the evaluation (see below) is assumed to be the algorithm's best alarm time. When performance with a given threshold is being evaluated, the algorithm should return the alarm metric and alarm time for each alarm. If an algorithm only returns a list of alarms with a window of time for the alarm without the alarm metric values in that window, then the evaluation must assume the alarm time is the center of the time window specified, and the alarm metric is the maximum k-sigma value (number of standard deviations over the background) found within the time window specified. Detection algorithms must be judged based on their value to the user and are best designed to output the optimal alarm metric and time of encounter (alarm time) for each source. The algorithm may optimize the alarm metric based on different aggregations of measurements and should consider any offsets or delays related to this accumulation time. Thus, the algorithm should tell the user when the closest encounter to the source occurred. As noted below, a generous alarm interval is used in the evaluation to allow for statistical variations that can lead to uncertainty in the exact offset between the sample time and the peak alarm metric.

3.1.2 Determining Detection Success/Failure

Defining what constitutes a successful source detection can play a significant role in the final evaluation. Therefore, every algorithm should be evaluated using the same definition and procedure for performing algorithm comparison. For this same reason, comparisons of algorithm results with performance metrics provided in literature reports should not be performed unless the same datasets are evaluated and said reports adequately documented the success/failure definition criteria and procedures such that they can be reproduced.

For new evaluations, the following simple criterion can be used to determine alarm success/failure. The algorithm must output the time the alarm occurred (alarm time); this time is assumed to yield the largest decision metric for a given event. The alarm time should correspond to the time of closest approach between the source and detector. To count as a detection, the alarm time must be within the alarm interval defined as follows:

- **Dynamic encounter** (source or detector moves continuously): The alarm interval is the time window during which the detector and source distance was within twice the distance of closest approach (DOCA).

- **Static encounter** (source and detector dwell at a fixed separation): The alarm interval is the duration of the dwell time plus any dynamic alarm interval time.

Algorithms sometimes calculate an estimate of the alarm interval, but this estimate is irrelevant to the detection scoring. If the source and detector motions are not well known for the test dataset, then the evaluator can define the alarm interval based on SNR, as defined in Section 2.2.1. This value is defined for each source encounter by finding the integration time window that achieves the maximum SNR. The start and end of this window will be close to the optimal integration time that corresponds to the time the detector was within $\sqrt{2}$ times DOCA from the source. The alarm interval is then $\sqrt{3}$ times the optimal integration time, as found to achieve maximum SNR (corresponding to the time the detector/source distance is within 2 times DOCA, as specified above). The exact specification of the alarm interval is not important, provided it is modestly longer than the integration time needed for maximum SNR and consistent for all sources and algorithms being evaluated.

An algorithm producing an alarm with an Alarm Time within the alarm interval is counted as a detection. Multiple detections within the alarm interval are counted as the same detection alarm. Alarms produced outside the alarm interval are false alarms.

3.1.3 Metrics

When evaluating source detection efficiency, two important metrics of algorithm performance are the true positive rate (TPR; fraction of sources detected) and the false positive rate (FPR; number false alarms per unit time). Two types of alarming definitions are used, and which is appropriate depends on the application. For small search teams run by experts in nuclear detection and evaluation, algorithms are often required to detect any nonbackground source. In this case, false alarms are the same as false detections (an alarm when only background is present). For larger or longer duration search and monitoring operations where the operators have many duties in addition to the nuclear search, and are not necessarily experts in radiation analysis, algorithms are required that alarm only on sources of concern, typically those from fissile materials or very high activity radiological sources. The algorithm must therefore discriminate sources from background and also from common, benign, “nuisance” sources such as those from NORM, medical radionuclides, or sources used in industrial applications.

The true positive and false positive metrics correlate well with the fieldability of the algorithm. Algorithms that fail to detect sources at sufficient source activities are not useful, and algorithms that yield too many false alarms can slow down operations or reduce operator trust [20].

For scenarios in which the source and/or detector are in motion, the TPR (also known colloquially as the probability of detection) is the probability of detecting a source per source encounter:

$$TPR = \frac{TP}{N}, \#(5)$$

where *TP* is the number of true positives (or alarms), and *N* is the number of source encounters. For specific encounter parameters (e.g., source activity, distance, and relative speed) the parameters can be varied to produce probability of detection parameters, such as those shown in Figure 12. Because TPR is binomial, Wilson confidence intervals are frequently used to estimate error [21, 22]. Another approach to error estimation that also accounts for systematic errors is to test many encounters with the same encounter parameters but different background conditions, detector responses, and other characteristics. Before generating the TPR curve, a desired FPR is set usually be adjusting an algorithm threshold.

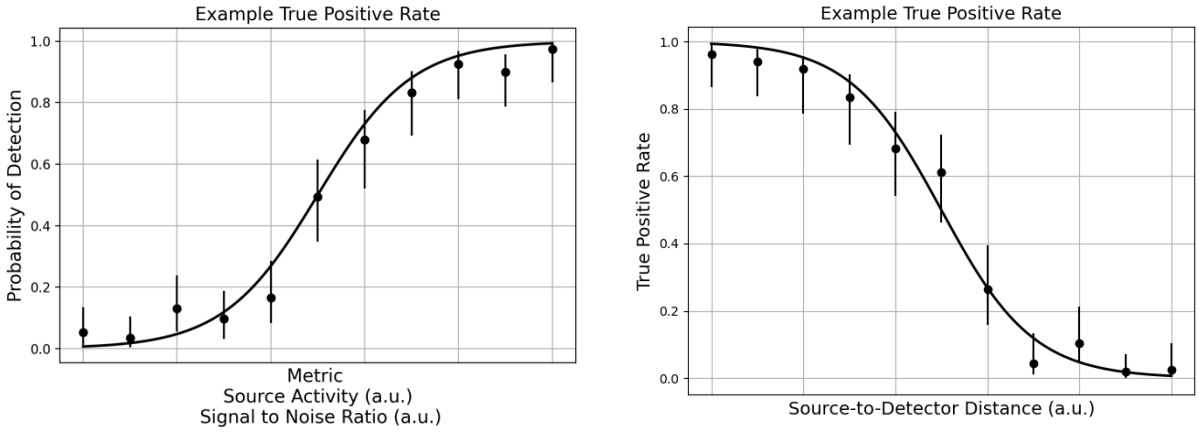


Figure 12. Illustration of probability of detection curves for source activity, SNR and source-to-detector distance. 95% Wilson confidence intervals are used for error estimation.

The FPR, also commonly known as false alarm rate, is defined differently as the number of false alarms per unit time. The FPR is defined as follows:

$$FPR = \frac{FP}{\Delta T}, \#(6)$$

where *FP* is the number of false positives (or background alarms), and ΔT is the length (in units of time) of the measurement that contains only background, or background with nuisance sources depending on the alarm definition in use. A common rule of thumb is for an algorithm to have no more than one false alarm per 8 h of background [23]. However, many standards are less stringent, such as 1–2 false alarms per hour for radiation backpack systems [24]. In some operations, particularly those involving many detectors, alarm rates as low as 1 per week may be required. Because the number of false positives is the number of false alarms in the background time interval, it follows Poisson statistics, and the error estimate, σ_{FPR}, is given by the following:

$$\sigma_{FPR} = \frac{\sqrt{(\mathrm{FP})}}{\Delta T}. \#(7)$$

Again, a Monte-Carlo approach to error estimating can also be used to include systematic errors.

If the algorithm is being tested at a specific detection threshold, then the algorithm is run on the evaluation dataset, and the number of detections and false alarms are counted and compared. A more general comparison can be made using receiver operating characteristic (ROC) curves, which compare both the TPR and FPR simultaneously. These curves show how the TPR and FPR change as an algorithm's threshold is varied. A typical ROC evaluation procedure is to run the algorithm on an evaluation dataset and collect the alarm metric for each data sample. The algorithm may aggregate a

range of samples to create this alarm metric. For the true positive calculation (y-axis of the ROC curve), the maximum alarm metric in each alarm interval is recorded to produce a list of sources with alarm metric scores. All the scores reported outside the alarm intervals are used for the false positive (false alarm) calculation (x-axis of the ROC curve, plotted on a log scale). The two lists are then ordered by detection metric and plotted such that true positive samples move the curve in the y direction, and false positive samples move the cure in the x direction.

An example set of ROC curves for algorithm performance for five sources is shown in Figure 13. The false positive rate is often shown on a log scale to help visualize the low alarm rate region typically required.

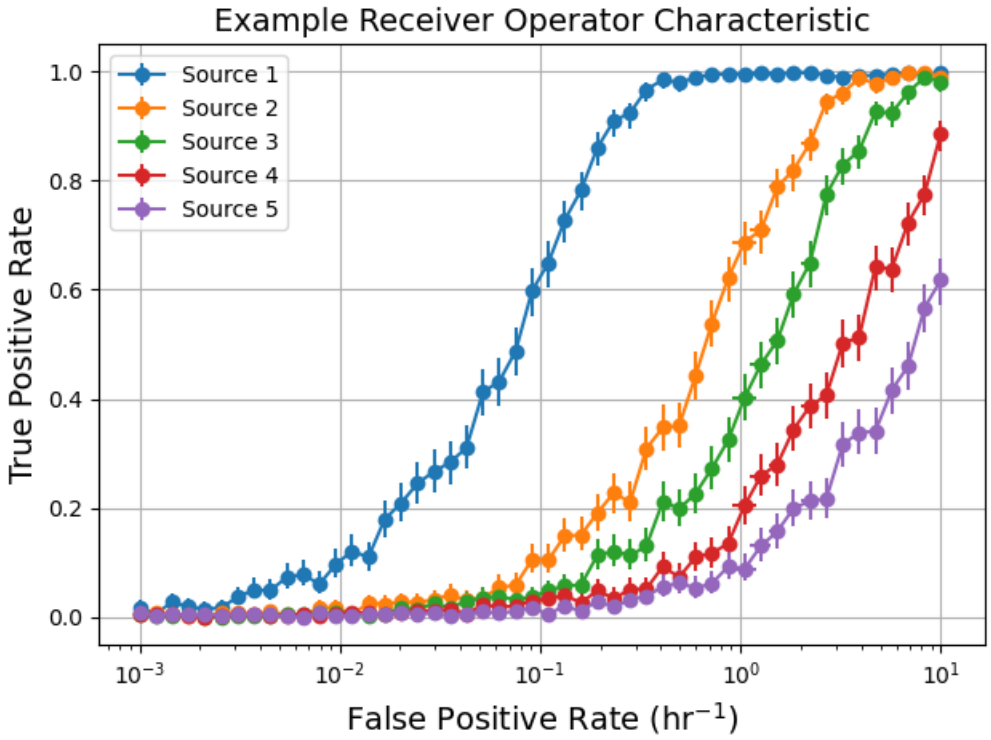


Figure 13. Example ROC curve for five evaluations of an algorithm, in this case five different sources. 95% Wilson confidence intervals are used for TPR error estimates and Poisson error estimates over a 48 h background window are used for FPR.

ROC curve analysis is a particularly powerful way to compare algorithms because it can be applied to comprehensive datasets that include the full range of sources and shielding combinations discussed in Section 2.2 (Inclusion Level 3) with a wide range of background conditions as discussed in Section 2.1. Datasets can also be created with a range of detector/source motions and separations to span the interest of the user typically by collecting or generating realistic background data and applying source injection for the user's chosen encounter geometries and dynamics. Source injection is accomplished by modeling or measuring a source's emissions, scaling them to the source/detector geometry at each time interval, including the impact of any intervening material, calculating the detector response to the source, and then adding counts to the background using Poisson statistics.

To make the comparison useful, sources should only be included that provide a challenging but measurable source signal when measured by the detector in the specified distance and dynamics. Sources too bright or close will always be detected by all algorithms and do not provide a useful comparison, and

sources that do not produce some counts in the detector will not be detected by any algorithm. Testing against a baseline algorithm such as k-sigma (detection metric is the number of standard deviations above background; Section 4.1.1) can be used to tune which source injections are included in the evaluation dataset. The result is a dataset that can test all the source encounters of interest to the user and the false alarm performance at the same time. An example of ROC curves comparing different algorithms using a large population of source types and source encounters is shown in Figure 14.

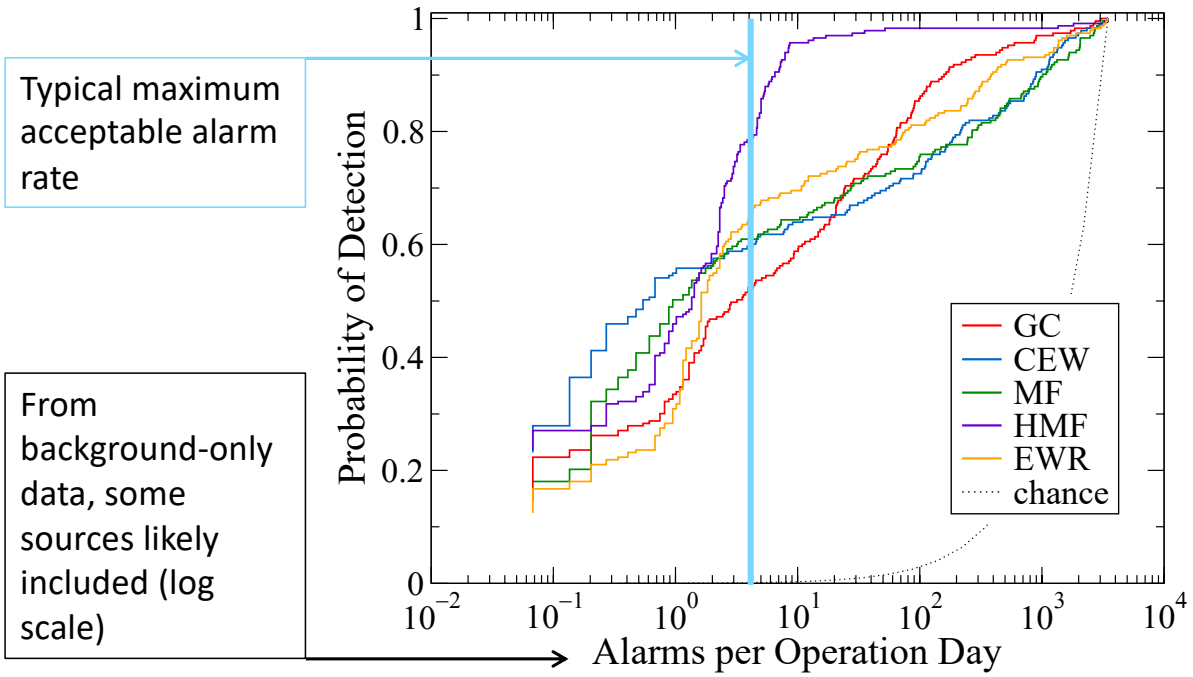


Figure 14. Sample ROC curves comparing different algorithms against a population of sources and a range of mobile detector speeds and source encounter distances.

In general, a ROC curve provides the most comprehensive detection algorithm comparison. To extract a single figure of merit from the ROC curves, the probability of detection (TPR) can be reported at a fixed false alarm rate. For a population of sources, the probability of detection relates to the fraction of test sources of interest that would have been detected. For measured datasets, unknown sources may be hidden in the background measurements, so the false alarm rate may include some real detections. To evaluate algorithm performance, the area under the ROC curve should not be used because it can be misleading: it will be dominated by the performance at very low alarm thresholds at which the false alarm rates are too high to be useful (right side of ROC curves), whereas the performance toward the left side of the ROC curves is most important.

Expert users operating in small groups will often want to include all nonbackground sources as detections. As the use of radiation detectors continues to expand, more operations are requiring nonexpert to rely on the detection algorithms to both detect sources and discriminate sources of interest from benign sources. For such an evaluation, the false positive tests can include measured and/or injected benign sources. ROC curves can then be generated.

3.2 RADIOISOTOPE IDENTIFICATION AND CLASSIFICATION

3.2.1 Metrics

Confusion matrices at both the isotope and class levels provide the best summary of radionuclide and source type classification performance. The confusion matrix shows which radionuclides were selected (usually in the columns) for each tested radionuclide (usually on the rows). The fraction selected correctly is called the recall. The weighted average of the recalls is sometimes used as a measure of total accuracy. With all the selections filled, the number of times each nuclide was selected correctly divided by the number of times it was selected is the precision for that nuclide. Recall describes how often a given nuclide will be identified correctly, and precision describes how often the algorithm is correct when it selects that nuclide.

$$\text{precision} = \frac{TP}{TP + FP}, \#(8)$$

$$\text{recall} = \frac{\text{TP}}{N} = \text{TPR}, \#(9)$$

where *TP* is the true positives, *FP* is the false positives, *N* is the number of trials, and *TPR* is the true positive rate. Good performance is indicated when the diagonal elements of the confusion matrix are high, the off-diagonal values are small, and the precision and recall values are close to 1. An example confusion matrix is shown in Figure 15.

SNR ~ 10	F1	SELECTED																																										Accuracy	0.897
	0.906	Am241	Ba133	Cd109	Co57	Co60	Cr51	Cs137	Eu152	F18	Ga67	Ge68	Ho166m	I123	I131	In111	Ir192	K40	Mn54	Mo99	Na22	Nb95	Np237	Pu239	Ra226	Sb124	Sc46	Se75	Sn113	Sr90	Tc99m	Tl201	U235	U238	UF6	Y88	Yb169	Total	Recall
TRUE	Am241	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	1.00
	Ba133	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	1.00
	Cd109	0	0	1	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0.01
	Co57	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	1.00
	Co60	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	1.00
	Cr51	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	1.00
	Cs137	0	0	0	44	0	0	56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0.56
	Eu152	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	1.00
	F18	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	1.00
	Ga67	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	1.00
	Ge68	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0.00
	Ho166m	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	1.00
	I123	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	1.00
	I131	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	1.00
	In111	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	1.00
	Ir192	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	1.00
	K40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	1.00
	Mn54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	1.00
	Mo99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	98	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	100	0.98
	Na22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	1.00
	Nb95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	1.00
	Np237	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	1.00
	Pu239	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	79	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0.79
	Ra226	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	100	1.00
	Sb124	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	100	1.00
	Sc46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	100	1.00
	Se75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	94	0	0	0	0	6	0	0	0	0	100	0.94
	Sn113	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	98	0	0	0	0	0	0	0	0	100	0.98
	Sr90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	1	0	0	0	87	100	0.12
	Tc99m	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	100	1.00
	Tl201	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	100	1.00
	U235	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	100	1.00
	U238	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	98	0	0	0	100	0.98
	UF6	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	94	0	0	100	0.94
	Y88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	100	1.00
	Yb169	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	100	1.00
	Total	122	100	2	248	100	100	56	100	200	100	0	100	100	100	100	100	100	100	98	100	100	100	79	100	100	100	98	98	13	102	100	107	98	94	100	187		0.897
	Precision	0.820	1.000	0.500	0.403	1.000	1.000	1.000	1.000	0.500	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.959	1.000	0.923	0.980	1.000	0.935	1.000	1.000	1.000	0.535	0.915	

Figure 15. Confusion matrix for a template matching algorithm for SNR of 10. The diagonal elements are shaded in green, and some of the off-diagonal values that are not low are shaded in red.

Many of the off-diagonal responses found in Figure 15 are expected because of degeneracies in the spectra. For example, both ^{19}F and ^{68}Ge are positron sources with a strong emission line at 511 keV. Another metric used to judge the accuracy of algorithms is called the F_1 score. The F_1 score is used in many fields and has become especially popular as ML algorithms become more popular. The F_1 score is defined as the harmonic mean of precision and recall:

$$F_1 = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.\#(10)$$

The F_1 score is best used to summarize the overall diagnostic performance because it reduces the entire confusion matrix to one number. In general, an F_1 score close to 1 means a high TPR and low FPR, and a low F_1 score means a low TPR and high FPR. The F_1 score is not recommended for evaluating detection performance because no direct operational interpretation exists. Furthermore, operational constraints such as an FPR of less than one FP per hour (or 1 per 8 h) are not incorporated in the metric, making it inappropriate to evaluate detection algorithm performance for field applications. Other degeneracies and issues with comparing isotope algorithm performance and producing F_1 scores are discussed elsewhere [25].

3.2.2 Guidance

A confusion matrix and its F_1 score should be computed and compared for different algorithms, including a baseline algorithm, so the performance of the algorithm under test can be understood in context. The same test set of spectra must be used for all algorithms to be compared. The SNR for these tests should be chosen to challenge the algorithms. A good starting point is the SNRs recommended for source detection, although slightly higher SNRs may be required to provide reasonable algorithm performance. An F_1 score less than 0.5 indicates considerable confusion, and an F_1 score more than 0.99 indicates the test set is not providing a challenge. The SNR for the test set should be chosen to provide F_1 scores in between these values. The test set should also include the range of nuclides and shielding materials as indicated for the different development levels in Section 2.2.

4. CASE STUDIES

This section discusses two case studies that investigate the feasibility of determining the SNR metric recommended by DRAG on two publicly available datasets related to a TopCoder challenge [26,27,28] and AIPT. To validate the metrics described herein, a case study was performed by authors not associated with creating these metrics. Each case study discussed in this section describe the two present datasets, provide results related to calculating SNR and JSD, and outline some of the major challenges. A notable difference between the two datasets is that the TopCoder dataset contains entirely synthetic spectra, whereas the AIPT dataset contains entirely measured spectra.

4.1 CASE STUDY 1: "DETECTING RADIOLOGICAL THREATS IN URBAN AREAS" TOPCODER CHALLENGE

4.1.1 Dataset Overview

The TopCoder challenge dataset was used to evaluate new algorithms for radiological search missions. Developed by Oak Ridge National Laboratory, this synthetic dataset is available via download through third-party software called Globus. Recognizing the potential application of ML, the dataset comes partitioned into training and testing files. Ground truth is provided for training only, and algorithm predictions on the testing data are submitted to a remote server for scoring. Also provided are an offline scoring script (Python-based) for use with the training data as well as reference spectral signatures for sources and background. Each data file contains a list of individual counts encountered while moving a detector through a cityscape. The source location, background count rate, and background composition all vary. For the training set, background samples numbered 4,900 and source samples numbered 4,800 spanning 6 classes: highly enriched uranium (HEU), weapons-grade plutonium, ^{131}I, ^{60}Co, ^{99}Tc, and a mixture of ^{99}Tc and HEU. Shielded variants of source samples are present, but specific details were not clear. Categorizing each of the classes produces the distribution in Figure 16. Examples of binned spectra over time are provided in Figure 17, and integrated background and event spectra are provided Figure 18 [27].

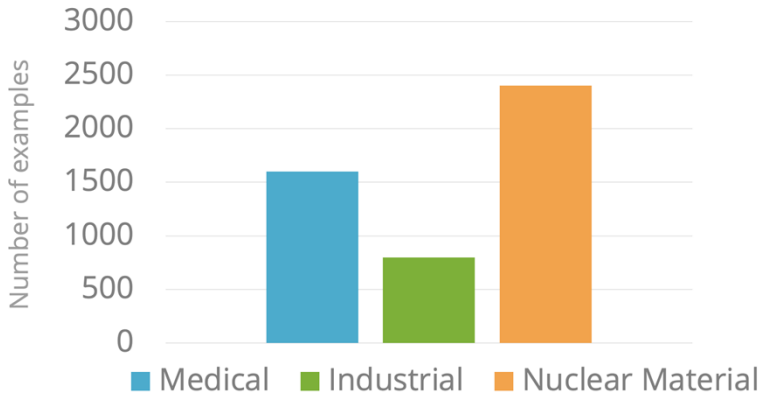


Figure 16. Category distribution of TopCoder runs with sources present.

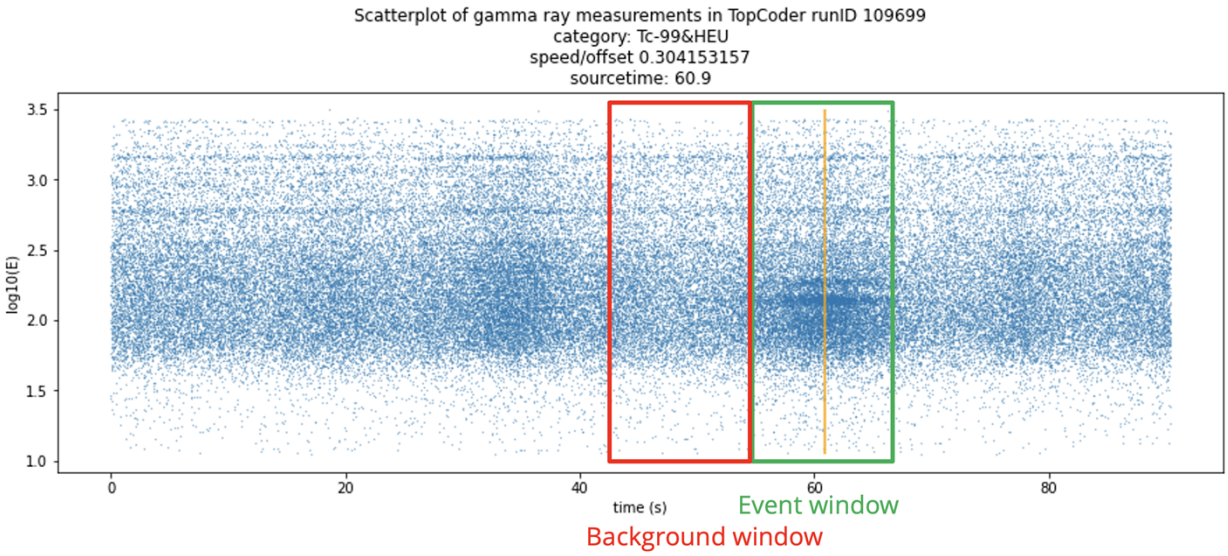


Figure 17. Scatterplot of gamma-ray measurements in TopCoder run #109699 containing ^{99}Tc and HEU mixture.

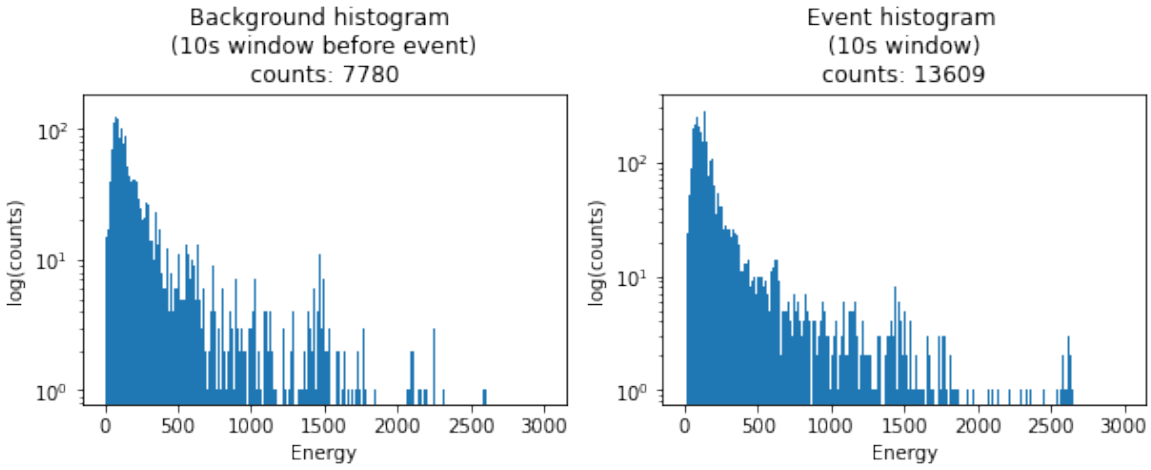


Figure 18. Spectra associated with TopCoder run #109699.

4.1.2 Calculating SNR

Given the variable nature of background during mobile search, which this dataset features, less-than-ideal background measurements were extracted from some amount of time immediately before each event. Such background measurements provide the background count rate estimate used in the denominator of the SNR calculation as well as to determine the source (i.e., net) count rate used in both the numerator and denominator. Thus, the most ideal background is one that accurately represents the background during an event in terms of both composition and count rate. Ensuring such a background was not feasible for either dataset.

To simply obtain as suitable a background as feasible, the SNR from backgrounds taken close in time to each event was maximized by performing an optimization problem that maximized SNR over various background durations from 0.1 to 10 s. Therefore, this per-event background is called the "optimal" SNR. The necessity of this approach was further justified by Figure 19, which reveals that varying time window

length, equally for both background and event, yielded sporadic SNR values across all sources and runs. Figure 19 shows the distributions of SNRs across all runs for extreme values of the integration time range. Lastly, Figure 20 shows the distribution of "optimal" SNRs that yielded 300 negative SNR events.

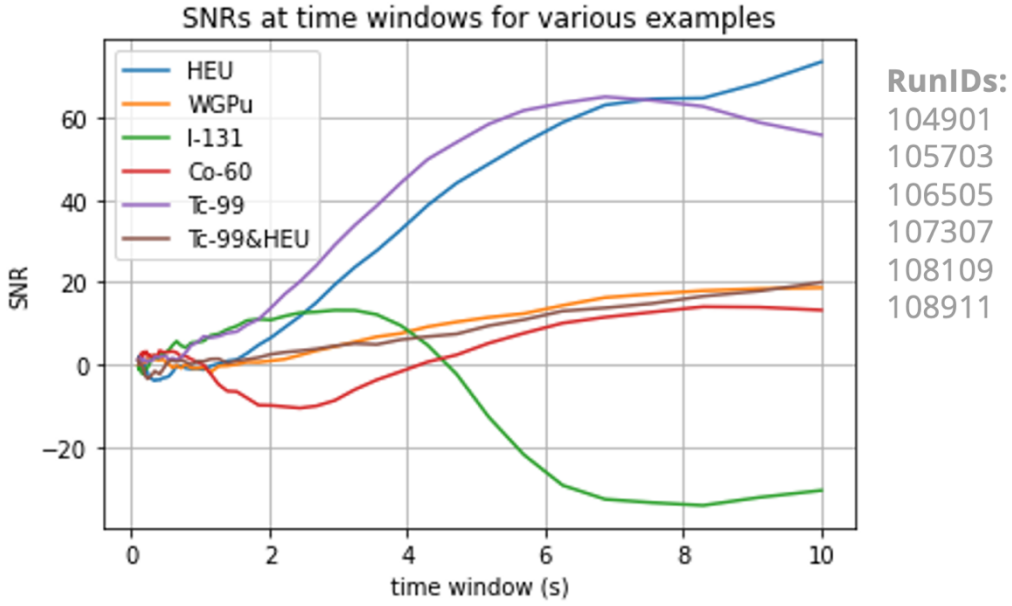


Figure 19. SNR of select events at various integration windows for the TopCoder dataset.

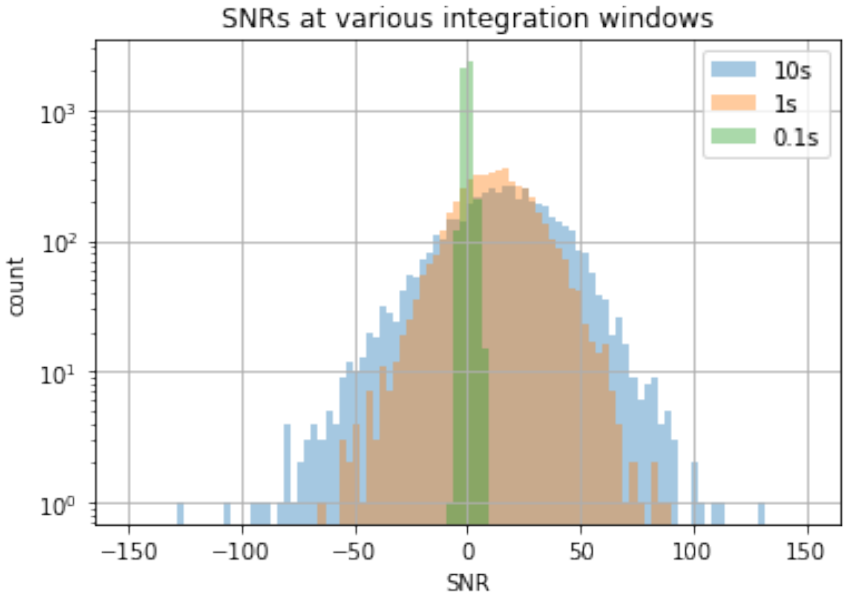


Figure 20. SNRs at various integration windows for the TopCoder dataset.

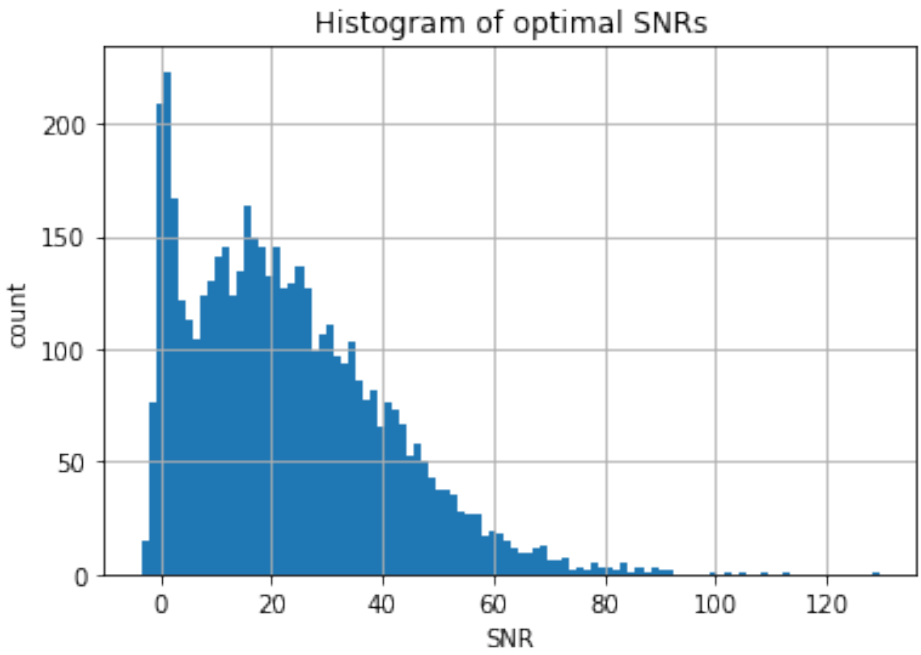


Figure 21. Histogram of "optimal" SNRs for the TopCoder dataset.

The optimal SNR values for the entire dataset were then compared with the ranges of interest specified by the DRAG inclusion levels. This comparison produced Table 19. Numbers do not add to 100% because of the presence of negative SNRs. This points to the difficulty of determining SNR coverage after a dataset has been created, where ground truth may not be available. Instead, SNR coverage should be determined while data is being generated where background and source terms can be measured directly. For this dataset, where backgrounds and source terms are known throughout the model, the SNR coverage could be calculated more accurately.

Table 19. Comparison of ideal signal coverage to actual for TopCoder dataset

		Ideal			TopCoder dataset			Score (ratio)		
		Source uniqueness								
		Low	Medium	High	Low	Medium	High	Low	Medium	High
Signal level	Low	23%	7%	3%	2%	8%	1%	8%	114%	33%
	Medium	23%	7%	3%	2%	8%	0%	8%	114%	∞
	High	23%	7%	3%	11%	46%	14%	46%	657%	466%

4.1.3 Calculating JSD

JSD was then used to obtain a cursory understanding of spectral variability in the dataset by comparing event samples and background samples. Figure 22 shows the JSD per source compared with the mean spectrum of all backgrounds, Figure 23 shows the JSD per source compared with the mean spectrum of all backgrounds as well as events.

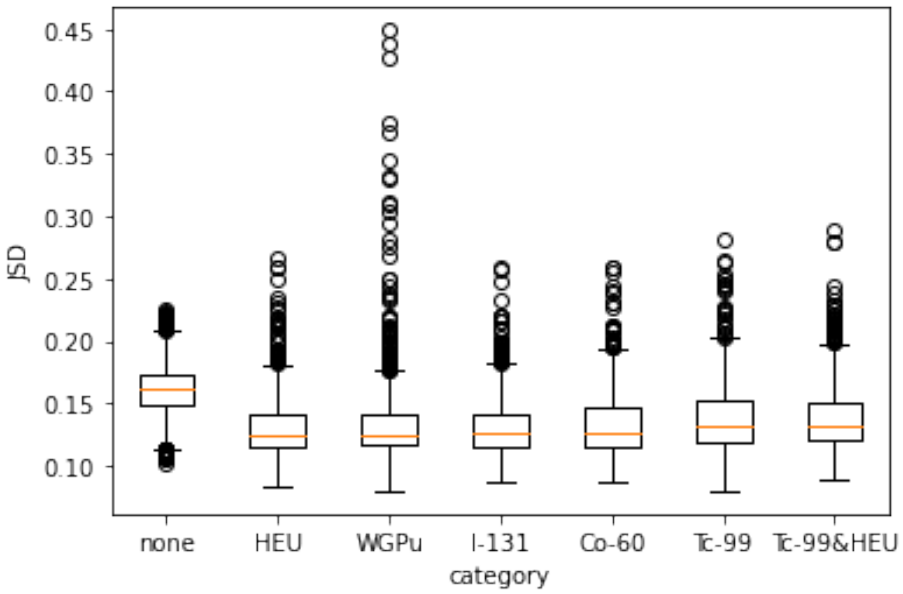


Figure 22. JSD distributions by category comparing each sample with the mean of all backgrounds.

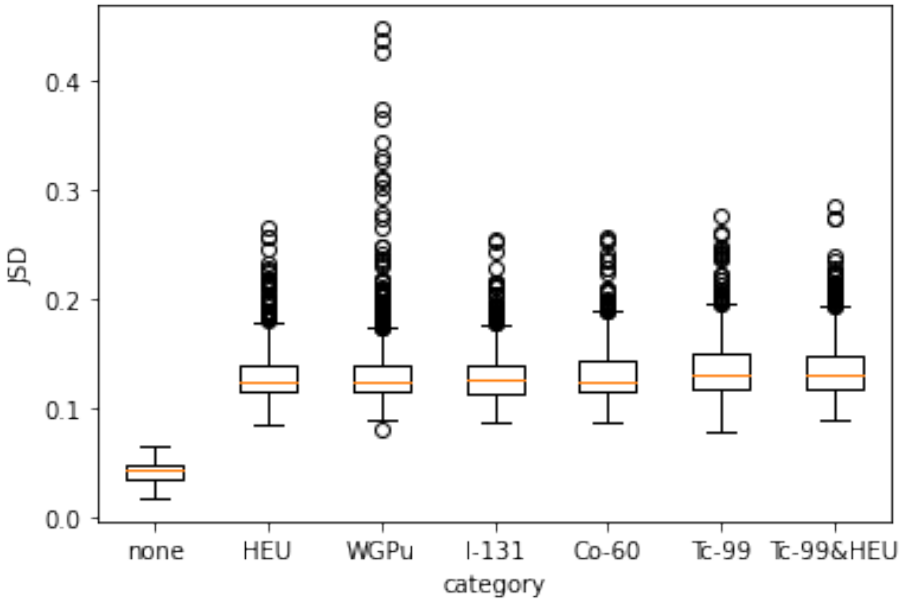


Figure 23. JSD distributions by source comparing each sample with the mean of all backgrounds and events.

4.2 CASE STUDY 2: ALGORITHM DEVELOPMENT RESOURCE STARTER KIT FROM THE ALGORITHM IMPROVEMENT PROGRAM TEAM

4.2.1 Dataset Overview

The Algorithm Development Resource (ADR) Starter Kit facilitates and evaluates new algorithms for search-related missions. The AIPT provides this dataset, or portions of it, upon request, and it comes with documentation and software utilities, notably including a Java-based data-viewing application, scoring application (pre-compiled), and a document explaining the data file formats. The dataset contains measured gamma spectra already binned into histograms and collected using four NaI logs running in

parallel and traveling around a test area where they pass a source at some point. To determine algorithm performance, the dataset uses an offline scorer to prevent bias in results. At the time of writing, the gamma spectra are provided as comma-separated values and are separated into source and background files. Source files represent repeated drive-bys, consisting of a time a series of spectra. Statistics on collected source files are as follows:

- 337,561 spectra in total
- ~64,000 spectra in which a source is present
- ~60,000 spectra "in zone" (refer to dataset documentation on meaning)
- ~52,000 spectra with source present and in zone
- 1,272 closest approaches

Separately, background files are drives on public roads on the East and West coasts that have no specific association with source files. Statistics collected on background files are as follows:

- 131,551 East Coast spectra
- 32,978 West Coast spectra

Categorizing each of the classes produces the distribution in Figure 24. Also, visualizations of the data are given in Figure 25 and Figure 26 showing spectral channels over time. These figures illustrate that the data vary in terms of their overall duration as well as the time step at which the source appears.

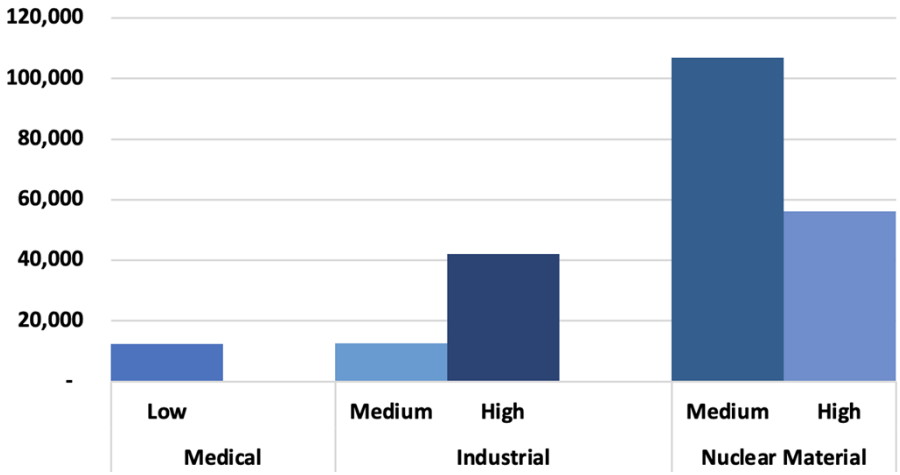


Figure 24. Category distribution of AIPT runs with sources present.

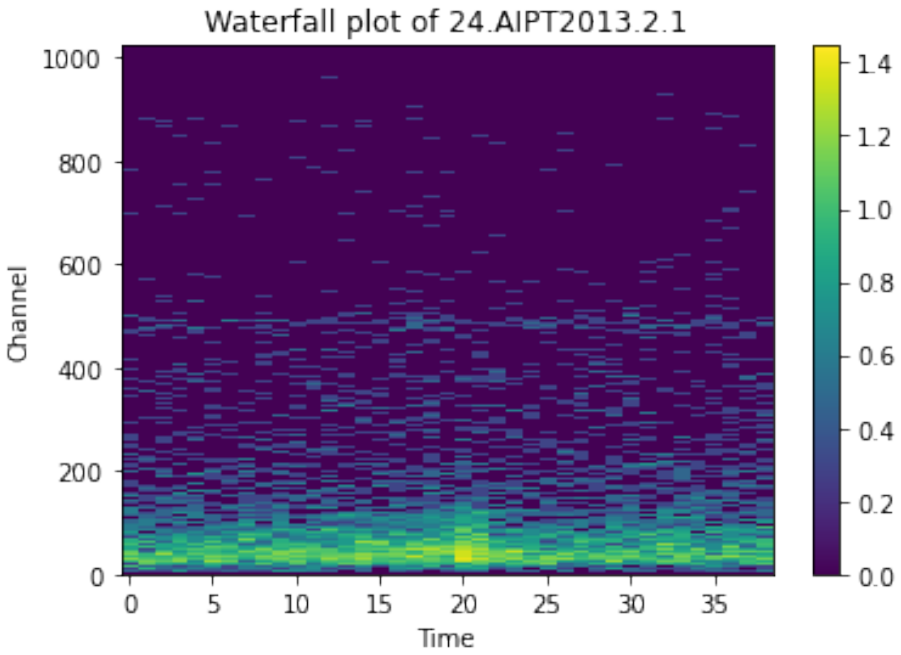


Figure 25. A drive-by with source around time step 20.

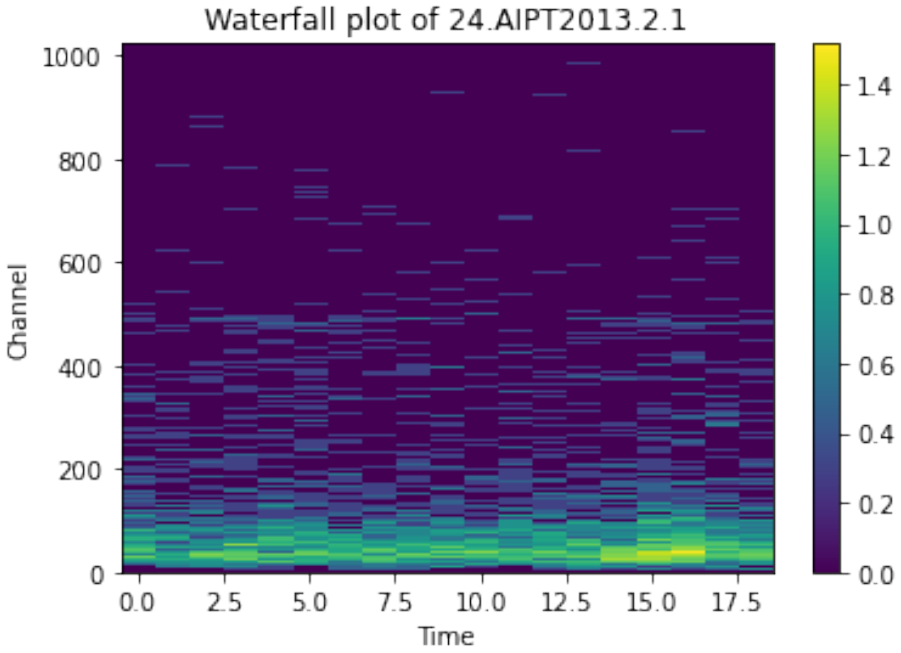


Figure 26. A drive-by with source around time step 15.

4.2.2 Calculating SNR

The AIPT data are like the TopCoder dataset in terms of its mobile search focus. As such, the challenges of selecting a quality background and localizing events in time are present. One difference between the two datasets is that the AIPT data are already binned into spectra (i.e., no initial list mode form), but the measurements from each of the four detectors still needed to be summed together. For selecting a

background, the “optimal SNR” method discussed in the previous section was used to maximize SNRs. The distributions of various integration windows are shown in Figure 27, and the distribution of the optimal SNRs (which still yielded 124 negative SNR events) is shown in Figure 28. Unfortunately, a table comparing ideal signal coverage with actual coverage was not obtained for this dataset.

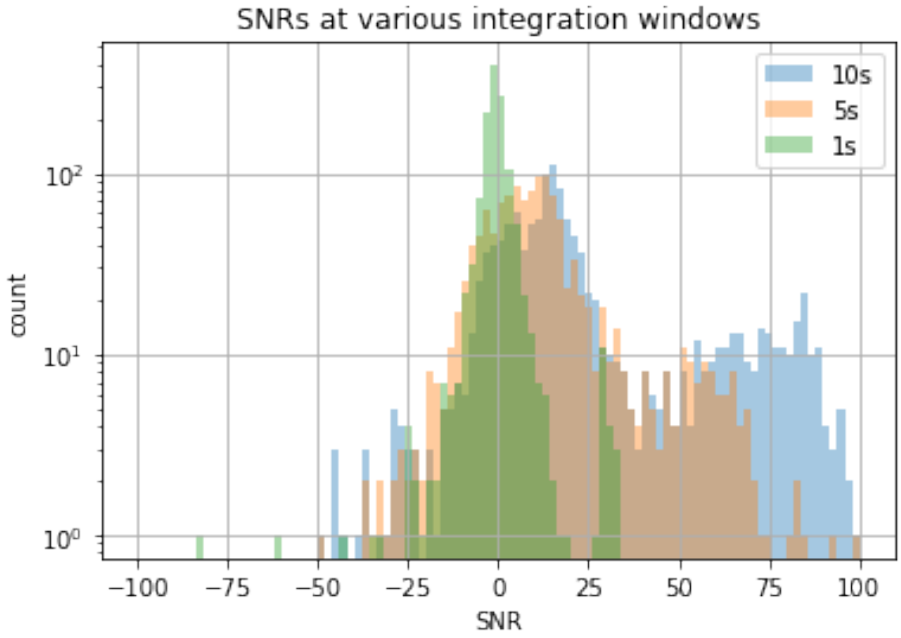


Figure 27. SNRs at various integration windows for the AIPT dataset.

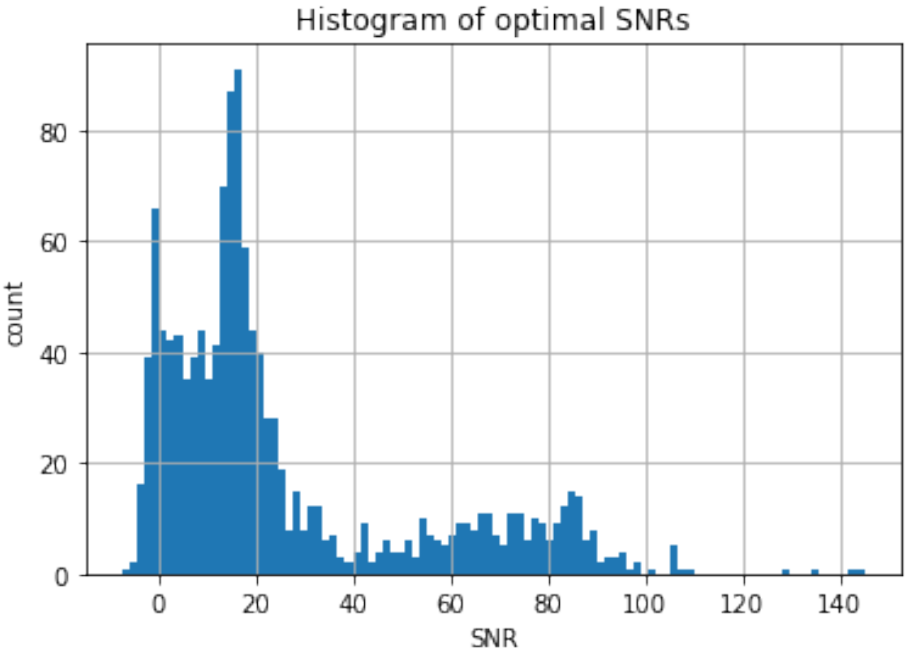


Figure 28. Histogram of “optimal” SNRs for the AIPT dataset.

4.2.3 Calculating JSD

To apply JSD to the AIPT dataset, only backgrounds samples from different coasts were compared with the mean of all background samples. As shown in Figure 29, the distributions of JSDs organized by coast are noticeably distinct. Drawing from details about the data collection, the differences in mean and variance are hypothesized to be related to the composition of building materials in the environment and proximity to the shoreline.

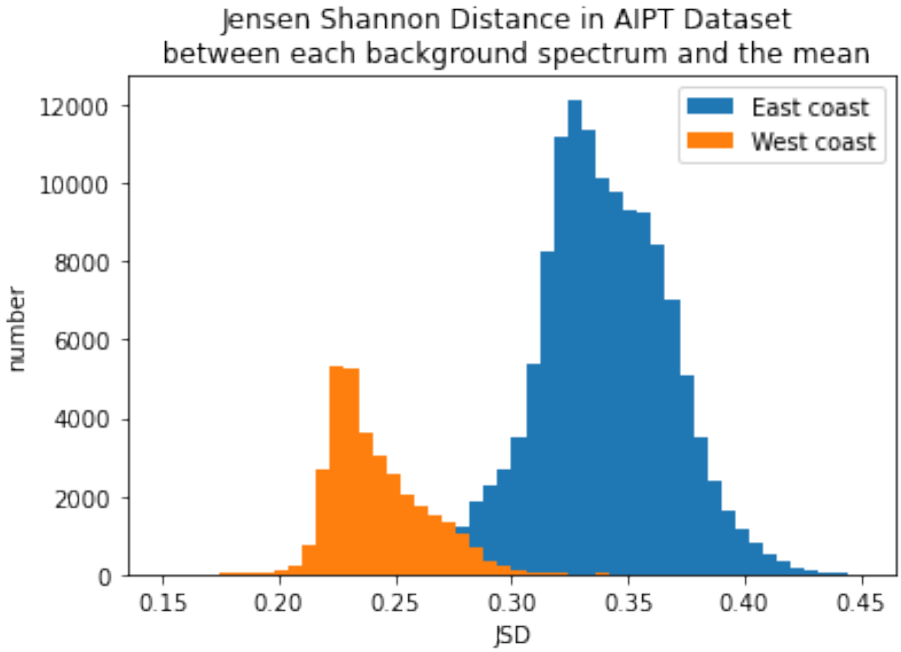


Figure 29. JSD distributions comparing each background sample to the mean of all backgrounds.

4.3 CASE STUDY CONCLUSIONS

4.3.1 Recommendation Coverage

With the resources available, SNR coverage was obtained for the TopCoder dataset only, as shown in Table 19. The results show that the dataset is imperfect with respect to coverage, but this result is expected for two reasons: (1) no recommendations existed to be followed when the dataset was created and (2) application-specific datasets are likely to occupy a subset of the space represented by DRAG recommendations, which are intentionally broad. The same is expected to be true for the AIPT dataset for the same reasons.

4.3.2 Dataset Availability

Although both datasets are considered "public," the AIPT dataset must be requested from the right people at JH-APL or obtained from someone who already has it. Moreover, at the time of writing, neither dataset is searchable directly, only referenced through related publications. This report should be able to offer a reference to a database containing all publicly available gamma spectrum datasets, but no such database exists, nor would such a database be feasible because each academic institution, laboratory, or government agency has their own rules. Instead, radiation detection practitioners are strongly urged to focus less on where data are hosted and more on ensuring information establishing a dataset's existence and acquisition procedure is indexed openly on data search engines. Therefore, even if the data are not

publicly available, they can still be found. This capability alone enables researchers to find and acquire data, regardless of their quality, but implementing community standards for dataset formats and metadata are natural next steps.

4.3.3 Background Selection

The most notable technical challenge with respect to calculating metrics was selecting an optimal background, which is even more challenging in the presence of varying background count rate and composition. Although real-world scenarios also feature these challenges, and datasets that reflect the challenge are rightfully intended to mirror this challenge, that same lack of understanding about background during an event will create inaccuracies in the calculated metrics. If possible, background should be well defined in the dataset if possible. Ground truth measurements should be collected in a measurement campaign, or in the case of virtual testbeds, would be to label background and source interactions individually. Algorithmic best-practices will always need to be employed when selecting a background for measured data, and synthetic datasets could easily separate the background and source-only data for events, if only for the purposes of checking DRAG metric coverage.

4.3.4 Software Utilities

Software tools are inevitably developed to facilitate working iteratively with datasets, but these tools (not unlike datasets themselves) can be kept in-house, creating a barrier to progress. As such, and when possible, supplemental software should be open source and accessible via standard package managers to avoid lengthy delays. Acquisition of utility software via industry-standard package managers, accompanied by permissive licensing, is especially important in the case of libraries that become dependencies of larger software applications. An exception to this recommendation would be scoring-related applications in which answers are compiled into the executable to prevent any cheating on evaluations. The upside of these scorers is that they can be run on classified networks without access to the internet. However, with respect to public scorers, a representational state transfer application programming interface (RESTful API) would be preferable. The TopCoder dataset likely used a RESTful API while the challenge was active because it avoids software approval by not requiring in-network execution of third-party software.

In the spirit of supporting these recommendations and assisting practitioners in the future with related work, converters for both datasets are available as part of the open-source Python package PyRIID [29]. The converters parse the proprietary formats of each dataset and enable export to hierarchical data format (HDF), JavaScript object notation (JSON), or profile configuration file (PCF). Even with converters, automated generation of coverage with respect to DRAG recommendations will be helpful once recommendations are finalized.

4.3.5 Metadata

The metadata provided by both datasets was sufficient for the task at hand, but integrating ground truth labels (such as isotope and/or configuration) directly into the metadata for at least a subset of the AIPT data would have enabled a more detailed understanding of coverage. However, given its use in algorithm evaluations, the obfuscation is unsurprising.

5. BASELINE ALGORITHMS

5.1 SOURCE DETECTION

5.1.1 K-Sigma

Perhaps the simplest statistical algorithm for radiation detection is the k-sigma algorithm, which has been used in a variety of applications such as portal monitors, mobile source search, and long-term radiation monitoring. The method was first described by L. Currie [30], and an implementation for this algorithm can be found in the literature [31]. For a background with B counts, estimated background variance (Poisson or normal distributions are commonly used) σ_B^2, and a foreground with F counts, the k-sigma metric, k, is calculated as follows:

$$k = \frac{(F - B)}{\sigma_B}. \#(11)$$

If k crosses a predefined threshold, then a detection is triggered. The values of B, σ_B^2, and F are usually estimated by rolling time windows, which can either be gross count windows over the entire gamma-ray energy spectrum or regions of interest centered around potential photopeak energies of interest. This algorithm works best when the background is well known, as in the case for measurements with backgrounds which vary much more slowly than the lengths of the background and foreground time windows. Consequently, this algorithm usually works best for static detector systems with near constant backgrounds and has lower performance in mobile systems with strongly varying backgrounds.

In general, the durations of the background and foreground time windows are not the same. Whereas the foreground time window should be chosen to maximize the SNR (and k), a sufficiently large background window can be advantageous to collect enough counts so that the relative uncertainty in B is small. When the duration of the foreground and background window are different, a more general and complete formula for k-sigma is required:

$$k = \frac{N_S - N_B \frac{t_S}{t_B}}{\sqrt{N_B \frac{t_S}{t_B} + (1 + N_B)\left(\frac{t_S}{t_B}\right)^2}}, \#(12)$$

where N_S is the number of counts in the source (foreground) window, N_B is the number of counts in the background window, t_S is the integration time in the source window, and t_B is the integration time in the background widow. This version of k-sigma also includes the noise term from the source (derived using the null-hypothesis test) and assumes the noise is only from the statistical Poisson distribution of the total number of counts measured. An example of the application of this approach is shown in Figure 30, which also shows some of the limitations of using k-sigma in a dynamic measurement such as is encountered with moving detectors. When the algorithm is run in postprocessing, it is sometimes useful to run the algorithm forward and backward, or with background integrated on either side of the source (foreground) window, to minimize the influence of dynamic background for mobile detectors. Because of these types of background variations, the k-sigma algorithm is not recommended for mobile detector systems.

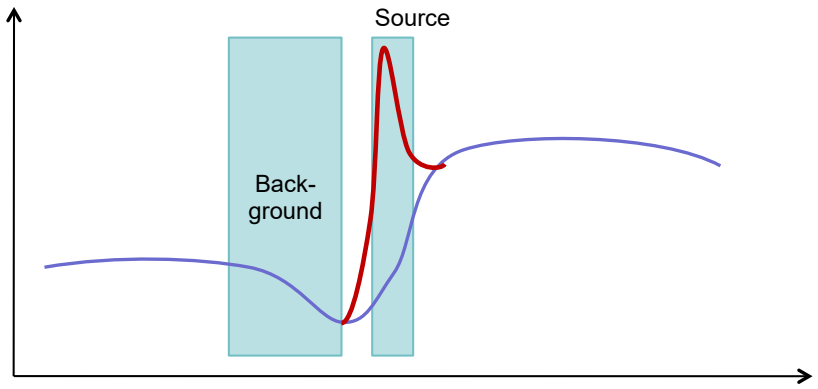


Figure 30. Source (foreground) and background windows of different duration with a dynamic background demonstrates the difficulty of using k-sigma in a dynamic environment.

Algorithm Availability:

- A Python 3 implementation of k-sigma for mobile search applications is available in the open-source RADAI Gitlab repository: https://gitlab.com/lbl-anp/radai/radai/-/tree/main/.

5.1.2 Sequential Probability Ratio Test

The sequential probability ratio test (SPRT) is a simple statistical algorithm that has been adapted for source detection in a variety of applications such as for portal monitoring [32, 33], safeguards [34], mobile source search [35], and long-term radiation monitoring. It is a statistical test with the following hypotheses:

- The null hypothesis H_0: The foreground count rate F is equal to the background count rate B.
- The alternative hypothesis H_1: F is composed of both B and source(s) count rate S.

For a dynamic background environment, F and B are calculated over rolling temporal windows of length T_F and T_B, and F_t and B_t are the foreground and background count rates at time step t. At each time step, the likelihoods of the null and alternative hypothesis given F_t are then defined as $\mathcal{L}(H_0|F_t)$ and $\mathcal{L}(H_1|F_t)$, respectively. Thus, the log likelihood ratio λ_t can be calculated using Eq. (2). Then λ_t is sequentially evaluated with a rolling sum to output a scoring metric: $\Upsilon_t = \Upsilon_{t-1} + \lambda_t$. The scoring metric can then be tested against a threshold τ_u such that a detection is triggered when $\Upsilon_t > \tau_u$.

$$\lambda_t = \log\left(\frac{\mathcal{L}(H_1 \mid F_t)}{\mathcal{L}(H_0 \mid F_t)}\right). \#(13)$$

Note that the integration times T_F and T_B should be optimized for each application. However, a general rule of thumb is that shorter integration times (on the order of 2 to 10 s) are optimal for fast-changing backgrounds such as in mobile search, and longer integration times (>30 s) are optimal for portal monitors or other applications for which the background tends to change rather slowly. As in the k-sigma example, this algorithm should not be used for strongly varying backgrounds.

Algorithm Availability:

A Python 3 implementation of SPRT for mobile search applications is available in the open-source RADAI Gitlab repository: https://gitlab.com/lbl-anp/radai/radai/-/tree/main/.

5.2 RADIOISOTOPE IDENTIFICATION AND CLASSIFICATION

5.2.1 Benchmark Algorithm for RadioNuclide Identification

The Benchmark Algorithm for RadioNuclide Identification (BARNI) is an open-source radionuclide identification software package that has been developed for use with a wide range of gamma-ray detectors ([36]; https://github.com/LLNL/barni). BARNI is designed to serve as a comparative tool against which all other vendor algorithms, such as those found in RIIDs, can be compared. The core identification algorithms in BARNI follow a typical peak search and match strategy, leveraging ML to generate an expert system customized for a specific detector.

The BARNI algorithm follows three steps from the input of a sample spectrum to the resulting radionuclide identification: (1) peak finding, (2) feature extraction, and (3) classification. Each piece of the algorithm is agnostic to the type of detector; instead, that information is carried in the algorithm configurations. This modular design allows the BARNI algorithm to function across a wide range of detectors and allows the algorithm itself to be open-source without revealing any proprietary or export-controlled information about specific commercial detectors. Several training tools are published with the BARNI software package. These tools enable users to create new algorithm configuration files. Examples of algorithm configurations for generic detectors are also published. A diagram that illustrates how these training tools, configuration files, and the identification algorithm fit together is shown in Figure 31.

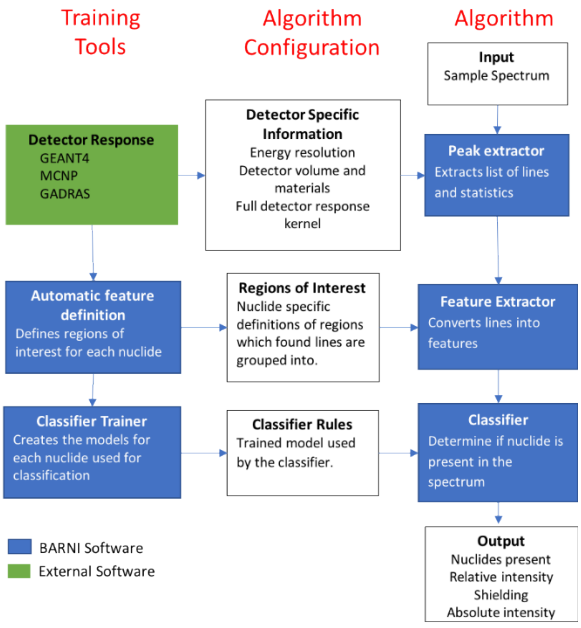


Figure 31. Overview of the BARNI algorithm, configuration, and training tools. The colored boxes denote software, and the white boxes indicate files

5.2.2 Non-Negative Matrix Factorization Template

Non-negative matrix factorization (NMF) with Poisson loss function is an additive, linear model that provides a very good approximation of the underlying physics and statistics of gamma-ray detection [37]. In this method, Kullback–Leibler NMF is used to learn a data-driven background model (usually 1–3 components determined by the Akaike information criterion) for a detector system or class. This model may be refined on a system-by-system basis if sufficient data are available. Templates for target sources—from simulation or learned using NMF—are used by appending each source individually to the

background model and fitting the target spectrum. A likelihood ratio test, comparing a background-only fit with a background plus source fit is computed as the alarm metric. Thresholds may be set analytically or empirically. Finally, multisource detection can be performed via recursion.

5.2.3 Gamma Detector Response and Analysis Software–Detector Response Function

GADRAS-DRF is a software application developed by Sandia National Laboratories for the simulation of gamma and neutron detectors [38]. It is the publicly available version of the limited-distribution GADRAS application, with export-controlled features removed. The distinguishing capabilities of GADRAS-DRF are the rapid high-fidelity gamma detector response function (DRF) and full-spectrum analysis (FSA). The DRF can simulate gamma spectra from almost any source in less than a second, allowing FSA to fit the entire spectrum for a variety of algorithms including quantification, shielding estimation, and identification.

The primary identification algorithm in GADRAS-DRF is called Isotope ID. It has gone by other names in the past, including DHS Isotope ID and HPGe FSA. It leverages the high-fidelity DRF to generate a database of spectral templates before the analysis. It also uses the DRF during the analysis to fit scatter in the measured spectrum not explained by mixing the precomputed templates. Approximately 70 radionuclides are included by default in the database along with varying shielding types to create just over 300 templates. It can be used with high-resolution detectors such as high-purity germanium (HPGe), low-resolution detectors such as NaI, and even plastic scintillators such as polyvinyl toluene (PVT).

The primary requirement for Isotope ID is a detector characterization in GADRAS-DRF. The characterization procedure involves collecting measured data with a single detector. Common calibration sources are measured, and GADRAS-DRF is used to fit a response function by comparing the simulated calibration source response with the measured response.

Isotope ID leverages several GADRAS-DRF tools to preprocess spectra, including energy calibration and an inverse pile-up model to linearize the spectrum with respect to radionuclide activities. Within the Isotope ID solver, subsets (or all) of the radionuclides are included one at a time. The subsets are called solution sets, and they are included based on the observation that multiple radionuclide events commonly share the same classification (e.g., all are medicals or industrials). Considering these solution sets independently reduces the probability that unusual mixtures (e.g., medical with nuclear material) are identified, although they are still possible in the “all” solution set. The solution sets are (1) background, (2) NORM, (3) medical, (4) industrial, (5) beta emitters, (6) SNM, and (7) all.

For each solution set, the correlation (activity divided by uncertainty) of each radionuclide with the measured spectrum is determined using weighted least squares regression. The different shielding configurations for each template are linearly mixed to determine the optimal (highest correlation) shielding fit. After iterating through the radionuclides, the one with the highest correlation is stripped from the spectrum, leaving a residual spectrum to fit. The process is then repeated and iterated upon until the residuals are negligible.

Several heuristics and filtering steps are then employed, including the following:

- Remove insignificant terms by comparing the goodness-of-fit with and without each radionuclide.
- Remove weaker sources that are spectrally similar to stronger sources present.
- Test the addition of a source if it is not present in the solution set but is commonly found in combination with another source present (e.g., test adding ^{235}U if ^{238}U is found) with a reduced threshold for inclusion.

- Reduce the threshold for inclusion of neutron-emitting sources if a neutron detector signal is included and the neutron counts above background are statistically significant.

The correlation value for each radionuclide in the solution set is then converted to a measure of confidence by taking the square root of the correlation.

The solver yields a list of radionuclides with confidences for each solution set. The solution sets are weighted by the inverse of the chi-squared fit, squared, and then normalized to the sum of solution set weights. The confidences for each radionuclide in each set are modified by these weights.

Radionuclides with very low confidence are removed from the fit. The solution set with the minimum chi-squared value is reported as the best fitting, and those radionuclides are reported as the final result. The full version of GADRAS also contains a search algorithm, which combines detection with identification for time-dynamic search data.

Benefits

FSA does not require photopeaks for source identification. Small inflections in the continuum from obscured sources can be readily fit and identified. Several examples are given in the following figures. FSA is especially powerful in low signal spectra.

Because the only requirement is a detector characterization, which is a well-defined process, GADRAS-DRF Isotope ID can work out of the box for almost any radiation sensor. It has been deployed and tested on numerous systems, including mobile search, portal systems, and handheld detectors. This extensive testing provides confidence in its capabilities as well as lessons learned with real measurement data that have been encoded into the algorithm. Moreover, the default radionuclide list is fairly comprehensive and includes all radionuclides specified in this document. The combination of deployment experience and large radionuclide list that can be made into arbitrary combinations makes it a robust baseline algorithm.

Disadvantages

The detector response function is based on a set of calibration measurements in an assumed scattering environment. In some scenarios, such as a portal configuration, the source-to-detector geometry is constrained, and the characterized scattering environment is representative of the measurement environment. In other mission spaces, such as mobile search, the geometry is unconstrained. Mitigations exist for scatter that deviates from the characterization, but the performance of the algorithm degrades as the measurement geometry deviates from the characterization geometry.

Examples

Figure 32 shows the Isotope ID analysis of a CsI spectrum of depleted uranium with roughly 250 net counts. Similarly, Figure 33 shows a spectrum of HEU with 200 net counts. Both are near the limit of detection. Even in the signal-starved regime, Isotope ID can leverage FSA to identify sources.

As another example, very good counting statistics enable this methodology to be used with PVT spectra, even with no visible photopeaks. Figure 34 shows an Isotope ID analysis of a shielded HEU spectrum. Small inflections from the Compton edges in the spectrum are exploited with FSA.

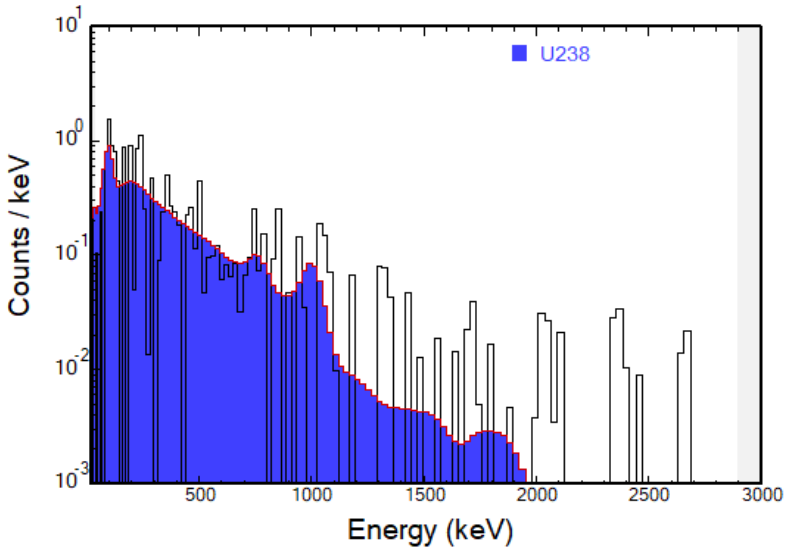


Figure 32. CsI Spectrum of depleted uranium and GADRAS-DRF Isotope ID Analysis.

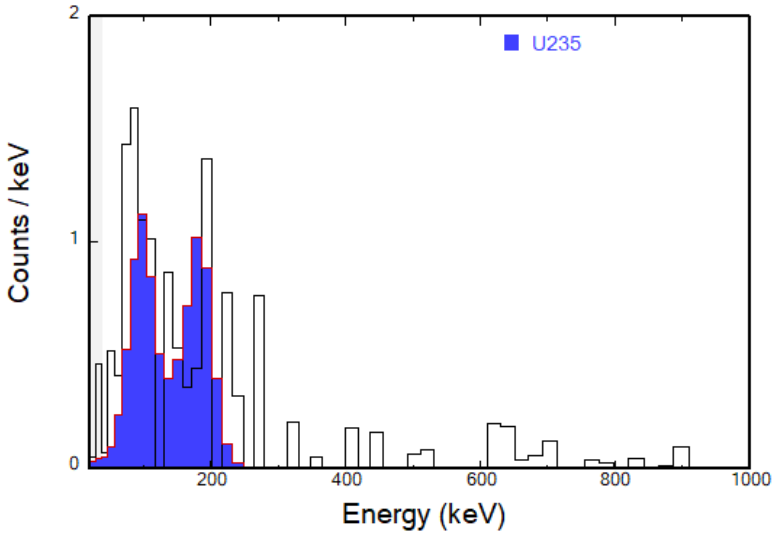


Figure 33. CsI Spectrum of HEU and GADRAS-DRF Isotope ID Analysis.

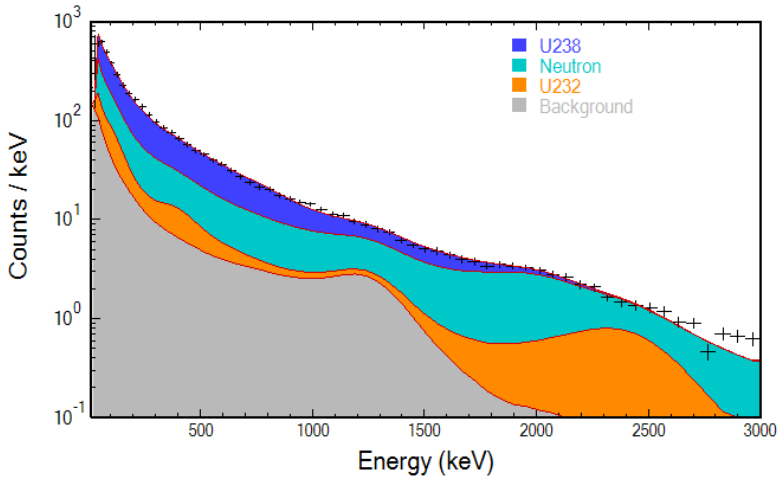


Figure 34. PVT Spectrum of Shielded HEU source with neutrons and GADRAS-DRF Isotope ID analysis.

6. MACHINE LEARNING CONSIDERATIONS

ML techniques are rapidly becoming highly relevant to algorithms for nuclear/radiological source search. Although none of the baseline algorithms described in this document are ML techniques, a survey of current state-of-the-art algorithms would likely include many ML techniques, and a survey of upcoming techniques would likely be dominated by a discussion of ML. ML techniques have many new considerations that are either not relevant to conventional techniques or become much more salient. This section briefly introduces a few of the most relevant concepts.

6.1 DATA STEWARDSHIP

Because ML models require training with data to become useful, proper stewardship of the data that the model is trained with is as important or more important than it is with comparable non-ML models. The findable, accessible, interoperable, and reusable (FAIR) guiding principles for scientific data management and stewardship [39] describe the important considerations:

- *Findable*: the data are uniquely identifiable, described by metadata, and indexed in a searchable manner.
- *Accessible*: A standard protocol exists for retrieving data and metadata. Even if the data are made inaccessible, the metadata will be retained and searchable.
- *Interoperable*: A standard form exists for analysis, storage, and processing of data and metadata.
- *Reusable*: Data and metadata have clear licensing and provenance. Data and metadata meet domain-relevant community standards.

6.2 REPRODUCIBILITY

Like many other fields, radiation detection is in the beginning stages of employing ML. Other fields have been met with many wild successes and advancements in the state of the art but have also encountered many reproducibility failures. Analysis from a Princeton study has produced hundreds of scientific studies that yielded misleading and wildly optimistic results [40].

The following examples of common errors can be prevented by adhering to ML best practices [40]:

- No train-test split
- Duplicates across train-test split
- Feature selection on test data
- Non-independence between train and test split
- Temporal leakage
- Preprocessing on train and test sets together
- Illegitimate features
- Sampling bias
- Test set not drawn from the distribution of interest

Examples of the same flaws found in the Princeton study are not absent from the radiation-detection literature. In fact, for reasons discussed in later sections, these errors are likely more common relative to other fields adopting ML rather than less common. With the current state of ML tools and resources, no comprehensive solution exists for ensuring that a model is free from all the issues identified (leave alone similar issues that were not identified). Adhering to standard ML practices is helpful, but these practices are not widely agreed upon and are often poorly understood by new practitioners. One suggestion from the study authors was to include a model data sheet that asks the authors questions about the conditions

for which they expect model results to continue to be valid. This step is likely helpful but not sufficient. Many additional caveats must be considered as part of radiation modeling, including caveats not yet identified in the literature.

6.3 TRANSFERABILITY

The goal of transferable algorithms is that they can be trained on one dataset and applied to another dataset without a significant degradation in the usefulness of the algorithmic predictions. Transferability can be attempted with or without fine-tuning, where a small amount of data is used to adjust the parameters learned during pre-training on a larger dataset. Specific considerations for ML algorithms will have significant implications for the ultimate transferability of the algorithm.

6.3.1 Synthetic Data/Simulated Data

There are many reasons to be interested in pursuing simulation for generating data for ML algorithm training and evaluation. ML algorithms that are fully data driven (as opposed to those incorporating varying degrees of assumed underlying structure) require large amounts of data to train owing to the vast number of model parameters that must be learned. Large amounts of radiation data can be experimentally generated rapidly for a single source–shielding–detector geometry. However, when the training dataset must include many different geometries or combinations of source/shielding/detector, the need for human intervention to change the setup makes collections unwieldy or impossible. Simulation is a very tempting solution to this problem because, compared with experimentation, rapid simulation of all these geometries and source/shielding/detector combinations is possible. Unfortunately, significant caveats come with using simulation. Despite the impressive performance of modern Monte Carlo codes, a perfect model of a detection scenario is impossible to achieve. Invariably, the modeled geometry will be different from the actual geometry (because no simulation includes the entire universe), and the modeled materials will be different from the true materials (large variability exists in common and important materials [41]). Consequently, the simulated data will not be drawn from the same distribution as the true data. Because of these differences, ML models trained on simulated data often fail to transfer to true data. A wide variety of methods can be used to help address this issue. A non-comprehensive list follows:

1. Generative adversarial networks [42]: To use a generative adversarial network, two networks are trained. The first network takes in noise and attempts to construct spectra that resemble real data. The second network attempts to discriminate between synthetic and real spectra. Iterative training yields progressively more realistic synthetic spectra.

2. Feature engineering: Training the algorithm on carefully constructed engineered features rather than on raw data may remove the meaningful differences between synthetic and real data from the feature space. Unfortunately, constructing such features or even verifying that they have the desired properties is very difficult.

3. Synthetic signal with real background: Often, the issues in modeling are much more pronounced in background than in signal. With a well-defined source, the major contribution to the error is just the transport of the photons. With background, both the source term and the transport are significant sources of error. By recording backgrounds and injecting simulated source data, the issue of simulating the background can be somewhat ameliorated.

4. Domain transfer algorithms [43]: Given a population of simulated data and a smaller population of real data, domain transfer algorithms attempt to construct features that are valid in both domains, often by trying to minimize the difference between the distributions in latent space of things that should classify similarly.

Common to many of these techniques is the need for large experimental datasets to have high confidence that synthetic data are incorporated correctly. Despite the many techniques that are employed to make synthetic data useful, no technique can make synthetic data sufficiently similar to real data for all applications.

6.3.2 Independent Variable Selection

A dataset is often constructed when a continuous range of an independent variable is being studied. To understand this range, a series of discrete values of the independent variable must be chosen to perform experiments (e.g., using 0, 2, 4, 6 in. of shielding to study the range from 0 to 6 in.). In this case, ML models tend to overfit to the values used in training. Thus, for values very near these test values, the ML model will overperform; for values far from the test values, the ML model will underperform.

If the data are obtained through simulation, then this problem can be resolved by varying the training values continuously rather than using discrete values. If the data are obtained experimentally, then continuous variation is impossible or impractical. In this case, the best practice it to exclude some values of the independent variable from the training set for use in the validation and test sets.

6.3.3 Extrapolation Checks

One way to check that an algorithm is transferring correctly is to take some of the target data and encode them in the model's latent space. Test data that encode similarly to the training data and look like they could be taken from one of the labeled populations of training data are the most likely to have been correctly characterized by the model. Data that are far away in latent space and clustered on an "island" are much more likely to be misclassified by the model. This test can be done via visualization [44] but can also be instantiated in distribution metrics such as the *d*-dimensional Kolmogorov–Smirnov (ddKS) test statistic [45] and the Mahalanobis distance [46].

To the extent that constructing inputs that embed between the island and the training distribution is possible, these inputs are useful in understanding the ability of the model to correctly categorize the data. Training a model on data with these intermediate values filled in will generally lead to better model generalization and better model behavior overall. One common source of islands in data is variation in the source and shielding strength. Desirable models will show monotonicity in accuracy/precision of prediction with changes in the source strength. For example, the model should be more able to identify and more confident in the prediction that it can find more source with less shielding than it can identify less source with more shielding. Testing the monotonicity of the model response with source strength and having intermediate source values in the training set are both potentially very beneficial to building good models.

6.4 EXPLAINABILITY

6.4.1 Nomeclature

Many different terms are used to convey similar concepts. This section discusses explainability, but other similar concepts would have been equally valid to discuss:

- Interpretability: The ability to determine cause and effect from a ML model.
- Explainability: The knowledge of both what a node in the model represents and its importance to the model's performance.
- Uncertainty quantification: The ability to understand the workings of a model under near counterfactual conditions

6.4.2 Why Explainability is Necessary

Many good reasons for explaining the outputs of neural networks include the following [47]:

- Debugging models: Helping ML developers know what is wrong so that they can fix the problem.
- Trust and adoption: Users and policymakers can decide whether the information the model relies on is reasonable enough to trust the model.
- Whether to intervene: Helping users to decide when to overrule the model in a specific case because the basis for the decision appears incorrect.
- Improving intervention assignments: Understanding the reasoning behind the decision to choose one of many possible options when acting based on the prediction.
- Recourse: Helping those affected by the decision to appeal it or to know what inputs can be changed to obtain a different output.

Recent regulation in Europe regarding the Right to an Explanation has put a great deal of attention on explainability as a means to leverage artificial intelligence models while still guaranteeing fairness.

6.4.3 Accuracy–Explainability Trade-Off

The field of ML generally acknowledges that there is often a trade-off (Figure 35) between the interpretability and the accuracy of the model, although this is not always the case [48]. In general, larger and more highly parameterized models can reflect underlying complex functions more accurately. However, the additional complexity of these large, highly parameterized models makes it more difficult to understand the inner workings. Therefore, selecting the correct model is a process of optimizing for the combination of accuracy and explainability that best fits the application.

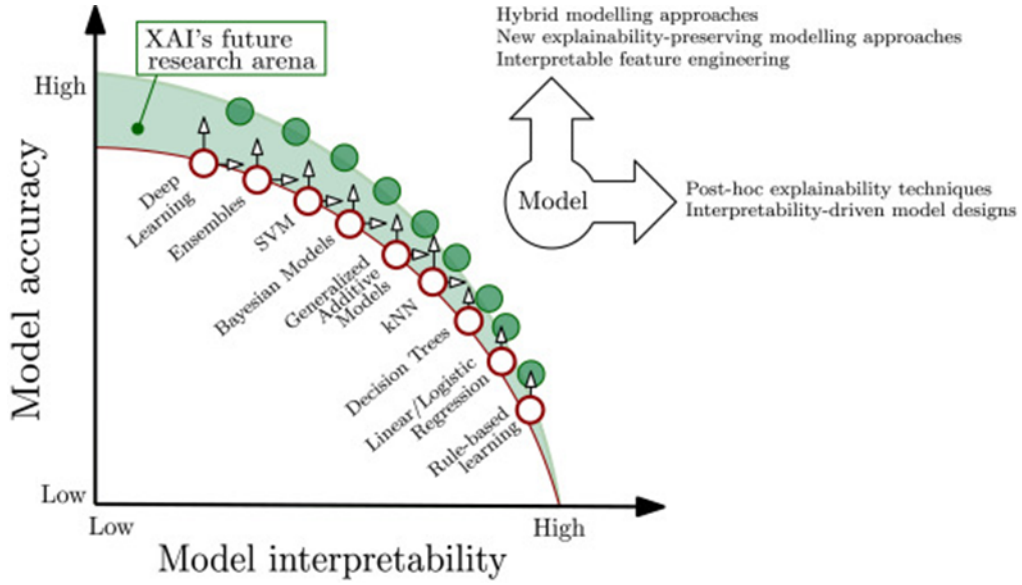


Figure 35. Accuracy–explainability trade-off.

Feature engineering is one way to increase model interpretability, trainability, and regularization, but these traits often come at the cost of model accuracy. With very highly parameterized models, the same feature-generation process happens via gradient descent rather than an engineer's specific guidance. Consequently, learned features can be generated in ways that humans would not have conceived, leading to very high accuracy. Many ML applications derive their very high accuracies from this phenomenon

and attempt to have as little human-guided featurization as possible. However, using ML-guided features relies on having other ways of providing explanations and regularization. Explainability methods can make the use of these large models tractable.

6.4.4 Types of Explainability

Many different taxa of explainability techniques have been suggested. Phillips [49] separates explainability methods into local or global scope and intrinsic model-specific or post hoc model-agnostic methods. Molnar uses a series of dichotomies [50]:

- Intrinsic or post hoc
- Result of the method
  - Feature summary statistic
  - Feature summary visualization
  - Model internals
  - Data point
  - Intrinsically interpretable model
- Model-specific or model-agnostic
- Local or global

Common techniques relevant to radiation detection include Local Interpretable Model-Agnostic Explanations (LIME) [51] and Shapley Additive Explanations (SHAP) [52]. These techniques are local, model-agnostic, post hoc methods. They use the approximation of model behavior around a single classification to attempt to elucidate the model's inner workings. Also very relevant are a host of intrinsic techniques specific to model architectures such as examining the structure of binary trees or analyzing the attention in a transformer.

Current efforts suffer from the fact that the methods are difficult to understand for non-ML practitioners. Few, if any, tools are available for making explainability results understandable to a lay person. Additionally, many tools output results that are qualitative rather than quantitative. The qualitative nature complicates the design of automated responses that use the results effectively.

7. CONCLUSION

This report presents recommendations and metrics for evaluating datasets and algorithms for radiation-detection missions. The authors hope that these tools are useful for those creating new datasets and evaluating algorithm performance and that they may guide future research in the field. Although this report is as comprehensive as possible, more work remains. In particular, more research and data are needed to quantify activity concentrations in radon washout, and more data are needed to better quantify variations caused by fast transients. Moreover, the algorithms presented in this report do not encompass all algorithm types—especially new algorithms developed using ML methods. As newer algorithms come online, more research will be required to quantify their performance.

8. REFERENCES

[1] American National Standard Performance Criteria for Backpack Based Radiation-Detection Systems Used for Homeland Security, ANSI Standard N42.53-2021.

[2] American National Standard Performance Criteria for HandHeld Instruments for the Detection and Identification of Radionuclides, ANSI Standard N42.34-2019.

[3] Trevisi, R., Risica, S., D'alessandro, M., Paradiso, D., Nuccetelli, C. (2012). "Natural radioactivity in building materials in the European Union: a database and an estimate of radiological significance." *Journal of Environmental Radioactivity* 105: 11–20.

[4] Swinney, M. W., Peplow, D. E., Nicholson, A. D., Patton, B. W. (2016). "NORM concentration determination in common materials in an urban environment." *Transactions of the American Nuclear Society* 114(1).

[5] Hannan, M., Wahid, K., Nguyen, N. (2015). "Assessment of natural and artificial radionuclides in Mission (Texas) surface soils." *Journal of Radioanalytical and Nuclear Chemistry* 305(2): 573–582.

[6] Isinkaye, M. O., Shitta, M. B. O. (2010). "Natural radionuclide content and radiological assessment of clay soils collected from different sites in Ekiti State, southwestern Nigeria." *Radiation protection dosimetry* 139(4): 590–596.

[7] Nicholson, A. D., Peplow, D. E., Ghawaly, J. M., Willis, M. J., Archer, D. E. (2020). "Generation of Synthetic Data for a Radiation Detection Algorithm Competition," *IEEE Transactions on Nuclear Science* 67(8): 1968–1975, https://doi.org/10.1109/TNS.2020.3001754.

[8] Livesay, R. J., Blessinger, C. S., Guzzardo, T. F., Hausladen, P. A. (2014). "Rain-induced increase in background radiation 495 detected by Radiation Portal Monitors," *J. Environ. Radioact.* 137: 137–141.

[9] Takeyasu, M., Yamasaki, K., Tsujimoto, T., Ogawa, Y., Urabe, I. (1993). "Radon-222 progeny in precipitation." *Proceedings of Asia congress on radiation protection.* (p. 751). China.

[10] Takeyasu, M., Iida, T., Tsujimoto, T., Yamasaki, K., Ogawa, Y. (2006). "Concentrations and their ratio of 222Rn decay products in rainwater measured by gamma-ray spectrometry using a low-background Ge detector." *Journal of Environmental Radioactivity* 88(1): 74–89.

[11] Paatero, J. (2000). "Wet Deposition of Radon-222 Progeny in Northern Finland Measured with an Automatic Precipitation Gamma Analyser." *Radiation Protection Dosimetry* 87(4): 273–280. https://doi.org/10.1093/oxfordjournals.rpd.a033008.

[12] Mercier, J. F., Tracy, B. L., d'Amours, R., Chagnon, F., Hoffman, I., Korpach, E. P., Johnson, S., Ungar, R. K. (2009). "Increased environmental gamma-ray dose rate during precipitation: a strong correlation with contributing air mass." *J. Environ. Radioact.* 100(7): 527–33. https://doi.org/10.1016/j.jenvrad.2009.03.002.

[13] Shaofei, Z., McCutchan, E. A. (2021). "Nuclear Data Sheets for A = 214", *Nuclear Data Sheets* 175: 1–149, https://doi.org/10.1016/j.nds.2021.06.001.

[14] Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., Houston, T. G. (2012). "An overview of the Global Historical Climatology Network-Daily Database." *Journal of Atmospheric and Oceanic Technology* 29: 897–910. https://doi.org/10.1175/JTECH-D-11-00103.1.

[15] Global Historical Climatology Network daily (GHCNd), National Climatic Data Center, NESDIS, NOAA, U.S. Department of Commerce [Accessed 05/01/2023]

[16] U.S. Hourly Precipitation Data, National Oceanic and Atmospheric Administration, DSI-3240 [Accessed 05/01/2023]

[17] Walker, M. I., Rose, K. S. B. (1990). "The radioactivity of the sea." *Nuclear Energy* 29(4): 267–278.

[18] Sandness, G. A., Schweppe, J. E., Hensley, W. K., Borgardt, J. D., Mitchell, A. L. (2009). "Accurate Modeling of the Terrestrial Gamma-Ray Background for Homeland Security Applications." *2009 IEEE Nuclear Science Symposium and Medical Imaging Conference Record*, 126–133.

[19] Bandstra, M. S., Quiter, B. J., Salathe, M., Bilton, K. J., Curtis, J. C., Goldenberg, S., Joshi, T. H. Y. (2021). "Correlations between panoramic imagery and gamma-ray background in an urban area." *IEEE Transactions on Nuclear Science* 68(12): 2818–2834.

[20] Leber, D. D., Pibida, L. (2020). "False Alarm Testing for Radiation Detection Systems." *NIST Technical Note* 2118.

[21] Brown, L. D., Cai, T. T., DasGupta, A. (2001). "Interval estimation for a binomial proportion." *Statistical Science* 16(2): 101–133.

[22] Agresti, A., Coull, B. A. (1998). "Approximate is better than 'exact' for interval estimation of binomial proportions." *The American Statistician* 52(2): 119–126.

[23] International Atomic Energy Agency. (2002). "Detection of radioactive materials at borders." IAEA-TECDOC-1312.

[24] Domestic Nuclear Detection Office. (2013). "Technical Capability Standard for Backpack Based Radiation Detection Systems." US Department of Homeland Security.

[25] Enghauser et al. (2015). *Nuclide Identification Algorithm Scoring Criteria And Scoring Application*. DHS/DNDO Algorithm Improvement Program Document Number: 600-AIP-124060v0.00.

[26] Nicholson, A. D., Peplow, D. E., Ghawaly, J. M., Willis, M. J., Archer, D. E. (2020) "Generation of synthetic data for a radiation detection algorithm competition." *IEEE Trans. Nucl. Sci.* 67(8): 1968-1975

[27] Topcoder. (2020) Detecting radiological threats in urban areas, https://www.topcoder.com/challenges/30085346

[28] Ghawaly, J. M., Nicholson, A. D., Peplow, D. E., Anderson-Cook. C. M., Myers, K. L., Archer, D. E., Willis, M. J., Quiter, B. J. (2020) "Data for training and testing radiation detection algorithms in an urban environment." *Scientific Data* 7(1): 328.

[29] Morrow, Tyler, Price, Nathan, & McGuire, Travis. (2021, April 28). *PyRIID v.2.0.0.* [Computer software]. https://github.com/sandialabs/pyriid. https://doi.org/10.11578/dc.20221017.2.

[30] Currie, L. A. (1968). "Limits for Qualitative Detection and Quantitative Determination: Application to Radiochemistry." *Anal. Chem.* 40: 586–593.

[31] Fehlau, P. E., Pratt, J. C., Markin, J. T., Scurry, Jr., T. (1983). "Smarter Radiation Monitors for Safeguards and Security." *Nuclear Materials Management* X11: 294.

[32] Fehlau, P. E., Coop, K. L., Markin, J. T. (1984). "Application of Wald's sequential probability ratio test to nuclear materials control." LA-UR-84-2782; CONF-8409170-1. Los Alamos National Laboratory. Los Alamos, New Mexico.

[33] Fehlau, P. E. (1993). “Comparing a recursive digital filter with the moving-average and sequential probability-ratio detection methods for SNM portal monitors.” *IEEE Transactions on Nuclear Science* 40(2): 143–146.

[34] Connolly, E. L., Martin, P. G. (2021). “Current and Prospective Radiation Detection Systems, Screening Infrastructure and Interpretive Algorithms for the Non-Intrusive Screening of Shipping Container Cargo: A Review.” *Journal of Nuclear Engineering* 2(3): 246–280.

[35] Jarman, K. D., Smith, L. E., Carlson, D. K., Anderson, D. N. (2003). “Sequential probability ratio test for long-term radiation monitoring.” *2003 IEEE Nuclear Science Symposium* 2: 1458–1462. (Conference Record IEEE Cat. No. 03CH37515).

[36] Monterial, M., Morton, A., Nelson, K., Labov, S., Hecht, A. (2019). “Benchmarking Algorithm for RadioNuclide Identification.” *2019 IEEE Nuclear and Science Symposium and Imaging Conference,* Machester, United Kingdom. October 26–November 2, 2019.

[37] Bilton, K. J., Joshi T. H., Bandstra, M. S., Curtis, J. C., Quiter, B. J., Cooper, R. J., and Vetter, K, (2019) “Non-negative Matrix Factorization of Gamma-Ray Spectra for Background Modeling, Detection, and Source Identification” *IEEE Transactions on Nuclear Science* 66(5): 827-837.

[38] Horne, S. M., et. al. (2019). *GADRAS-DRF Version 18 User’s Manual*. SAND2019-14655. Sandia National Laboratories, Albuquerque, New Mexico.

[39] Wilkinson, M. D., et al. (2016). “The FAIR Guiding Principles for scientific data management and stewardship.” *Scientific Data* 3: 160018.

[40] Kapoor, S., Narayanan, A. (2023). “Leakage and the Reproducibility Crisis in ML-based Science.” *Patterns* 4(9): 100804. https://doi.org/10.1016/j.patter.2023.100804.

[41] Detwiler, R. S., McConn, R. J., Grimes, T. F., Upton, S. A., Engel, E. J. (2021). *Compendium of Material Composition Data for Radiation Transport Modeling*. PNNL-15870-Revision-2 TRN: US2216118. Pacific Northwest National Laboratory. Richland, Washington.

[42] Goodfellow, I. J., et al. (2014) “Generative Adversarial Networks.” arXiv:1406.2661.

[43] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H. Laviolette, F., Marchand, M., Lempitsky, V. (2016). “Domain-Adversarial Training of Neural Networks.” *Journal of Machine Learning Research* 17: 1–35.

[44] McInnes, L. Healy, J., Saul, N., Grossberger, L. (2018). “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.” *Journal of Open Source Software* 3(29): 861.

[45] Hagen, A., Jackson, S., Kahn, J., Strube, J., Haide, I., Pazdernik, K., Hainje, C. (2021). “Accelerated Computation of a High Dimensional Kolmogorov–Smirnov Distance.” arXiv:2106.13706.

[46] Mahalanobis, P. C. (1936). “On the generalized distance in statistics.” *Proceedings of the National Institute of Sciences of India* 2(1): 49–55.

[47] Amarasinghe, K. Rodolfa, K. T., Lamba, H., Ghani, R. (2023). “Explainable Machine Learning for Public Policy: Use Cases, Gaps, and Research Directions.” *Data & Policy* 5: e5. https://doi.org/10.1017/dap.2023.2.

[48] Rudin, C. (2018). “Please stop explaining black box models for high stakes decisions.” arXiv:1811.10154v1.

[49] Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., Przybocki, M. A. (2020). *Four Principles of Explainable Artificial Intelligence*. NISTIR 8312.

National Institute of Standards and Technology. Gaithersburg, Maryland. https://doi.org/10.6028/NIST.IR.8312.

[50] Molnar, C. (2022). *Interpretable Machine Learning: A Guide For Making Black Box Models Explainable*. Munich, Germany: Christoph Molnar.

[51] Ribiero, M. T., Sing, S., Guestrin, C. (2016). "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 1135–1144. https://doi.org/10.1145/2939672.2939778.

[52] Lundberg, S. M., Lee, S.-I. (2017). "A Unified Approach to Interpreting Model Predictions." *NIPS17: Proceedings of the 31st International Conference on Neural Information Processing Systems*: 4768–4777.