# Domain-Aware ML-Driven Predictive Analytics for Real-Time Proliferation Detection in Urban Environments

Ellyn Ayton
Sannisth Soni
ellyn.ayton@pnnl.gov
soni.sannisth@pnnl.gov
Pacific Northwest National Laboratory
Richland, Washington, USA

Mark Banstra
Brian Quiter
Reynold Cooper
msbandstra@lbl.gov
bjquiter@lbl.gov
rjcooper@lbl.gov
Lawrence Berkeley National Laboratory
Berkeley, California, USA

Svitlana Volkova
svitlana.volkova@pnnl.gov
Pacific Northwest National Laboratory
Richland, Washington, USA

## ABSTRACT

This work addresses the national security non-proliferation mission by enhancing traditional methods for real-time nuclear proliferation detection. First, by adding a novel predictive component that allows us to move from a reactive to a proactive posture. Second, by adding the ability to mitigate the operational burden posed by nuisance alarms during the deployment of unattended radiological sensors in urban environments. We demonstrate how to successfully operationalize the state-of-the-art machine learning (ML) and natural language processing (NLP) models to quantitatively estimate (1) to what extent historical radiological sensor data is useful to anticipate future isotope signatures across sensors and locations, and (2) whether contextual data e.g., language extracted from construction permits can inform and explain future nuclear sensor and isotope signatures. Our models are trained on real-world data collected from seven sensors located in Washington, DC and Fairfax, VA during the time of seven to nine months in 2019 and 2020. Our sensor data includes alerts from three medical Tc-99m, I-131 and 511 from Positron Emission Tomography (PET) and one industrial Cs-137 isotope. Our experimental results show clear ability of ML models to anticipate isotope signatures from historical data across locations and sensors, and show strong predictive power of linguistic terms extracted from construction permits to classify and explain industrial alerts. We found that when learning from historical data detecting isotopes in Fairfax is easier than in DC, learning from longer historical windows e.g., month is better than days and weeks, and 511 signatures are more difficult to predict than Tc-99m and I-131 across locations.

## CCS CONCEPTS

**• Information systems → Sensor networks**; **• Computing methodologies → Natural language processing**.


*Conference'17, July 2017, Washington, DC, USA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

## KEYWORDS

real-time proliferation detection, distributed sensor networks, machine learning, natural language processing

**ACM Reference Format:**
Ellyn Ayton, Sannisth Soni, Mark Banstra, Brian Quiter, Reynold Cooper, and Svitlana Volkova. 2022. Domain-Aware ML-Driven Predictive Analytics for Real-Time Proliferation Detection in Urban Environments. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

The national security community looks to expand threat detection capability through the deployment of urban sensing systems that must operate without continuous reliance on human subject matter expertise. The nature of these systems makes it harder to detect elicit proliferation activities in the city compared to rural and industrial areas because of the necessity to be discrete and not interfere with daily city life. To enable that, we present a pioneering study that demonstrates and rigorously validates the value of interpretable ML-driven analytics to anticipate radiological isotope signatures in real-world urban sensing environments. Traditional approaches for urban-scale proliferation detection have been reactive in nature and relied on simulated data when incorporating contextual information. In this work, we are the first to forecast radiological alerts in urban environments in order to identify anomalous signals that indicate the presence of suspicious or nefarious activity. In our experiments, we model isotope time series under three settings ranging from the most general granularity to the most specific: aggregated within each location (location-specific), aggregated within each sensor (sensor-specific), and individual isotope time series (isotope-specific).

*AI for Distributed Sensor Networks.* Machine learning applied to distributed sensor networks primarily focuses on solving the task of anomaly detection, rather than classification or regression. SOTA deep learning methods such as encode-decoder models [18] and variational autoencoders with convolutional layers [2] require large amounts of training data and explode in complexity with additional layers and parameters that adds strain to deployed systems in an online setting. Other methods developed using dimentionality reduction techniques e.g., principal component analysis (PCA) [22]

or least square-support vector machines and Gaussian process regression [17] reduce this strain. However, automated, unsupervised feature selection still suffers from a lack of interpretability of model decisions which is vital to support the nonproliferation mission.

*AI for Proliferation Detection.* Machine learning and natural language processing approaches have not been widely applied to support non-proliferating mission by developing and deploying domain-aware real-time analytics. Some recent examples include anomaly detection [21, 23] and ML to analyze gamma ray spectra [13, 26], the application of ML models to perform nuclear reactor core diagnosis [14], enhance nuclear energy systems behavior and decision making [9], and analyze distributed and mobile sensor networks [4, 7]. However, there have been an increased interest to take advantage of publicly available information and combine it with machine learning to discover elicit proliferation activities, as described in 2021 Nuclear Threat Initiative report [12]. The 2021 report on nuclear proliferation and arms control monitoring, detection, and verification highlights the role of ML-driven analytics and open data sources to discover and prevent global proliferation [20].

Unlike any previous work, this paper focuses on interpretable ML models that can be easily operationalized and deployed to advance the national security mission by analysing real-world radiological sensor data to anticipate future isotope signatures across sensors and locations [25], and incorporate real-world contextual data to disambiguate future physical sensor signatures in urban environments [24].

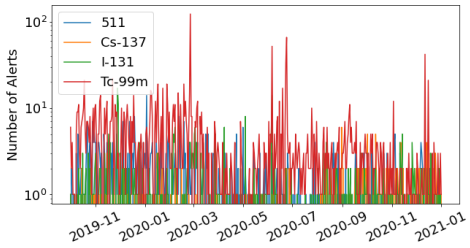
Unlike any other efforts, our data-driven approach is:

- Taking advantage of both historical pattern-of-life (*PoL*) data from urban sensors and open-source data (e.g., publicly available construction permits).
- Capable of anticipating radiological isotope signatures of three medical Technetium-99m, radioactive iodine (I-131) and 511 from Positron Emission Tomography(PET) and one industrial isotope Caesium-137 (Cs-137).
- Making reliable predictions with confidence levels reported across multiple sensors in multiple locations – Washington, DC and Fairfax, VA.

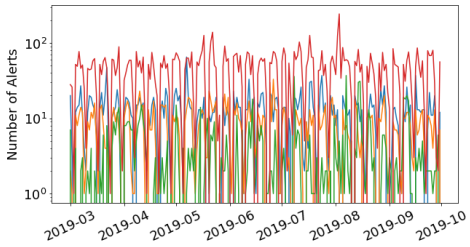
## 2 DATA

### 2.1 Radiological Sensor Data

We collected radiological sensor data of four isotopes from static sensors in two locations – Washington, DC [6] and Fairfax, VA [11]. DC data was collected from 5 sensors between October 2019 and January 2021. Fairfax data was collected from 4 sensors between March 2019 and October 2019. All sensors were placed in or near fire stations. Though the exact detection range is dependent on factors such as the strength, type, and speed of the source, the urban sensors used for our data collection are typically capable of detecting gamma-ray sources within a radius of several tens of meters from each sensor. To separate gamma-ray spectra from background noise and to classify which isotope generated that spectra, sensor alerts were summed into 3-second-long increments. A spectrum is distinguished from background noise if its gross count rate was an anomalously large increase over the mean rate obtained from a moving average. The spectra thus flagged were normalized and clustered using k-means [16] with Kullback-Leibler




**(a) Washington, DC**




**(b) Fairfax, VA**

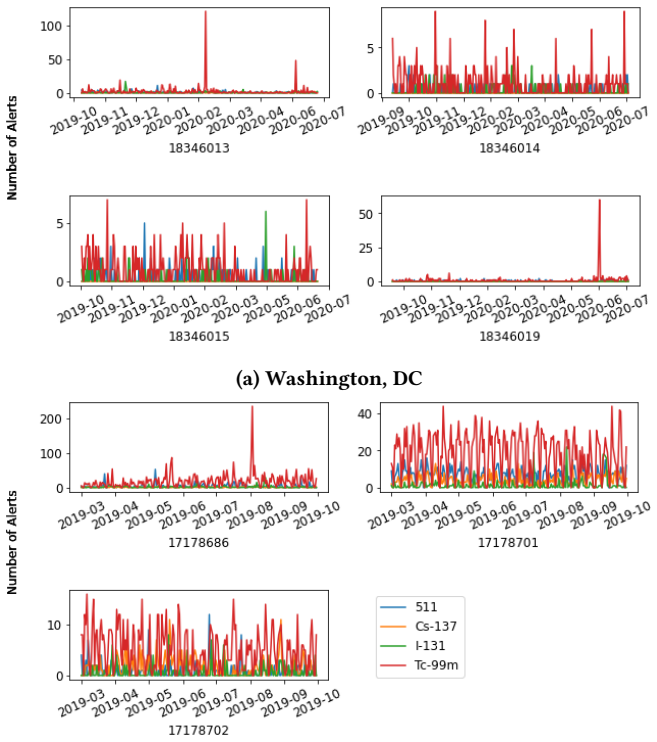
**Figure 1: Location-specific radiological sensor data aggregated over 5 sensors in DC and 4 sensors in Fairfax represented as the daily number of alerts (log-scale) over several months for four isotopes: Tc-99m, I-131, 511, and Cs-137.**

divergence [10] as the distance metric. The resulting clusters were culled and hand-labeled by subject matter experts (SMEs) with the isotope responsible for that cluster of anomalies[1]. Non-anomalous spectra were also used to characterize the background signal [15]. Then the identified isotopes were searched for in all of the spectra according to the method described in [3]. During the time period in which the sensors were recording data, no isotopes associated with nuclear proliferation efforts *e.g.,* plutonium, highly-enriched uranium, or tritium were detected. Thus, we focus on modeling and forecasting only the isotopes present in the data.

Figure 1 shows the number of daily of alerts (log-scale) across the two locations. Three of the four detected isotopes, Technetium-99m (Tc-99m), radioactive iodine (I-131), and 511 gamma photons (511), are typically associated with medical procedures. For example, an alert of 511 might mean a person passing close to a sensor had a recent positron emission tomography (PET) scan. The fourth isotope, Cs-137, is commonly associated with construction work. Across all sensors, isotope Tc-99m is the most abundant with 10,283 alerts between both locations. We analyzed weekly patterns for all isotopes with the majority of activity occurring during weekdays.

For our experiments we construct two datasets of radiological sensor data split by time and sensor. For location-specific and sensor-specific experiments, we include the entire time period of Fairfax detections and include up to July 2020 of DC detections (*Sensor Data A*). Note, due to sparse signal from these four isotopes in some sensors, we dropped one sensor from Fairfax and one sensor from DC from *Sensor Data A*. The daily frequency of alerts from the remaining sensors are shown in Figure 2. We see sensor 17178701

[1]We acknowledge there are limitations with this approach and possibly erroneously clustered spectra, however, without knowing the true isotopes to pass by the sensors we must trust our SME labels.




(a) Washington, DC

(b) Fairfax, VA

**Figure 2: Sensor-specific data in DC and Fairfax represented as the daily number of alerts for four isotopes: Tc-99m, I-131, 511, and Cs-137.**

(a) Washington, DC (b) Fairfax, VA

**Figure 3: Construction permit signal coverage around radiological sensors (marked as red pins) in both locations. Red regions represent areas with higher density, and lighted green areas - with lower density.**

from Fairfax has the most number of alerts (6,290), and the 18346019 sensor from DC has the least number of alerts (307) during the time period. For isotope-specific experiments, we include the full time period for both cities, but filter to only Cs-137 isotope alerts. With the full time frame included, the two sensors previously dropped from *Sensor Data A* contain many more Cs-137 alarms (1,935) and thus are kept in *Sensor Data B*. However, three of the DC sensors continued to have no Cs-137 detections. *Sensor Data B* consists of Cs-137 alerts from two sensors located in DC and four sensors in Fairfax.

## 2.2 Open Source Contextual Data

We collected contextual data, specifically construction permits from two county websites for Washington, DC[2] and Fairfax, VA[3]. Overall, we processed 46,057 permit records for Fairfax and 41,580 permit records for Washington, DC. Construction permits data is semi-structured and contains spatial e.g., latitudes, longitudes, and temporal information e.g., start and end dates of construction permits, and the metadata e.g., construction type, permittee name, contractor name, applicant name, permit type, in addition to natural language description of the construction work. The example of construction permit descriptions for DC (top) and Fairfax (bottom) are shown below. Figure 3 presents construction permit signal density around radiological sensors in DC and Fairfax.

DC: *Installation of a Class B telecom facility on a replacement DDOT owned streetlight pole.*

FF: *Office exact replacement 100,000 BTU gas furnace and 5 ton air conditioner. Installing sump pump. 3-lights, 7 plugs, 2 switches, 1-120 volt circuit 2 GFI outlets.*

We associate construction permit data with sensors from *Sensor Data B* based on spatial and temporal characteristics. For example, for each day we collect active permits within 0.25 mile radius around three sensors 18346013, 18261333, and 17178701, a 0.5 mile radius for sensor 17178686, and a 1 mile radius for sensors 17178694 and 17178702. We tune the radius parameter to ensure variability in the number of active construction permits around each sensor. To gain more insights about the topics covered by construction permits, specifically topic coverage and topic diversity, we visualize topics extracted from construction permit documents using the topic model from [1] in Figure 4. As we can see from the UMAP projection [19] of topic vectors, construction topics extracted from permits in DC include street work and telecommunications work; whereas topics extracted from Fairfax permits include interior repairs, electric work, installation, renovation, replacement type of work, excavation, paving etc.

# 3 METHODOLOGY

We develop and validate novel predictive models capable of anticipating isotope signatures from historical and contextual data across locations and sensors. For that, we designed three types of classification experiments (two multi-class and one binary) to perform (a) location-specific, (b) sensor-specific and (c) isotope-specific predictions. Our detailed experimental setup is presented in Table 1.

**Table 1: Detailed experimental setup for next day medical and industrial isotope prediction.**

| Experiment | Inputs | Outputs | History (*PoL*) |
|---|---|---|---|
| Location-specific<br>Sensor-specific | PoL | Tc-99m, I-131<br>511, Cs-137 (4-way) | Ensemble, day,<br>week, month |
| Isotope-specific | PoL, OSD<br>PoL + OSD | Cs-137 (2-way) | Day |

We rely on the SOTA machine learning models – Support Vector Machines (SVM), Random Forest (RF), Logistic Regression (LR) and AdaBoost (only for isotope-specific experiments) to predict next

[2]https://opendata.dc.gov/
[3]https://ldip.fairfaxcounty.gov/page/search

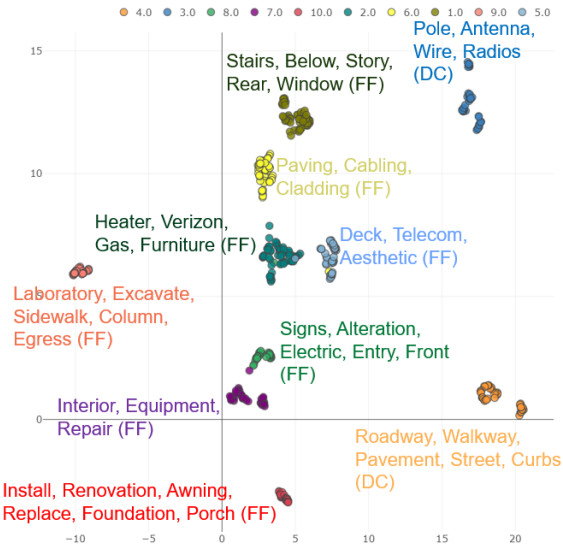



**Figure 4: Topic modeling results on construction permits from Fairfax and DC. Permits are aggregated on daily basis before fitting the topic model e.g., if there are three permits active on a day, all are aggregated into a single document (one data point on the plot). Topics are numbered from 1 to 10; top representative tokens per topics are reported.**

day isotope presence in each location and each sensor. Unlike deep learning (DL) approaches, that are black-box models, ML-driven analytics powered by our model are interpretable, faster, and more secure to deploy. We frame our experiments as classification tasks which make traditional time series model *e.g.,* ARIMA unsuitable. We learn location-specific and sensor-specific models from historical *pattern of life data (PoL)* by considering an $n$-size historical window in the past. In these experiments, we vary the window size of historical data for training between one day, one week, and one month. In addition, we also experiment with the ensemble models of the three windows. In our isotope-specific experiments, where we focus on anticipating the presence an industrial Cs-137 isotope, we consider a historical window size of one day in addition to different contextual data representations. For every model, we perform a grid search of the hyperparameter space[4]. We compare our models to a weighted classifier where predictions are sampled from the training distribution weighted by the class prevalence.

Our experiments aim to answer four research questions:

RQ1 How much historical *PoL* signal is required to train best-performing models *i.e.,* a day, a week, or a month?

RQ2 How predictive performance vary among isotopes *i.e.,* are some isotopes easier to predict than others? What are the most predictive models, and what are their limitations ?

RQ3 How does predictive performance vary with the type and the amount of training data *i.e.,* are models more accurate with sensor-specific or location-specific data?

RQ4 How much can model performance be improved when the historical signals are replaced or augmented with contextual open-source information?

[4]SVM hyperparameters include the kernel (linear, poly, rbf, sigmoid) and the regularization term (0.01, 0.1, 1.0), RF hyperparameters include the number of estimators (100, 200, 500), max depth (10, 50), and criterion function (gini, entropy). LR hyperparameters include the penalty term (L1 and L2), tolerance (0.0001, 0.001, 0.00001), and regularization term (0.01, 0.1, 1.0).

## 3.1 Location-Specific Models

In our location-specific experiments, we classify next-day alerts of four isotopes from signals aggregated over all Fairfax sensor data (three sensors) and all DC sensor data (four sensors). We reserve the last 40% of our data for testing. From Fairfax data, the start date of the test set is 07-16-2019. In DC, the test set start date is 02-26-2020. This splitting scheme does not create a 60/40 division of days, but rather a 60/40 division of alerts. Thus, the test set may cover more than 40% of the time period. In total, we train 12 (3 ML models x 4 time windows) models for each location.

## 3.2 Sensor-Specific Models

In our sensor-specific experiments, we classify next-day alerts of four isotopes from sensors in *Sensor Data A*. However, we do not aggregate isotope signals across sensors as in our location-specific experiments. Instead, each sensor is treated as a distinct dataset on which separate models are trained to classify next-day isotope signatures. To evaluate the effect on the type and the size of the training data and to get insights about the specificity vs. generalizability of each model, we use three settings:

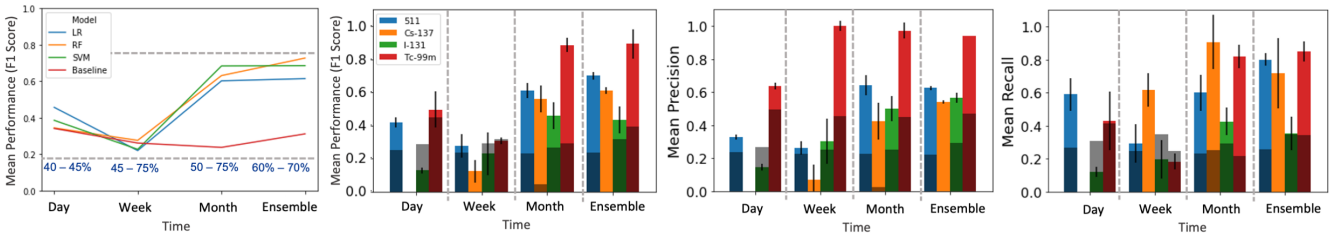
- *My Sensor Data*: We only learn from sensor-specific historical data (e.g., correspond to high specificity).
- *My Location Sensor Data*: We learn from historical data collected from all sensors in a specific location.
- *All Sensor Data*: We learn from historical data collected from all sensors across two locations (e.g., correspond to high generalizability).

Similarly to location-specific experiments, we use a 60/40 train/test splits. However, our total number of models for this experimental setup has increased from 24 (12 x 2 locations) to 12 models per sensor yielding 84 models for each setup (in total 232 sensor-specific models).

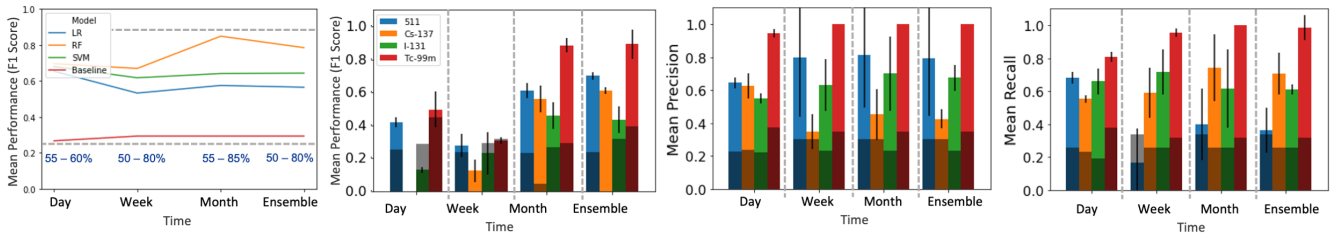
## 3.3 Isotope-Specific Models

In addition to learning from historical *PoL* data, we experiment with isotope-specific models to predict next-day industrial isotope presence (Cs-137) from five types of representations learned from construction permits as described below.

- Metadata: 67-dim vectors that encode construction company names and construction types e.g., Verizon, Wash Gas & Light Co., excavation, fixture, paving etc.
- Embeddings: 10-dim embedding vectors (reduced from the original 512 dimensions using UMAP) learned using the model from [5].
- Topics: 10-dim topic vectors encoded using the model from [1].
- Part-of-speech-tags (PoS): 50-dim token vectors that include nouns and verbs extracted from construction permits using the AllenNLP model from [8] (reduced from original 515 dimensions).
- TFIDF: 48-dim token vectors filtered based on frequency.




**(a) Washington, DC**




**(b) Fairfax, VA**

**Figure 5: Location-specific model performance with signals aggregated over multiple sensors per location for 4-way isotope classification task. Top figures demonstrate how model performance depends on the size of the historical window (day, week, month, ensemble). Bottom plots report model performance with confidence ranges for each isotope (511, Cs-137, I-131, Tc-99m). The shaded regions represent the baseline model. Model confidence ranges reported as percentages.**

**Table 2: Isotope-specific experimental setup: train and test time periods per sensor. Total training samples across all sensors is 294 and testing samples is 131. Detectors that start with 18 represent two DC detectors and those that start with 17 represent four Fairfax detectors.**

| Detector | Train Period | Test Period |
|---|---|---|
| 18261333 | 2019/10/03, 2020/09/27 | 2020/10/10, 2020/12/25 |
| 18346013 | 2019/10/17, 2020/05/10 | 2020/05/11 2020/06/28 |
| 17178686 | 2019/03/02, 2019/07/09 | 2019/07/10, 2019/09/28 |
| 17178694 | 2019/03/04, 2019/06/21 | 2019/06/22, 2019/08/02 |
| 17178701 | 2019/03/07, 2019/06/27 | 2019/07/12, 2019/09/12 |
| 17178702 | 2019/03/07, 2019/07/31 | 2019/08/01, 2019/09/27 |

For each sensor, we aggregate construction permits for every day of the sensor's lifespan and indicate if there was a Cs-137 alert or not. We train binary models for two locations, experiment with three type of ML models (RF, SVM and AdaBoost) and train models using contextual data only vs. contextual data plus the *PoL*. We report the details about the number of data samples and the train and test periods for each sensor across locations in Table 2.

# 4 EXPERIMENTAL RESULTS

This section presents key findings from three types of analytics proposed in this work – location-specific, sensor-specific, and isotope-specific to predict next day isotope prevalence for medical and industrial isotopes in DC and Fairfax. Our results provide quantitative estimates on (1) how historical sensor data can be leveraged to predict next-day isotope signatures, and (2) how much the incorporation of real-world contextual data e.g., language extracted from construction permits informs and explains sensor signatures.

## 4.1 Location-Specific Results

In Figure 5, we present the performance of our ML models *i.e.*, model confidence ranges and F1 scores over the 4-way isotope classification task from aggregated sensor signals for DC (5a) and Fairfax (5b). In DC, all day and week models perform considerably worse than their Fairfax equivalents. The best model trained with DC sensors is the Random Forest ensemble with an F1 score of 0.72. We see from the confusion matrix that isotope 511 has the highest rate of classification (90.5%) followed by both Tc-99m and Cs-137 (79.8% and 79.3%). The I-131 class has the most misclassifications with a success rate of 44.2%. Across both locations, any model trained with a month of history or an ensemble of all time granularities significantly outperforms the baseline. In particular, the Fairfax monthly Random Forest model has the highest F1 score of 0.85. From the confusion matrix we infer that the Fairfax model correctly classifies Tc-99m 100% of the times, has a 75% correct prediction rate for I-131 isotope, a 52% correct prediction rate for 511, and an 87.5% rate for Cs-137. Interestingly, Cs-137 is frequently over-predicted, most often mistaking 511 alerts and I-131 alerts. We also note that the predictions from Fairfax models are generally more confident (approximately 10% greater) than the DC models. Model confidence allows us to measure how close a prediction is to the decision boundary. A more confident model lends greater trust to the user when encountering unseen data.

Next, we analyze how predictions from these models vary among the class labels (isotopes) in Figures 5a and 5b. Almost unanimously, Tc-99m is the easiest isotope to classify from sensor signals in both DC and Fairfax. The single exception comes from the DC weekly models where the random baseline outperforms the trained models. In Figure 5a, isotopes Cs-137 and I-131 are the most difficult to predict especially for the daily models. We see vast improvement

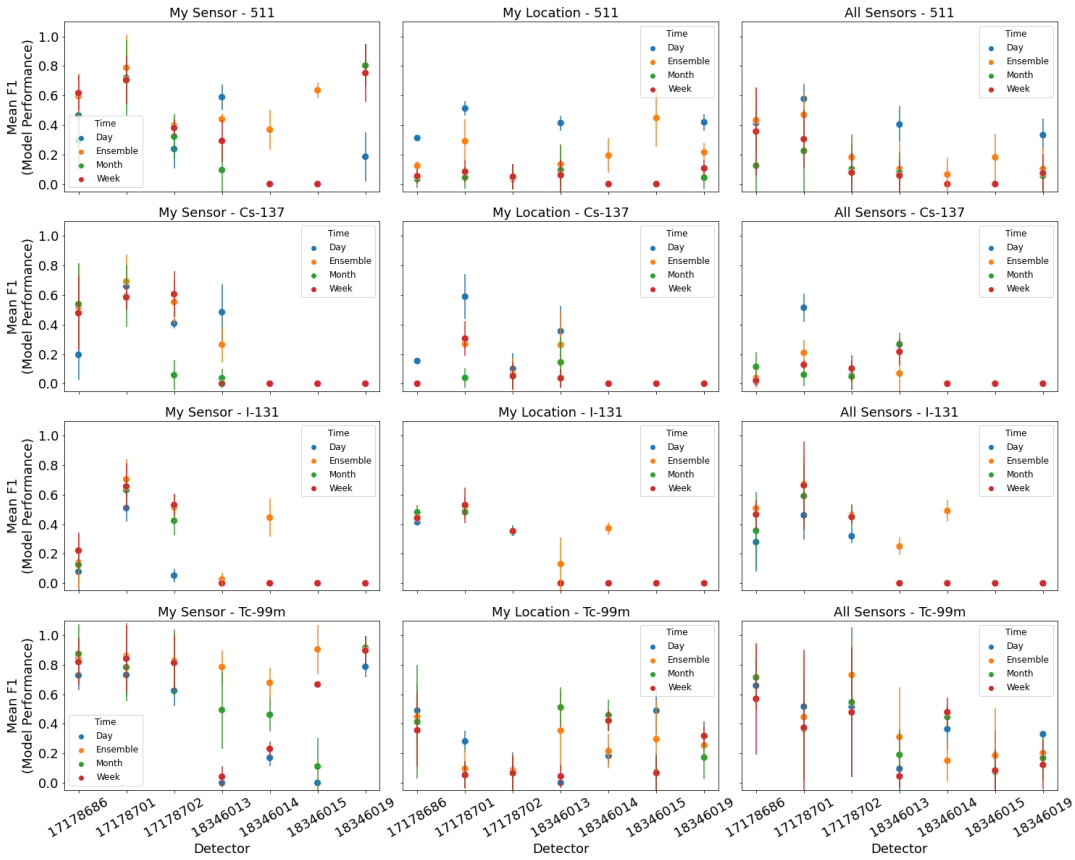



**Figure 6: Model performances for the 4-way isotope classification task. Each point indicates the mean F1 score (averaged over three models - SVM, RF and LR) and 95% confidence interval. Rows indicate isotopes; columns indicate the type and the amount of training data used. In the leftmost column (*My Sensor*), models are trained only on individual sensor data. In the middle column (*My Location*), the training data includes alerts from all sensors in a given city. In the rightmost column (*All Sensors*), models are trained on data from both cities. In each experimental set up, models are tested on only individual sensor data. Sensors 18346014, 18346015, and 18346019 have no Cs-137 alerts. Additionally, 18346015 and 18346019 have no I-131 alerts.**

in these classes from the Fairfax models with a trade-off in 511 class performance. Figure 5b illustrates this trade-off showing that Cs-137 and I-131 have higher individual F1 scores from most models than 511. *We conclude that a larger PoL window e.g., a month or an ensemble leads to more accurate models (RQ1) and that Tc-99m is the easiest isotope to classify while others are location-dependent (RQ2).*

## 4.2 Sensor-Specific Results

Figure 6 shows how predictive performance varies by sensor by reporting mean F1 scores from the three model types per sensor under three training conditions: *My Sensor Data, My Location Sensor Data*, and *All Sensor Data.*

**Predictive performance variability across sensors and isotopes** We observe that signal from the Fairfax sensor, 17178701, yields models with the highest performance with a mean F1 score of 0.51 across isotopes and training conditions. Models trained on the *PoL* data from this sensor have higher performance in the *My Sensor Data* setting (when only trained on this sensor's data).

As more data is included in train, we see large decreases in model performance. For example, in classifying 511, the 17178701 models always achieve an F1 score $\geq 0.6$ (see the top leftmost plot in Figure 6). However, in the *My Location Sensor Data* and *All Sensor Data* settings, 0.6 is the upper bound on model performance (see top row plot in Figure 6). Similar observations can be made of the other isotopes classified by the 1718701 models. On average, the lowest performing models receive data from sensor, 18346013, originating from DC and has a mean F1 score of 0.31. However, the overall worst model (F1=0.0) is the SVM weekly model from sensor 18346015 trained with *All Sensor Data.* In Fairfax, the worst model (F1=0.08) is the Logistic Regression model from sensor 17178702 trained with *My Location Data.*

For all three Fairfax sensors, the best model is an SVM with a historical window size of one week or ensemble under the *My Sensor Data* conditions *e.g.*, F1=0.89 from SVM ensemble. These models are proficient at detecting Tc-99m and Cs-137 alerts with classification rates between 80% and 100%. Similarly to location-specific models, isotopes I-131 and 511 are more challenging to predict and have a

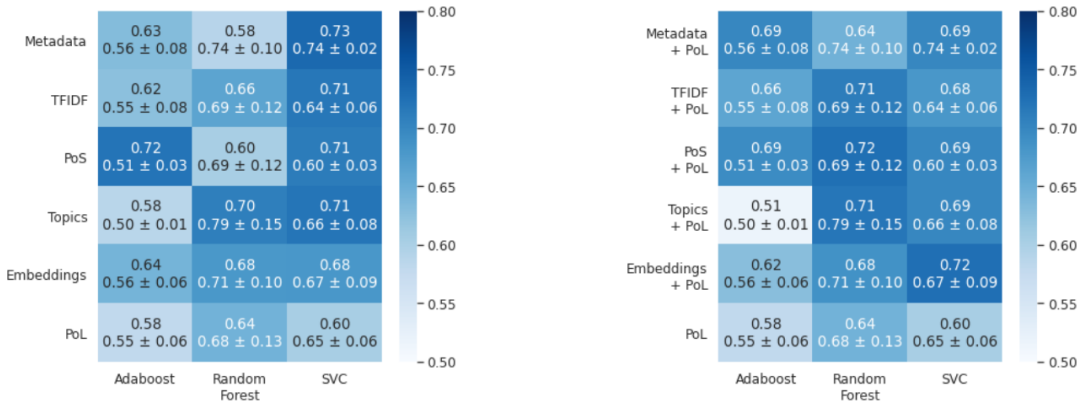



**Figure 7: Isotope-specific model accuracy for next-day industrial Cs-137 isotope prediction in Washington DC. We report results for three ML models - RF, SVM and AdaBoost and five types of representations that encode construction permits language – metadata, TFIDF, topics, embeddings and part-of-speech tags. ML models trained on construction permits data only are shown on the left. ML models that combine contextual data with the *PoL* data are shown on the right. Mean and std. dev. of ML model confidences are reported below accuracy value for each model-data representation combination.**

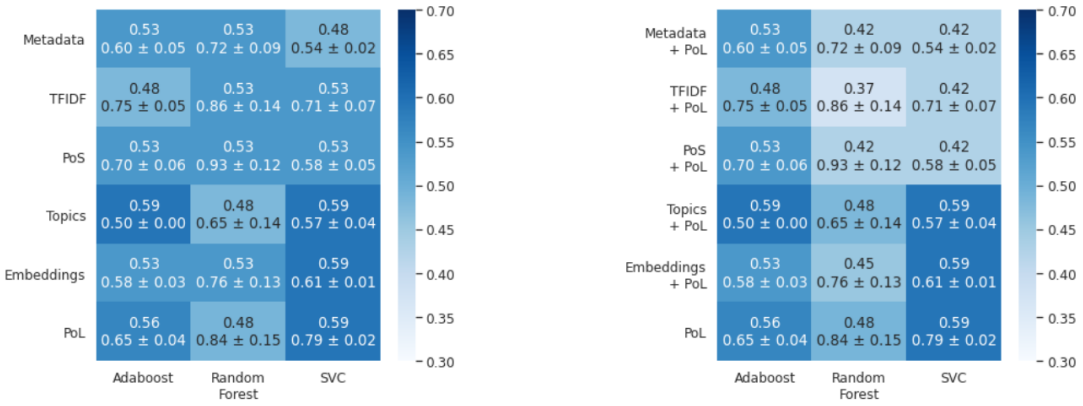



**Figure 8: Isotope-specific model accuracy for next-day Cs-137 isotope prediction in Fairfax, VA for different model-data representation combinations.**

wider range of classification rates; as low as 13% for I-131 and 45% for 511. Across the DC sensors, the best models are the Logistic Regression classifiers with a window size of one month or ensemble under the *My Sensor Data* conditions *e.g.,* F1=0.95 from the Logistic Regression monthly model. Likewise, Tc-99m and Cs-137 alerts are the easiest to classify. It should be noted, however, that only one DC sensor detects Cs-137. Furthermore, the classification rates for I-131 and 511 vary greatly from sensor to sensor *e.g.,* no errors (100% correct) to only errors (0% correct). To investigate *how predictive performance varies with the inclusion of additional training data*, we rank the three experimental setups by average F1 score as follows: *My Sensor Data* (mean F1=0.65); *All Sensor Data* (mean F1=0.33); *My Location Sensor Data* (mean F1=0.25).

We find this ranking interesting since our intuition would place *All Sensor Data* at the bottom of the list *i.e.,* training data from the most diverse sources. One explanation for the poor performance from the *My Location Sensor Data* models is that the majority of the test time period covers the beginning months of the COVID-19 pandemic where urban activity vastly changed and so might the patterns of life of isotope alerts.

We discover that models trained with exclusively Fairfax data outperform models trained on DC data in both the *My Sensor Data* (FF F1=0.67 > DC F1=0.63) and *My Location Sensor Data* (FF F1=0.27 > DC F1=0.22) cases. *In summary, we confirmed that sensor-specific models achieve higher F1 scores than models trained on additional sensor data (RQ3). These findings confirm that individual detectors have isotope-specific signatures that are not easily generalizable to other detectors.*

*Predictive performance variability by temporal granularity* Considering all sensors, isotopes, and training configurations, the daily models have the largest mean F1 (0.44 ± 0.16). The ensemble models are almost equivalent, but have a higher standard deviation (0.43 ± 0.27). And lastly, the weekly model has the lowest mean F1 and highest deviation (0.38 ± 0.28). When we only consider *My Sensor*

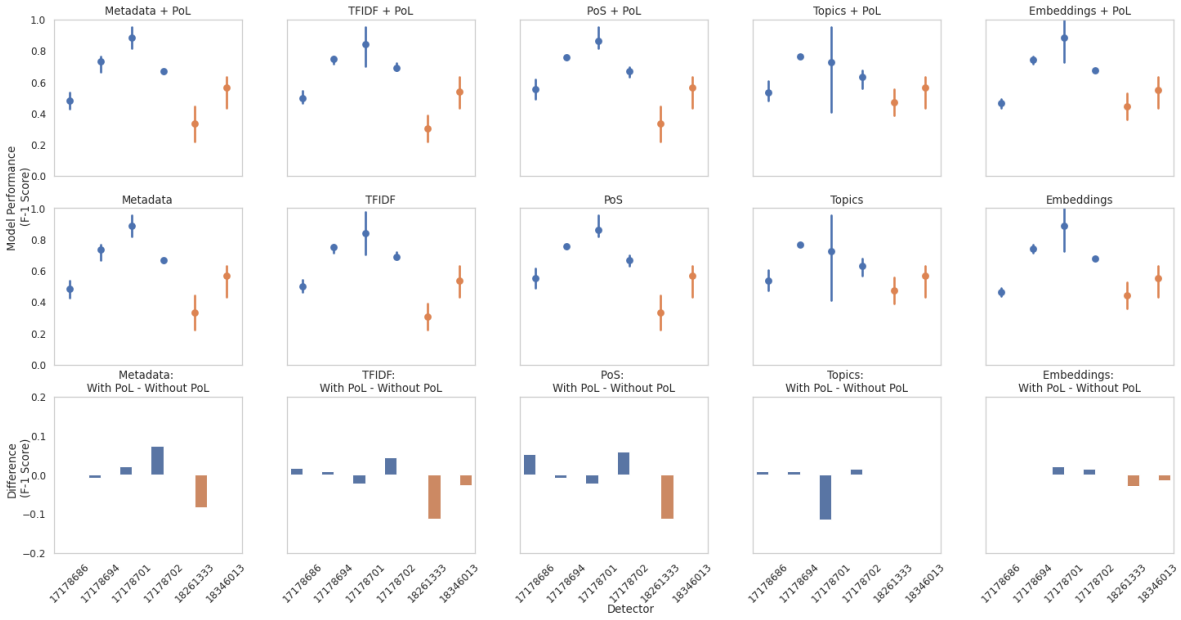



**Figure 9: Isotope-specific model performance differences trained with and without the *PoL* data. F1 scores are averaged across model types and reported for each individual sensor-text representation combinations in DC (orange) and Fairfax (blue).**

*Data* models, *i.e.,* the best performing overall sensors, the largest mean F1 (0.65 ± 0.17) is produced by the ensemble models. The monthly models have a mean F1 of 0.60 and a standard deviation of 0.18. This is consistent with our previous findings from the location-specific results (RQ1). One hypothesis for this agreement is that in both the *My Sensor Data* and original location-specific set ups the training distribution matches the test distribution more closely than in the *My Location Sensor Data* or *All Sensor Data*. For completeness, the time window with the highest mean F1 for *My Location Sensor Data* and *All Sensor Data* is one day with 0.35 and 0.37 mean F1.

## 4.3 Isotope-Specific Results

Figures 7 and 8 present model accuracy results when trained on text features extracted from construction permits with and without the inclusion of the historical *PoL* data. Across all feature combinations, model performances for DC are always better than for Fairfax. For example, model accuracy for DC are in the range between 0.53 and 0.7; whereas model accuracy for Fairfax range between 0.37 and 0.59. We find that all models (except RF and SVM when *PoL* data is combined with metadata, TFIDF and topic representations) outperform the baseline model that always predicts no alerts (the majority class) - DC accuracy=0.2 and Fairfax accuracy=0.44 (RQ4). For all text representations, SVM models outperform Random Forest and AdaBoost classifiers.

*The most predictive text representations* for DC models are metadata, TFIDF, topic and embedding vectors that yield the best results, 0.73 and 0.71 F1 scores, respectively. For Fairfax, the top performing data representations are topic and embedding vectors, that yield F1 score of 0.59. We observe that model confidences for Random Forest are significantly higher (about 20%) than SVM and AdaBoost, even though Random Forest is not the best performing model across both locations. Model confidences for Fairfax are generally higher than model confidences for Washington, DC.

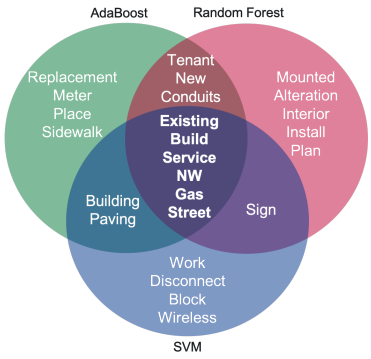



**Figure 10: Linguistic feature importances shown for unigrams shared across isotope-specific models trained on TFIDF representations.**

Most DC models trained exclusively on permits data perform as good as or better than models trained exclusively on historical data only except Random Forest with the PoS and metadata representations. Unlike DC, Fairfax models trained on the *PoL* data mostly outperform models that are trained on contextual data only. Figure 9 presents a more detailed analysis of model performance differences with and without the *PoL*. We find that combining open source and the *PoL* data is only beneficial for Fairfax but not for DC models. In DC, contextual data is more predictive than historical data with best F1=0.73 (SVM model) vs. F1=0.64 (RF model).

To interpret isotope-specific model predictions, we present text feature importances across ML models Figure 10. We observe that

multiple models rely on terms like 'existing', 'new', 'gas', 'street', 'build', etc. when predicting industrial Cs-137 alerts. In addition to feature importances, we perform a correlation analysis between the metadata features and Cs-137 signatures. We identify correlations between linguistic terms and alerts for company names: Wash Gas and Light, AT&T in DC; demolition electrical and residential terms, and people names in Fairfax; and construction terms and types of work in both locations.

## 5 CONCLUSIONS AND FUTURE WORK

This work presents a novel capability that could transform national security mission by advancing traditional approaches for real-time nuclear proliferation detection in urban environments. First, it will add a non-existing predictive component that would allow nuclear analysts to move from a reactive to a more proactive posture. Second, it will empower nuclear analysts with the ability to mitigate the operational burden posed by nuisance alarms during the deployment of unattended radiological sensors in urban settings.

Our novel ML-driven predictive analytics demonstrates the ability to anticipate radiological isotope signatures across multiple locations, sensors and isotopes by taking advantage of both historical and open-source data. Our quantitative results show that location specific models are as accurate as 80% with predictions across isotopes ranging between 60% and 80% with Tc-99m isotope being the easiest to anticipate followed by I-313 and 511 isotopes. Sensor-specific model performances vary across sensors with DC sensors being more difficult to predict than Fairfax sensors on average, with best Fairfax model F1=89% and best DC model F1=95%. Finally, in the absence of historical data, we demonstrated how open data sources can be leveraged to predict next-day industrial Cs-137 isotope signatures. The best predictive models rely solely on embedding and topic representations learned from construction permits yield accuracy of 71% for DC and 59% for Fairfax.

To extend our predictive analytics, we will move from classification to regression tasks to anticipate the number of alerts (in addition to predicting whether there will be an alert of a specific type or not) from each isotope at a specific hour (rather than a day) across locations up to 24 hours in advance. Given our success using construction permit data, we plan to investigate additional open data sources e.g., (a) traffic data between hospitals, construction sites and sensors, (b) precipitation, temperature and pressure sensor signals, (c) lidar data, and (d) video to incorporate them into ML models. Finally, we will build visual analytics to allow nuclear analysts interact with model prediction in real time across locations and sensors.

## REFERENCES


[1] Dimo Angelov. 2020. Top2Vec: Distributed Representations of Topics. *arXiv preprint arXiv:2008.09470* (2020).

[2] Mulugeta Weldezgina Asres, Grace Cummings, Pavel Parygin, Aleko Khukhunaishvili, Maria Toms, A. J. Campbell, Seth I. Cooper, David Ren-Hwa Yu, Jay Richard Dittmann, and Christian W. Omlin. 2021. Unsupervised Deep Variational Model for Multivariate Sensor Anomaly Detection. *2021 IEEE International Conference on Progress in Informatics and Computing (PIC)* (2021), 364–371.

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

[4] Sean M Brennan, Angela M Mielke, David C Torney, and Arthur B Maccabe. 2004. Radiation detection with distributed sensor networks. *Computer* 37, 8 (2004), 57–59.

[5] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175* (2018).

[6] DARPA. [n.d.]. DARPA's SIGMA Program Transitions to Protect Major U.S. Metropolitan Region. https://www.darpa.mil/news-events/2020-09-04. Accessed: 2021-05-23.

[7] Robert R Flanagan, Logan J Brandt, Andrew G Osborne, and Mark R Deinert. 2021. Detecting Nuclear Materials in Urban Environments Using Mobile Sensor Networks. *Sensors* 21, 6 (2021), 2196.

[8] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. AllenNLP: A Deep Semantic Natural Language Processing Platform. arXiv:arXiv:1803.07640

[9] Mario Gomez Fernandez, Akira Tokuhiro, Kent Welter, and Qiao Wu. 2017. Nuclear energy system's behavior and decision making using machine learning. *Nuclear Engineering and Design* 324 (2017), 27–34. https://doi.org/10.1016/j.nucengdes.2017.08.020

[10] John R Hershey and Peder A Olsen. 2007. Approximating the Kullback Leibler divergence between Gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, Vol. 4. IEEE, IV–317.

[11] Nathan Hoteling, Eric T. Moore, William P. Ford, Thomas McCullough, and Lance McLean. 2021. An analysis of gamma-ray data collected at traffic intersections in Northern Virginia. arXiv:2104.04137 [physics.soc-ph]

[12] Arterburn J., E. D. Dumbacher, and P. O. Stoutland. 2021. Preventing Nuclear Proliferation with Machine Learning Publicly Available Information. (2021).

[13] Mark Kamuda and Clair J. Sullivan. 2019. An automated isotope identification and quantification algorithm for isotope mixtures in low-resolution gamma-ray spectra. *Radiation Physics and Chemistry* 155 (2019), 281–286. https://doi.org/10.1016/j.radphyschem.2018.06.017 IRRMA-10.

[14] Hanjoo Kim, Dongmin Yun, H Shin, S Moon, and Deokjung Lee. 2020. Feasibility study on machine learning algorithm in nuclear reactor core diagnosis. In *Proceedings of the Transactions of the Korean Nuclear Society Virtual Spring Meeting, Korea (online)*. 9–10.

[15] Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (1999), 788–791.

[16] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. 2003. The global k-means clustering algorithm. *Pattern recognition* 36, 2 (2003), 451–461.

[17] Liansheng Liu, Qing Guo, Datong Liu, and Yu Peng. 2019. Data-Driven Remaining Useful Life Prediction Considering Sensor Anomaly Detection and Data Recovery. *IEEE Access* 7 (2019), 58336–58345.

[18] Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam M. Shroff. 2016. LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection. *ArXiv* abs/1607.00148 (2016).

[19] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[20] Engineering National Academies of Sciences, Medicine, et al. 2021. Nuclear Proliferation and Arms Control Monitoring, Detection, and Verification: A National Security Priority: Interim Report. (2021).

[21] Shiven Sharma, Colin Bellinger, Nathalie Japkowicz, Rodney Berg, and Kurt Ungar. 2012. Anomaly detection in gamma ray spectra: A machine learning perspective. In *2012 IEEE Symposium on Computational Intelligence for Security and Defence Applications*. 1–8. https://doi.org/10.1109/CISDA.2012.6291535

[22] Hui Yie Teh, I Kevin, Kai Wang, and Andreas W Kempa-Liehr. 2021. Expect the Unexpected: Unsupervised Feature Selection for Automated Sensor Anomaly Detection. *IEEE Sensors Journal* 21, 16 (2021), 18033–18046.

[23] Matthias Tuma, Valdemar Rørbech, Mark K. Prior, and Christian Igel. 2016. Integrated Optimization of Long-Range Underwater Signal Detection, Feature Extraction, and Classification for Nuclear Treaty Monitoring. *IEEE Transactions on Geoscience and Remote Sensing* 54, 6 (2016), 3649–3659. https://doi.org/10.1109/TGRS.2016.2522972

[24] Svitlana Volkova, Ellyn Ayton, Sannisth Soni, , Mark Bandstra, Brian Quiter, Nico Abgrall, and Ren Cooper. 2021. AI-driven Analytics for Radiological Source Detection and Localization. (2021).

[25] Svitlana Volkova, Sannisth Soni, Ellyn Ayton, Ren Cooper, Mark Bandstra, and Brian Quiter. 2021. Open-Source Data Analytics Value Quantification to Inform and Explain Radiological Source Detection, Localization, and Tracking. (2021).

[26] Eiji Yoshida, Kiyoshi Shizuma, Satoru Endo, and Takamitsu Oka. 2002. Application of neural networks for the analysis of gamma-ray spectra measured with a Ge spectrometer. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 484, 1 (2002), 557–563. https://doi.org/10.1016/S0168-9002(01)01962-3