Proceedings of the INMM & ESARDA Joint Virtual Annual Meeting
August 23-26 & August 30-September 1, 2021

FEATURE ENGINEERING: A CASE STUDY FOR RADIATION SOURCE LOCALIZATION IN COMPLICATED ENVIRONMENTS

Matthew Durbin
Department of Nuclear Engineering
Pennsylvania State University

Ryan Sheatsley
Department of Computer Science
Pennsylvania State University

Patrick McDaniel
Department of Computer Science
Pennsylvania State University

Azaree Lintereur
Department of Nuclear Engineering
Pennsylvania State University

ABSTRACT

Machine learning is a powerful data analysis technique; however, many facets must be optimized to reap the greatest benefits. Model selection and hyper-parameter tuning are important areas for optimization, but for certain scenarios, domain-aware feature engineering may lead to the greatest increase in model utility. Feature engineering can greatly contribute to increased model performance, and can also provide insight into the machine learning process, thus enhancing model explainability. To illustrate this, feature engineering was explored for the challenge of source localization in complicated and obstructed environments. Datasets were simulated with various gamma ray sources located up to five meters away from a four detector NaI array for three cases: no background or obstructions, with background, and with obstructions. A random forest model was used and tasked with predicting the angle at which the source was located utilizing only a static measurement from an array of NaI detectors. The simplest choice of input features for this scenario was the total counts received in each detector. Additional features were also explored, including counts of photopeak and Compton continuum regions, and simple spectral binning schemes. Results show that domain aware feature engineering can improve model performance, even when the detection data includes complications from background contributions or obstructions. Across scenarios, accuracy was improved by close to a factor of two, and the mean absolute error of the angular predictions was improved by several degrees.

INTRODUCTION

Machine learning (ML) has shown great promise as an analysis method in a variety of radiation detection applications including pulse shape discrimination [1, 2], isotope identification [3, 4], and radiation source localization [5, 6]. Many software packages and libraries allow for simple implementation of standard ML algorithms [7, 8] which can be optimized in a number of ways, including hyper-parameter tuning [9]. In certain cases, however, domain-aware feature engineering can complement the data-driven approach of ML and lead to greater increases in model performance and utility.

Broadly speaking, feature engineering focuses on extracting physically motivated features from the raw data and presenting those features to the ML algorithm in an optimized manner.

One benefit of ML is that it is often times no harder to implement an algorithm that analyzes high dimensional (or many feature) data than lower dimensional data. While conventional or analytical approaches can also analyze higher dimensional data, it may not be intuitive how each dimension should be incorporated or weighted. Through feature engineering, the user can, in essence, transfer the domain knowledge to the ML algorithm, which will incorporate and optimally weight the dimensions. This process can allow for more effective representation of the data and increase ML performance. While feature engineering leverages the physical domain, it is also possible that further insight can be gained by analyzing the usefulness of individual features to the ML model's decision making process. This in turn can lead to increased performance, as well as enhanced model explanability.

This work investigated the benefits of feature engineering data from a static source localization application with an array of NaI detectors. The physical basis of the angular prediction capability from a static acquisition using the detector array arises from two simultaneous phenomenon: the differences in the solid angle subtended by each detector from the source, and differing amounts of partial attenuation or array self-occlusion experienced by each detector. For a fixed distance, and with sufficiently high statistics, each angular position will have a unique combination of normalized detector responses because of the solid angle and self-occlusion effects. Similar directional detection problems have been discussed in the literature making use of a least-squares based comparison to a prepopulated reference table to make the angular predictions [10, 11]. Here, a random forest (RF) algorithm is used. Utilizing Monte Carlo n-Particle (MCNP6) [12] simulations to generate training and testing data, feature engineering is explored for this application in a simple case (no background, no obstructions), a case were background is present, and a case where obstructions are introduced.

METHODS

Array

The detector array used in this work consisted of four 5.08 cm by 10.16 cm by 40.64 cm (2"x4"x16") NaI detectors arranged in a 30 cm square with respect to their inner faces, with the long axis orthogonal to the source plane. A gamma ray emitting point source was placed up to five meters away from the detector array center, and a single, static acquisition was taken. Based on the subtle differences between the counts received in each detector, the angular component of the source location was predicted. Figure 1 illustrates the simulated responses of the detectors in the array for a source at a fixed distance as a function of angle.

Algorithm

The ML algorithm used in this work, Random Forest (RF), consists of an ensemble of individual decision trees. Each decision tree acts as a system of optimized if-then type pathways that aim to sufficiently separate the data based on the input features. RFs are computationally efficient, easy to implement, and possess the additional benefit of having a convenient way of quantifying feature importance, called the mean decrease in impurity. This approach essentially notes the effectiveness of a certain feature in acting as a decision node compared to the other features.

Input Features

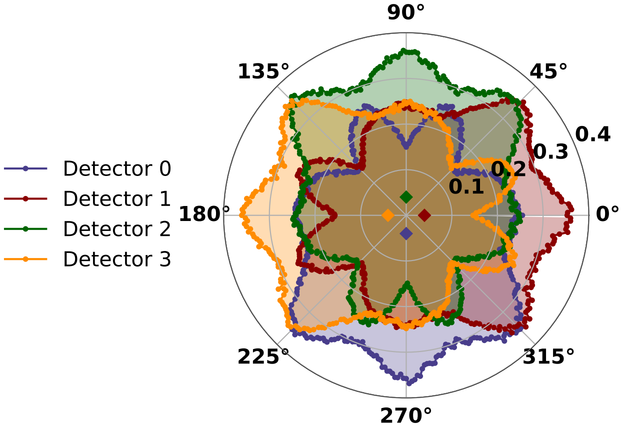

Figure 1. Simulated angular array response to a ^{60}Co source 3 m from array center. Counts are normalized to unity, and diamonds represent detector position.

Most of the input features used came from the spectra each detector produced during an acquisition. Some features made use of additional contextual information available in certain applications, such as those based on isotope identification that could be performed on the collected spectra. The simplest input feature scheme consisted of the total counts from each detector, which were normalized to unity as shown in Figure 1. For each detector, if isotopic information was made available, counts could be split into photopeak (P) and Compton continuum regions (C) to be used as seperate features. The motivation for this selection was that these two feature types are associated with the two facets of the modality that provide directional information — solid angle and self-occlusion. Counts under the photopeak region are nominally unattenuated, and are related to the differences in solid angle of the detectors. By contrast, many counts under the Compton region have been attenuated by the array itself, and thus loosely correspond to the effects of self-occlusion. In reality, both solid angle and self-occlusion contribute to changes in photopeak and Compton counts, however the magnitude of the contributions were different. It was hypothesized that separating the total counts into distinct features corresponding to different phenomenon in the physical domain will provide more relevant degrees of freedom for the model to *learn*.

Extending this concept, energy bins can also be used as input features given the spectroscopic qualities of NaI, essentially acting as down sampled versions of a raw detector spectra. This could allow the algorithms to further leverage the energy dependent phenomena of the scenario, without using the entire spectra, which may be computationally inefficient or filled with low signal-to-noise-ratio bin. A feature that captured the isotope (I) was also used, numerically representing either ^{60}Co, ^{137}Cs, or ^{192}Ir. Additionally, input features were created that put the detectors in order based on the counts received (D). While this detector order feature could be redundant, it may provide utility in that it effectively narrows down the quadrant the source was likely to be located in.

Stationary detector arrays could be deployed with additional contextual sensors, such as LIDAR or video cameras, which could allow the system to know if obstructions were present within the range of interest. To emulate this, a binary input feature was also created to capture whether or not obstructions were present (O). The features described in this section were used in various combinations for the different experiments. While other scenarios may benefit from additional features, the features used here were highlighted to demonstrate the utility of feature engineering, not to make a case for any one specific feature.

Experiments

To investigate the effects of implementing engineered features, datasets of individual ^{60}Co, ^{137}Cs, and ^{192}Irmeasurements were simulated using MCNP6. The detector array, supporting electronics and a concrete floor were modeled. For each trial a point source was simulated at a random angular position a distance of 0.5-5 m from the array center on a two dimensional plane. Ten million source particles, emitted isotropically, were simulated for all scenarios. An F8 pulse height tally was captured for each trial, and Gaussian energy broadening was applied based on the measured energy resolution of an analogous laboratory system[6].

Three scenarios were considered in this work. The first, the simple case, did not include a background source or obstructions. This was intended to isolate the effects of feature engineering and capture the full potential in an idealized scenario. For this experiment, a 30,000 trial data set was used with 10,000 trials each of ^{60}Co, ^{137}Cs, and ^{192}Ir. The performance of the RF was quantified using the mean absolute error in angular predictions (MAE), as well as the accuracy (ACC), which was defined as the percentage of trials that were correct within one degree. To study the effects of feature engineering, three sets of input features were used. The first set consisted of the simple input feature scheme (totals), the second consisted of P, C, I, and D, and the third consisted of the energy bins.

The second experiment investigated the effects of background in the scenario that the background was known or constant, only fluctuating according to nominal Poisson statistics. For these trials, a two minute background spectra acquired in the laboratory was injected into each MCNP6 F8 tally and Poisson randomly sampled. This roughly emulated the typical Poisson variation of a constant background. The MCNP tally was correlated to represent a two minute count time of a 50 μCi ^{60}Co source. This process was done for ^{60}Co, ^{137}Cs, and ^{192}Ir. An example spectra for ^{60}Co is shown in Figure 2.

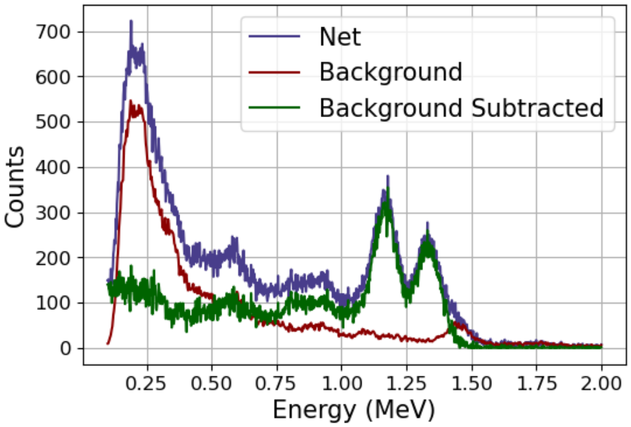

Figure 2. Simulated ^{60}Co spectra used in the background experiment.

Background was addressed in two ways. The first was simple subtraction, shown in green in Figure 2, in which a background spectra was subtracted channel wise from each detector. Then, the same procedure was used as in the simple case. The second way of handling background was to keep the net counts for each spectra, and create additional input features based on the background spectra, analogous to the existing input features. Specifically the total counts, P, C, and energy bins.

The last experiment investigated the effects of obstructions. The same procedure was used for the simple case, with the exception that half of the simulated trials included various concrete obstructions modeled after cinder blocks. It is worth noting that this does not mean half the trials were obstructed, just that obstructions were present for half the trials. These obstructions were located in the 45°-135° quadrant relative to Figure 1. A simple schematic of the obstructions is shown in Figure 3. In addition to the input features of the simple case, the binary input feature that captured whether or not a source was present was used emulating the discussed scenario in which additional contextual awareness would be available. Obstructions with random locations are reserved for future work.

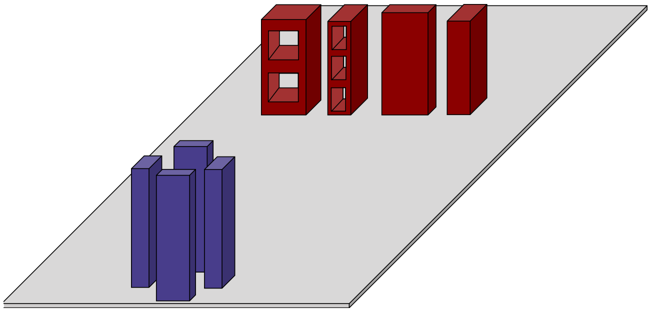

Figure 3. Schematic of the obstructed scenario. NaI crystals blue, concrete obstructions shown in red.

While some decline in performance was expected for the second two experiments due to the increased complexity, the experiments were designed to showcase the utility of feature engineering, even in complex environments.

RESULTS

Before the individual experiments were analyzed, the optimal number of equally spaced energy bins to be used as input features had to be determined. To do this, the RF was tested and trained using a range of energy bins from 3-21. The results of this study are shown in Figure 4.

The ACC and MAE both plateaued above 15 bins. Thus, to optimize these parameters without increasing overall algorithm complexity, 15 bins were used for the remainder of this work. While more energy bins would lead to more information, the signal-to-noise per bin would also decrease. Thus, the true optimal number of bins would depend on the specific data and algorithmic approaches, the work presented here serves as a proof of concept for this method.

Simple Case Expirement

Shown in Figure 5 is the true angle plotted against the predicted value for the three sets of input features on the simple case test data. While the entire range of angles were tested, for visual clarity only one quartile is shown as it is representative of the entire range due to the array symmetry.

The dominant element in Figure 5 is the diagonal, corresponding to predictions close to the true value. The cross-like artifact centered at 90° was a result of the symmetry about each detector (located

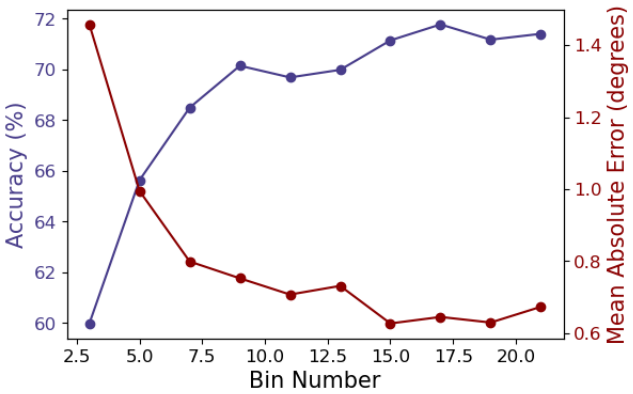

Figure 4. RF performance as a function of the number of equally spaced energy bins used as input features.

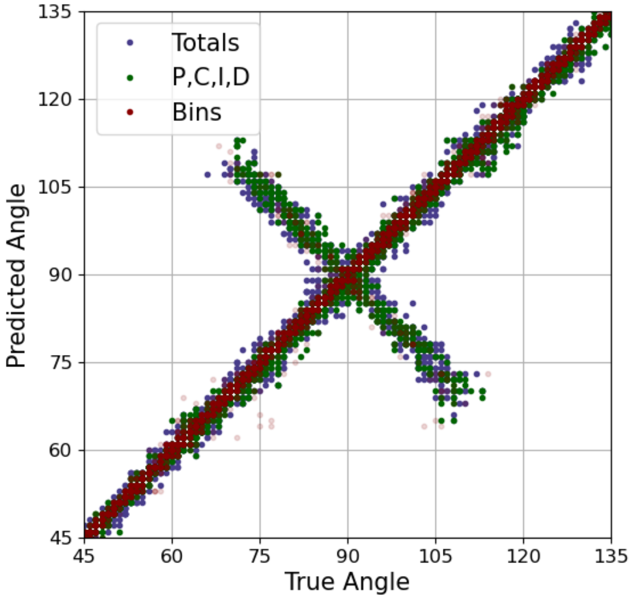

Figure 5. True versus predicted angular values for the various input feature selections with the simple case experiment. Here, P, C, I, and D correspond to photopeak, Compton continuum, isotope, and detector order.

at 0°, 90°, 190°, and 270°). On close inspection of Figure 1, the symmetry becomes clear. Immediately surrounding each detector (±15°), there was a reflective symmetry that causes certain *reflected* points to have similar responses. The benefit of feature engineering can be seen in the tighter distribution for the P,C,I,D features about the diagonal compared to the totals distribution, and the even tighter distortion for the bins. Additionally, the effect of the symmetry related cross feature was notably less severe for the energy bins distribution. These benefits are quantified in Table 1.

Figure 6 shows the feature importance for the P, C, I, and D features. Features extracted from the photopeak and Compton region were all considered important, which implies they contributed to the decision making process of the RF. The D feature has near zero importance, revealing that it did

Table 1. Results for the Simple Case

Input Features	MAE	ACC
Totals	3.05°	33.4%
P,C,I,D	1.78°	40.7%
Bins	0.80°	62.7%

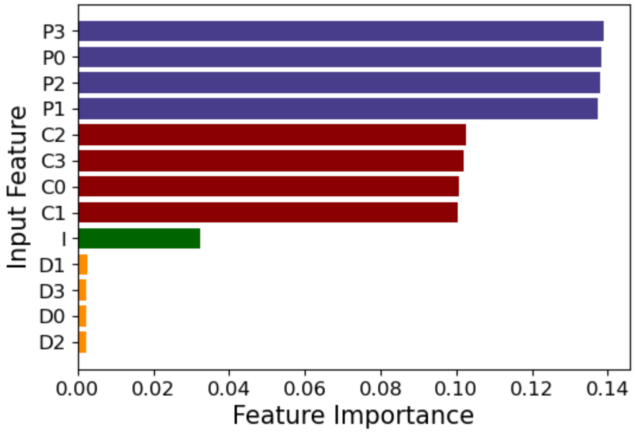

Figure 6. Feature importance for P, C, I, and D features used in the simple case experiment.

not contain much (or any) information helpful for distinguishing different angles. The 1.2° reduction in MAE and 8% increase in ACC showed the clear benefit of feature engineering, and the feature importance analysis showed that some features were more useful than others. It was likely the case that the information provided by the I and D features was largely captured within the other features. It was hypothesised that as P and C features point to physically different processes, there will be some information gain not captured by total features alone. The feature analysis confirmed this, and also points to those features alone being able to capture isotopic or energy information, despite not providing any explicit energy information. Using the energy binning scheme led to an additional 1° decrease in MAE and 18% increase in ACC, further showing how domain-informed choices of features can greatly improve ML performance.

Background

Results for the background experiment are given in Table 2, where BGS indicates the data that was background subtracted and BG + indicates the data in which the background was input via various features in lieu of subtraction. The distribution of true versus predicted angle are not included due to their similarity to Figure 5. While a slight dip in performance was seen when compared to the simple case, the benefits of feature engineering are still clear. While there was no notable difference in performance between the two approaches (subtracting the background verses using the background as features), in situations where the background in unknown or highly variable, incorporation of background based input features could be of use. The feature importance analysis for these cases looked similar to that of Figure 6, with low importance given to the background features.

Obstructions

Table 2. Results for the Background Experiment

Input Features	MAE	ACC
BGS: Totals	3.14°	33.0%
BGS: P,C,I,D	1.84°	40.2%
BGS: Bins	0.94°	59.4%
BG + Totals	3.19°	31.3%
BG + P,C,I,D	1.73°	40.0%
BG + Bins	1.65°	52.9%

The results for the obstruction experiment are given in Table 3, and the predictions are plotted in Figure 7.

Table 3. Results for the Obstruction Experiment

Input Features	MAE	ACC
Totals	3.46°	37.1%
P,C,I,D	1.51°	45.5%
Bins	0.46°	91.61%

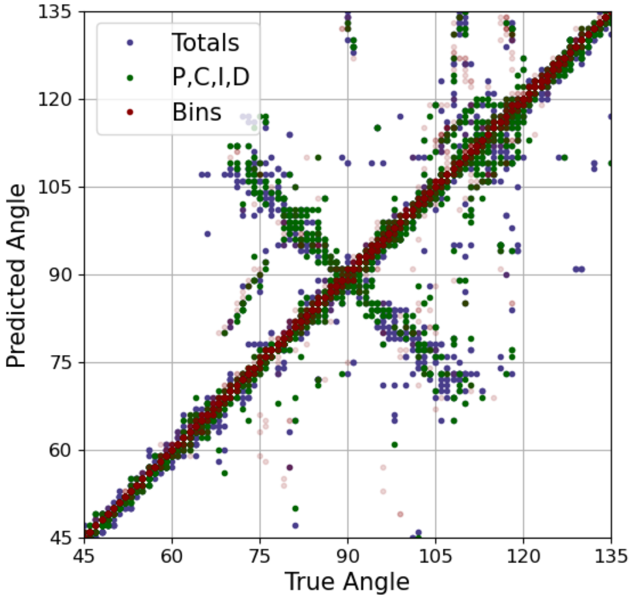

Figure 7. True versus predicted angular values for the various input feature selections with the obstruction experiment.

The inclusion of obstructions had two very different effects on the data arising from a combination of distance and energy related effects. In certain cases, the obstructions lead to low or almost zero counts in some detectors. This drastically distorts the angular response, especially in the cases where different detectors experience notably different attenuation effects. Sometimes, these distortions can lead to the response at one angle looking like the response at a different angle. In these cases, the RF was effectively confused, and may not have provided a good prediction. On the other hand, as the specific obstructions were included in the training data (emulating training on a known environment), these distortions can act as a special signature, which the RF will flag as a certain angle. In these cases, the RF can actually do extremely well, as evident by Table 3.

Here again, the benefit of feature engineering is clear. By using the same data, just presented to the model in a different way, a 3° improvement in the MAE was seen. The feature importance analysis looked similar to Figure 6, with low importance given to obstruction indicator feature.

CONCLUSION

Feature engineering was explored for the application of static gamma ray source localization using an array of NaI detectors. This work demonstrated the utility of domain driven feature selection. A random forest algorithm was used with the task of predicting the angular component of the source's location. Simulated datasets were created to emulate complicated environments, including Poisson variable background, and concrete obstructions. The raw data from these simulations consisted of gamma ray counts for each of the detectors. The simplest, non-engineered, choice of input features was taken as the normalized detector totals. Additional input features investigated included the photopeak area, the Compton continuum, the isotope class, and options to capture the background and obstructions. Using these features improved accuracy by over 7% across datasets, while reducing the MAE in angular predictions by close to 1.3°. Through an analysis of feature importance, it was seen that photopeak and Compton continuum based features were able to sufficiently capture isotopic or energy information relevant to the problem. Using an energy binning scheme as the input features led to further improvements, increasing accuracy by over 20% and reducing the(MAE) by over 2° compared to using the totals. It is important to note that these improvements were all derived from the same data processed by the same exact algorithm, the only difference being how the data was feature engineered. With non-engineered features, complications such as background and obstructions can lead to a decrease in algorithm performance. With engineered features, this decrease can be partially mitigated. As an example, using only the totals as input features led to an MAE of 3.04° with no obstructions and increased to 3.46° with obstructions. Using energy bins as input features actually led to an increase in performance under the same conditions, with an MAE of 0.80° without obstructions and 0.46° with obstructions. While other scenarios may benefit from different specific input features, it is hypothesised that all scenarios would benefit from feature engineering. The experiments shown in this work highlight not only the utility of machine learning for radiation detection applications, but also illustrate that domain aware feature engineering can increase performance and robustness while using the same data and algorithms.

Future work will include investigating scenarios in which the background is variable or unknown, and in which obstruction locations are not well defined. Additionally, steps will be taken to quantify the performance of input features such as those presented here on real word detection data.

REFERENCES

[1] C. Fu, A. Di Fulvio *et al.*, "Artificial neural network algorithms for pulse shape discrimination and recovery of piled-up pulses in organic scintillators," *Annals of Nuclear Energy*, vol. 120, pp. 410–421, 2018.

[2] M. Durbin, M. Wonders *et al.*, "K-Nearest Neighbors regression for the discrimination of gamma rays and neutrons in organic scintillators," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, p. 164826, 11 2020.

[3] P. E. Keller, L. J. Kangas *et al.*, "Nuclear Spectral Analysis via Artificial Neural Networks for Waste Handling," *IEEE Transactions on Nuclear Science*, vol. 42, no. 4, pp. 709–715, 1995.

[4] M. Kamuda, J. Zhao, and K. Huff, "A comparison of machine learning methods for automated gamma-ray spectroscopy," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 954, no. October, 2020. [Online]. Available: https://doi.org/10.1016/j.nima.2018.10.063

[5] A. D. Nicholson, D. E. Peplow *et al.*, "Detection Algorithm Competition," vol. 67, no. 8, pp. 1968–1975, 2020.

[6] R. Sheatsley, M. Durbin *et al.*, "Improving Radioactive Material Localization by Leveraging Cyber-Security Model Optimizations," *IEEE Sensors Journal*, vol. 21, no. 8, pp. 9994–10 006, 2021.

[7] F. Pedregosa, V. Gael *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 1, pp. 2825–2830, 2011.

[8] M. Abadi, A. Agarwal *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," 2016. [Online]. Available: http://arxiv.org/abs/1603.04467

[9] A. Geron, *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*. Sebastopol, CA: O'Reilly Media, Inc., 2017.

[10] D. Hanna, L. Sagnières *et al.*, "A directional gamma-ray detector based on scintillator plates," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 797, pp. 13–18, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.nima.2015.06.019

[11] C. Schrage, N. Schemm *et al.*, "A low-power directional gamma-ray sensor system for long-term radiation monitoring," *IEEE Sensors Journal*, vol. 13, no. 7, pp. 2610–2618, 2013.

[12] T. Goorley, M. James *et al.*, "Initial MCNP6 release overview," *Nuclear Technology*, vol. 180, no. 3, pp. 298–315, 2012.