Algorithmically Secure Classification of Weapons-Grade Nuclear Material for Treaty Verification

Heidi Komkov[1], Adam Hecht[2], Ryan Kamm[1], Eduardo Padilla[1,2], Christopher Siefert[1] and Kyle Weinfurther[1]

[1]Sandia National Laboratories, Albuquerque, NM USA

[2]The University of New Mexico, Albuquerque, NM USA

Submitted to the INMM & ESARDA Joint Annual Meeting, May 2023, Vienna, Austria

***Abstract*— If nuclear-weapon states agree to continue nuclear arms reduction treaties, there will be a pressing need for high confidence nuclear warhead verification techniques. Such a verification method must simultaneously satisfy two competing objectives: each party requires their weapons design information to be protected from disclosure, but also needs to allow the collection of sufficient information to reliably confirm treaty accountable items in each other's inventory. We propose an inherently information-limited neural algorithm to verify whether the gamma-ray emissions from a nuclear weapon can be associated with a class of treaty accountable items. The algorithm, which we call the *buffered classifier* model, never stores the full gamma-ray spectrum, which can relay sensitive nuclear weapon information. Instead, it processes gamma-ray energy information pulse by pulse, storing a reduced, irreversible representation of the gamma-ray spectrum. This reduced representation is processed with a network that classifies the measurement as being from a warhead-like object or not. Both the classifier accuracy and the reconstruction error of the reduced representation are simultaneously optimized through gradient descent, training on a non-sensitive dataset encompassing many configurations of radioisotopes and shielding. The buffered classifier has the potential to serve in future arms control treaties as a transparent yet secure and trustworthy nuclear warhead verification method.**

Keywords—treaty verification, machine learning, gamma-ray spectrum, information barrier

I. INTRODUCTION

Despite a principal stated objective of the Treaty on the Non-Proliferation of Nuclear Weapons (NPT) to achieve complete disarmament, it is estimated that many global nuclear powers are currently increasing their inventories of nuclear warheads [1]. The US-Russian New START treaty, which limits warheads and delivery vehicles, is in effect only until 2026, and currently on-site inspections have been suspended [2]. Treaties often take years to negotiate, and a successor to New START may involve additional nuclear-weapons states, which could add time and complexity to the negotiations.

The challenge in warhead verification lies in the fact that weapons design information is considered highly sensitive, with nuclear-weapons states making great efforts to safeguard this information from general dissemination, adversaries, and even allies in most cases. Therefore, verification must be done without revealing design information while providing a high measure of confidence to treaty partners that the object under inspection is in fact the item it is purported to be. Current and historical proposed methods for warhead verification include attribute and template-based approaches, usually involving some form of an information barrier (IB) [3]. Information barriers are designed such that sensitive information is processed and/or modified in such a way that allows non-sensitive information to be relayed to an inspector without the risk of revealing

sensitive host information such as nuclear warhead design. An example of an electronic IB is one that processes gamma spectra using a dedicated algorithm, but only relays a confirmation message (red light/green light) to the inspector. Another form of IB are methods of physical encryption that involve slightly altering the intrinsic or induced radiation signatures to obfuscate sensitive details about the treaty-accountable item, while still allowing for high confidence confirmation of a template match or warhead attribute [4-7].

Arms control verification in general spans several regimes, including, but not limited to: absence verification, presence confirmation, warhead counting and dismantlement verification. Each of these regimes requires tailored methods and all are likely to be implemented in gradually more intrusive order. Certainly, all methods require high levels of trust between parties, encounter chain of custody issues, and may leave the opportunity for information leakage through backdoor exploits. Thus, methods that *intrinsically minimize information collected and stored result in increased trust and higher probability of adoption.*

Our goal in this paper is not to address all the practical issues within a specific comprehensive treaty verification regime. Rather, we develop a minimally invasive algorithm—or an algorithmic information barrier—that has the possibility to be applied as part of a larger, potentially multi-approach, nuclear arms control verification system. The principal technical objectives for the algorithm are to demonstrate irreversibility of the reduced representation and to find the parameters for which classification accuracy is maximized.

II. Methods

A. Dataset Generation

Public, collaborative nuclear arms control verification research requires the use of non-sensitive datasets that allow for researchers and academics from around the globe, including from non-nuclear-weapons states, to engage. Thus, the dataset involved in this study consists of many arbitrary configurations of nuclear materials and shielding generated according to a published algorithm and does not involve any expert knowledge about weapons design.

Gamma-rays are characteristic fingerprints of nuclear isotopes, arising from the relaxation of a nucleus from an excited state after a radioactive decay. Determining the radioisotopes that created a gamma-ray spectrum is generally done by a human analyst using template-matching and/or linear regression techniques, and is complicated by the presence of shielding, variations in radiation background and scattering environments, and other details of a particular measurement configuration. Furthermore, the detector type influences the resolution (peak shape) of the spectrum, with the highest-quality detectors clearly resolving photopeaks, and other detectors broadening them. Example gamma-ray spectra from our dataset are shown in Figure 1. For further information about gamma-ray energy spectra, nuclear processes, and detection methods, we refer the reader to Knoll [4].

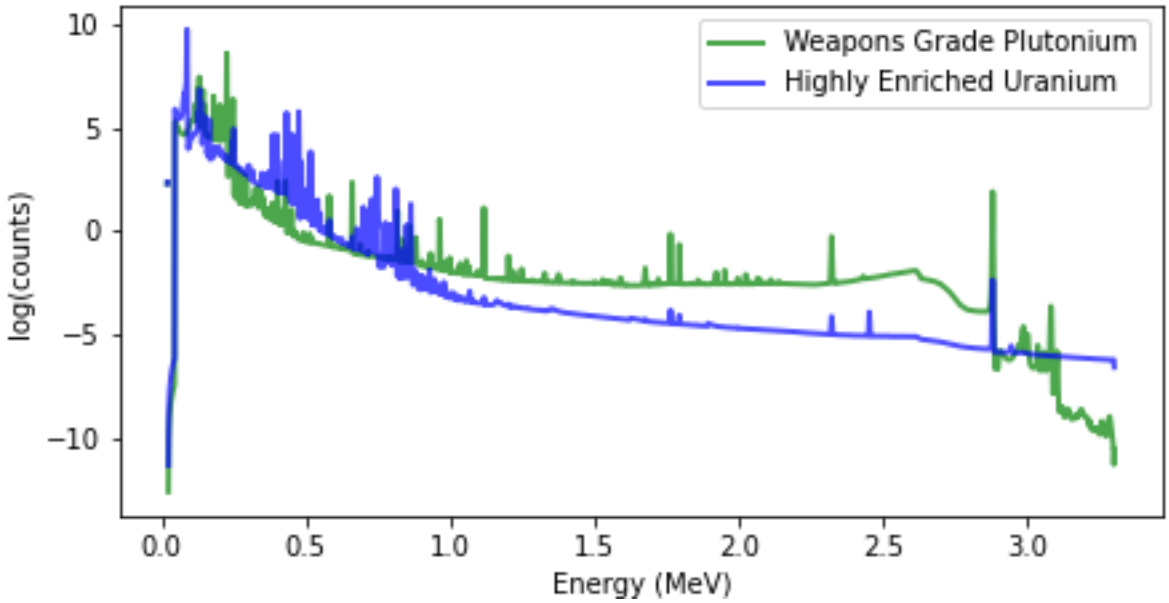

Figure 1: Example gamma-ray energy spectra from our dataset.

To generate simulated gamma-ray spectra for our training and validation dataset, we use the Gamma Detector Response and Analysis Software (GADRAS) package from Sandia National Laboratories [5]. GADRAS performs gamma-ray detector response modeling, computes radiation transport through shielding, and is a state-of-the-art toolset used by analysts to perform isotope identification from spectra. Using the workflow shown in Figure 2, we used GADRAS to generate data in two classes: weapons-grade nuclear sources, and non-weapons-grade nuclear sources.

B. Weapons-grade nuclear material radiation sources

Nelson and Sokkappa [6] outline a non-sensitive method for generating sources containing various forms of nuclear material. Avoiding the use of nuclear weapon device designs, their model is based upon physics principles and generalized bounding cases. Using GADRAS, we simulated many 1-dimensional spherical models consisting of fissile material (one or two layers) surrounded by various layers of shielding. The materials and thicknesses of the layers are selected from probability distributions that span the range of possibilities outlined by Nelson and Sokkappa. For this study, we have defined weapons-grade plutonium as 94% Pu-239 and weapons-grade uranium as 90% or greater U-235 (weight percentages).

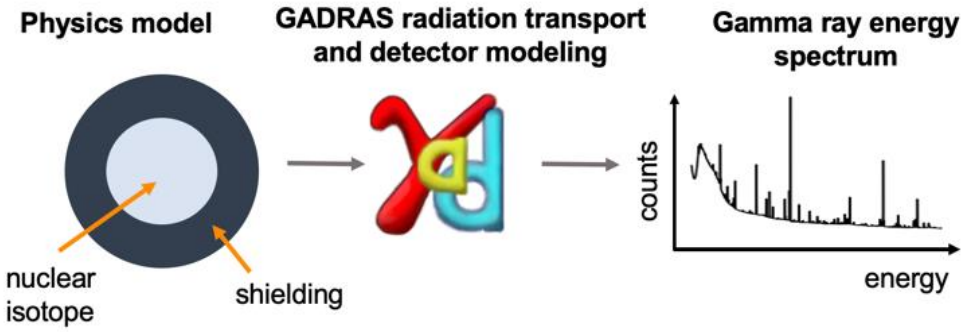

Figure 2: The Nelson and Sokkappa algorithm is used to generate a large range of plausible combinations of nuclear material and radioisotopes surrounded by shielding. GADRAS is used to simulate the gamma-ray emissions from the physical model and the resulting detector output.

C. *Non-weapons-grade radiation sources*

The GADRAS built-in radioisotope library contains 184 natural, medical, and industrial radionuclides, which we placed into the non-weapons-grade class. These radionuclides were randomly selected and grouped up to three at a time, in varying activities from 10μCi to 1mCi. Modeled as point sources, these radioisotope mixtures were then placed inside layers of shielding with random material composition and thickness as prescribed by the Nelson and Sokkappa algorithm.

Furthermore, as defined by the algorithm, uranium and plutonium below the thresholds of 94% Pu-239 and 90% U-235 (weight percentages) are defined in this study as non-weapons-grade and assigned to this class as well. Future studies can be conducted to assess the sensitivity and specificity of this approach as a function of enrichment levels.

D. *Radiation detector*

For this study, the standard detector response function for an ORTEC Detective EX-100 high purity germanium (HPGe) detector was used, with all spectra including default Albuquerque, NM natural background radiation. All spectra generated for model training were ideal, without Poisson noise; the impacts of varying background and counting statistics were not considered as part of this initial study. All simulations were screened to ensure that count rates at the detector are statistically significant, exceeding three standard deviations above background, and detector dead times were less than 25% to avoid pulse pileup effects.

III. Concept of Operation

In a treaty verification scenario, a spectroscopic gamma detector system would run exclusively in pulse-by-pulse mode, also called list-mode operation, in which data streams in as a list of individual pulses without being binned into an energy spectrum. To normalize for relative spectral importance independent of source strength, a pre-defined number of gamma-ray photons would be measured sequentially. Administrative controls for minimum and maximum count rates would be necessary to guard against highly shielded sources or detector saturation, respectively. To avoid storing a sensitive gamma-ray energy spectrum of the nuclear weapon, data would be immediately converted to a reduced, irreversible form, as depicted in Figure 3 and described in the next section. The only output visible to the inspector would be the final output class determination by the neural algorithm.

IV. Neural Algorithm

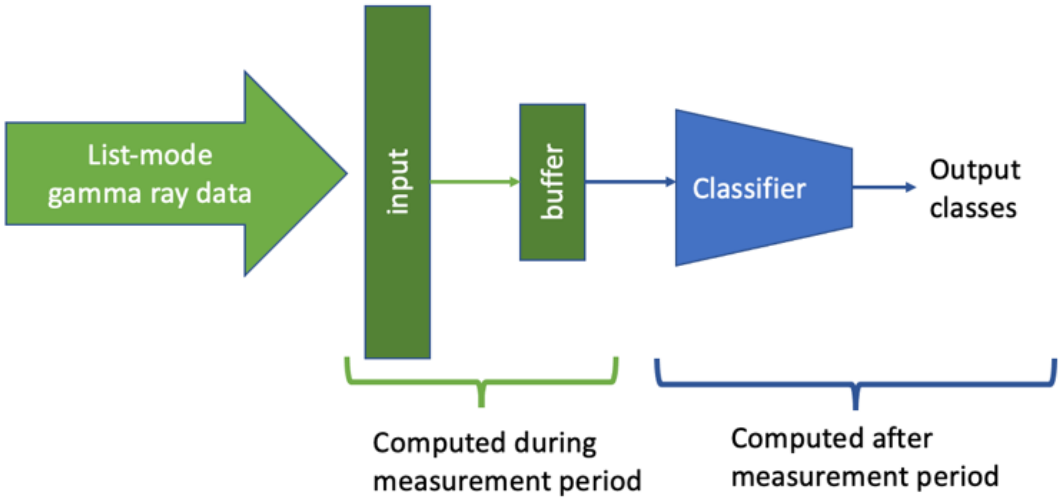

Figure 3: Schematic of the buffered classifier concept. Gamma-ray data in list-mode is fed through a dense neural network and accumulated in a buffer. The buffer is a reduced, irreversible representation of the input. At the end of a measurement run, the buffer state is fed through a classifier to arrive at output classes representing the class of the item measured (e.g. weapons grade/non-weapons grade).

A. *The Buffered Classifier Concept*

Gamma-ray photons are measured by a detector as individual pulses in time, with pulse heights corresponding to photon energy, as shown in Figure 4. This data is typically immediately binned into a histogram to form an energy spectrum. The shape of the spectrum reveals information about radioisotopes within the radiation source and shielding that surrounds it. To avoid the collection of a full gamma-ray spectrum of a weapon, which is sensitive information, the buffered classifier takes in data photon by photon, immediately converting it to a reduced, irreversible representation called a buffer, using a linear, fully connected neural network layer. The front end projects the data onto a smaller space (the buffer), from which the input cannot be practically recovered. The buffer is followed by a machine learning classifier including nonlinearities, such as a neural network with nonlinear activations.

Separating the linear front end from the nonlinear classifier allows the algorithm to process pulse-by-pulse data, and also take advantage of the complex decision functions afforded by a neural network or other classifier containing nonlinearities. Training the linear front-end weights and the classifier simultaneously allows the algorithm to learn how to perform the down-sampling optimally for class separation.

B. *The Front-End Buffer*

In more detail, list-mode data can be represented by a temporal series of photons arriving at the detector, denoted P, comprising impulses $E_i\delta_t$ where time is in the range $t \in [0, T]$ and energies (E_i) are discretized into bins with $i \in [0, N]$. In practice, N is the number of channels in a multichannel analyzer, or conversion gain. We're assuming perfect energy calibration, the practical effects of which can be the subject of future studies.

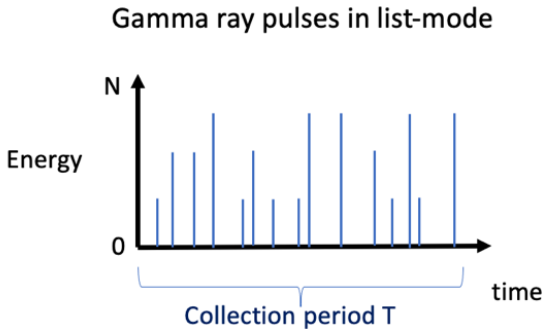

Figure 4: Gamma-ray pulses in list mode. This pulse train P is comprised of impulses $E_i\delta_t$.

A gamma-ray spectrum ($\vec{x}$) is the histogram of this pulse train in time over the collection period T, in which each feature is the total number of counts per energy bin:

$$x_i = \int_{t=0}^{T} P dt$$

Utilizing a spectrum input, the linear front end evaluates the following full expression to arrive at buffer states x_j. Weights between input feature i and buffer activation j for N pulses are denoted $w_{i,j}$, and the bias term is b.

$$a_j = \sum_{i=0}^{N} w_{i,j} \int_{t=0}^{T} P dt + b$$

With list-mode input instead, the buffer state is incremented upon every discrete incoming pulse:

$$a_{running}{}_{j,t} = \sum_{i=0}^{N} w_{i,j} P_i + b$$

The output at the end of the collection period is:

$$a_j = \int_{t=0}^{T} a_{running}{}_{j,t} = \int_{t=0}^{T} \sum_{i=0}^{N} w_{i,j} P_i \, dt + b,$$

which is equivalent to the prior expression due to the independence of the summation over energy bins and the integral over time. Because of the mathematical equivalence of taking the running sum of counts before

or after the linear reduction, the algorithm can be trained on spectra and ultimately used on list-mode data. For simplicity, we train and test exclusively on spectra.

C. Dual Optimization

Reflecting the competing objectives present in the nuclear warhead verification scenario – that each party wishes to confirm information about another party's declared treaty accountable items while protecting information about their own—we minimize classification accuracy and maximize reconstruction error. The two objectives being simultaneously optimized are:

(1) Maximizing reconstruction error, which we define as mean squared error between the input and its reconstruction from the stored buffer values.

(2) Minimizing cross-entropy loss between the computed output class (weapons-grade/non-weapons-grade) and the ground truth.

The reconstruction of the input spectrum ($\overrightarrow{x_{inv}}$) is computed using the Moore-Penrose pseudo-inverse of the linear front end's weight matrix ($\boldsymbol{W}^{+}$), the buffer activations $\overrightarrow{x_{buf}}$, and the bias term b.

$$\overrightarrow{x_{inv}} = \boldsymbol{W}^{+}(\overrightarrow{x_{buf}} - b)$$

Figure 5 shows an example input and its reconstruction, using a variety of models with varying buffer sizes.

D. Experiment

Our training data consists of 72,857 examples, of which 10% are used for validation, and our testing set consists of 8096 examples. About 60% of the examples are in the non-weapons-grade class. There are 8127 features, spanning energies from 20keV – 3.27 MeV. Every spectrum in the dataset is normalized so that the features sum to 1.

For the classifier, we used a two-layer neural network, with each layer containing a batch norm [7], a number of hidden units, and exponential linear unit (ELU) activations. Buffer sizes of 64 or greater had 64 and 32 hidden units in the classifier layers, respectively. Buffer sizes below 64 had the first classifier layer having an equal number of hidden units as the buffer size, and the second classifier layer having half the number of hidden units as the buffer size. The weights are initialized to a normal distribution about 0.5, with a standard deviation of 1.

Networks were trained for 1000 epochs, which was sufficient for cross-entropy loss to achieve convergence. The networks did not achieve convergence of reconstruction error, which monotonically rose with epoch number. Considering that the networks appear to be uninterpretable with present results, we do not consider convergence of the reconstruction error to be important.

Both objectives described in section IIIc are class-weighted to balance the uneven numbers of examples in each class, and optimized using the Adam optimizer [8]. The learning rate for objective (2) is chosen to maximize the rate at which the loss function decreases according to the method by Smith, in a version of the network not maximizing objective (1) [9]. Otherwise, the learning rate for objective (1) was held at a constant value of 0.001.

E. Results & Analysis

Our primary goal was to demonstrate irreversibility of the reduced representation, and secondly to find the parameters for which classification accuracy was high. Figure 5 shows examples of original spectra in comparison to those that were back-solved from the buffer. As part of this study, a large selection of inverted spectra was shown to human analysts, who were unable to identify any of them due to high levels of noise and distortion present from the downsampling and reconstruction. While a rigorous proof of non-invertibility will be a subject of a future publication, this initial study indicates that through the linear front end, the buffered classifier substantially obscures the sensitive input.

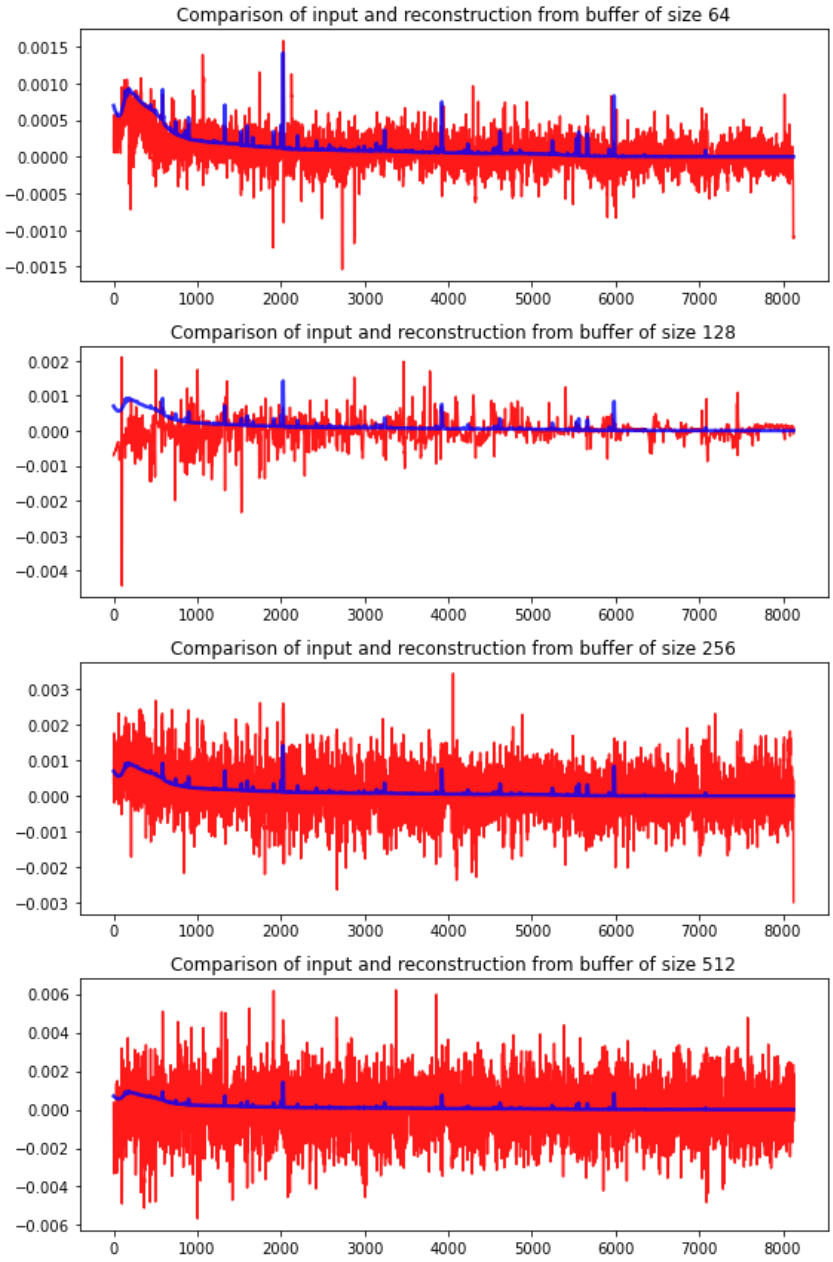

Figure 5: A spectrum input to a model (blue) and its reconstruction (red) from networks with various buffer sizes. In no observed cases across many observed examples was the reconstruction interpretable by a human expert analyst, due to high levels of noise and distortion.

Figures 6 and 7 show reconstruction error and classification accuracy as a function of buffer size, with the best model performing with 86% classification accuracy. A model without a linear front-end layer performs classification with 89% accuracy, meaning that the addition of the irreversible linear front-end buffer comes at the cost of only a modest decrease in accuracy. The results show that even small buffer sizes perform quite well for classification, and large buffer sizes (up to our tested 512) do not preserve information well enough for a human expert analyst to be able to reconstruct the data. Despite the lack of convergence of reconstruction error during training, requiring the network to maximize this error pushes the reconstruction

to be farther from the input, promoting noise and distortion in the reconstruction. Finally, Figure 9 shows the confusion matrix of the top-performing model.

A challenge in performing machine learning on gamma-ray spectra is the large variation in magnitudes between features, with the majority of feature magnitudes being very close to zero. Furthermore, due to the necessary linearity of the front end, we are unable to perform any nonlinear transforms that would bring the features into a more uniform range, which would improve the model performance.

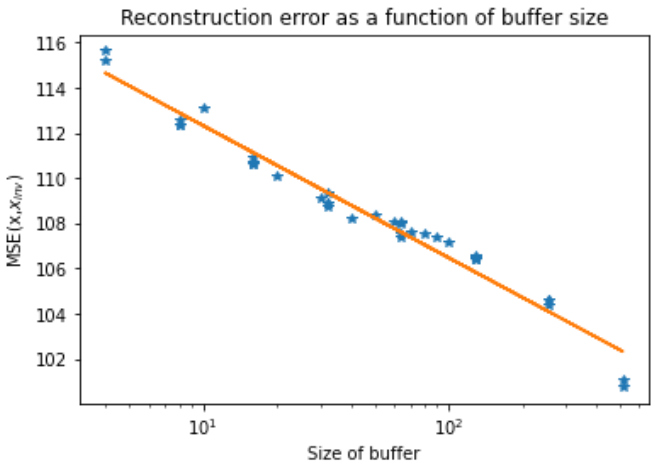

Figure 6: Reconstruction error (mean squared error) as a function of buffer size shown with best fit line illustrating the trend across multiple buffer sizes and various random seeds affecting network initialization.

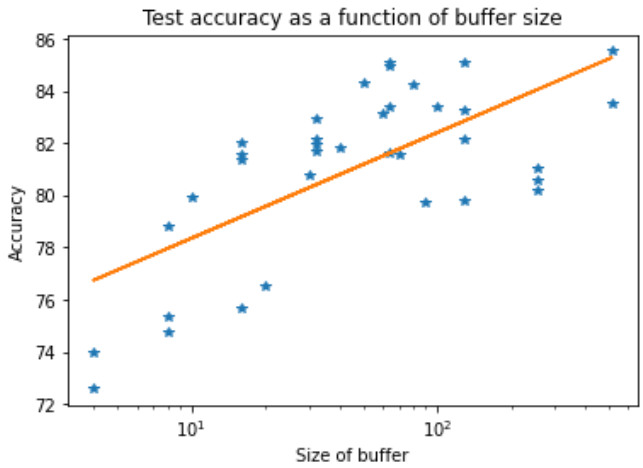

Figure 7: Classification accuracy as a function of buffer size shown with best fit line illustrating the trend across multiple buffer sizes and various random seeds affecting network initialization.

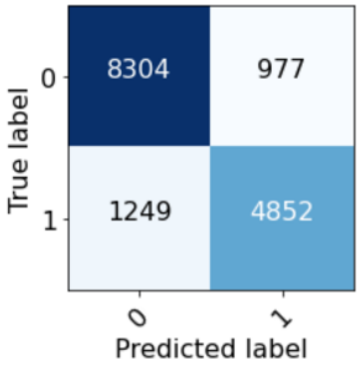

Figure 8: Confusion matrix of the best-performing result, a network with 1024 buffer neurons, resulting in a class-balanced accuracy of 86%. Class 0 is non-weapons-grade material, and Class 1 is weapons-grade material.

V. Conclusion

In a nuclear arms control verification scenario, each party wants to confirm information about the other's nuclear inventory while not revealing sensitive information about its own, using a method approved by all. In this paper, we present a secure neural algorithm called the *buffered classifier*, which is inherently limited in the information it collects, for classifying gamma-ray signatures of nuclear materials. Trained on a non-sensitive dataset that we generated that spans a very large variety of configurations of nuclear materials and shielding, the buffered classifier showed 85.5% classification accuracy in distinguishing weapons-grade

nuclear material sources from non-weapons-grade radioactive sources, which is a modest reduction in accuracy from a model with no front-end buffer, for the substantial benefit of security.

Beyond our use case, the buffered classifier may find an application in any scenario in which a detector measures single photons at a time and a reduced representation needs to be stored to preserve privacy or be compatible with memory constraints. One such example may be single-photon cameras in high-security systems with significant memory or power limitations.

The buffered classifier model is a promising, secure, and trustworthy algorithm that has the potential for adoption in future arms control treaties. Further additions to the very broad training dataset, as well as algorithmic improvements, hold the promise for higher classification accuracies.

REFERENCES

[1] "Status of World Nuclear Forces." Federation of American Scientists. https://fas.org/issues/nuclear-weapons/status-world-nuclear-forces/ (accessed.

[2] D. Evans, "Strategic Arms Control Beyond New Start: Lessons from Prior Treaties and Recent Developments," 2021.

[3] M. C. Hamel, "Next-Generation Arms-Control Agreements based on Emerging Radiation Detection Technologies," Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2018.

[4] G. F. Knoll, *Radiation detection and measurement*. John Wiley & Sons, 2010.

[5] *Gamma Detector Response and Analysis Software (GADRAS) v. 16.0*. (2009). United States. [Online]. Available: https://www.osti.gov/biblio/1231259

[6] K. Nelson and P. Sokkappa, "A Statistical Model for Generating a Population of Unclassified Objects and Radiation Signatures Spanning Nuclear Threats," United States, 2008-10-29 2008. [Online]. Available: https://www.osti.gov/biblio/947761

[7] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," presented at the Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, Lille, France, 2015.

[8] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," p. arXiv:1412.6980. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2014arXiv1412.6980K

[9] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE winter conference on applications of computer vision (WACV)*, 2017: IEEE, pp. 464-472.