IMPACT OF SAFEGUARDS MEASUREMENT ERRORS ON DEEP NEURAL NETWORKS

PROCEEDINGS OF THE INMM & ESARDA JOINT VIRTUAL ANNUAL MEETING
AUGUST 23-26 & AUGUST 30-SEPTEMBER 1, 2021

Nathan Shoman*
Sandia National Laboratories†
nshoman@sandia.gov

Tom Burr
Los Alamos National Laboratory
tburr@lanl.gov

ABSTRACT

Machine learning (ML) has seen many successes in a variety of domains from image recognition to natural language processing. There has been a general call to apply these concepts to safeguards in order to develop more cost-effective solutions. The International Atomic Energy Agency (IAEA) has specifically solicited next-generation approaches to help remain effective in the face of continued growth in the nuclear industry. One specific priority is to identify, evaluate, and test promising applications of machine learning and/or artificial intelligence to improve safeguards. While some applications such as image recognition are relatively straight-forward, application to other safeguards tasks presents unique challenges. Nuclear material accountancy (NMA), for example, would benefit greatly from efficient algorithms that could utilize unattended measurement systems, which are often accompanied by higher uncertainties, to reduce costs. Current practices for large nuclear facilities employ frequent use of destructive assay (DA) measurements, which while very accurate, are time consuming and expensive. This work outlines some unique challenges associated with applying deep learning to improve NMA. It can be shown that often traditional accountancy approaches will outperform unsupervised ML approaches when used with datasets contaminated with measurement errors. ML can remain competitive to traditional accountancy, however, special care must be taken to mitigate the impact of measurement errors that disproportionately affects ML-based approaches.

1 Introduction

The goal of international nuclear safeguards is the timely detection of diversion of significant quantities (SQ) of nuclear material for weapons purposes and deterrence of such diversion by the risk of detection. The IAEA [1] has several guidelines for achieving these safeguards goals which are based on open literature estimates for weaponization of nuclear material. These guidelines are informed by material type, form, and quantity [2, 3]. One cornerstone of traditional safeguards is nuclear material accountancy (NMA). Here, physical measurements are combined with statistical tests to verify nuclear material has not been diverted.

NMA can be thought of as an audit that verifies a facility's reported quantities of material and aims to ensure that it is present and has not been diverted. Several measurement techniques are employed such as sampling, weighing, and item counting. NMA is particularly useful in that it requires no knowledge of loss pathways and has a straightforward statistical framework. However, one drawback is that high precision measurements are often required for NMA at facilities with large inventories of material.

*Corresponding author

†Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND2021-8054C

2 Traditional Statistics for Material Accountancy

True values of material properties, such as mass, can never be known due to the presence of measurement error. Safeguards measurements are often characterized by a multiplicative error model as described in Equation 1 below.

$$\begin{aligned} &M_{i,t} = G_{i,t}(1 + S_i + R_{i,t}) \\ &\text{where} \\ &S_i \sim \mathcal{N}(0, \delta_S^2) \\ &R_{i,t} \sim \mathcal{N}(0, \delta_R^2) \end{aligned} \tag{1}$$

$M_{i,t}$ is the measured quantity at location i and time t and $G_{i,t}$ is the true quantity of interest. This error model usually applies to all of the individual terms in the MB.

The short-term systematic error, S_i arises from measurement conditions or settings such as calibration curves which are not changed for some period of time and vary in an unpredictable way and is difficult to reduce. In contrast, the random error $R_{i,t}$ varies in an unpredictable way under repeatable conditions and can be reduced through repeated measurements.

The random and systematic errors ($R_{i,t}, S_i$) are assumed to be independent Gaussian random variables with zero mean and variances δ_R^2 and δ_S^2. The specific value of the variances δ_R^2 and δ_S^2 depend on the measurement technology that is used. This results in an observed measurement that is distributed according to Equation 2.

$$M_{i,t} \sim \mathcal{N}(G_{i,t}, {G_{i,t}}^2(\delta_R^2 + \delta_S^2)) \tag{2}$$

Careful analysis is required to detect material losses in the presence of these measurement errors. Traditional anomaly detection algorithms found in the literature will fail to perform adequately on these error contaminated datasets. The following sections briefly describe some principle definitions and properties used in statistics for traditional nuclear safeguards.

2.1 Material Balances

The material balance (MB), also called material unaccounted for (MUF) or inventory difference (ID), is a fundamental quantity used in NMA. The MB is a straightforward calculation [4] described below in Equation 3.

$$\text{MB}_t = \left(\sum_{i=1}^{n_\text{I}} I_{i,t-1} + \sum_{i=1}^{n_\text{in}} Tin_{i,t} - \sum_{i=1}^{n_\text{out}} Tout_{i,t} \right) - \sum_{i=1}^{n_\text{I}} I_{i,t} \tag{3}$$

Terms in Equation 3 include the total input transfer ($\sum_{i=1}^{n_\text{in}} Tin_{i,t}$), the total output transfer ($\sum_{i=1}^{n_\text{out}} Tout_{i,t}$) and total inventory ($\sum_{i=1}^{n_\text{I}} I_{i,t}$) at time t.

Large facilities often have multiple material balance areas (MBA), each of which have their own MB that is calculated at regular intervals of time. The MBA is a physical area that is designated by an subject matter expert and is based on several properties such as quantity and type of nuclear material present. The material balance period (MBP) is the interval of time at which the MB is calculated for each MBA. The MBP is also chosen by a subject matter expert and is often a balance between accountancy goals and ease of measurement for individual material locations.

Under normal conditions it is expected that $\text{MB}_t = 0$ as all material would be accounted for. However, as described in Equation 1, the individual components of the MB have measurement error resulting in $\text{MB}_t \neq 0$ even under normal conditions. This can complicate efforts to detect losses using the material balance directly.

As a consequence of measurement error the material balance can be expressed as a normal distribution in Equation 4 where MB_t has a mean, μ_t, that is directly related to any material loss and a variance, σ^2_{MB}, which is a function of measurement error.

$$\text{MB}_t \sim \mathcal{N}(\mu_t, \sigma^2_{\text{MB}}) \tag{4}$$

2.2 Sequential Material Balances

It is desirable to extend the concept of the material balance to a sequence of values for two reasons. First, the IAEA has requirements for the detection time of a loss of nuclear material [2]. More frequent material balance calculations, which would result in a sequence of material balances, could potentially detect a loss sooner than a single yearly material balance.

Second, consider a material loss which can be thought of as a shift in the mean of the material balance from μ_t to μ_t^* in Equation 4. It follows that the probability of detection for a material loss directly relies on σ_{MB}. As the variance in the material balance increases the probability of detection will approach the false alarm probability (i.e. random chance of detection). More formally, $\lim_{\sigma_{\text{MB}} \to \infty} PD(\mathcal{N}(\mu_t \to \mu_t^*, \sigma_{\text{MB}}^2)) = \text{FAP}$ where FAP is the false alarm probability. This is demonstrated in Figure 1 where the probability of detection is shown as a function of the material balance variance with a constant detection threshold and material loss.

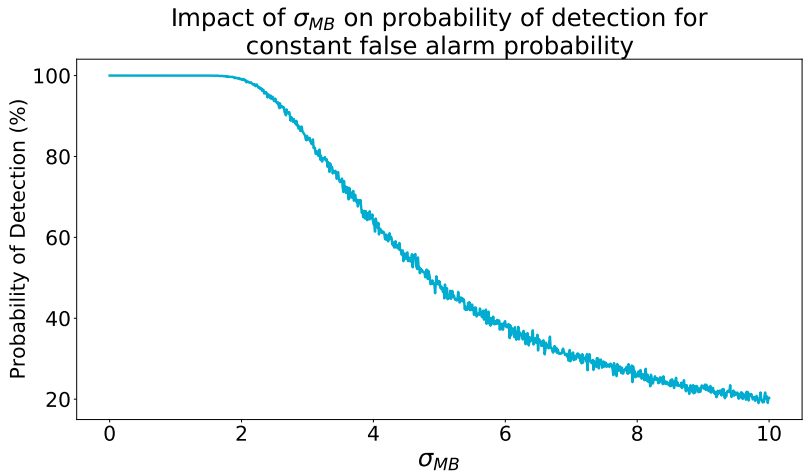

Figure 1: Probability of detection as a function of σ_{MB} for a constant false alarm probability

One of the ways to reduce the material balance variance is through more frequent material balance closures. Although not described in depth here, note that there is a limit to the beneficial impacts of increased material balance closures [5].

Now that the benefits of the material balance sequence are clear, Equation 3 is extended to a more general case describing a temporal sequence of material balances in Equation 5. Note that time is denoted as t (where $t \geq (2 * \text{MBP})$) and the MB sequence ($\overrightarrow{\text{MB}_t}$) is expressed as a multivariate normal distribution with mean μ_t and covariance Σ_t. As with the singular material balance, measurement uncertainty forms the basis for the covariance of the material balance sequence.

$$
\begin{aligned}
&\overrightarrow{\text{MB}_t} \sim \text{MVN}(\overrightarrow{\mu_t}, \Sigma_t) \\
&\text{where} \\
&\overrightarrow{\text{MB}_t} = (\text{MB}_1, \text{MB}_2, ..., \text{MB}_t) \\
&\overrightarrow{\mu_t} = (\mu_1, \mu_2, ..., \mu_t)^{\text{T}}
\end{aligned}
\tag{5}
$$

Briefly note that MB sequence exhibits some degree of correlation between sequential MBs. That is, inventories for previous material balance are used to calculate the current material balance. Although not discussed in depth here, it is possible to estimate the covariance matrix to produce a uncorrelated sequence of MBs. This is commonly referred to as the SITMUF transform, which was introduced by Picard [6]. This uncorrelated sequence of material balances results in better detection probabilities when used with sequential testing methods. A key feature of the SITMUF transform is the explicit treatment of the known error structure.

3 Machine Learning for Loss Detection

Machine learning (ML) has seen incredible success in several domains and has been considered as a candidate to improve overall detection probability for material losses while cutting costs associated with measurements. It would be desirable if ML could improve detection of material loss through use of process monitoring (PM) and non destructive assay (NDA) that often have higher uncertainties that prohibit direct use in a MB calculation. Instead of direct actinide quantification seen with traditional approaches for NMA, the ML approach would attempt to develop a simple binary metric (i.e. normal or off-normal) to determine if a process is operating as expected.

It is important to note that this work only represents one approach for applying ML to NMA. Specifically, focus is restricted to unsupervised neural-based ML approaches, which are a class of neural networks that do not require labeled examples of anomalous behavior. This is a reasonable analog to traditional approaches to NMA which do not require examples of material loss pathways. This work also exclusively considers matters concerning substitution losses where material removed is replaced with an equal mass of less desirable material. Direct losses, where material is not replaced, is not considered as detection becomes trivial through the use of a bulk mass measurement that relies on relatively high precision unattended measurements.

3.1 Problem Statement

This work proposes an unsupervised machine learning approach for NMA where it is assumed that a neural network can be used to learn a function $y = f(x, \theta)$, when given some PM data, x, measured from an unattended system and set of parameters, θ, to accurately predict some other quantity resulting from normal facility behavior, y. Under off-normal conditions, such as a material loss, the original function, $y = f(x, \theta)$, should provide poor predictions as it no longer reflects the current state of operation at the facility. Thus, the residual between the prediction, $\hat{y}$, and observed value, y, should be small for normal conditions and large for off-normal conditions. An example of this approach applied to a large reprocessing facility might have head end NDA measurements as the input data, x, with product NDA measurements as the response y which is used to train an arbitrary ML algorithm to understand the process by which the input material, x, is transformed to a final product, y, such that $y = f(x, \theta)$. A graphical representation of this approach is shown in Figure 2.

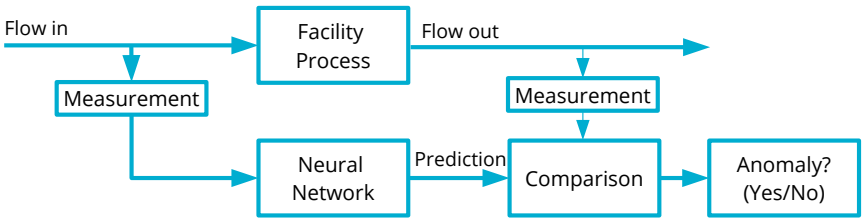

Figure 2: Proposed setup for applied ML for NMA

3.2 Machine Learning Introduction and Training

ML algorithm are often initialized according to some strategy that is problem specific. Initially, the algorithms themselves produce nonsensical results. Training is used to adjust weights of an algorithm to give predictions that are closer to the training dataset. Specifically, gradient-based optimizers are used to minimize some objective function, which is problem dependent. For example, the training objective could be categorical (e.g. distinguish between cat and dog pictures) or continuous (e.g. predict house prices given square footage and number of rooms).

This proposed framework in the previous section uses a mean squared error (MSE) loss function, $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$, which is appropriate given the goal is to predict facility behavior in a continuous space. During training, the MSE loss function is minimized in order to find a set of parameters, $\hat{\theta}$, such that $f(x, \hat{\theta})$ provides a good prediction of y when applied to the training dataset.

It is important to note that training can often be expressed in statistical terms. For example, minimization of the MSE loss function is equivalent to maximization of the Gaussian log-likelihood. Effectively, during training the optimizer is

adjusting the parameter estimate $\hat{\theta}$ (i.e. weights and biases) to maximize the likelihood (i.e. distribution similarity) of the predicted distribution $p(\hat{y})$ and the training distribution $p(y)$.

Data requirements often scale with the size of the model which necessitate large training datasets. Many real-world ML problems are overparameterized, that is, there are more parameters in the model than there are training examples. This leads to multiple possible solutions to the optimization problem which can lead to suboptimal performance. Consider the equation $5 = 3x^2 + 8y + 10z^4$ given that $20 = 10x$. The variable x can be solved for, however, y and z are unconstrained leading to many possible solutions. Consequently, increasing the size of the training dataset will almost always improve performance as the training distribution is better characterized.

4 Machine Learning Limitations for Loss Detection

The dataset used for training a machine learning algorithm must meet certain requirements. First, the dataset must adequately represent the process of interest (i.e. training and testing distributions match) and secondly, the dataset must be large enough to adequately train the machine learning algorithm. The context of a sufficiently large training dataset depends on the size of the algorithm used. It could be difficult to collect large quantities of training data especially from nuclear facilities. However, it is reasonable to assume that if the data could be obtained that it would originate from multiple measurement campaigns. That is, even if the same process or feature were observed repeatedly, the collected data would have different calibration curves (and thus different systematic biases).

Consequently, the training dataset would then represent the total range of expected normal conditions which include not only process variation, but also measurement variations. When multiple sets of data from different campaigns are aggregated together it will increase the distribution variance as seen in Figure 3. This is an important consequence to note as increased variance in normal behavior will reduce performance of any anomaly detection algorithm, including the proposed ML approach.

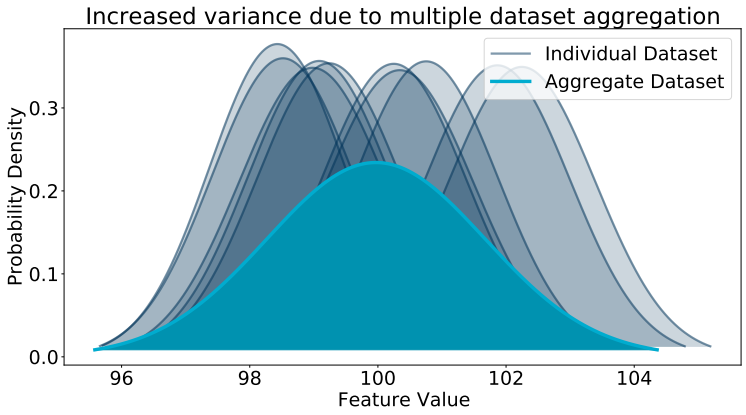

Figure 3: Probability density functions for multiple normal datasets

As another example, consider the task of predicting a simple sine function, $y = \sin(x) + 10$, with an applied safeguards error structure (Equation 1). This example can be more formally expressed by Equation 6. Here, the objective is to minimize the MSE by learning $y_{\text{observed},t} = f(x_{\text{observed},t}, \theta)$.

$$\begin{aligned} x_{\text{true}} &\in [-3\pi, 3\pi] \\ y_{\text{true}} &= \sin(x_{\text{true}}) + 10 \\ x_{\text{observed},t} &= x_{\text{true}}(1 + R_t + S) \\ y_{\text{observed},t} &= y_{\text{true}}(1 + R_t + S) \end{aligned} \tag{6}$$

A Bayesian neural network that consists of Flipout layers to learn from the training data was used for this next example. This network contrasts with a traditional neural network in that the Bayesian network has weights that are expressed as distributions. Consequently, the Bayesian network can be sampled multiple times resulting in different predictions for the same input. This example considers the performance of this Bayesian network as the level of random and

systematic error in x from Equation 6 are increased. Figure 4 shows the predictions of the network when trained on $1\%,\ 3\%,$ and 5% random and systematic errors. Increasing error increases the variance of the weights of the neural network, which in turn, also increases the variance in the prediction.

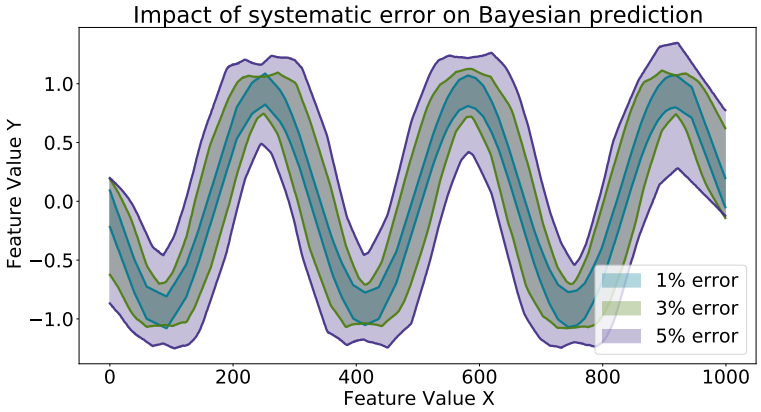

Figure 4: Bayes neural network prediction for sine wave with $\pm 2\sigma$.

Recall that the goal of the proposed ML approach is to use a neural network to detect changes in some learned baseline to detect loss of nuclear material. As the uncertainty in the training data (driven by measurement) increases, so does the uncertainty in predictions. This is clearly visible in Figure 4 where a Bayesian network is used, but note this occurs in all deep neural networks. For example, a traditional feed forward neural network would tend to learn the mean of the training distribution (essentially the average value of a feature at a specific location and time when also factoring in measurement error) which would also provide poor per-run predictions. Such prediction uncertainty is a manifestation of the problem as currently framed (i.e. material losses will appear as deviations from expected baseline measurements). This problem cannot be mitigated through traditional preprocessing methods (e.g. scaling, whitening, standardization) operating on the entire dataset or through selection of specific architecture (CNN [7], LSTM [8], fully-connected, etc). The next several section seeks to bound the problem and compare this specific ML approach to traditional statistical methods.

4.1 Establishing lower limit of detection

Previous discussion framed material loss as a mean shift in the distribution of some feature. Leveraging methods that require knowledge of the underlying distribution will prove difficult in practice due to limitations on data availability. It is useful nonethelessess to conduct several thought experiments to bound the material loss problem in terms of systematic error and inventory size as both will have a significant impact on detection of a material loss.

Although similar experiments can be performed using techniques such as ANOVA (analysis of variance), first consider a simple Bayes ratio test used to compare distributions. The use of Bayes theorem allows establishment of the odds ratio as in Equation 7, which describes the relative odds of some data point D belonging to distribution $\theta_{loss} = \mathcal{N}(\mu_{\text{loss}}, {\sigma_{\text{loss}}}^2)$ vs $\theta_{normal} = \mathcal{N}(\mu_{\text{loss}}, {\sigma_{\text{loss}}}^2)$. Note that Equation 7 assumes that $\frac{P(\theta_{loss})}{P(\theta_{normal})} = 1$, which is assumed for illustrative purposes but is impossible to quantify in practice due to the unknown distribution $P(\theta_{loss})$.

$$P(\theta_{\text{loss}}|D) = \frac{P(D|\theta_{\text{loss}})P(\theta_{\text{loss}})}{P(D)}$$
$$P(\theta_{\text{normal}}|D) = \frac{P(D|\theta_{\text{normal}})P(\theta_{\text{normal}})}{P(D)}$$

$$\frac{P(\theta_{\text{loss}}|D)}{P(\theta_{\text{normal}}|D)} = \frac{P(D|\theta_{\text{loss}})P(\theta_{\text{loss}})}{P(D|\theta_{\text{normal}})P(\theta_{\text{normal}})} = \frac{P(D|\theta_{\text{loss}})}{P(D|\theta_{\text{normal}})}$$

where (7)

$$P(D|\theta_{\text{loss}}) = \frac{1}{\sigma_{\text{loss}}\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{D-\mu_{\text{loss}}}{\sigma_{\text{loss}}})^2}$$

and

$$P(D|\theta_{\text{normal}}) = \frac{1}{\sigma_{\text{normal}}\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{D-\mu_{\text{normal}}}{\sigma_{\text{normal}}})^2}$$

It can be shown using Equation 7 that, assuming equal odds of drawing from each distribution, the odds of belonging to one distribution over another decreases with increasing variance. This is an important observation that has consequences for applied ML for NMA. This reinforces intuitions from previous sections that variance in an observed signal will decrease anomaly detection performance, even for a machine learning approach. That is, even applied ML approaches cannot escape the limitation $\lim_{\sigma \to \infty} PD(\mathcal{N}(\mu_t \to \mu_t^*, \sigma^2) = \text{FAP}$.

This example can be extended to consider the impact of both inventory size and systematic error on the odds of belonging to the normal distribution over the (in practice unknown) loss distribution as illustrated in Figure 5. Two key assumptions were made in Figure 5. First, it was assumed that the evaluation observation D was precisely the mean of the loss such that $D = \mu_{\text{loss}}$. Secondly, it was assumed that there was a sufficiently large dataset to capture the variation due to measurement errors. The majority of Figure 5 shows an odds ratio of 1 meaning that there is an equal probability that D belongs to θ_{normal} or θ_{loss}. The only regions with a reasonable expectation of loss detection (i.e. odds > 1 for $\frac{P(\theta_{\text{loss}}|D)}{P(\theta_{\text{normal}}|D)}$) are areas with low uncertainty values.

It should be stressed that the use of a likelihood ratio is a simple toy example that is not used in safeguards often as the true loss distribution $P(\theta_{\text{loss}})$ is unknown. However, this example provides some intuition behind the lower limits of detection for any algorithm attempting to detect anomalies at a nuclear facility. Regardless of the algorithm or approach, an increase in measurement uncertainty and total inventory will result in a larger variance and lower probability of detection for a specific loss pattern. The next section will describe why the proposed ML approach is particularly sensitive to this effect.

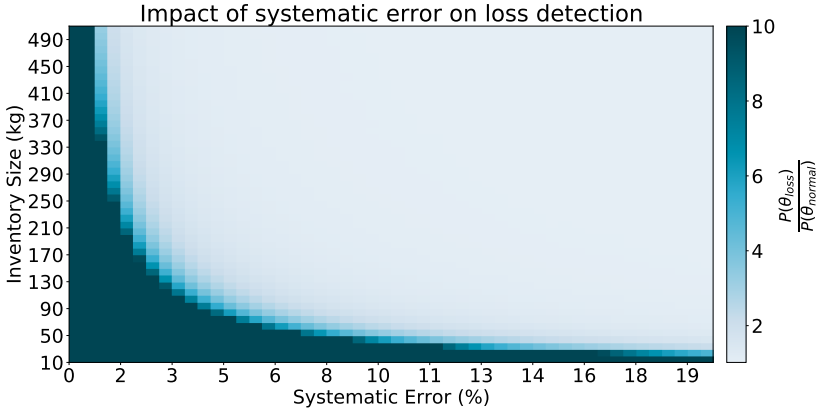

Figure 5: Odds ratio of a loss of 8kg of material assuming that D is the loss mean. Higher odds indicate increased change that D belongs to the loss distribution and not the normal distribution

5 Comparison of Machine Learning and Traditional Statistics

Traditional safeguards have often focused on lowering the variance of measurements precisely due to reasons described in the previous section. Often, destructive assay measurements are required to reach safeguards goals for facilities with large inventories which incur a significant cost. However, concepts like the material balance are well understood and permit operations such as the SITMUF transform that removes correlations in the MB sequence. The estimate of variance in the MB sequence improves over time as successive measurements are made. This leads to an increasing probability of detection for a given loss pattern until the covariance estimate stops improving. It is also important to note that the covariance of the MB sequence utilizes observations from a single set of measurement calibrations (i.e. systematic errors).

The proposed ML approach has two key disadvantages compared to traditional statistics for NMA that arise from the conventional training scheme used for ML algorithms. First, it is reasonable that a sufficient large training dataset for any ML algorithm for NMA would be comprised of multiple measurement campaigns each with a different sensor calibration. The composite training dataset would capture changes that not only arise from facility operation but also measurement error. Consequently, this will lead to a larger variance than a dataset that captures facility operation variance alone. This ultimately leads to a more difficult change detection problem. Second, there is no explicit treatment of the underlying error structure in the proposed ML method. This contrasts with the SITMUF transform which does account for measurement error through the covariance estimate. This results in a ML-based approach with a larger variance in the training dataset and no regard for measurement error which almost all but guarantees a worse probability of detection for any loss pattern as compared to traditional statistics for NMA.

For a supporting thought experiment, return to the Bayes ratio described in Section 4.1 to compare the traditional and ML-based approaches when a composite training dataset is used (i.e. dataset with multiple measurement campaigns). The traditional approach is assumed to have a standard deviation $\sigma = 1$, however, the ML-based approach will have to aggregate multiple sets of measurement data resulting in a higher standard deviation of $\sigma = 1.4$). The distributions for this thought exercise are summarized in Equation 8. Subscripts T, M refer to distributions associated with the traditional and machine learning approaches respectively. Superscript * refers to a mean shift after some material loss.

$$\begin{aligned} \theta_{\mathrm{T}} &= \mathcal{N}(100, 1.0) \\ \theta_{\mathrm{T}}^* &= \mathcal{N}(92, 1.0) \\ \theta_{\mathrm{M}} &= \mathcal{N}(100, 1.4) \\ \theta_{\mathrm{M}}^* &= \mathcal{N}(92, 1.4) \end{aligned} \tag{8}$$

Clearly the odds of belonging to the loss distribution, for the same observation, are higher when considering the lower variance distributions (i.e. traditional NMA) than the higher variance distribution (i.e. ML approach with aggregated training dataset). Note that the total difference between standard deviations of the traditional safeguards and ML-based approaches (roughly 0.4 in this thought example) may vary depending on several factors, however, the variance of the ML training distribution will almost always be higher than the error adjusted SITMUF sequence.

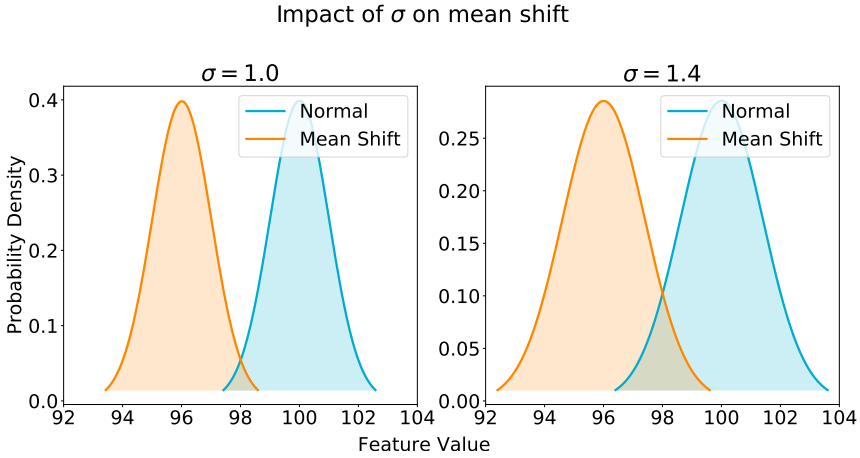

Figure 6: Systematic error, when incorporated into a large training set, increases overall variance making detection of a mean shift more difficult.

6 Discussion

There are several possible mitigation strategies that could be considered to improve the performance of ML-based algorithms when training on datasets with a safeguards-like error structure (i.e. Equation 1). The traditional statistical tools for NMA provide better detection probabilities as a result of limited systematic error impact and explicit treatment of underlying error structure. The first issue, constraining the systematic error at test time, has two potential solutions. First, a synthetic dataset that is large enough for training but only contains a single set of calibrations could be generated. However, this would necessitate significant quantities of historical data or a robust process model, both of which could be challenging. Perhaps more desirable would be to develop a few-shot learning scheme where training is performed using a smaller dataset. General focus of few-shot learning has been on computer vision tasks, however, some recent work suggests that few-shot learning could work for regression tasks as well.

The second potential area for improvement is direct incorporation of the measurement error structure. Traditional statistics for safeguards are able to capture measurement error through prior knowledge and propagation of variance. However, the error structure of the ML prediction and associated residual is unclear. Estimates of ML variance could be performed using empirical approaches, but this could require large amounts of data as well. An estimate of the covariance in the ML residual sequence could enable a SITMUF like transform which would improve performance.

7 Conclusions

This work presented a hypothesis that ML algorithms could learn facility behavior and complex but subtle changes that might reveal anomalies in order to improve safeguards at large throughput nuclear facilities. Specifically, a neural network would learn normal facility behavior to predict future measurements. Under abnormal conditions, the neural network would give poor prediction resulting in high prediction error. As machine learning algorithms often require large amounts of data, it was assumed that a training dataset would consist of data collected over multiple measurement campaigns which would lead to higher variances in the training dataset. This increase variance, which makes it more difficult to detect loss of material, arises from the traditional methods used to train ML algorithms and will persist regardless of the architecture used and will not be improved by most standard preprocessing techniques.

There remains some promise for ML for NMA if the variance in the training data could be reduced. The most straightforward approach would be to reduce the size of the training dataset itself using strategies such as few-shot learning or direct inclusion of the safeguards error structure. This work showed issues relating to variance and training dataset sizes, but only for deeply parameterized models. Non-parametric approaches could help resolve some of the shortcomings listed here.

8 Acknowledgments

The coauthors would like to acknowledge the open-source packages Tensorflow [9] and Scikit-learn [10] that made this work possible in addition to Benjamin Cipiti, Philip Honnold, Richard Fields (Sandia National Laboratories) who contributed to this work.

References

[1] United Nations, “IAEA Statute,” Oct 1956.

[2] International Atomic Energy Agency, *IAEA Safeguards Glossary, 2001 Edition*.

[3] International Atomic Energy Agency, *Guidance for States and Implementing Comprehensive Safeguards Agreements and Additional Protocols*, 21 ed., May 2016.

[4] A. Goldman, R. Picard, and J. Shipley, “Statistical methods for nuclear materials safeguards: An overview,” *Technometrics*, vol. 24, no. 4, pp. 267–275, 1982.

[5] R. Avenhaus and J. Jaech, “On subdividing material balances in time and/or space,” *Journal of Nuclear Materials Management*, vol. 10, 1981.

[6] B. Jones, “Near real time material accountancy using sitmuf and a joint page’s test: comparison with muf and cumuf tests,” 1988.

[7] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, May 2015.

[8] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, 1997.

[9] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.