SANDIA REPORT
SAND2022-12069
Printed September 2022

The Power of Priors: Improved Enrichment Safeguards

Nathan Shoman, Philip Honnold

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185
Livermore, California 94550

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology & Engineering Solutions of Sandia, LLC.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@osti.gov
Online ordering:	http://www.osti.gov/scitech

Available to the public from

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.gov
Online order:	https://classic.ntis.gov/help/order-methods

ABSTRACT

International safeguards currently rely on material accountancy to verify that declared nuclear material is present and unmodified. Although effective, material accountancy for large bulk facilities can be expensive to implement due to the high precision instrumentation required to meet regulatory targets. Process monitoring has long been considered to improve material accountancy. However, effective integration of process monitoring has been met with mixed results. Given the large successes in other domains, machine learning may present a solution for process monitoring integration. Past work has shown that unsupervised approaches struggle due to measurement error. Although not studied in depth for a safeguards context, supervised approaches often have poor generalization for unseen classes of data (e.g., unseen material loss patterns). This work shows that engineered datasets, when used for training, can improve the generalization of supervised approaches. Further, the underlying models needed to generate these datasets need only accurately model certain high importance features.

ACKNOWLEDGMENTS

The authors would like to thank National Nuclear Security Administration's Office of International Nuclear Safeguards for funding this work.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

Material accountancy is an important cornerstone of international safeguards that helps to determine if nuclear material is present and unmodified. However, implementing material accountancy at large scale bulk facilities can be expensive as high precision equipment is often required to meet regulatory targets. Process monitoring, or the use of non-nuclear data streams (e.g., tank level measurements, temperatures, pressures, etc.) has been considered since at least the 1980s [2] as a way to improve material accountancy. Efforts to integrate these signals have been met with mixed results, but have been used in production facilities to some extent.

Machine learning has seen successes in several domains for a variety of tasks and may be suitable for anomaly detection tasks for safeguards applications. While machine learning has been applied successfully for safeguards tasks where there is a close analog in the commercial domain (e.g., video or text applications) [3, 4], its' suitability for process monitoring applications is not yet proven. Prior work that considered unsupervised machine learning showed that performance was ultimately limited by measurement uncertainty and would likely perform worse than existing material accountancy techniques for most cases (although direct comparisons are challenging) [5, 6]. Supervised learning is more powerful than unsupervised, but requires labeled examples of anomalous conditions. This is problematic for safeguards applications where it is impossible to know all potential anomalies. Limited efforts have shown that supervised approaches can reliably identify labeled anomalies, however, their ability to generalize has not been well studied within a safeguards context [7].

This work is focused on the generalization of supervised approaches to unseen material loss patterns by using gaseous centrifuge enrichment facilities as a use case. A computational model simulates features observable by the OnLine Enrichment Monitor [8, 9]. This data is then used by a supervised classification algorithm, specifically state-of-the-art InceptionTime algorithm [1] to attempt detection of material loss scenarios.

Three different material loss scenarios of increasing difficulty are modeled; scenario 1 is the easiest to detect whereas scenario 3 is the most difficult. The baseline case trains the classification algorithm on one of the scenarios with the goal of classifying the others. For example, InceptionTime is trained on scenario 1, and then classification performance is evaluated on all three scenarios. This baseline case is analogous to a situation wherein only a single high consequence loss pattern is known and available for training.

Results are reported based on the classification performance for each training scenario. Specifically, the F1-score, which is a balance between the precision and recall of a classification algorithm, is used to explore the performance of InceptionTime when trained on different scenarios. Note that these values are not directly comparable to probability of detection and are employed to compare the relative impact of different training strategies (e.g., trying to generalize to multiple unseen scenarios when only training on single scenario). The results for the baseline case are shown in Figure 0-1 below.

The baseline case performs as expected; poor generalization (i.e., low F1-score) is observed for all scenarios not seen in training. For example, when trained on scenario 1, InceptionTime performs well on scenario 1 and the no loss case only (i.e., high F1-score). It would be advantageous to leverage supervised algorithms for anomaly detection in safeguards, but the algorithms must be able to generalize, as the baseline performance shown in Figure 0-1 is insufficient. Prior work indicated that

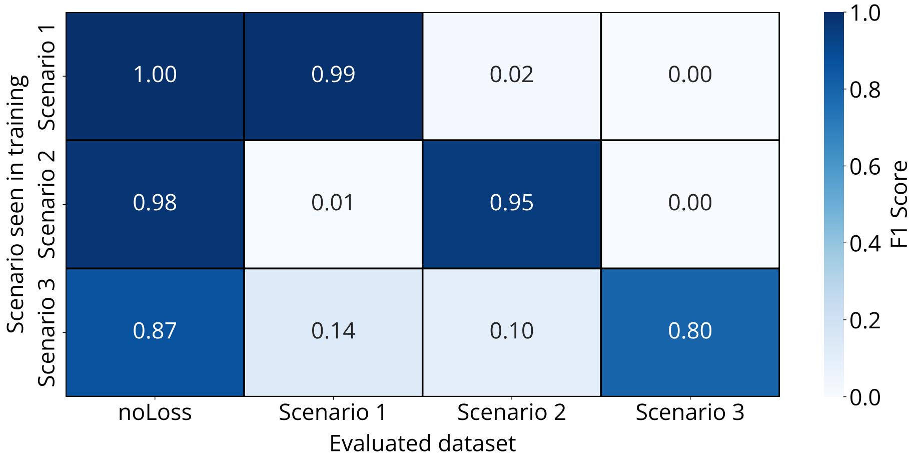

Figure 0-1 Summary of baseline performance expressed as a binary F1-score

InceptionTime might generalize to unseen loss patterns if a loss pattern at the same facility location was previously observed [10].

Dataset engineering, wherein unrealistic training datasets are generated to improve generalization, is considered. This approach serves as a prior by injecting domain knowledge into training of the machine learning algorithm through labeled examples. Specifically, five different dataset engineering schemes are explored. Each of these schemes refers to a different genernated anomaly present in the training dataset. Uniform refers to a uniform loss at all locations whereas random refers to a random loss at all locations.The performance of InceptionTime, when trained on these engineered datasets and evaluated on the baseline material loss scenarios is shown in Figure 0-2.

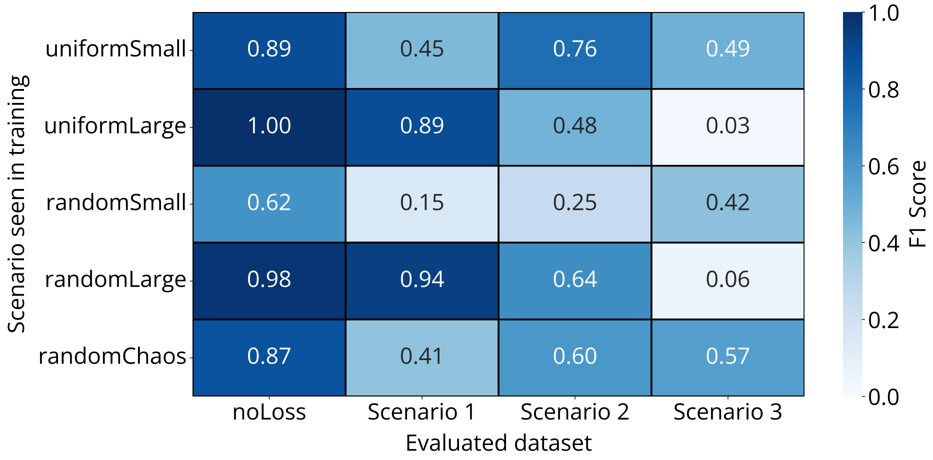

Figure 0-2 Summary of InceptionTime performance trained on engineered datasets as expressed by a binary F1-score

Training with the engineered datasets significantly improve the ability of InceptionTime to generalize to material losses not observed during training. Using an explanability technique called SHAP [11, 12] to probe feature importance, it can be shown that training on these engineered datasets result in models that more heavily weigh anomaly-relevant features, thus improving generalization. However, one drawback of dataset engineering is that synthetic datasets are needed. In practice, dataset engineering could require individual high fidelity talored to specific facilities, which could be difficult to create.

A final experiment was conducted to determine the impact of incomplete facility knowledge on data generated from computational models. Datasets are generated where some feature means during normal operation are shifted from their true value (i.e., a bias is applied to normal behavior). This should negatively impact algorithm performance as the training dataset would no longer represent real-world facility behavior as a consequence of modeling biases.

Contrary to the initial hypothesis, experimentation shows that the model bias has little impact on performance of the classification algorithm. This results from the minimal weight that the shifted features have in the trained model (i.e., the model relied very little on these features to make accurate classification predictions). This shows that an engineered dataset approach might not require a high precision model for all facility behavior as long as important features are modeled accurately. Identification of these features could be accomplished through modeling and simulation or expert opinion. Future work should further investigate the dataset engineering approach by considering other parameters not considered in depth here, such as quantity of training data or time series window length.

NOMENCLATURE

GCEP Gaseous Centrifuge Enrichment Plant

IAEA International Atomic Energy Agency

MUF Material Unaccounted For

NRTA Near Real Time Accounting

OLEM OnLine Enrichment Monitoring

PCA Principal Component Analysis

SHAP SHapely Additive exPlanation

SITMUF Standardized Independent Transformed Material Unaccounted For

1. INTRODUCTION

The International Atomic Energy Agency (IAEA) is tasked with implementing international safeguards in non-proliferation signatory countries. This can be difficult given the increasing number of nuclear facilities and flat budgets. The IAEA has solicited new technologies to help reduce costs of safeguarding facilities to remain effective despite these challenges. One potential area for improvement is safeguards at large throughput facilities. Material accountancy, the practice of establishing the quantities of nuclear material present in defined areas, often requires expensive, high precision measurements to reach regulatory goals (see Appendix A). It has long been theorized that the inclusion of process monitoring data, which refers to measurements that do not directly quantify nuclear material, but often have relatively high precision, could improve safeguards. Examples of process monitoring data include signals such as temperatures, pressures, tank level measurements, and more. Deep learning, which is a subset of machine learning characterized by the use of neural networks, is considered here to detect anomalous patterns in process monitoring data given its' success in similar tasks from other domains.

Deep learning for anomaly detection [13] has been shown to perform well in a variety of domains. However, past work [14, 15] has demonstrated that unsupervised deep learning, which is desirable for safeguards applications as it requires no explicit example of abnormal behavior, offers worse performance for safeguards material loss detection than traditional material accountancy. This low performance is likely due to a lack of strong assumptions (i.e., priors) about the underlying process data (see Appendix B). Simpler models considered in [14] offered competitive anomaly detection performance compared to traditional material accountancy approaches but required calibration of numerous alarm thresholds. Previous work [7] showed that supervised approaches could outperform traditional material accountancy for anomaly detection. The increase in performance over unsupervised approaches is due to a direct description of anomalous conditions rather than specification of some proxy problem as is the case for unsupervised. However, [7] did not perform a robust analysis on generalization, which is a known problem for supervised methods and the target of this work.

Contributions: The goal of this paper is to better define the limitations of deep learning for safeguards anomaly detection by exploring the generalization of supervised approaches to unseen anomalies. A simulated enrichment facility was used as an exemplar due to the number of observable process monitoring signals using existing technologies such as the OnLine Enrichment Monitor (OLEM) [8, 9]. Specific contributions are as follows:

- **Evaluation of baseline supervised performance (section 4.1):** This section establishes baseline performance using the state-of-the-art supervised classification algorithm InceptionTime [1]. These results show the out-of-the-box generalization performance to InceptionTime to unseen anomalies. Baseline performance in this section reflects a scenario wherein very few high consequence material loss scenarios are known, simulated, and incorporated into training.
  - This corresponds to a situation wherein only a single high consequence loss scenario is known and available for training.
- **Impact of scenario engineering (section 4.2):** Past work [10] suggests that generalization of InceptionTime could be improved through manual engineering of loss scenarios. These scenarios do not necessarily correspond to plausible real-world scenarios and instead focus on improving generalization of the learned model parameters by avoiding overfitting to a few known scenarios.

  - This section seeks to develop effective priors to train the model by developing additional examples of material loss. However, these priors are not formed based on specific high consequence material loss and are instead formed based on randomized patterns that might improve generalizability.

- **Imperfect system knowledge (Section 4.3):** Although good results were observed in Section 4.2, it was assumed that facility patterns seen in training reflected reality at test time. This section evaluates the impact of imperfect knowledge about facility processes might impact performance.
  - Although training on randomize loss scenarios improved performance, high fidelity models might be required to obtain sufficient training data as real-world data is scarce.

2. RELATED WORK

The use of process monitoring for improving safeguards has been considered since at least the 1980s. However, most safeguards literature focuses on the development and application of "near-real-time" accounting [2] (i.e., traditional material accountancy) rather than process monitoring based strategies. Near real time accounting involves the calculation of statistical quantities such as Material Unaccounted For (MUF) [16, 17], the standardized independent transformed material unaccounted for (SITMUF) [18], and Page's trend test [19, 20, 21, 22, 23, 24]. The NRTA approach requires direct accountancy of nuclear material through high precision measurements.

In contrast, process monitoring based approaches aim to leverage measurements that do not directly quantify nuclear material. These measurements could include spectral data from gamma or neutron sensors, flow rates, temperatures, or pressures. Process monitoring data streams often have lower uncertainties, which is beneficial as this could result in higher detection probability for anomalous conditions [25, 26]. The majority of previous efforts to incorporate process monitoring have largely focused on using residual signals calculated from models [27, 28], event marking [26, 29, 30] or through simple statistical models [31, 14]. These approaches have been limited by required inclusion of domain knowledge, difficulties setting multi-variate alarm thresholds, and limitations from using simplistic models. Yet other attempts have tried to develop new measurement technologies to improve the ability of NRTA to directly quantify material [32, 33].

Machine learning for anomaly detection [34] has been a recent topic of interest for safeguards given the potential for improved performance over simple statistical models. Principle component analysis (PCA) was applied to gamma spectroscopy for anomaly detection [35, 36, 37] with some limited success. However, PCA has limited expressive power and a detailed study of measurement error was not conducted.

Deep learning for anomaly detection [13], an extension of machine learning that relies on neural networks, has had broader success in other areas of safeguards for tasks such as video surveillance [3, 4] and prediction of burnup in spent fuel [38]. Efforts to apply deep learning to process monitoring data has been met with mixed results. Measurement error has been shown to severely reduce anomaly detection performance in unsupervised algorithms [5, 6, 14]. Supervised algorithms have been shown to perform well [7], but common challenges, such as generalization to unseen classes was not explored. However, some recent work [10] has suggested that supervised algorithms might be pre-trained with domain knowledge through curated training datasets to improve generalization.

3. METHODOLOGY

3.1. Algorithms

The goal of this work is to explore the application of supervised deep learning to process monitoring data. If successful, this approach would reduce costs to the IAEA and would be potentially applicable to other large throughput facilities. Enrichment facilities were used as an exemplar due to the process monitoring streams collected by existing measurement systems such as OLEM (e.g., pressure and temperature).

The data from enrichment facilities could be represented to an algorithm in a number of ways, but perhaps the most obvious would be in a time series format. That is, given streaming data from a OLEM measurement system, segments of facility operation would be consumed by an algorithm with the goal of classifying individual time series segments. There are a multitude of possible algorithms, each with their own trade-offs, that could be implemented to perform this task. A more mature state-of-the-art algorithm based on the Inception network [39] rather than a cutting edge Transformer-based architecture was selected based on improved reproducibility and ease of implementation. Specifically, this work implements and applies the InceptionTime algorithm [1] which has demonstrated high accuracies on the UCR dataset archive [40].

The InceptionTime algorithm, shown in Figure 3-1, is comprised of a few key components. First, the algorithm consists of several Inception modules that are stacked together and connected by shortcut connections which help mitigate issues that arise when neural networks try to learn from long time series segments (e.g., vanishing gradients). Next, a global average pooling layer averages the output over the temporal dimension. Finally, a fully connected neural network with a softmax layer is used to predict the time series class, which for this application, is either 0 for normal (e.g., no loss) or 1 for anomalous (e.g., loss). It is important to note that both the Inception modules themselves and the use of multiple modules have an important impact on the overall algorithm's ability to classify time series, which is particularly relevant given the temporal nature of process monitoring data.

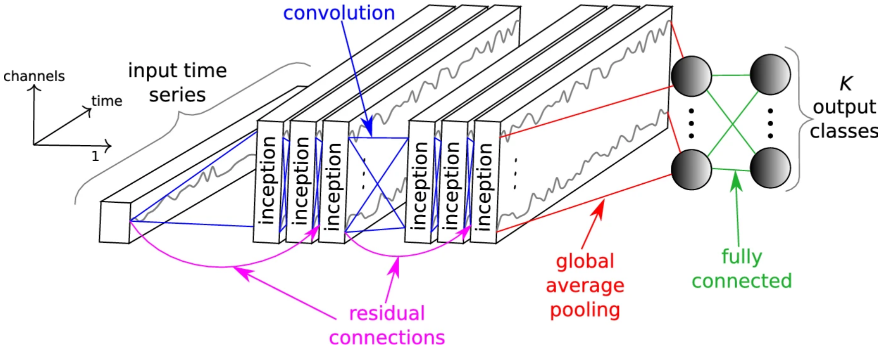

Figure 3-1 Overview of InceptionTime algorithm [1]

Inception modules start with a bottleneck layer wherein sliding filters significantly reduce the dimensionality of the time series in order to reduce model complexity and potential for overfitting. The next component of the Inception model is comprised of sliding filters of different lengths to capture features of different lengths. A separate MaxPool layer followed by a bottleneck layer is applied in order to reduce impact of small perturbations. Finally, the MaxPool and bottleneck layer are concatenated with the multi-length sliding filters to form the module output. Stacking multiple Inception modules allows for extraction of features of multiple resolutions due to specification of filters with various lengths.

There are a number of hyperparameters that must be set for the InceptionTime algorithm. The default values specified by the algorithm's authors are used for most hyperparameters, except batch size, as they provided good empirical performance. Many of the hyperparameters that required tuning varied with number of classes and time series length, as discussed by the original authors. The relatively short length of the input time series and few number of classes further reinforce the choice of the default hyperparameters for this work.

InceptionTime Hyperparameters	
Parameter	**Value**
Batch Size	128
Bottleneck	True
Residual connections	True
Depth	6
Filter length	$\{10, 20, 40\}$
Number of filters	32

Table 3-1 InceptionTime hyperparameters used in this work.

Machine learning algorithms are often described as "black-box" and can be difficult to interpret as learned features can often make little sense to a human. To better understand decisions made by the InceptionTime algorithm, analyses are performed using SHAP (SHapely Additive exPlanation) [11, 12]. The basic concept behind SHAP is that features in an input dataset have different level of contribution to the output of a machine learning algorithm, which can be used to obtain some local understanding. SHAP unified six popular methods and is currently the state-of-the-art explanability framework for deep learning models.

3.2. Data

Real-world data from nuclear facilities is very challenging to obtain and will not include examples of adversarial anomalies (i.e., material loss, excess production, higher than declared enrichment, etc.). As such, synthetic data from a detailed process model formed the basis of this work. A generic systems-level gaseous centrifuge enrichment plant model was developed in MATLAB Simulink to simulate fundamental activities to support this analysis. The model simulates feed and withdrawal activities, flow of material through cascades, OLEM measurements, and more.

The generic model is comprised of 8 parallel cascade halls. Each hall is assigned a dedicated feed station, though product and tails stations are shared between them. This enables each hall to operate semi-autonomously from the others, similar to how a real facility may operate. Facility-level characteristics such as size in tSWU/yr, feed enrichment, product enrichment, and tails enrichment can all be specified by the user. Pressure (kPa), temperature ($^{\circ}$C), and several material masses derived from radiation signatures (kg) are measured at each timestep. Key facility parameters that were selected to generate data used in this work are listed below in Table 3-2.

GCEP Model Parameters	
Parameter	**Value**
Throughput	$3200\ \frac{tSWU}{yr}$
Feed enrichment	0.711 wt% ^{235}U
Product enrichment	4.5 wt% ^{235}U
Tails enrichment	0.2 wt% ^{235}U

Table 3-2 GCEP model parameters used in analysis.

The GCEP model was designed to support development of strategies that could be feasibly implemented by the IAEA. As such, the model uses simple empirical relationships to calculate enrichment of material flows and does not consider facility details such as cascade configuration, number of stages, centrifuge design, etc. OLEM measurements are only simulated in locations that could be potentially accessed by the IAEA (i.e., outside the cascade halls).

4. EXPERIMENTAL RESULTS

The first two experiments in this work seek to bound the performance of supervised algorithms for identifying material losses using a set of simulated facility features. The first of these considers a baseline case wherein only one high consequence material loss sequence is known and is evaluated against other, unseen scenarios. The second experiment examines the impact of dataset engineering on overall performance. Here, a variety of different loss scenarios are provided during training, after which the generalization of the algorithm is evaluated.

Experiments conducted here rely on simulated data; however, a real-world deployment scenario might rely on synthetic training data and real-world evaluation data. It is possible that the synthetic data will not accurately represent facility processes, which will impact the ability of a supervised approach to detect anomalies. The final experiment considers the possibility of imperfect system knowledge and the impact on classification performance.

All experiments are conducted using datasets generated from the GCEP model described in the previous section. Specifically, datasets of size $[n, m, 144]$ are generated where 144 refers to the total number of thermophysical and derived mass features simulated at various facility locations; n is the number of runs, and m is the number of samples. Generally, $n > 1$ to allow for expression of different systematic biases that might occur and $m \approx 8760$ to represent one operational year in hours.

These datasets are processed by performing the following steps:

- **Scaling:** Datasets are scaled such that $X \in [0, 1]$ as a standard preprocessing step to enable learning during training. The scaling was based on a set of simulated normal data.

- **Windowing:** Simulated GCEP features have a temporal dependence that must be captured. This is done by binning the datasets into 200-hour windows that are later classified by the InceptionTime algorithm.

  - Windows are classified as either normal, no loss (0) or abnormal, loss (1) and are not divided into finer categories (i.e., abrupt and protracted losses are labeled as the same class).

  - The 200-hour window was based on expert judgement based on facility operation characteristics. Windowing and the accompanying labeling approach could have an impact on performance but was not explored in-depth during this work. For example, should a window have a fully anomalous segment or only partially anomalous segment before being classified as abnormal?

- **Class balancing:** Each dataset run (i.e., for each run n) is simulated to be an operational year, however, the material loss interval could potentially be small compared to the total run length. This leads to a class imbalance that could incentivize the classification algorithm to simply predict everything as normal. A class balancing procedure is performed to generally maintain a 50/50 ratio of normal and abnormal segments for training. In some cases the class balance was adjusted from 50/50 to 25/75, depending on the quantity of available data.

- **Trimming:** The final step is only applicable in some cases where class balance is very skewed, which leads to a large dataset required to reach a specified class balance. The trimming step simply reduced the amount of training data by a fixed fraction. This is conducted to enable reasonable training time within the limits of currently available computational resources.

Each experiment was evaluated against a static set of anomalous scenarios that are generically labeled 'Scenario 1-3'. These scenarios consist of a material loss at a single, but varied, location with various levels of intensity. The ratio of the material loss rate to the overall measurement uncertainty decreases with increasing scenario number leading to a more difficult detection problem (e.g., scenario 1 is easier to detect than scenario 2). The overall measurement uncertainty was held constant for each scenario, and only the material loss rates and locations are varied. Equation 4.1 below provides a general description of the loss scenario.

$$\frac{d\mu_1}{dt}\frac{\mu_1}{\sigma} > \frac{d\mu_2}{dt}\frac{\mu_2}{\sigma} > \frac{d\mu_3}{dt}\frac{\mu_3}{\sigma} \tag{4.1}$$

Unless otherwise stated, all scenarios discussed are designed to simulate the removal of one significant quantity. Therefore, Equation 4.1 can be reduced to Equation 4.2.

$$\frac{d\mu_1}{dt}\mu_1 > \frac{d\mu_2}{dt}\mu_2 > \frac{d\mu_3}{dt}\mu_3 \tag{4.2}$$

The results of these experiments are reported on the basis of a binary F1-score with respect to the abnormal class for diversion cases and binary F1-score with respect to the normal class unless otherwise stated[1]. F1-score is a better metric than accuracy alone as it expressed the balance between precision (i.e., positive predictive value) and recall (i.e., sensitivity). This score is calculated based on the abnormal class, rather than both classes, as the evaluation often involved imbalanced classes. Put simply, the binary F1-score describes performance in detecting anomalous segments. A higher F1-score indicated better performance where $f1 \in [0, 1]$ and is used here to compare relative performance of different training strategies.

NOTE: F1-score is a measure of classification performance and is not the same as probability of detection. While these measures are correlated (e.g., higher F1-score leads to improved probability of detection), additional work is required to relate the two quantities.

4.1. Evaluation of baseline supervised performance

This first experiment represents a baseline performance level where only a single high consequence loss scenario is known. Several different potential loss scenarios with a range of material loss rates and durations are considered. For each case, InceptionTime is only trained on a single scenario and is then evaluated against the other baseline sets. The purpose of this experiment is to establish the baseline

[1] Binary F1-score with respect to the abnormal class is undefined for no loss cases, so the binary F1-score with respect to the normal class is used instead for no loss cases.

generalization performance of InceptionTime. Results, reported as binary F1-score with respect to the anomalous class, are shown below in Figure 4-1.

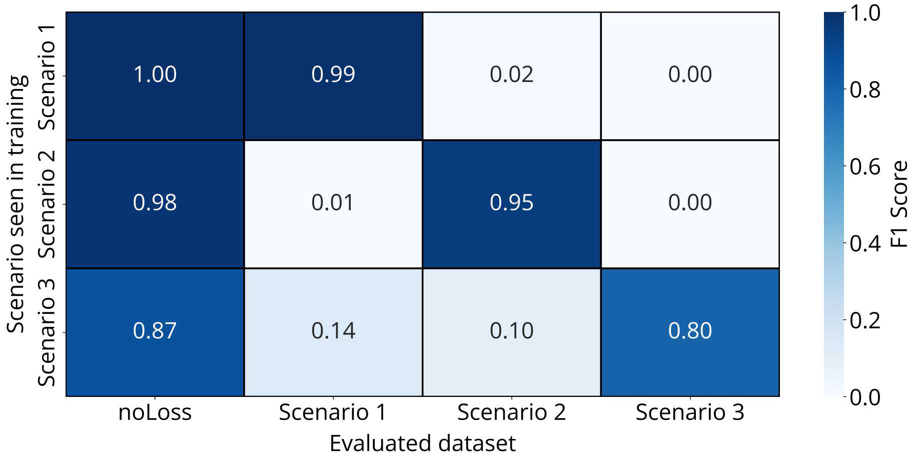

Figure 4-1 Summary of baseline performance expressed as a binary F1-score

The baseline results show that InceptionTime generalizes poorly to unseen cases, but performs well on classes seen in training. This is further illustrated by considering the SHAP values for a high and low performance case. First, consider the SHAP values of Scenario 1 when trained on Scenario 1 (i.e., a high performance case) in Figure 4-2 below. Features are ordered by average SHAP value magnitude (i.e., impact on model prediction).

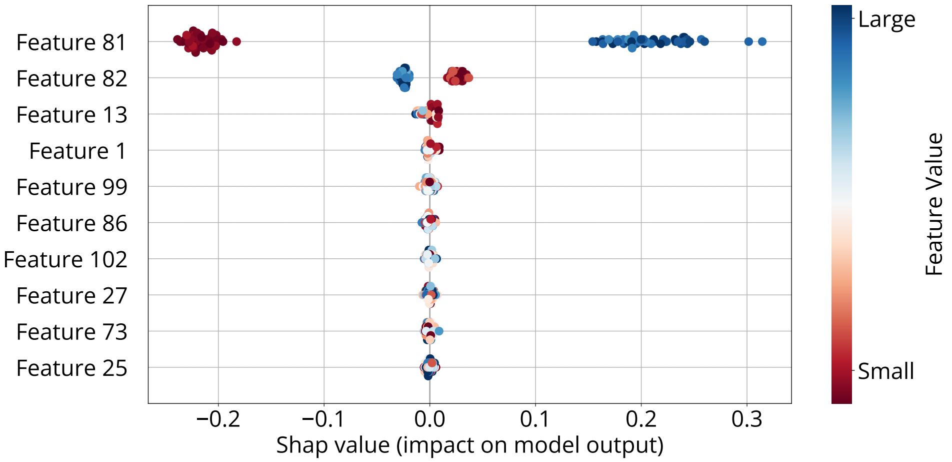

Figure 4-2 SHAP values for a high performance case where the evaluated scenario has been observed during training

First, note the relatively high SHAP values for feature 81, which is associated with inferred mass. Smaller values of feature 81 push the prediction in a negative direction whereas larger values push the prediction in a positive direction. Feature 82, which is associated with a thermophysical feature, is similar in its' impact on the model prediction, but to a lesser extent. This is expected behavior as feature 81 and 82 are directly related to an OLEM sensor impacted by scenario 1. In contrast to feature 81 and

82, the SHAP values of the other features vary little during scenario 1. This indicates that although the model places importance on these features, they are largely unaffected by changes induced by scenario 1. Next, SHAP values for the low performance scenario where a model trained on scenario 1 is evaluated on scenario 2, shown in Figure 4-3.

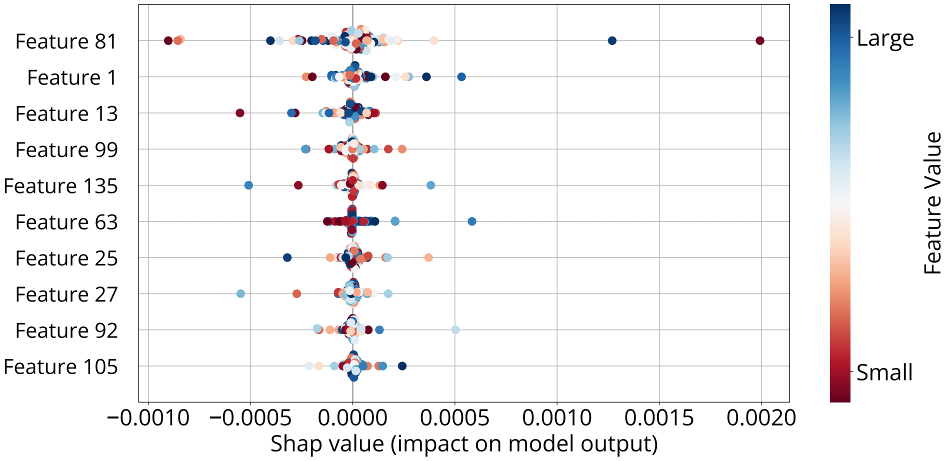

Figure 4-3 SHAP values for a low performance case where the evaluated scenario has been observed during training

Unlike the high performance scenario, SHAP values for all feature values are much smaller. This indicates that changes in feature value have little impact on the model's prediction of normality for scenario 2 for all features. The invariance of feature value on SHAP values shows the inability for the baseline approach to generalize to unseen losses.

Poor supervised generalization for supervised algorithms is well documented in the machine learning literature [41, 42]. This behavior represents one of the key challenges with using supervised learning for safeguards applications. Supervised approaches can learn more powerful representations of the data for classification tasks, leading to good performance, but require explicit examples for each class. In safeguards, this is particularly challenging as it is impossible to know all anomalies (i.e., classes), or to even known all anomalies of a particular type (e.g., material diversion, sabotage, etc.).

4.2. Impact of scenario engineering

Previous work [10] has suggested that the generalization of InceptionTime improves if a diversion is observed at a specific location, even if only a single diversion pattern is observed. The experiments in this section are aimed at improving generalization by providing a range of randomized loss scenarios for training. A number of different scenarios were considered for training InceptionTime:

1. **Fixed Large:** Specifies a material loss rate of 3σ that occurred at all observed locations simultaneously
   a) Note that although the material loss was 3σ, that does not imply the simulated features change by 3σ
   b) $3\sigma \leq 1$ as loss flow rates cannot exceed 1

2. **Fixed Small:** Similar to fixed loss, this scenario Specifies a material loss rate of 1σ that occurred at all observed locations simultaneously
3. **Randomized Small:** Specifies a randomized loss $\in [0, \sigma]$ at a random location
    a) Loss was only simulated at one location per run n
    b) Location, loss, and duration were randomized
    c) $\sigma \leq 1$ as loss flow rates cannot exceed 1
4. **Randomized Large:** Specifies a randomized loss $\in [0, 3\sigma]$ at a random location
    a) Loss was only simulated at one location per run n
    b) Location, loss, and duration were randomized
    c) $3\sigma \leq 1$ as loss flow rates cannot exceed 1
5. **Randomized Chaos:** Specifies randomized losses $\in [0, 1]$ at all locations simultaneously
    a) Duration and loss were randomized
    b) Occurred at all locations, but each location has a different duration and loss
    c) Unlike other scenarios described in this work, the randomization here was unconstrained by a target loss of one significant quantity

A summary of classification performance of each of the training schemes, when evaluated on the baseline loss scenarios, is described below in Figure 4-4.

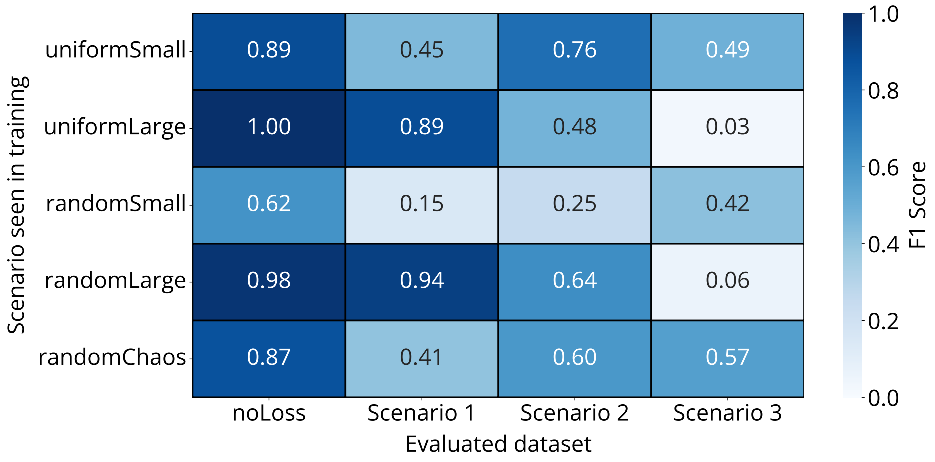

Figure 4-4 Summary of InceptionTime performance trained on engineered datasets as expressed by a binary F1-score

Training on engineered scenarios substantially increases performance over the baseline case. The small engineered scenarios (uniformSmall, randomSmall) and chaos scenario (randomChaos) provide better performance overall than the larger scenarios (uniformLarge, randomLarge). While the large cases perform well on the more abrupt losses and are more confident in the no loss case, there is a steep performance dropoff for scenario 3. The random chaos case and uniform small provide the best performance across all scenarios with random chaos providing the best performance on scenario 3, the most difficult to detect. SHAP values for the random chaos case when evaluated on the baseline loss cases are shown in Figures 4-5, 4-6, and 4-7. These figures provide intuition for how InceptionTime responds when evaluated on the baseline scenarios.

The random chaos SHAP values respond strongly in both Scenarios 1 and 2 with large separation in several features. In scenario 1, there is clear separation in features 81 and 82 indicating that they both have strong impacts on the model's predictions. Similarly in scenario 2, feature 63, and to a lesser extent feature 64, both also have a strong impact on the prediction. The other features listed in Figures 4-5 and 4-6 are features that are weighted heavily by the model, but do not change much in the currently evaluated scenario. For example, feature 87 has high importance in both scenario 1 and scenario 2 but vary little in these scenarios. If the material loss modeled in scenario 1 occurred at a different location, then less response would be observed in features 81 and 82 and more response would be observed in others.

NOTE: SHAP values only describe feature importances for a given subset of the data. Feature responses here reflect importance for the *specific scenario considered*. If multiple scenarios were evaluated at once using the SHAP algorithm, then even more features would exhibit behavior demonstrated by Feature 81 and 82 in scenario 1.

Feature 87 corresponds to an input feature at another location in the facility. In fact, most of the heavily weighed features by the model trained on the random chaos dataset reflect locations that have the largest changes due to modeled anomalies. Further, these features often correspond to inferred mass (e.g., from gamma observed by OLEM). This shows that the random chaos set helps the model learn important features for detecting anomalies regardless of location but does not effectively utilize changes in thermophysical properties due to material loss.

Importantly, the random chaos scenario provides the best performance on the subtle scenario 3. SHAP values in Figure 4-7 show that Feature 57, a feature close to the anomaly location, is important for classification of scenario 3. Note that the separation in feature 57 is still present despite the relative difficulty of detection scenario 3 which shows that the model places heavy importance on the feature, but the change in that feature is relatively small. The relatively small change is due to the subtly of modeled loss pattern in scenario 3.

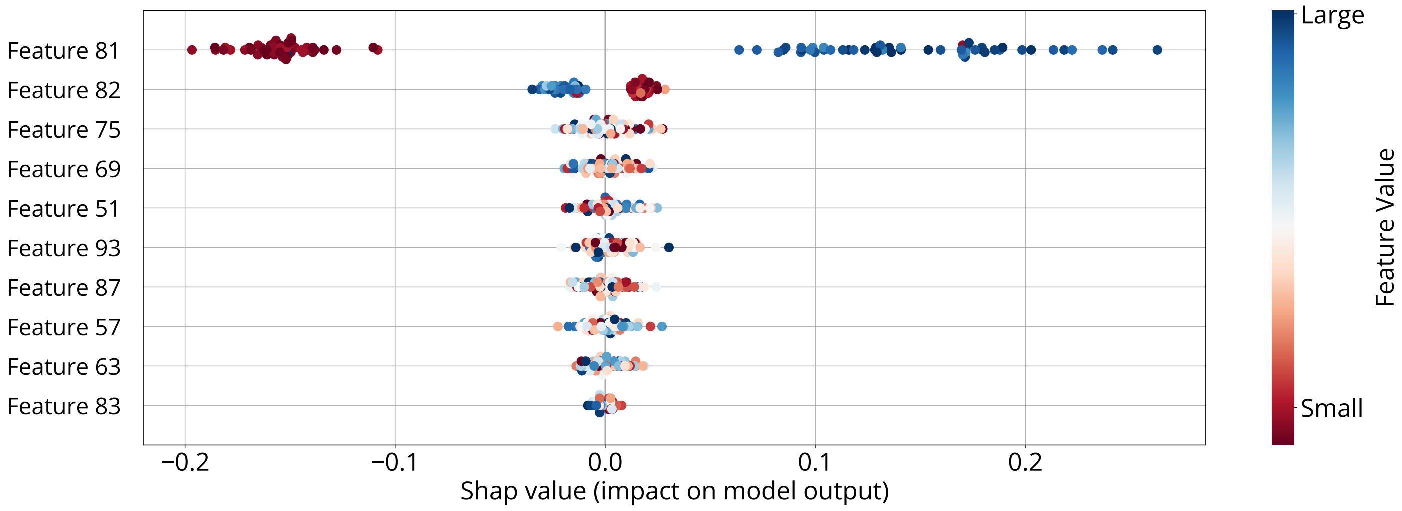

Figure 4-5 SHAP values for random chaos trained model evaluated on Scenario 1

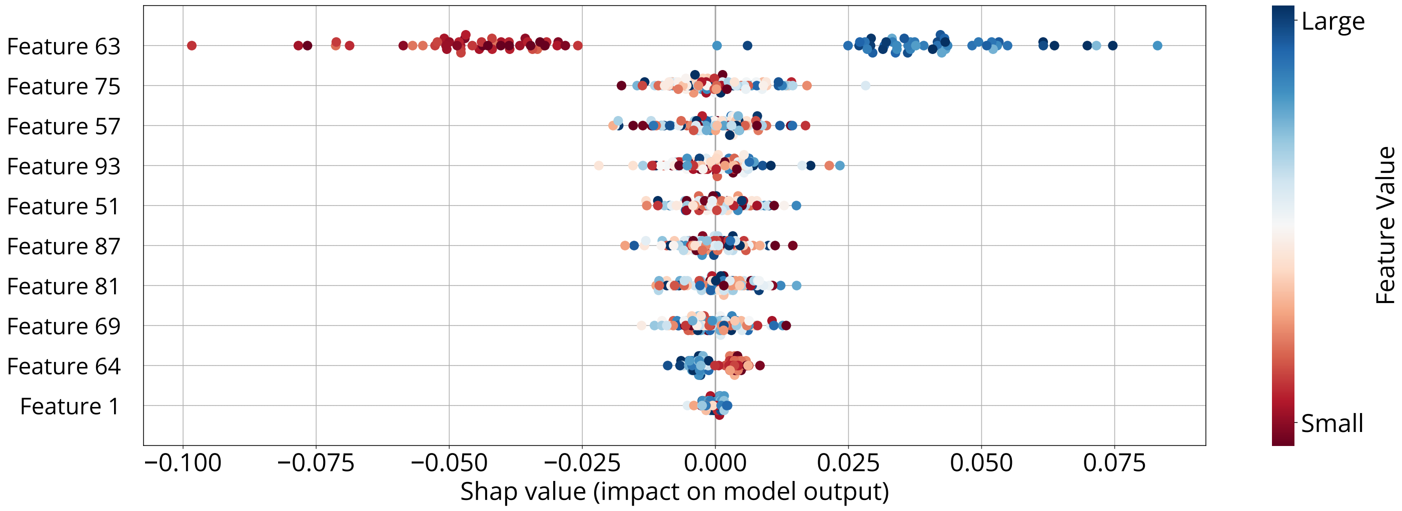

Figure 4-6 SHAP values for random chaos trained model evaluated on Scenario 2

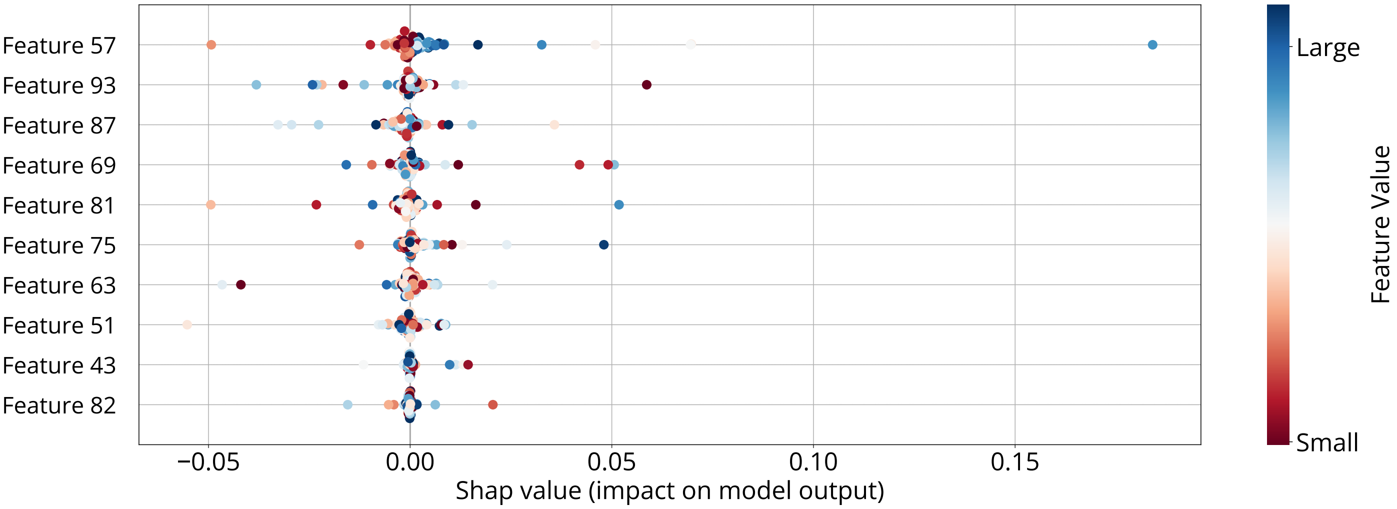

Figure 4-7 SHAP values for random chaos trained model evaluated on Scenario 3

4.3. Imperfect system knowledge

This final experiment considers imperfect system knowledge. Real-world deployment of a classification algorithm to aid material accountancy might require training on synthetic data and evaluation on real-world data. Supervised machine learning often assumes that the training distribution match the evaluation distribution [43], but this might not be available in a safeguards use case. This experiment seeks to provide some intuition regarding the performance reduction when the training data does not perfectly reflect the real-world dataset. Specifically, this experiment trains on a dataset wherein the normal thermophysical features have different means than the evaluation dataset. Consequently, the *behavior* is modeled correctly, however the feature mean is shifted.

The evaluation of the imperfect system knowledge experiment follows the baseline performance case; InceptionTime is trained on a single material loss and evaluated on all others. However, the thermophysical features of the training dataset differ from the evaluation dataset (i.e., features are mean shifted). Results are summarized in Figure 4-8 below.

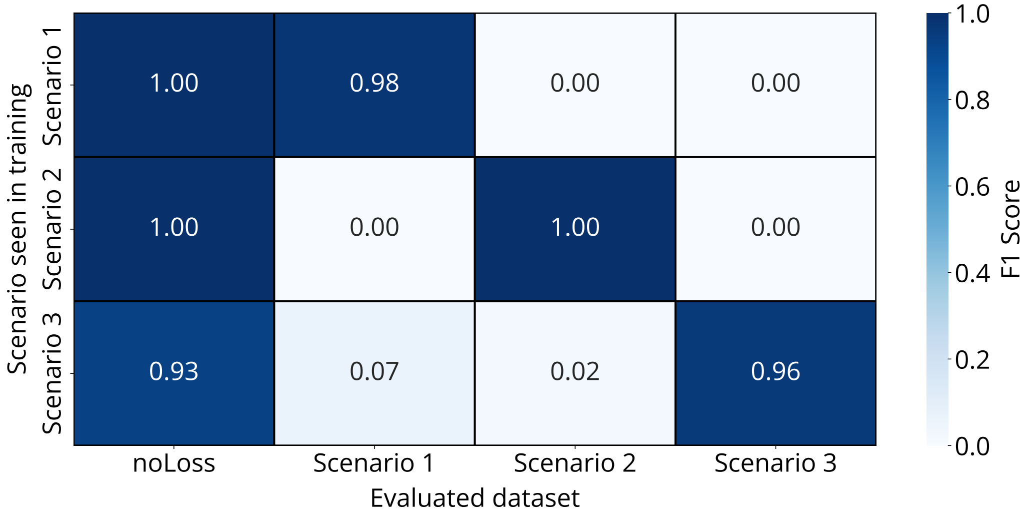

Figure 4-8 Baseline performance when InceptionTime is trained on a mean shifted dataset

Despite the change in normal mean for thermophysical properties, the supervised classification algorithm still performs comparable to the baseline case where the mean was not shifted. This is due to the model's low weighting of thermophysical features for classification. A follow-on experiment where the normal mean of the mass features, which are more heavily weighted by the model, would likely show the expected performance degradation. However, this indicates that high fidelity modeling, which is required to generate synthetic training data, might only be needed for certain key features.

5. CONCLUSION

This work progresses the understanding limitations associated with applied deep learning for safeguards anomaly detection by establishing baseline performance and exploring the impact of dataset engineering on algorithm generalization. The baseline performance, wherein only a single material loss pattern was used during training, is relatively poor due to poor generalization to unseen loss conditions. However, dataset engineering improved the generalization and substantially improved performance over the baseline case. This improvement was largely driven by improved weighing of relevant features caused by exposure to a wider range of randomly generated diversion scenarios. The prospect of dataset engineering to boost supervised performance for safeguards anomaly detection is promising, however, this approach likely requires generation of synthetic training data given the difficulties of obtaining real world data.

This work also showed that even imperfect synthetic datasets might still help boost the generalization of supervised anomaly detection algorithms provided important features were modeled correctly. These features could be discovered through additional analyses on synthetic data or identified by subject matter expertise. Future work should focus on further exploring the data engineering approach by considering the impact of other parameters not considered here such as quantity of training data.

REFERENCES

[1] H. Ismail Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.-A. Muller, and F. Petitjean, "Inceptiontime: Finding alexnet for time series classification," *Data Mining and Knowledge Discovery*, vol. 34, no. 6, pp. 1936–1962, 2020.

[2] H. A. Dayem, A. L. Baker, D. D. Cobb, E. A. Hakkila, and C. A. Ostenak, "Demonstration of near-real-time accounting: the agns 1980-81 miniruns," 1 1984.

[3] Y. Cui, "Using deep machine learning to conduct object-based identification and motion detection on safeguards video surveillance," 11 2018.

[4] Y. Lin, X. Zhang, J. H. Park, S. Yoo, Y. Cui, M. Thomas, and M. Moeslinger, "Using machine learning to track objects across cameras," 8 2021.

[5] N. Shoman and T. Burr, "Impact of safeguards measurement errors on deep neural networks.," 8 2021.

[6] N. Shoman, B. Cipiti, T. Grimes, B. Wilson, and R. Gladen, "Insights from applied machine learning for safeguarding a purex reprocessing facility.," 8 2021.

[7] R. Gladen, T. Grimes, B. Wilson, J. Dermigny, B. B. Cipiti, and N. Shoman, "Neural assessment of non-destructive assay for material accountancy," in *Annual Meeting Proceedings of the Institute of Nuclear Material Management*, 2021.

[8] L. E. Smith and A. R. Lebrun, "Design, modeling and viability analysis of an online uranium enrichment monitor," in *2011 IEEE Nuclear Science Symposium Conference Record*, pp. 1030–1037, 2011.

[9] J. A. March-Leuba, J. Garner, J. Younkin, and D. W. Simmons, "On line enrichment monitor (olem) uf6 tests for 1.5" sch40 ss pipe, revision 1," 1 2016.

[10] N. Shoman and P. Honnold, "Limitations for data-driven safeguards at enrichment facilities," in *Annual Meeting Proceedings of the Institute of Nuclear Material Management*, 2022.

[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 4765–4774, Curran Associates, Inc., 2017.

[12] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, *et al.*, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature Biomedical Engineering*, vol. 2, no. 10, p. 749, 2018.

[13] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM Comput. Surv.*, vol. 54, mar 2021.

[14] N. Shoman, P. Honnold, and B. Cipiti, "Pattern and motif recognition for improved enrichment safeguards," *SAND Report*, vol. SAND2021-11235, 2021.

[15] N. Shoman, P. Honnold, and B. Cipiti, "Fy20 final report on developing signatures-based safeguards for enrichment facilities," *SAND Report*, vol. SAND2020-9461, 2020.

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

[17] J. Doyle, *Nuclear safeguards, security and nonproliferation*. Oxford, England: Butterworth-Heinemann, Jun 2008.

[18] B. Jones, “Near real time material accountancy using SITMUF and a joint page’s test: comparison with MUF and CUMUF tests,” 1988.

[19] B. Jones, “Calculation of diversion detection using the sitmuf sequence and page’s test: application to evaluation of facility designs,” in *Proceedings of the 7th ESARDA Symposium on Safeguards and Nuclear Material Management*, 1985.

[20] B. Jones, “Calculation of diversion detection using the sitmuf sequence and page’s test,” in *Nuclear safeguards technology 1986*, 1987.

[21] B. Jones, “Comparison of near real time materials accountancy using sitmuf and page’s test with conventional accountancy,” in *Proceedings of the 9th ESARDA Symposium on Safeguards and Nuclear Material Management*, 1987.

[22] B. Jones, “Near real time materials accountancy using sitmuf and a joint pages test: improvement of the test,” *ESARDA Bulletin*, vol. 16, pp. 13–19, 1989.

[23] E. S. Page, “Continuous inspection schemes,” *Biometrika*, June 1954.

[24] T. Burr and M. S. Hamada, “Revisiting statistical aspects of nuclear material accounting,” *Science and Technology of Nuclear Installations*, March 2013.

[25] T. L. BURR, C. A. COULTER, J. HOWELL, and L. E. WANGEN, “Solution monitoring: Quantitative and qualitative benefits to nuclear safeguards,” *Journal of Nuclear Science and Technology*, vol. 40, no. 4, pp. 256–263, 2003.

[26] T. Burr, M. Hamada, M. Skurikhin, and B. Weaver, “Pattern recognition options to combine process monitoring and material accounting data in nuclear safeguards,” *Statistics Research Letters*, vol. 1, no. 1, pp. 6–31, 2012.

[27] T. Burr, M. S. Hamada, L. Ticknor, and J. Sprinkle, “Hybrid statistical testing for nuclear material accounting data and/or process monitoring data in nuclear safeguards,” *Energies (Basel)*, vol. 8, 1 2015.

[28] J. B. Coble, S. E. Skutnik, S. N. Gilliam, M. Cooper, and J. Mitchell, “Integrating data sources for improved safeguards and accountancy of electrochemical fuel reprocessing systems. final report,” 3 2020.

[29] H. Garcia, W.-C. Lin, and R. Carlson, “Evaluating safeguards benefits of process monitoring as compared with nuclear material accountancy,” 7 2014.

[30] H. E. Garcia, “Integrated process monitoring based on systems of sensors for enhanced nuclear safeguards sensitivity and robustness,” 7 2014.

[31] T. Burr, M. S. Hamada, L. Ticknor, and B. Weaver, "Model selection and change detection for a time-varying mean in process monitoring," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 751, pp. 79–87, 2014.

[32] S. A. Bryan, T. G. Levitskaia, A. J. Casella, J. M. Peterson, A. M. Lines, E. A. Jordan, D. E. Verdugo, and F. N. Skomurski, "On-line monitoring for control and safeguarding of radiochemical streams at spent fuel reprocessing plant," 7 2011.

[33] L. E. Smith, K. A. Miller, B. S. McDonald, J. B. Webster, M. A. Zalavadia, J. R. Garner, S. L. Stewart, S. J. Branney, L. C. Todd, N. S. Deshmukh, H. A. Nordquist, J. A. Kulisek, and M. T. Swinhoe, "An unattended verification station for uf6 cylinders: Field trial findings," *Nuclear Instruments and Methods in Physics Research. Section A, Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 874, 8 2017.

[34] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, jul 2009.

[35] C. R. Orton, S. A. Bryan, J. M. Schwantes, T. G. Levitskaia, C. G. Fraga, and S. M. Peper, "Advanced process monitoring techniques for safeguarding reprocessing facilities," 11 2010.

[36] N. Shoman, J. Coble, and D. Meier, "Experimental performance of the multi isotop process monitor," in *Transactions of the American Nuclear Society*, vol. 113, pp. 483–485, 2015.

[37] J. Coble and D. Meier, "Monitoring aqueous reprocessing systems for detection of facility misuse," *IEEE Transactions on Nuclear Science*, vol. 66, 2 2019.

[38] Y. Cui, W. Stern, L.-Y. Cheng, C. Soto, O. Dim, M. Todosow, M. Gemmill, J. H. Park, and S. Yoo, "Machine learning in safeguards at pebble bed reactors," 4 2021.

[39] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2015.

[40] Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista, "The ucr time series classification archive," July 2015. `www.cs.ucr.edu/~eamonn/time_series_data/`.

[41] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. S. Yu, "Generalizing to unseen domains: A survey on domain generalization," 2021.

[42] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2022.

[43] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

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

[45] R. R. Picard, "Sequential analysis of material balances," *Journal of Nuclear Materials Management*, vol. 15, 1987.

APPENDIX A. LOWER LIMIT OF DETECTION FOR UNIVARIATE NORMAL DISTRIBUTION

Generally, high precision measurements are required to meet regulatory targets for high throughput bulk facilities. These regulatory targets are often based on σ_{MUF}. This section discusses the lower limit of detection (LLD) for MUF in terms of σ_{MUF}, but note the LLD can be generalized to shifts in any univariate Gaussian distribution. Further background for MUF and statistics for safeguards can be found in [17, 16, 18, 44].

NOTE: The below discussion is focused on one-sided testing for a single material balance period (i.e., testing for material loss only); however, the procedure for determining thresholds for two-sided (i.e., testing for material loss and gain) testing is similar.

A single material balance instance at a specified time is defined by a distribution such that MUF $\sim \mathcal{N}(\mu, \sigma_{\text{MUF}})$. Under loss conditions, the mean of MUF will shift based on the magnitude of the loss such that $\text{MUF}_{\text{loss}} = \mathcal{N}(\mu^*, \sigma_{\text{MUF}})$. This is illustrated below in Figure A-1.

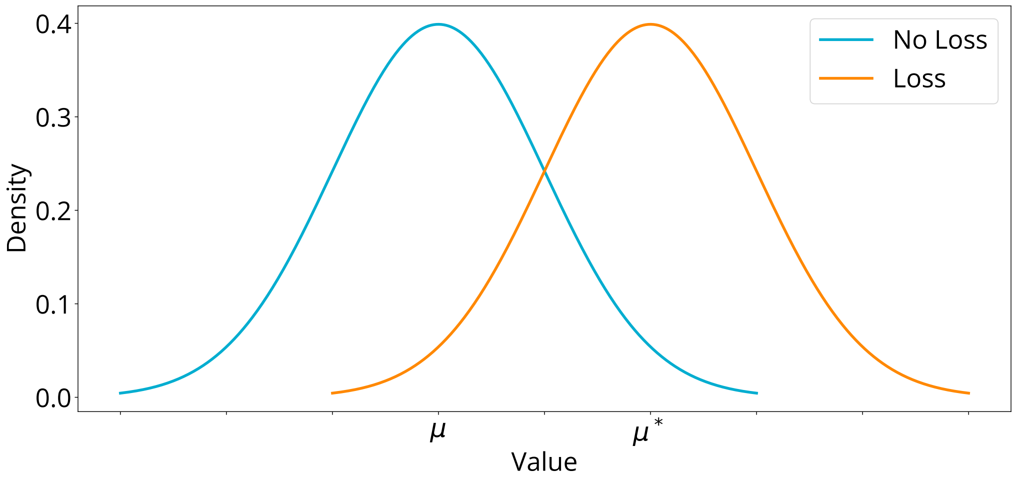

Figure A-1 Shift in MUF distribution due to material loss

One common goal for material accountancy is to set system requirements such that the probability of detection for a material loss be 95% with a false alarm probability of 5%. A lower limit of detection probability for a single material balance instance can be established that relates a mean shift due to a material loss $(\mu - \mu^*)$ to σ_{MUF}. These constraints will be expressed as follows:

$$\begin{aligned} P(x > h \mid \mathcal{N}(\mu, \sigma_{\text{MUF}})) &\leq 0.05 \\ P(x > h \mid \mathcal{N}(\mu^*, \sigma_{\text{MUF}})) &\geq 0.95 \end{aligned} \tag{A.1}$$

Where h denotes some threshold, μ is the average MUF under normal conditions, and μ^* is the average MUF under loss conditions. For simplicity, assume that $\mu = 0$ and $\sigma_{\text{MUF}} = 1$. This leads to an updated set of constraints that can be used to develop a relationship between μ^* and σ_{MUF}:

$$P(x > h \mid \mathcal{N}(0,1)) \leq 0.05 \tag{A.2}$$

$$P(x > h \mid \mathcal{N}(\mu^*,1)) \geq 0.95 \tag{A.3}$$

Specifically note the normal cumulative distribution function and normal quantile function:

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] \tag{A.4}$$

$$F^{-1}(p) = \mu + \sigma\Phi^{-1}(p) = \mu + \sigma\sqrt{2}\text{erf}^{-1}(2p-1),\ p \in (0,1) \tag{A.5}$$

First, determine h by combining the constraint in Equation A.2 with the expression for the quantile function in Equation A.5 to find $h = F^{-1}(0.95) \approx 1.64$ for $\mathcal{N}(0,1)$.

Next, use the constraint from Equation A.4, the expression for the quantile function in Equation A.5, and the previously determined value for $h \approx 1.65$. Solving Equation A.5 as $F^{-1}(p = 0.05; \sigma_{\text{MUF}} = 1) = 1.65$ for μ^* leads to $\mu^* \approx 3.28$.

An expression for the relationship between μ^* and σ_{MUF} subject to the general performance constraints of 95% detection probability and 5% false alarm probability:

$$\begin{aligned}
F^{-1}(p) = \mu + \sigma\Phi^{-1}(p) = \mu + \sigma\sqrt{2}\text{erf}^{-1}(2p-1),\ p \in (0,1) \\
F^{-1}(p = 0.95|\mathcal{N}(0,\sigma_{\text{MUF}})) = F^{-1}(p = 0.05|\mathcal{N}(\mu^*,\sigma_{\text{MUF}})) \\
1.64\sigma_{\text{MUF}} = \mu^* - 1.64\sigma_{\text{MUF}} \\
3.28\sigma_{\text{MUF}} = \mu^* \\
\sigma_{\text{MUF}} = \frac{\mu^*}{3.28}
\end{aligned} \tag{A.6}$$

Equation A.6 refers to the case of fixed probabilities, it can be expanded to a more general case of $\sigma_{\text{MUF}} \leq \frac{\mu^*}{3.28}$ by nothing that $F^{-1}(p|\mathcal{N}(0,\sigma_1) \leq F^{-1}(p|\mathcal{N}(0,\sigma_2)$ where $\sigma_1 < \sigma_2$.

APPENDIX B. MSE LOSS, EXPECTATIONS, AND PRIORS

B.1. Traditional material accountancy includes prior knowledge

The workhorse of traditional material accountancy is Page's trend test [23, 24] on SITMUF [18]. Rather than testing on a single material balance, as described in Appendix A, this approach utilizes trend testing on a transformed sequence of values.

A wide range of existing literature has shown this approach to be versatile and perform well on a wide range of loss patterns [19, 20, 21, 22]. SITMUF is the standardized independent transformed MUF. With knowledge that MUF *should* be zero under the no loss condition, but that it is non-zero due to measurement error, and knowledge of previously observed MUF values, it is possible to transform MUF to the SITMUF sequence. Specifically, where $\text{MB}_{\mathbf{t}} \sim \text{MVN}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ and $\text{SITMUF}_{\mathbf{t}} \sim \mathcal{N}(0, 1)$ (see [16, 45] for further details). The SITMUF transformation is achieved by subtracting the conditional expectation from the observed MUF value.

Consider the MB residual, $i(t)$ at time t, which can be formulated as $i(t) = \text{MB}_t - \boldsymbol{\sigma}^T_{\mathbf{t-1}} \boldsymbol{\Sigma}^{-1}_{t-1} \mathbf{MB}_{t-1}$. The term $i(t)$, sometimes called Independent MUF (ITMUF), is the residual between the observed MB, (i.e. MB_t) and the conditional expectation $\boldsymbol{\sigma}^T_{\mathbf{t-1}} \boldsymbol{\Sigma}^{-1}_{t-1} \mathbf{MB}_{t-1}$ given a particular covariance estimate ($\hat{\boldsymbol{\Sigma}}$), previously observed MUFs (e.g., $\text{MB}_{t-1}, \text{MB}_{t-2}, \ldots, \text{MB}_1$), and the assumption no loss of material ($\mu = 0$).

Picard [45] showed a convenient way to calculate this quantity by way of Cholesky decomposition. Specifically, given the Cholesky decomposition of the covariance such that $\boldsymbol{\Sigma} = \mathbf{LU}$ where $\mathbf{L}$ is a lower-triangular square matrix and $\mathbf{U} = \mathbf{L}^T$ is a upper-triangular transpose of $\mathbf{U}$. Then $\text{SITMUF}_t = \mathbf{L}^{-1}\mathbf{MB}_t$ where $\text{SITMUF} = \frac{i(t)}{\widetilde{\sigma}_l}$ and $\widetilde{\sigma_l}$ is the standard deviation of the ITMUF. One particular benefit is that the variance in the transformed, uncorrelated sequence (SITMUF) decreases over time as the conditional expectation improves which results in increased probability of detection for material loss. Put simply, the conditional expectation under loss conditions will more closely align to the no-loss case, which will lead to a larger residual that can be detected through a trend test.

It is difficult to bound the performance of this approach as was done in Appendix A for the single material balance. Some complexities include generalizing to sequence length, expression of the cumulative distribution function, expression of the inverse cumulative distribution function, and more. There have been some attempts to bound performance of Page's trend test on SITMUF [44, 24], however, they often have simplifications. Generalized performance bounds could be a target for future work.

Key takeaway: Trend testing on SITMUF has been shown to perform well for most material loss patterns, but analytical performance bounds are difficult to determine. The covariance used in the conditional expectation on MUF uses domain knowledge to calculate an analytical estimate as empirical in-field estimates can be difficult.

B.2. Unsupervised deep learning has less powerful assumptions about process monitoring data

Unsupervised methods for anomaly detection are attractive candidates for safeguards process monitoring as they do not require specific examples of anomalies. Instead, unsupervised methods use some intermediate proxy methods to estimate data abnormality. There are a wide range of techniques to estimate abnormality; probabilistic, distance-based, reconstruction-based, and information-theoretic approaches [34]. Unsupervised methods are limited by constraints described in Appendix A. Namely, their performance will be limited by the overall measurement error of the underlying process monitoring data.

For example, consider a reconstruction-based unsupervised anomaly detection approach called autoencoders. The goal of this family of algorithm is to learn compressed (i.e., lower-dimensional) representations of data that can effectively represent the input dataset. This lower-dimensional representation, if learned using normal data, should poorly represent anomalous data leading to a larger-than-normal reconstruction error.

Concretely, an autoencoder consists of an original dataset $\mathbf{X} = \mathbb{R}^m$ and latent representation $\mathbf{Z} = \mathbb{R}^n$ where $m > n$ with an encoder $E_\phi : \mathbf{X} \to \mathbf{X}$ and decoder $D_\theta : \mathbf{Z} \to \mathbf{X}$. For any $x \in \mathbf{X}$ and $z \in \mathbf{Z}$ the encoder are usually expressed as $z = E_\phi(x)$ and $x' = D_\theta(z)$ respectively as $x \neq x'$ in practice due to imperfect encoder and decoder functions.

The training objective to learn E_ϕ and D_θ is often described using the mean squared error loss function:

$$\begin{aligned} \min_{\theta,\phi} L(\theta,\phi), \text{where } L(\theta,\phi) &= \frac{1}{N}\sum_{i=1}^{N} \|x_i - x_i'\|_2^2 \\ &= \frac{1}{N}\sum_{i=1}^{N} \|x_i - D_\theta(E_\phi(x_i))\|_2^2 \end{aligned} \tag{B.1}$$

The mean squared loss function described above, which is used to train many unsupervised training algorithms, can be expressed in terms of distributions. In fact, it can be shown under i.i.d. assumptions that minimization of the mean squared error loss function is equivalent to maximization of the conditional log likelihood [43]:

$$\underset{\theta,\phi}{\operatorname{argmax}} \sum_{i=1}^{N} \log P(x_i' \mid x_i; \theta, \phi) = \underset{\theta,\phi}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N} \|x_i - x_i'\|_2^2 \tag{B.2}$$

This is important to note as the use of mean squared loss functions, often used for training unsupervised algorithms, is equivalent to finding the set of parameters to maximize the conditional log likelihood. Put simply, this common training procedure tries to find a set of parameters that results in an output distribution most closely matches the training target distribution.

Autoencoders will be limited by constraints applicable normally distributed random variates described in Appendix A. Similar arguments can be made for other unsupervised algorithms as their optimization target is also impacted by process monitoring measurement error.

Meaningful gains in anomaly detection for process monitoring over testing of single material balance periods will be dependent on specific responses of process monitoring streams to anomalies. The shift in the mean process monitoring value die to an anomaly must be significant compared to the relative change in MUF under the same conditions:

$$\frac{|\mu_{\mathrm{PM}} - \mu^*_{\mathrm{PM}}|}{\sigma_{\mathrm{PM}}} > \frac{|\mu_{\mathrm{MUF}} - \mu^*_{\mathrm{MUF}}|}{\sigma_{\mathrm{MUF}}} \tag{B.3}$$

Comparing unsupervised methods to trend testing on SITMUF is more difficult given the challenges to analytically bound performance of trend testing on SITMUF. Currently, comparison of machine learning on process monitoring must be directly compared to SITMUF trend testing via empirical means. This is a potential area for future work.

Key takeaway: Direct analytical comparisons of unsupervised machine learning to trend testing on SITMUF is difficult and must be performed empirically. However, comparisons between unsupervised machine learning to non-sequential material balances can be developed analytically. Performance gains using unsupervised machine learning will only be realized if the relative change in the process monitoring signal is greater than the change in MUF.

DISTRIBUTION

Email—External [redacted]

Name	Company Email Address	Company Name
Wayne Mei	wayne.mei@nnsa.doe.gov	NNSA

Email—Internal [redacted]

Name	Org.	Sandia Email Address
Nathan Shoman	8845	nshoman@sandia.gov
Philip Honnold	8845	phonnol@sandia.gov
Scott Sanborn	8845	sesanbo@sandia.gov
Technical Library	1911	sanddocs@sandia.gov

Sandia National Laboratories
National
Laboratories

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & 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.