Sandia National Laboratories

Exceptional service in the national interest

Limitations for Data-driven Safeguards at Enrichment Facilities

Nathan Shoman[1], Philip Honnold[1]

[1]Sandia National Laboratories

July 26, 2022

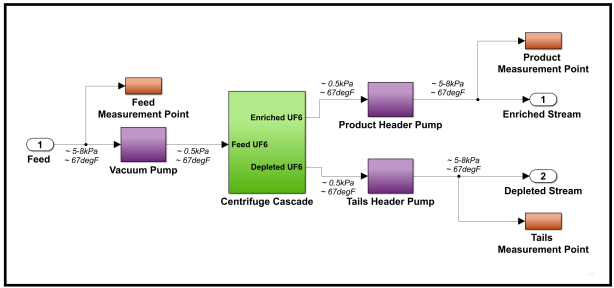
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Motivation: Safeguards at enrichment facilities are expensive and considered in isolation

- Multiple potential class of safeguards anomalies at bulk facilities
  - Over enrichment, excess production, material diversion, etc.
- Existing safeguards effective, but address safeguards anomalies in isolation
  - Weigh scales for cylinders are not used to help monitor for excess production
- **Key question:** Can a unified analysis of facility signals provide better safeguards anomaly detection than existing approaches that consider anomalies individually?
  - Bonus: Could this unified analysis be performed cheaper (both in terms of labor and direct capital costs) than the existing traditional approach?

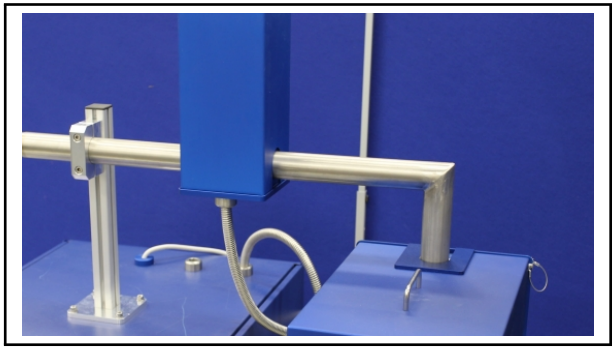
Context: gaseous enrichment overview

- Centrifuges spin very fast to separate homogeneous mixture of UF_6
- Gaseous centrifuge enrichment plants (GCEPs) consists of hundreds to thousands of individual centrifuges
- Specific arrangements of centrifuges are often classified or proprietary
  - Safeguards often conducted outside the cascade hall (with exception to design verification activities).



Context: State-of-the-art systems to measure enrichment - OLEM

- Online Enrichment Monitor (OLEM) can verify enrichment levels
- Utilizes several data modalities to estimate enrichment
  - Temperature, pressure, gamma spectra
- Highly effective, but not widely deployed
- Does not provide information on other classes of safeguards anomalies



V. Fournier/IAEA

Context: material accountancy relies on counting everything

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

where

- $\sum_{i=1}^{n_\mathrm{I}} I_{i,t}$ is the total inventory at time t across all locations
- $\sum_{i=1}^{n_\mathrm{in}} Tin_{i,t}$ is the total input at time t across all locations
  - If inputs are flows they should be summed over the time period of interest (i.e. $t-1$ to t)
- $\sum_{i=1}^{n_\mathrm{out}} Tout_{i,t}$ is the total output at time t across all locations
  - Similar to the inputs, if outputs are flows, they should be summed over the time period of interest

Machine learning can solve all our problems!

Pros:

- Powers many modern conveniences
- Demonstrated beyond human levels of performance on several tasks
- Accessible state-of-the-art approaches and frameworks

Cons:

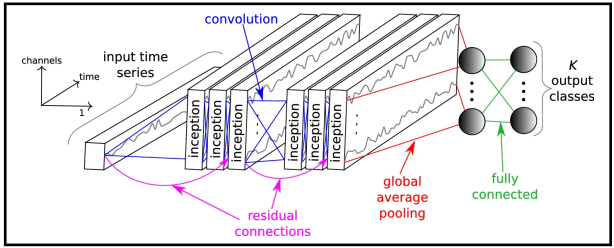
- Documented issues with unsupervised approaches for safeugards tasks
- Large quantities of training data needed in some cases
- Labeled examples required to leverage full power of machine learning

Problem overview

- **Key question:** Can data science, combined with the totality of signals currently observed at enrichment facilities, outperform traditional safeguards approaches?
- **Key question:** What are the limitations for supervised methods? Could documented remedies in machine learning literature for common problems (e.g., small training datasets, lack of labels) or domain knowledge reduce the impact of known limitations?
- **Key question:** Could important streams be implemented using unattended or autonomous approaches?
- **Experiment:** Apply a state-of-the-art time-series classification model to safeguards data from a synthetic enrichment model to evaluate performance and search for remedies to common machine learning limitations.
  - Requiring examples of all possible anomaly pathways is infeasible for safeguards applications.

Methodology: algorithm

- InceptionTime algorithm demonstrated state-of-the-art performance on UCR dataset archive (supervised time series classification)
  - Using a benchmarked algorithm should eliminate architecture as a reason for poor performance
- Several algorithmic features to address challenges with time series (e.g., vanishing gradients from long sequences)
- Somewhat more intuitive than cutting edge Transformer based architectures



Fawaz et al., "InceptionTime: Finding Alexnet for time series classification"

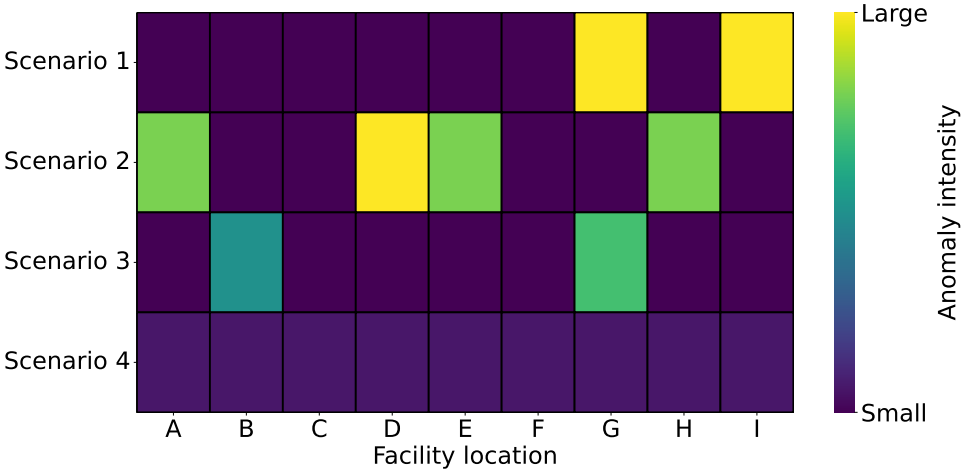
Methodology: data generation

- Based on synthetic data generated from GCEP model
- Generic balance-of-plant model with 8 parallel cascades
- Designed to support simulation of measurements that could feasibly be obtained by IAEA
- Simulates thermophysical feedbacks in facility resulting from operations
- Ignored detailed questions around time series off-normal classification interval

GCEP Model Parameters	
Parameter	**Value**
Throughput	$600\ \frac{tSWU}{yr}$
Feed enrichment	0.711 wt% ^{235}U
Product enrichment	4.5 wt% ^{235}U
Tails enrichment	0.2 wt% ^{235}U

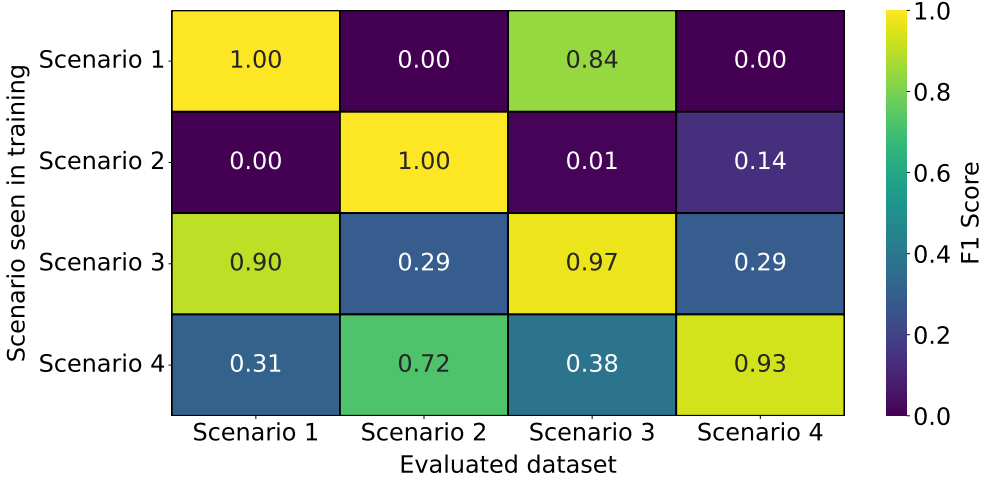
Experiments: scenario description

Work considered several scenarios at different locations within GCEP facility.



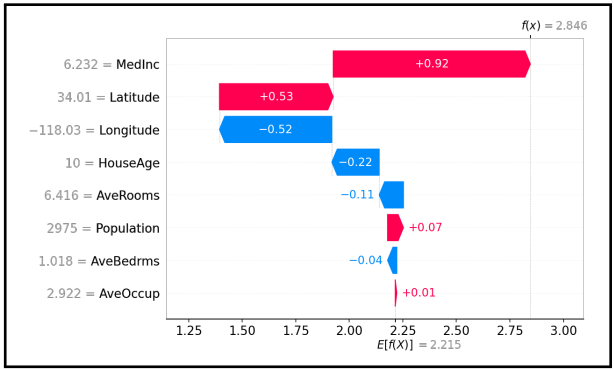
Experiments: empirical performance

Performance of InceptionTime classification algorithm when trained of different scenarios.



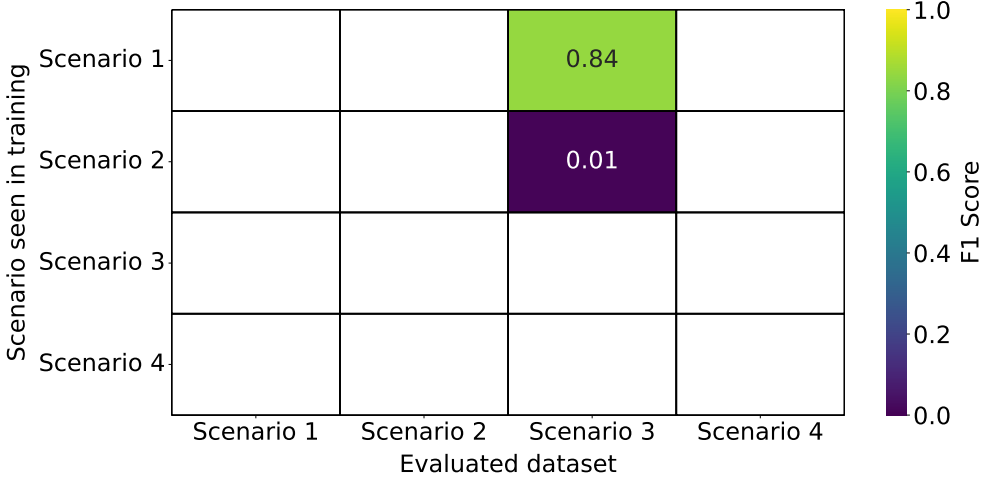
Explainability: SHAP Values

- “The core idea behind Shapley value based explanations of machine learning models is to use fair allocation results from cooperative game theory to allocate credit for a model’s output among its input features”
- Shap values explain difference between specific observation and the model average prediction
- Model-agnostic approach to approximate feature impact on model
- Local only explanations
- Useful for understanding model decisions



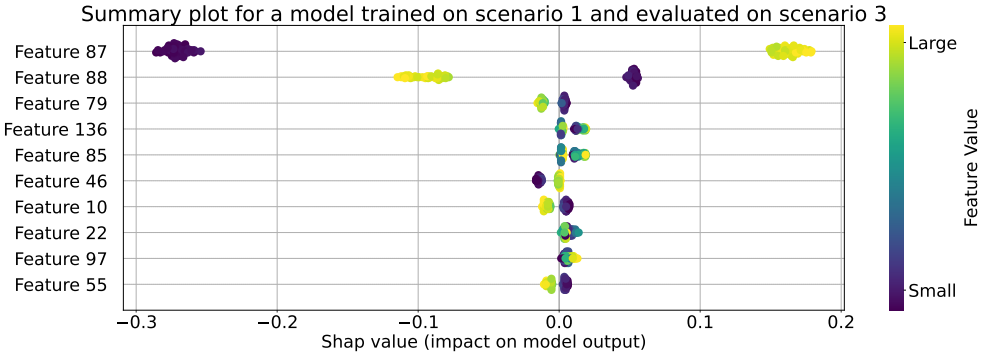
Slundberg et al., Shap values documentation: Example on house pricing

Explainability: focus on comparison of two scenarios; one high performance and low performance



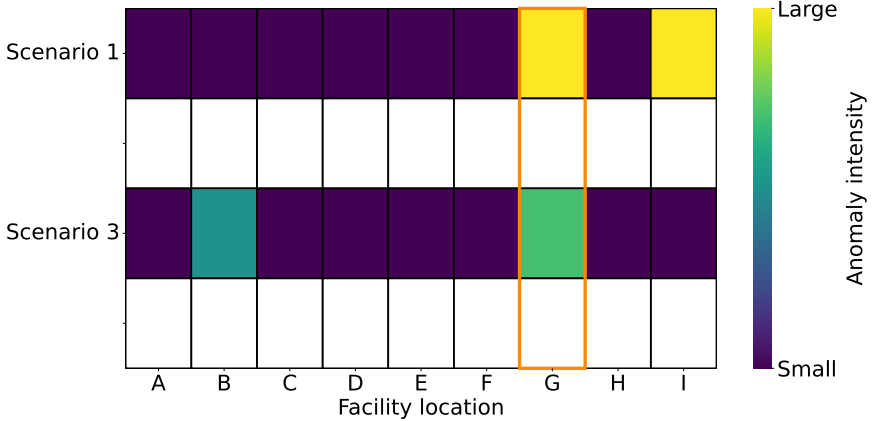
Explainability: why the performance discrepancy? Shap values of high performance case.

Response of InceptionTime to scenario 3 when trained on scenario 1 (**F1 = 0.84** – high performance case). Note the strong Shap value response to different feature values for feature 87 and 88.



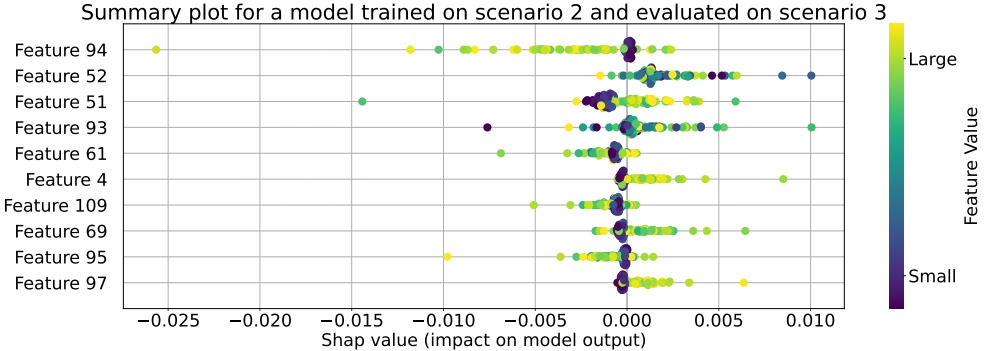
Explainability: Contributing factors to high performance case.

Good performance on unseen anomaly (scenario 3) largely arises from common features in unseen anomaly. Although the anomaly pattern is new/different, anomalous behavior had been previously observed at that location causing the algorithm to learn about those relevant features.

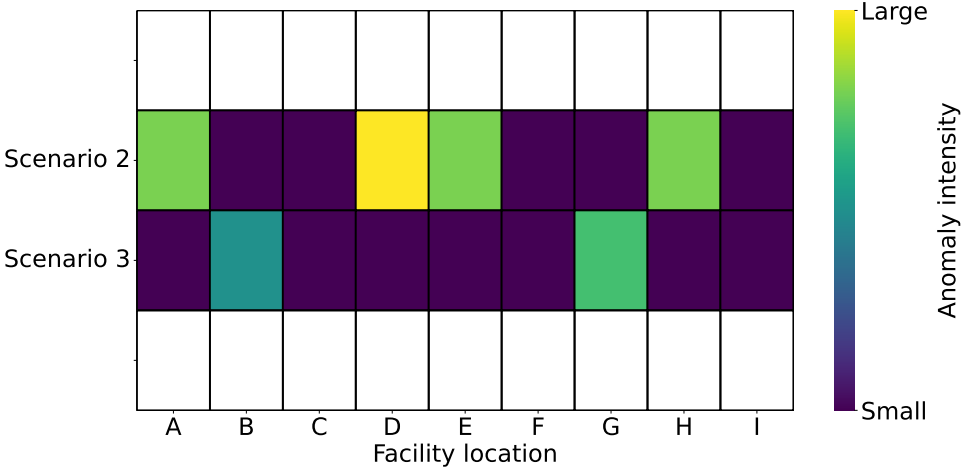


Explainability: SHAP values for low performance case

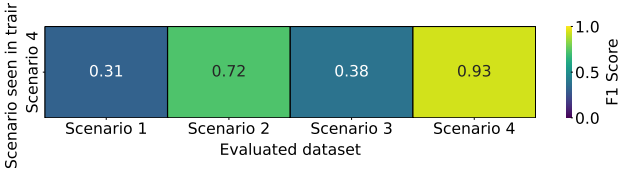
Response of InceptionTime to scenario 3 when trained on scenario 2 (**F1 = 0.01** – low performance case). Note the smaller Shap value response to top features.

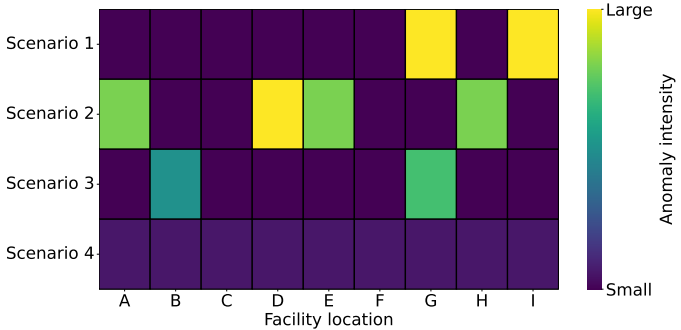


Explainability: features of the lower performance case seen during training



Explainability: performance scaling with number of anomaly locations.





Conclusions

- Presented work sought to improve enrichment safeguards by utilizing existing features combined with machine learning for enhanced anomaly detection
- Good performance observed on anomalies seen during training
- Generally, supervised approaches tended to exhibit poor generalization on scenarios with anomalous features not seen in training
  - Non-zero detection was observed for new anomalous patterns that contained features previously observed with different anomalies
- Possible engineering of training datasets could lead to better supervised performance, even on unseen datasets

Acknowledgements

This work was funded through the National Nuclear Security Administration's Office of International Nuclear Safeguards.